frontiers | Research Topics

Enhanced biological mechanism study, drug discovery and individualized medicine with single-cell multiomics data and integrative analysis

Edited by
Panwen Wang, Mulin Jun Li, Feng Xu and Jing Qin

Published in
Frontiers in Genetics

ISSN 1664-8714
ISBN 978-2-8325-2965-2
DOI 10.3389/978-2-8325-2965-2

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

Enhanced biological mechanism study, drug discovery and individualized medicine with single-cell multiomics data and integrative analysis

Topic editors

Panwen Wang — Mayo Clinic Arizona, United States
Mulin Jun Li — Tianjin Medical University, China
Feng Xu — University of British Columbia, Canada
Jing Qin — Sun Yat-sen University, China

Citation

Wang, P., Li, M. J., Xu, F., Qin, J., eds. (2023). *Enhanced biological mechanism study, drug discovery and individualized medicine with single-cell multiomics data and integrative analysis*. Lausanne: Frontiers Media SA. doi: 10.3389/978-2-8325-2965-2

Table of contents

TYPE Editorial
PUBLISHED 26 June 2023
DOI 10.3389/fgene.2023.1236126

Check for updates

OPEN ACCESS

EDITED AND REVIEWED BY
Maxim B. Freidin,
Queen Mary University of London,
United Kingdom

*CORRESPONDENCE
Panwen Wang,
✉ pwwang@pwwang.com
Jing Qin,
✉ qinj29@mail.sysu.edu.cn

RECEIVED 07 June 2023
ACCEPTED 19 June 2023
PUBLISHED 26 June 2023

CITATION
Zhang X, Wang P and Qin J (2023), Editorial: Enhanced biological mechanism study, drug discovery and individualized medicine with single-cell multiomics data and integrative analysis. *Front. Genet.* 14:1236126. doi: 10.3389/fgene.2023.1236126

Editorial: Enhanced biological mechanism study, drug discovery and individualized medicine with single-cell multiomics data and integrative analysis

Xinying Zhang[1], Panwen Wang[2]* and Jing Qin[1]*

[1]School of Pharmaceutical Sciences (Shenzhen), Shenzhen Campus of Sun Yat-sen University, Shenzhen, China, [2]Department of Quantitative Health Sciences, Mayo Clinic Arizona, Scottsdale, AZ, United States

KEYWORDS

individualized medicine, multi-omics, prognostic biomarker, therapeutic targets, computational approach

Editorial on the Research Topic
Enhanced biological mechanism study, drug discovery and individualized medicine with single-cell multiomics data and integrative analysis

Individualized medicine, also known as personalized medicine or precision medicine, is a field in which medical decisions concerning disease prevention, diagnosis, and treatment are tailored to individual patients based on their genetic information (König et al., 2017). The personal genetic information of patients could be measured by multi-omics technologies, including single-cell (SC) multi-omics. This Research Topic consists of nine manuscripts, covering a diverse range of topics from the computational analyses on various bulk and SC omics data to their applications to uncover underlining disease mechanisms, identify optimal diagnostic and prognostic biomarkers, and discover therapeutic targets and corresponding drugs for individual patients.

The review *RNA-seq data science: From raw data to effective interpretation* (Deshpande et al.) introduced basic concepts of RNA-seq data and defined discipline-specific jargon. Various RNA-seq technologies and their advantages as well as limitations were discussed in this review. Moreover, it described the major steps of computational analysis of RNA-seq data, beginning from the processing of raw data to the uncovering of biological insights, which is helpful to explore the mechanism of disease and thus identify related biomarkers at the molecular level.

Understanding the biomarkers of diseases is vital in identifying cell diversity and molecular classification, and single-cell RNA-seq (scRNA-seq) data is an effective tool for this purpose. Zhao et al. analyzed scRNA-seq data from 23 colon cancer patients to discover biomarker genes for various cancer-associated fibroblast (CAF) subtypes. This helped classify colon cancer patients into six groups and provided new insights into the significant role of CAF in cancer treatment. The CAF-related signature genes were also utilized to develop a prognostic model for colon cancer patients using LASSO Cox regression. The model surpassed traditional clinical feature-based models in predicting the prognosis of colon cancer patients. It is worth noting that other researchers have also developed prognostic models for pancreatic cancer (Tao et al.), hepatocellular carcinoma (Liu et al.), and acute myeloid leukemia (Shi et al.) using a similar bioinformatic framework.

Apart from molecular classification and prognostic model construction, multi-omics analyses were often conducted to investigate the biological functions of biomarkers or therapeutic target genes. Zhu et al. studied the complex functions of REV1, a member of the translesion synthesis DNA polymerase Y family, at multi-omics levels. The authors combined single nucleotide polymorphisms (SNPs), gene expression, and drug sensitivity information to explore the role of REV1 in carcinogenesis and prognosis as well as predicting drug sensitivities of specific signaling pathways. The study demonstrated that REV1 had the potential to be a novel prognostic biomarker for various cancers.

Genes play a vital role in various biological processes, with both protein-coding (mRNAs) and non-coding (ncRNAs) genes being integral components. Among the latter, microRNAs (miRNAs) are short endogenous RNAs consisting of 20–25 nucleotides (Ambros, 2001). They regulate gene expression post-transcription and are currently being studied for their potential to serve as biomarkers for premature ovarian failure (POF). Zhang et al. have identified miRNA-190a-5p as a promising biomarker after conducting bioinformatics analysis with miRBase and TargetScan, as well as animal experimental validation. In rats, this miRNA activates primordial follicles by targeting the expression of PHLPP1 and key proteins in the AKT-FOXO3a and AKT-LH/LHR pathways. These findings highlight miRNA-190a-5p's potential as a therapeutic target for POF and call for further research in this area.

Screening or designing optimal drugs based on biomarkers and targets is a significant step in drug discovery. Zhang et al. developed a deep learning framework that uses convolutional neural networks (CNNs) and attention mechanisms to predict drug-protein interactions (DPIs). Attention mechanisms help identify relevant information features and improve DPI prediction performance. Guan et al. created a computational model, BNEMDI, to identify miRNA-drug interactions (MDIs) through drug substructure fingerprint, miRNA sequence, and MDIs bipartite graph. BNEMDI is stable and effective, as demonstrated by identifying miRNAs that potentially interact with 5-fluorouracil (5-FU). Understanding drug-miRNA relationships is crucial for investigating drug function mechanisms and developing treatments.

In summary, these manuscripts involved reviews, new computational pipelines, and new models with multi-omics data, helping to enhance biological mechanism study, drug discovery, and individualized medicine. Altogether, they provide a broad overview of the current status of biological and pharmaceutical integrative analysis, showing promising advances in multi-omics-based research through the application of computational approaches.

Author contributions

XZ drafted the editorial. PW and JQ reviewed all manuscripts in the Research Topic and revised the editorial. All authors contributed to the article and approved the submitted version.

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

Ambros, V. (2001). microRNAs: tiny regulators with great potential. *Cell* 107 (7), 823–826. doi:10.1016/s0092-8674(01)00616-x

König, I. R., Fuchs, O., Hansen, G., von Mutius, E., and Kopp, M. V. (2017). What is precision medicine? *Eur. Respir. J.* 50 (4), 1700391. doi:10.1183/13993003.00391-2017

ORIGINAL RESEARCH
published: 28 June 2022
doi: 10.3389/fgene.2022.900380

Machine Learning Screens Potential Drugs Targeting a Prognostic Gene Signature Associated With Proliferation in Hepatocellular Carcinoma

Jun Liu[1,2†], Jianjun Lu[3†], Wenli Li[4], Wenjie Mao[5] and Yamin Lu[1]*

[1]Department of Clinical Laboratory, Yue Bei People's Hospital, Shantou University Medical College, Shaoguan, China, [2]Medical Research Center, Yue Bei People's Hospital, Shantou University Medical College, Shaoguan, China, [3]Department of Medical Affairs, First Affiliated Hospital of Sun Yat-sen University, Guangzhou, China, [4]Reproductive Medicine Center, Yue Bei People's Hospital, Shantou University Medical College, Shaoguan, China, [5]Emergency Department, Yue Bei People's Hospital, Shantou University Medical College, Shaoguan, China

Edited by:
Andrey Ivanov,
Emory University, United States

Reviewed by:
Hyun Uk Kim,
Korea Advanced Institute of Science and Technology, South Korea
Lingming Kong,
The First Affiliated Hospital of University of Science and Technology of China Anhui Provincial Hospital, China

***Correspondence:**
Yamin Lu
luyamin303@163.com

†These authors share first authorship

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

Received: 08 April 2022
Accepted: 13 June 2022
Published: 28 June 2022

Citation:
Liu J, Lu J, Li W, Mao W and Lu Y (2022) Machine Learning Screens Potential Drugs Targeting a Prognostic Gene Signature Associated With Proliferation in Hepatocellular Carcinoma. Front. Genet. 13:900380. doi: 10.3389/fgene.2022.900380

Background: This study aimed to screen potential drugs targeting a new prognostic gene signature associated with proliferation in hepatocellular carcinoma (HCC).

Methods: CRISPR Library and TCGA datasets were used to explore differentially expressed genes (DEGs) related to the proliferation of HCC cells. Differential gene expression analysis, univariate COX regression analysis, random forest algorithm and multiple combinatorial screening were used to construct a prognostic gene signature. Then the predictive power of the gene signature was validated in the TCGA and ICGC datasets. Furthermore, potential drugs targeting this gene signature were screened.

Results: A total of 640 DEGs related to HCC proliferation were identified. Using univariate Cox analysis and random forest algorithm, 10 hub genes were screened. Subsequently, using multiplex combinatorial screening, five hub genes (FARSB, NOP58, CCT4, DHX37 and YARS) were identified. Taking the median risk score as a cutoff value, HCC patients were divided into high- and low-risk groups. Kaplan-Meier analysis performed in the training set showed that the overall survival of the high-risk group was worse than that of the low-risk group ($p < 0.001$). The ROC curve showed a good predictive efficiency of the risk score (AUC > 0.699). The risk score was related to gene mutation, cancer cell stemness and immune function changes. Prediction of immunotherapy suggetsted the IC50s of immune checkpoint inhibitors including A-443654, ABT-888, AG-014699, ATRA, AUY-922, and AZ-628 in the high-risk group were lower than those in the low-risk group, while the IC50s of AMG-706, A-770041, AICAR, AKT inhibitor VIII, Axitinib, and AZD-0530 in the high-risk group were higher than those in the low-risk group. Drug sensitivity analysis indicated that FARSB was positively correlated with Hydroxyurea, Vorinostat, Nelarabine, and Lomustine, while negatively correlated with JNJ-42756493. DHX37 was positively correlated with Raltitrexed, Cytarabine, Cisplatin, Tiotepa, and Triethylene Melamine. YARS was positively correlated with Axitinib, Fluphenazine and Megestrol acetate.

NOP58 was positively correlated with Vorinostat and 6-thioguanine. CCT4 was positively correlated with Nerabine.

Conclusion: The five-gene signature associated with proliferation can be used for survival prediction and risk stratification for HCC patients. Potential drugs targeting this gene signature deserve further attention in the treatment of HCC.

Keywords: hepatocellular carcinoma, proliferation, gene signature, immune checkpoint inhibitors, antitumor drugs

INTRODUCTION

As one of the most common cancers worldwide, hepatocellular carcinoma (HCC) is currently the third leading cause of cancer-related death (Cronin et al., 2018). In the past few decades, the effects of drug resistance and long-term toxicity of systemic therapy on overall survival (OS) have limited its application, making systemic therapy only used for advanced HCC. Before 2017, the anti-angiogenic tyrosine kinase inhibitor sorafenib was almost the only option for systemic treatment for advanced HCC patients. Subsequently, several molecularly targeted therapeutic agents, including lenvatinib, regorafenib, and ramucirumab, have broadened the treatment options for advanced HCC. In recent years, the important role of immune system regulation in HCC has made immunotherapy the focus of HCC research efforts.

Immune checkpoint inhibitors (ICIs) are monoclonal antibodies that block the interaction of checkpoint proteins with their ligands, thereby preventing T cell inactivation. The antitumor effects of immunotherapy drugs are based on immune checkpoint-mediated inhibition of programmed cell death-1 (PD-1), programmed cell death ligand 1 (PD-L1), and cytotoxic T lymphocyte-associated protein 4 (CTLA-4). Previous studies have shown that immune checkpoint inhibitors, including anti-PD-1, anti-PD-L1, and anti-CTLA-4 antibodies, have shown potential therapeutic promise for advanced HCC (Zongyi and Xiaowu, 2020). The combination of the anti-PDL1 antibody atezolizumab and the vascular endothelial growth factor-neutralizing antibody avastin is about to become the standard treatment for HCC. Compared with sorafenib, the immunotherapy combination regimen based on atezolizumab and avastin showed a clear advantage in improving the survival rate of patients with unresectable HCC. In addition, the anti-PD1 drugs nivolumab and pembrolizumab began to be used after the use of anti-angiogenic tyrosine kinase inhibitors. Currently, the combination of HCC checkpoint immunotherapy with other systemic or local treatments is considered the most promising treatment option for HCC. And immunotherapy is expected to be integrated into early and mid-stage treatment regimens.

However, on the one hand, the severe toxicity of systemic drugs has slowed the development of new HCC drugs over the past decade (Busato et al., 2019). On the other hand, the predictive power and accuracy of traditional pathological staging have been shown to be insufficient due to the marked heterogeneity of HCC. The lack of predictive biomarkers makes the choice of immunotherapy over kinase inhibitors an empirical treatment decision that balances antitumor efficacy and drug toxicity (Fulgenzi et al., 2021). The identification and validation of predictive biomarkers and the screening of more effective immunotherapeutic drugs or drug combinations are urgently needed for HCC immunotherapy (Sangro et al., 2021).

As we know, HCC cells are characterized by fast growth and strong invasiveness. Therefore, proliferation-related gene signatures are potential prognostic biomarkers for HCC. Previous researches suggest that DEPDC1 can promote the occurrence and proliferation of HCC (Qu et al., 2019). High expression of E2F1 can promote cancer cell proliferation by activating PKC-α phosphorylation in HCC (Lin et al., 2019). YTHDF2 can inhibit the proliferation of cancer cells by destroying the stability of EGFR mRNA in HCC (Zhong L. et al., 2019). In addition, in terms of microRNA, miR-424-5p can inhibit the proliferation and invasion of HCC cells by targeting TRIM29 (Du et al., 2019). MiR-125a-5p can inhibit the growth and metastasis of liver cancer cells by targeting TRIAP1 and BCL2L2 (Ming et al., 2019). MiR-490-5p inhibits the proliferation, migration and invasion of cancer cells by directly regulating ROBO1 in HCC (Chen et al., 2019). MiRNA-217 can inhibit the proliferation of cancer cells by regulating KLF5 in HCC (Gao et al., 2019). MiR-664 may target SIVA1 to promote proliferation, migration and invasion in HCC (Wang X. et al., 2019). In terms of long non-coding RNAs (lncRNAs), LncRNAs A1BG-AS1 can inhibit the proliferation and invasion of HCC cells by targeting miR-216a-5p (Bai et al., 2019). While LncRNA 01123, LncRNA HAGLROS, LncRNA MNX1-AS1, LncRNA CRNDE, and LncRNA RNA CCAT2 can promote the proliferation and metastasis of HCC cells (Ji et al., 2019a; Ji et al., 2019b; Liu et al., 2019; Wei H. et al., 2019; Xiao et al., 2020). Therefore, the above genes have potential value as prognostic biomarkers in HCC.

In this study, we used the CRISPR Library and the Cancer Genome Atlas (TCGA) database to screen for important genes related to the proliferation of HCC cells. Then, hub genes most relevant to the prognosis of HCC patients were identified and used to establish a gene signature for survival prediction. Subsequently, the prognostic values of the gene signature were confirmed both in the training set and validation set. Time-dependent receiver operating characteristic (t-ROC) curve was used to verify the prediction accuracy of the survival model. Associations of risk scores with genetic mutations, cancer cell stemness and immune function were analyzed, respectively.

Finally, drugs targeting this proliferation-related gene signature were identified. In conclusion, this study comprehensively analyzed the prognostic value of a new proliferation-related gene signature in HCC. This gene signature can not only be used for prognostic assessment and risk stratification of HCC patients, but also is expected to be a therapeutic target for HCC. Furthermore, therapeutic drugs targeting this gene signature may have potential therapeutic prospects.

MATERIALS AND METHODS

Data Source and Identification of Proliferation-Related Differentially Expressed Genes

The RNASeq data and clinical information used to construct the prognostic gene signature were downloaded from the TCGA HCC dataset (n = 365). The RNASeq data and clinical information used to verify the gene signature were downloaded from the International Cancer Genome Consortium (ICGC) HCC dataset (n = 232). The limma package was used to perform differentially expressed gene analysis between tumor and matched normal tissues. Candidates with false discovery rate (FDR) <0.05 and multiple of change >1 were considered to be significantly upregulated in tumor tissues. The genome-wide CRISPR screening of HCC cells was downloaded from the DepMap portal (https://depmap.org/portal/download/). The CERES algorithm was used to calculate the dependency scores of candidate genes (Meyers et al., 2017). Candidate genes were defined as proliferation-related genes. The above three databases are public. Therefore, this study did not require the approval of the local ethics committee.

Candidate Gene Selection and Gene Signature Establishment

Random forest is a machine learning algorithm based on decision tree, which is a nonlinear classifier and can be used for sample classification or regression tasks. The method of random forest to evaluate the importance of features is to calculate how much each feature contributes to different decision trees in random forest, then take the average value, and compare the contribution of different features. In this study, using univariate Cox regression with a p value < 0.01, the candidate genes that are most relevant to the prognosis of HCC patients were identified. Next, we used random forest to rank the importance of genes and selected the top 10 hub genes. Subsequently, we identified a gene signature with a smaller number of genes and a more significant p value from multiple combinations of 10 hub genes to construct a survival model. The single-sample gene set enrichment analysis (ssGSEA) algorithm was used to quantify the performance of proliferation-related pathways and transcription factors. In addition, gene mutations, cancer cell stemness and immune function changes can affect tumor proliferation and the prognosis of HCC, so we explored the correlations between the gene signature and gene mutations/mRNSsi/immune functions.

Survival Analysis Based on Risk Score

Taking the median risk score as the cut-off value, we divided HCC patients into high- and low-risk groups. Then the prognosis of the two groups was compared in the training set and the validation set, respectively. Kaplan-Meier method was used for survival analysis. ROC curve was used to evaluate the predictive accuracy of the risk score. And t-ROC was used to evaluate the predictive ability (R package "survival-ROC") (Heagerty et al., 2000). Cox proportional hazard regression model was used to evaluate the importance of each parameter to OS. In addition, a two-factor survival analysis combining risk score and proliferation-related pathways was also performed to evaluate the impact of risk score and proliferation-related pathways on the prognosis of HCC patients.

Establishment and Evaluation of Nomogram for Predicting OS of HCC Patients

Nomogram is an effective tool for predicting the prognosis of cancer patients by simplifying complex statistical prediction models into maps that assess the probability of individual patients' OS (Park, 2018). In this study, we constructed a nomogram based on the five-gene signature to evaluate the probability of OS in HCC patients at 1-, 3-, and 5-year. Meanwhile, the predicted probability of the nomogram was compared with the measured probability by the calibration curve to verify the accuracy of the nomogram. In addition, t-ROC curve was used to evaluate the survival prediction ability of the nomogram. Decision curve analysis (DCA) curve was used to evaluate the clinical benefit of the nomogram.

Drug Discovery Based on Risk Score

In order to find candidate drugs that show potential efficacy in the high-risk group, we used the half-maximum inhibitory concentration (IC50) of each HCC patient to evaluate their treatment response on Genomics of Drug Sensitivity in Cancer (GDSC) (https://www.cancerrxgene.org/) (Geeleher et al., 2014).

Drug Sensitivity Analysis of Five Hub Genes

The drug sensitivity data was downloaded from the CellMiner™ database (version: 2020.3, database: 2.4.2, https://discover.nci.nih.gov/cellminer/home.do) (Reinhold et al., 2012). The R packages "impute," "limma," "ggplot2," and "ggpubr" were used for data processing and visualization.

Bioinformatics and Statistical Analysis

IBM SPSS Statistics 20 (IBM Corp., Armonk, NY, United States) and R software (version 3.5.2, https://www.r-project.org) were used to analyze data and draw graphs. Z-score were used to normalize the ssGSEA score. Principal component analysis was conducted by using the Rtsne R package. The log-rank test was used to assess the differences. The "wilcox.test" function was used to compare the risk scores between groups.

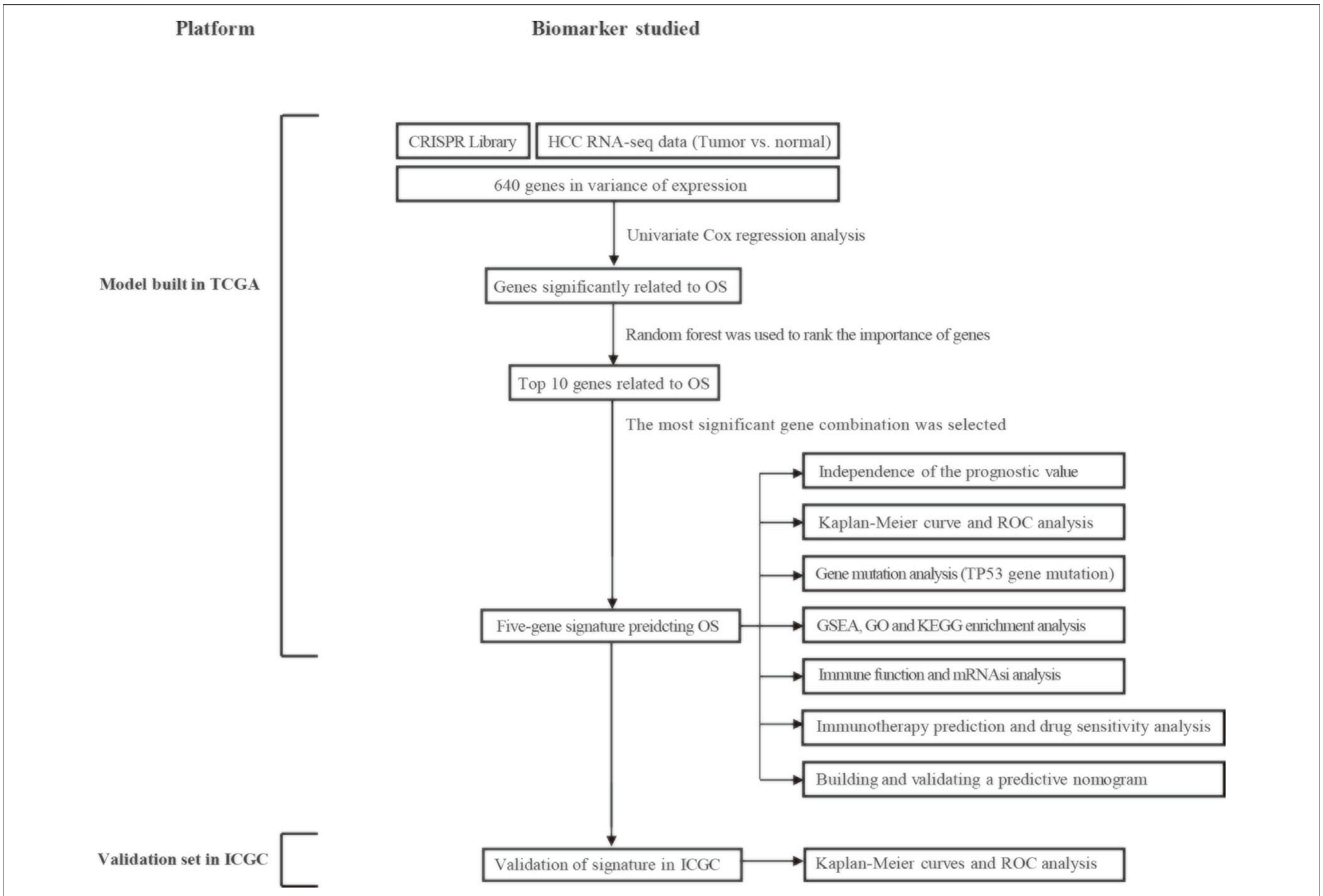

Platform
Biomarker studied
Model built in TCGA
CRISPR Library
HCC RNA-seq data (Tumor vs. normal)
640 genes in variance of expression
Univariate Cox regression analysis
Genes significantly related to OS
Random forest was used to rank the importance of genes
Top 10 genes related to OS
The most significant gene combination was selected
Independence of the prognostic value
Kaplan-Meier curve and ROC analysis
Gene mutation analysis (TP53 gene mutation)
Five-gene signature preidcting OS
GSEA, GO and KEGG enrichment analysis
Immune function and mRNAsi analysis
Immunotherapy prediction and drug sensitivity analysis
Building and validating a predictive nomogram
Validation set in ICGC
Validation of signature in ICGC
Kaplan-Meier curves and ROC analysis

FIGURE 1 | Overall flowchart of this study. HCC, hepatocellular carcinoma; OS, overall survival; ROC, receiver operating characteristic; GSEA, gene set enrichment analysis; GO, gene ontology; KEGG, kyoto encyclopedia of genes and genomes; mRNAsi, mRNA expression-based stemness index.

RESULTS

Schematic Diagram of Research Design

Figure 1 shows the entire workflow of this research. Firstly, using the CRISPR Library and TCGA HCC dataset, differentially expressed genes (DEGs) related to HCC proliferation were screened out. Then, univariate Cox regression analysis was used to screen promising candidates. Next, the random forest algorithm and multiple combinatorial screening methods were used to establish a prognostic gene signature. Specifically, we screened genes associated with overall survival in HCC by univariate COX regression, and then used random forests to rank the importance of these survival-related genes and listed the top 10 genes. We then randomly combined these 10 genes and constructed a risk model by multivariate COX regression. Subsequently, we calculated and ranked the *p*-values for each model by K-M survival analysis. Furthermore, we screened out the risk model with the smallest *p* value and the relatively small number of genes. Finally, the prognostic values of the gene signature were evaluated in the training set and validation set, respectively.

Establishment of Proliferation-Related Prognostic Gene Signature

A total of 640 DEGs in HCC were identified, with |log2FC| > 1 and FDR < 0.05 as the thresholds. The heat map shows the expression profiles of some DEGs related to proliferation in HCC (**Figure 2A**). As shown in **Figures 2B,C**, biological processes significantly enriched by 640 DEGs included ribosomal subunit, U2-type spliceosomal complex, spliceosomal complex, cytosolic part and cytosolic ribosome; Significantly enriched cell components included mRNA splicing, *via* spliceosome, RNA splicing *via* transesterification reactions with bulged adenosine as nucleophile, RNA splicing, viral transcription, and translational initiation; Significantly enriched molecular function included structural constituent of ribosome, catalytic activity acting on RNA, helicase activity, nucleotidyltransferase activity and rRNA binding. In addition, significantly enriched pathways included spliceosome, ribosome, RNA transport, cell cycle and spinocerebellar ataxia. Using $p < 0.01$ as the threshold for univariate Cox regression, candidate genes related to the prognosis of HCC patients were identified (**Figure 2D**). Subsequently, we used random forest ranking to rank candidate

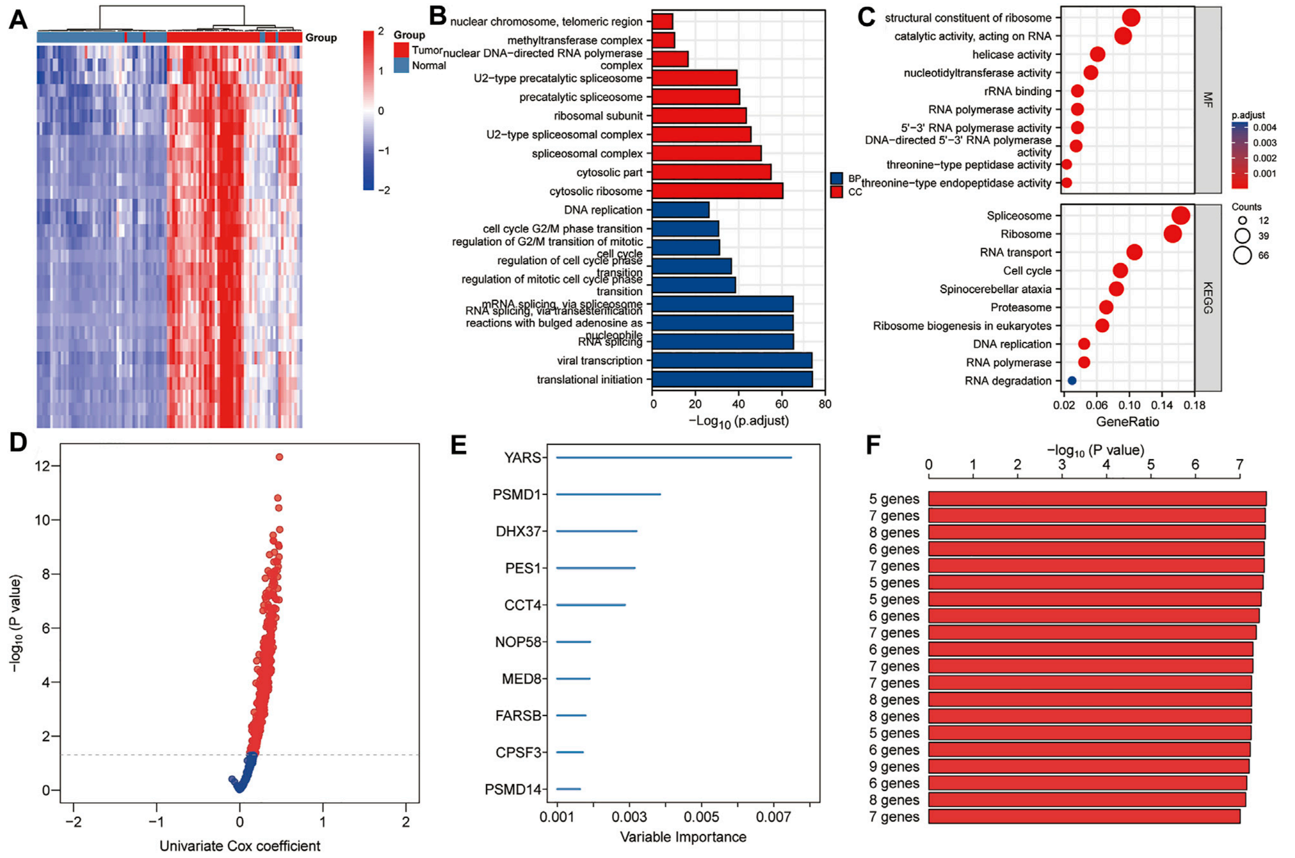


FIGURE 2 | Establishment of prognostic gene markers related to cancer cell proliferation in HCC. **(A)** Heat map showing significantly DEGs in HCC related to cancer cell proliferation. Using the RNA sequencing data of the TCGA HCC cohort and the CRISPR Library, 640 DEGs related to proliferation in HCC were screened out (|log2FC| > 1 and FDR < 0.05). **(B,C)** GO and KEGG analysis revealed important biological processes, cell components and KEGG pathways enriched by 640 DEGs. **(D)** Using univariate Cox analysis, candidates related to prognosis were identified (p < 0.05). **(E)** Using random forest, the top 10 most characteristic genes are screened out. **(F)** A combination with a relatively small number of genes and a relatively significant p value was selected to construct a survival prediction model from a variety of combinations of 10 genes. DEGs, differentially expressed genes; FDR, false discovery rate; BP, biological processes; CC, cell components; MF, molecular function.

genes and screened out the top ten relatively important genes (**Figure 2E**). Next, we selected a gene combination with a smaller number of genes and a more significant p value from multiple combinations of ten hub genes to construct a survival prediction model (**Figure 2F**). Finally, five hub genes were used to construct a prognostic model of HCC: risk score = 0.010 * FARSB + 0.07 * NOP58 + 0.001 * CCT4 − 0.026 * DHX37 + 0.022 * YARS.

Risk Score Based on the Five-Gene Signature Was an Independent Prognostic Factor for HCC

The TCGA HCC dataset was used as a training set to evaluate the prognostic values of this five-gene signature. As shown in **Figure 3A**, Kaplan-Meier analysis showed that the prognosis of the high-risk score group was worse than that of the low-risk score group (p < 0.001). The high- and low-risk score groups were defined by risk scores based on the five-gene signature. The median risk score calculated from the risk model was 0.867. Taking the median risk score of HCC patients as a cutoff value, we divided HCC patients into high-risk and low-risk groups. Patients with a risk score higher than 0.867 were classified as high-risk group, while those with a risk score lower than 0.867 were classified as low-risk group. Subsequently, in order to evaluate the relationship between the five-gene signature and the prognosis of HCC patients, we took the median of the risk scores of 338 HCC patients from the training set as the cut-off value, divided these patients into high- and low-risk groups, and compared the survival status and the expressions of the five hub genes between the two groups. The results showed that the prognosis of the high-risk group was worse than that of the low-risk group, and the expression levels of five hub genes in the high-risk group were higher than that of the low-risk group (**Figure 3B**). Next, Principal component analysis suggested that risk score could be used as a new dimension to assess the prognosis of HCC patients (**Figure 3C**). The ROC curve showed that the AUCs of the risk score for predicting 1-year, 3-year, and 5-year survival rates were 0.744, 0.699, and 0.743, respectively, indicating that the risk score was a good model for predicting the survival rate of HCC patients (**Figure 3D**). Univariate and multivariate Cox regression analysis showed that risk score based on five-gene signature (HR = 2.48, p <

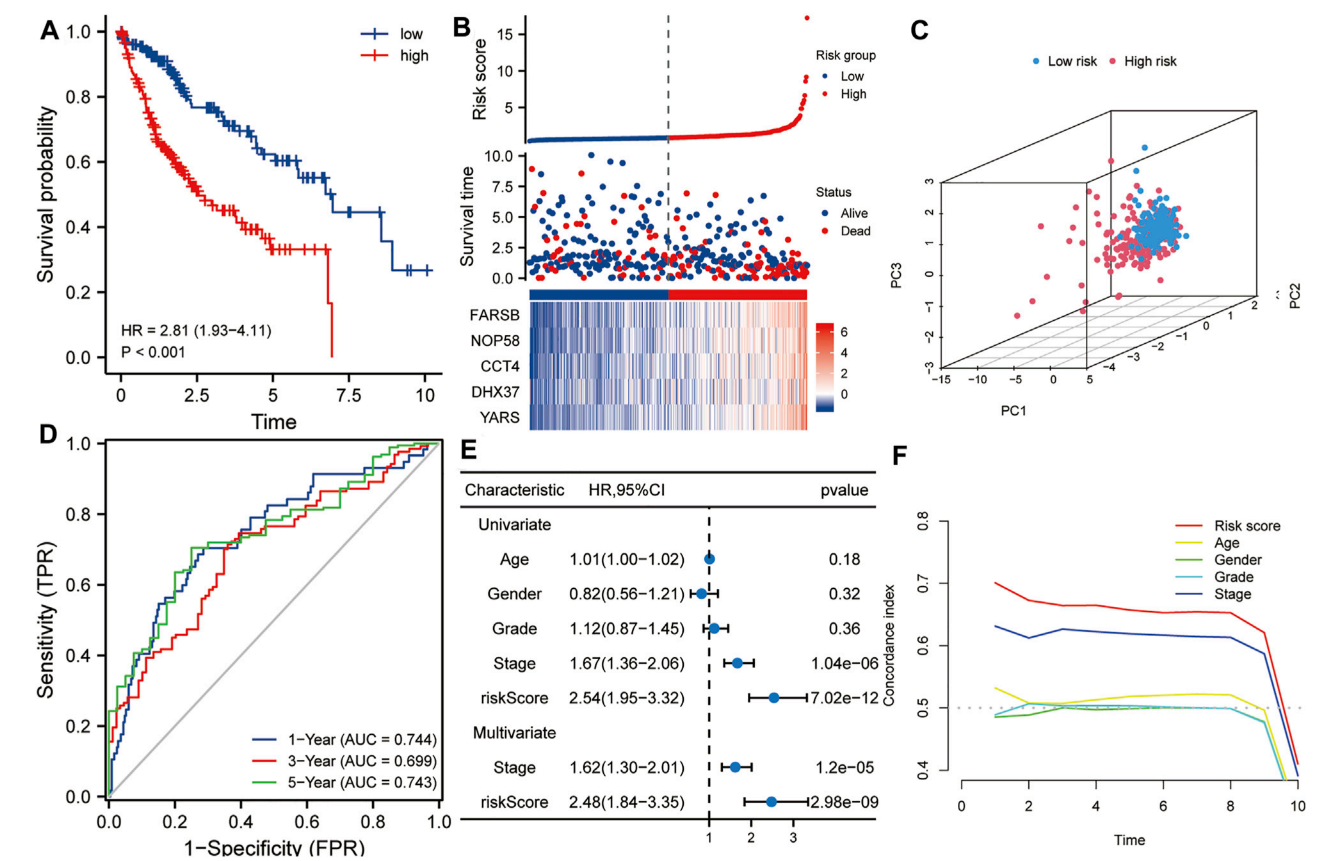


FIGURE 3 | The risk score predicts poor survival in the training set. **(A)** Kaplan-Meier analysis showed that HCC patients with higher risk scores had a worse overall survival rate. **(B)** The risk score distribution, survival profile and heat map of patients in the high- and low-risk groups in the training set. **(C)** Principal component analysis suggested that risk score could be used as a new dimension to evaluate the prognosis of HCC patients. **(D)** The ROC curve showed the prediction efficiency of the risk score in the training set (AUC > 0.699). **(E)** Univariate and multivariate Cox regression analysis showed that risk score was an independent risk factor for OS in HCC patients. **(F)** The tROC analysis showed that the predictive power of risk score was significantly higher than that of other clinical characters. HR, hazard ratio; OS, overall survival; tROC, time-dependent receiver operating characteristics.

0.001) and pathological stage (HR = 1.62, p < 0.001) were independent risk factors affecting OS in HCC patients (**Figure 3E**). Besides, tROC analysis showed that the survival predictive ability of risk score was significantly higher than other clinicopathological characteristics (**Figure 3F**).

Verifying the Prognostic Values of the Five-Gene Signature in the Validation Set

The ICGC HCC dataset was used as a validation set to verify the robustness of this five-gene signature. Kaplan-Meier analysis showed that the prognosis of the high-risk group was worse than that of the low-risk group (p < 0.001, **Figure 4A**). Similarly, taking the median of the risk scores of 232 HCC patients from the validation set as the cutoff value, we divided these patients into high- and low-risk groups, and compared the survival status and the expression levels of five hub genes between the two groups. The results showed that the prognosis of the high-risk group was worse than that of the low-risk group, and the expression levels of the five hub genes in the high-risk group were higher than that of the low-risk group (**Figure 4B**). Principal component analysis also suggested that risk score could be used as a new dimension to assess the prognosis of HCC (**Figure 4C**). The ROC curve showed that the AUCs of the risk score for predicting 1-year, 3-year, and 5-year survival rates were 0.747, 0765, and 0.852, respectively, which further indicated that the risk score was a good model for predicting the survival rate of HCC patients (**Figure 4D**). Univariate and multivariate Cox regression analysis showed that risk score (HR = 2.29, p < 0.001) and pathological stage (HR = 1.57, p < 0.05) were independent risk factors affecting OS in HCC patients (**Figure 4E**). In addition, tROC analysis showed that the survival predictive ability of risk score was significantly higher than that of other clinicopathological characteristics (**Figure 4F**).

Correlations Between Risk Score and Proliferation-Related Pathways and Corresponding Two-Factor Survival Analysis

Using the ssGSEA algorithm, the Z-scores of some proliferation-related pathways and some proliferation-related transcription

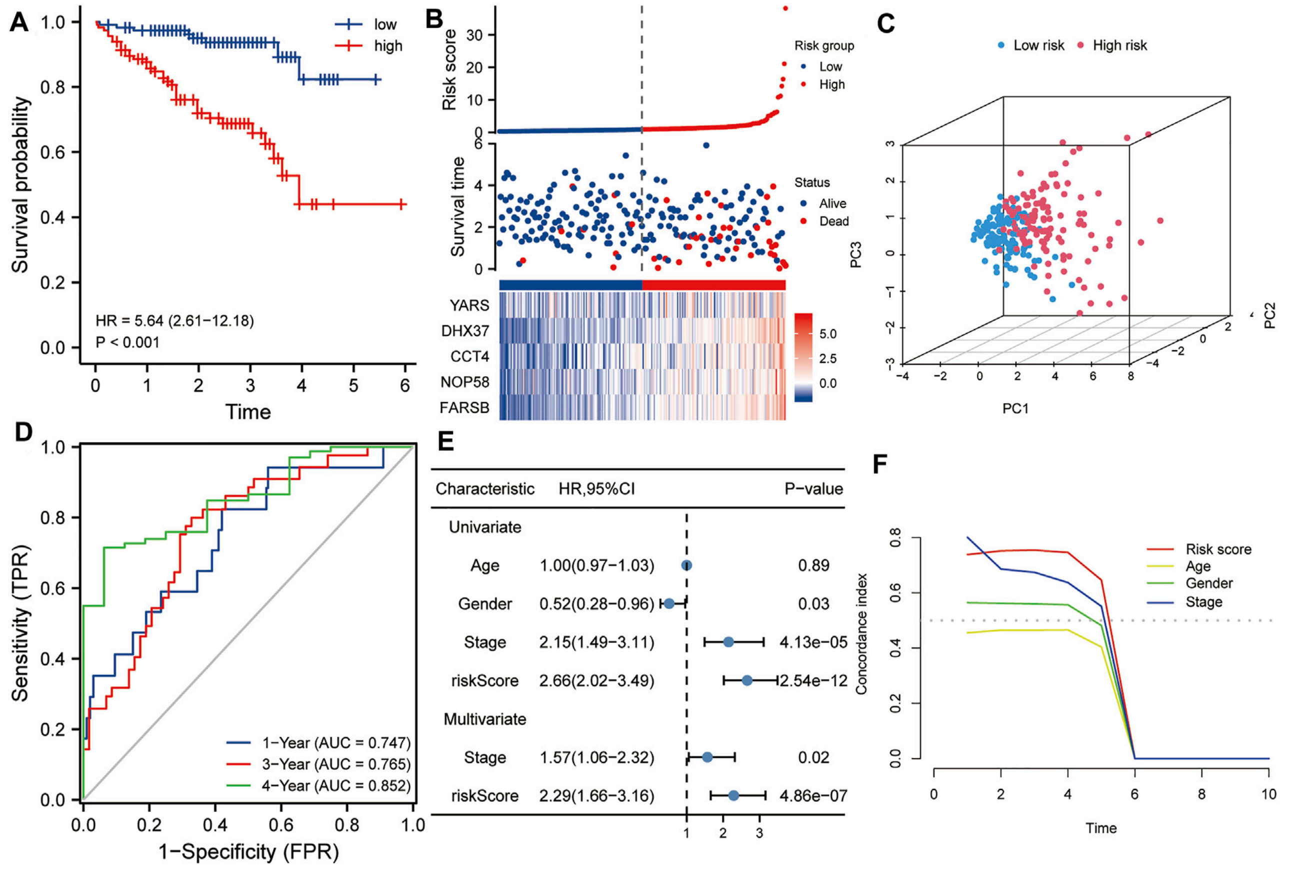


FIGURE 4 | Validation of the risk score in the ICGC dataset. **(A)** Kaplan-Meier analysis showed that HCC patients with higher risk scores had a worse overall survival rate. **(B)** The risk score distribution, survival profile and heat map of patients in the high- and low-risk groups in the ICGC dataset. **(C)** Principal component analysis suggested that the risk score in the ICGC data set was a new dimension for evaluating the prognosis of patients. **(D)** The ROC curve shows the survival prediction efficiency of the risk score in the ICGC data set (AUC > 0.74). **(E)** Univariate and multivariate Cox regression analysis showed that risk score is an independent risk factor for OS in HCC patients. **(F)** tROC analysis showed that the predictive power of risk score was significantly higher than other clinical characters. HR, hazard ratio; OS, overall survival; tROC, time-dependent receiver operating characteristics.

factors were calculated. Subsequently, the Z-scores of proliferation-related pathways and the Z-scores of proliferation-related transcription factors between the high and low-risk groups were compared, respectively. As shown in **Figure 5A**, the Z-scores of the proliferation-related pathways in the high-risk group were higher than those in the low-risk group. Meanwhile, as shown in **Figure 5B**, the Z-scores of the proliferation-related transcription factor of the high-risk group were higher than those of the low-risk group. Subsequently, a two-factor survival analysis combining risk score and proliferation-related pathway Z-scores showed that high risk score and high proliferation-related pathway Z-scores predicted the worst prognosis (**Figures 5C–G**).

Differences in Gene Mutations Between High- and Low-Risk Groups

The gene mutation data of HCC patients in TCGA was downloaded to compare the gene mutation status between the high- and the low-risk groups. The results showed that there were some differences in gene mutation frequency between the two groups. TP53 gene mutation status between the two groups was significantly different (**Figures 6A,B**). The risk score of the TP53 mutant group was higher than that of the TP53 wild group ($p < 0.001$, **Figure 6C**). TP53 mutation rate of the high-risk group was higher than that of the low-risk group ($p < 0.001$, **Figure 6D**). In addition, the mRNAsi of the high-risk group was higher than that of the low-risk group ($p < 0.001$, **Figure 6E**). There were statistically significant differences between the high- and low-risk groups in immune function of Type IL IFN Reponse, MHC class I and Cytolytic activity ($p < 0.01$, **Figure 6F**).

Correlations Between Risk Score and Tumor Progression in HCC Patients

To explore the correlations between the risk score and tumor progression, the mortality and pathological stage of the high- and low-risk groups were compared. The results suggested that in the TCGA dataset, the mortality of the high-risk group was higher than that of the low-risk group ($p = 0.001$, **Figure 7A**).

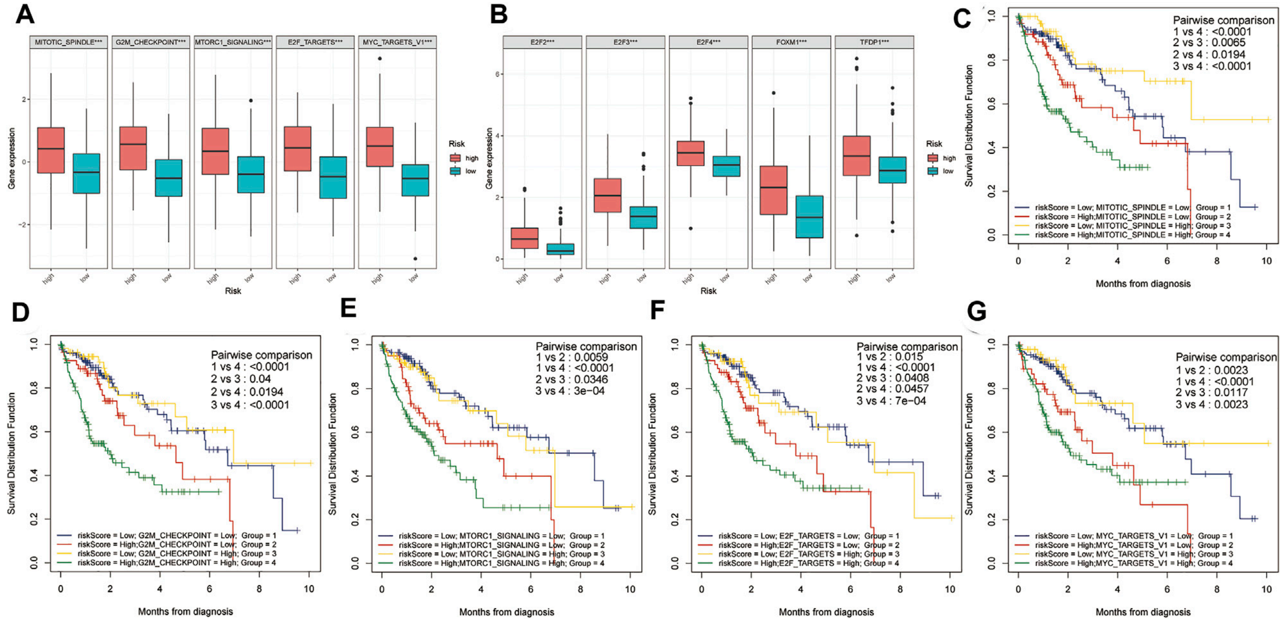


FIGURE 5 | Two-factor survival analysis combining proliferation-related pathways and risk scores. **(A)** The proliferation-related pathways Z-scores in the high-risk group were significantly higher than those in the low-risk group. **(B)** The proliferation-related transcription factor Z-scores of high-risk patients were significantly higher than those of low-risk patients. **(C–G)** Two-factor survival analysis combining risk score and proliferation-related pathway Z-scores showed that high-risk score and high proliferation-related pathway Z-Scores predicted the worst prognosis.

Meanwhile, the proportions of patients with advanced pathological stages (or pathological grades) in the high-risk group were higher than those of the low-risk group (p = 0.001, **Figures 7B–D**). In the ICGC dataset, the mortality rate of the high-risk group was also higher than that of the low-risk group (p = 0.001, **Figure 7E**). At the same time, the proportions of patients with advanced pathological stages (or pathological grades) in the high-risk group were also higher than those of the low-risk group (p = 0.002, **Figure 7F**).

Risk Score Was an Indicator of Poor Prognosis in the Subgroups Divided by Various Clinicopathological Characteristics

Clinicopathological characteristics including age, gender, grade, and pathological stage were used to divide multiple subgroups. As shown in **Figures 8A–H**, risk scores based on five-gene markers can distinguish high-risk patients with poor prognosis in these subgroups ($p < 0.001$).

Enrichment Analysis Based on the Risk Score

Taking the median of the risk scores of all HCC patients from the TCGA dataset as the cut-off value, we divided these samples into high- and low-risk groups. GSEA analysis was conducted to identify the significant enrichment pathways of the high and low risk groups, respectively. Significantly enriched pathways in the high-risk group included cell cycle, cytokine-cytokine receptor interaction, DNA replication, ECM receptor interaction, and hematopoietic cell lineage (**Figure 9A**). And significantly enriched pathways in the low-risk group included drug metabolism cytochrome p450, fatty acid metabolism, glycine serine and threonine metabolism, metabolism of xenobiotics by cytochrome p450 and peroxisome (**Figure 9B**). Subsequently, we performed GO and KEGG analysis on DEGs between the high and low risk groups. The results suggested that significantly enriched BP included chromosome segregation, organelle fission, mitotic sister chromatid segregation, nuclear division and mitotic nuclear division; Significantly enriched CC included chromosomal region, chromosome, centromeric region, spindle, condensed chromosome, centromeric region, and condensed chromosome (**Figure 9C**); Significantly enriched MF includes oxidoreductase activity, acting on CH or CH2 groups, DNA replication origin binding, steroid hydroxylase activity, arachidonic acid monooxygenase activity, and arachidonic acid epoxygenase activity. Besides, significantly enriched KEGG pathways included metabolism of xenobiotics by cytochrome P450, ECM-receptor interaction, central carbon metabolism in cancer, retinol metabolism, and cell cycle (**Figure 9D**). These results suggested that the five-gene signature may play an important role in tumorigenesis and development.

Prediction of Immunotherapy Based on Risk Score

In order to select appropriate checkpoint inhibitors for HCC patients, we performed immunotherapy predictions based on risk scores. The results showed that the high-risk group had lower IC50s for six immunotherapy drugs including A-443654, ABT-

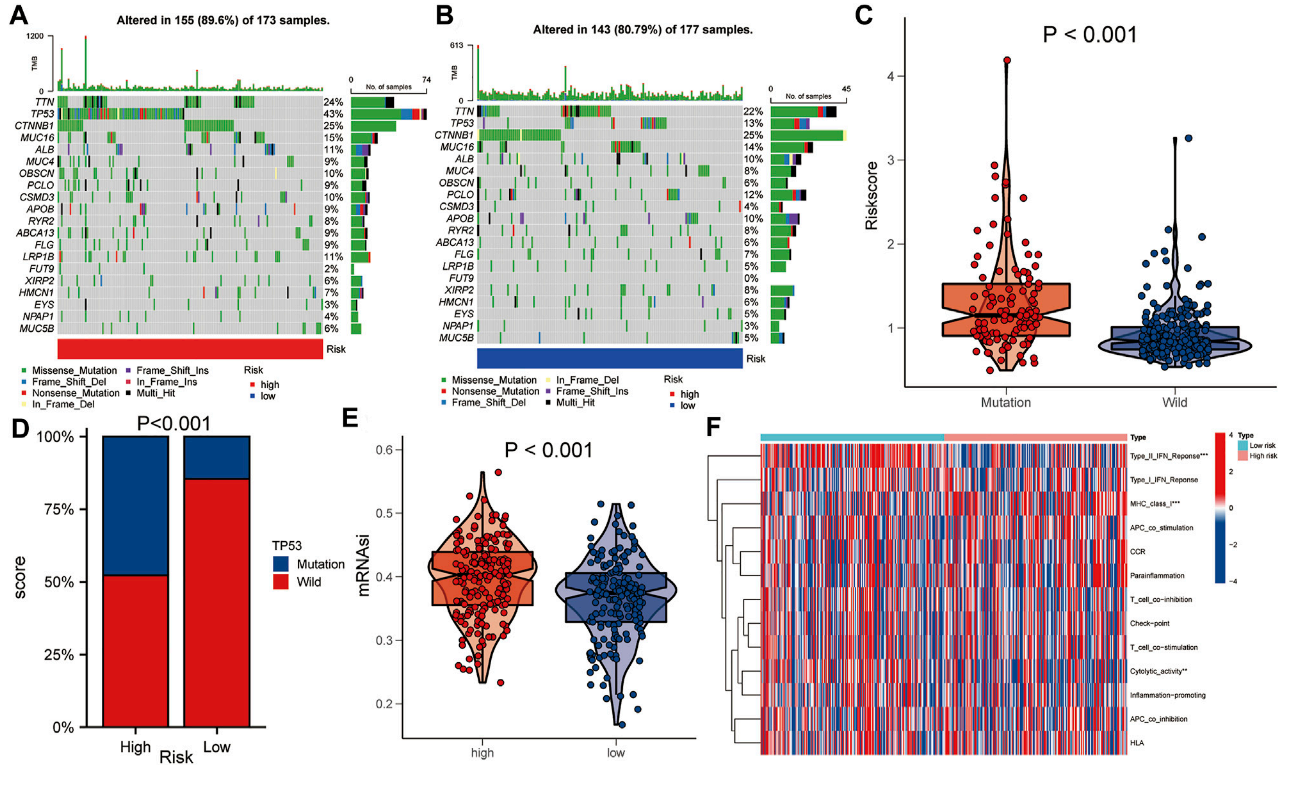


FIGURE 6 | Comparison of gene mutations between the high- and low-risk groups. **(A,B)** Comparison of gene mutation rate between high- and low-risk group. **(C)** Comparison of risk score between TP53 mutation group and TP53 wild group. **(D)** Comparison of TP53 mutation ratio between high- and low-risk group. **(E)** Comparison of mRNAsi between high- and low-risk group. **(F)** Heat map showed the difference in immune function between the high- and low-risk group. Asterisks indicate statistical significance at: *: $p < 0.05$; **: $p < 0.01$, and ***: $p < 0.001$.

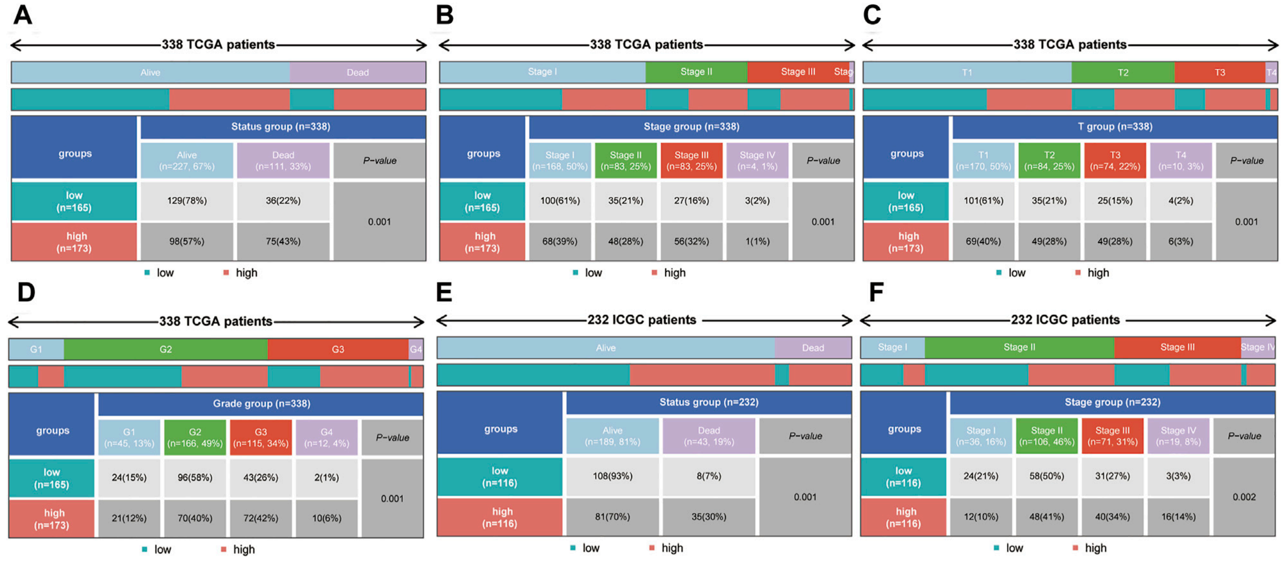


FIGURE 7 | The correlation between the risk score and mortality (or pathological stage) in the training set and validation set. **(A–D)** The high-risk group had higher mortality, more advanced pathological stage, a higher T stage and a higher tumor grade in the training set. **(E,F)** The high-risk group had higher mortality and more advanced pathological stage in the validation set.

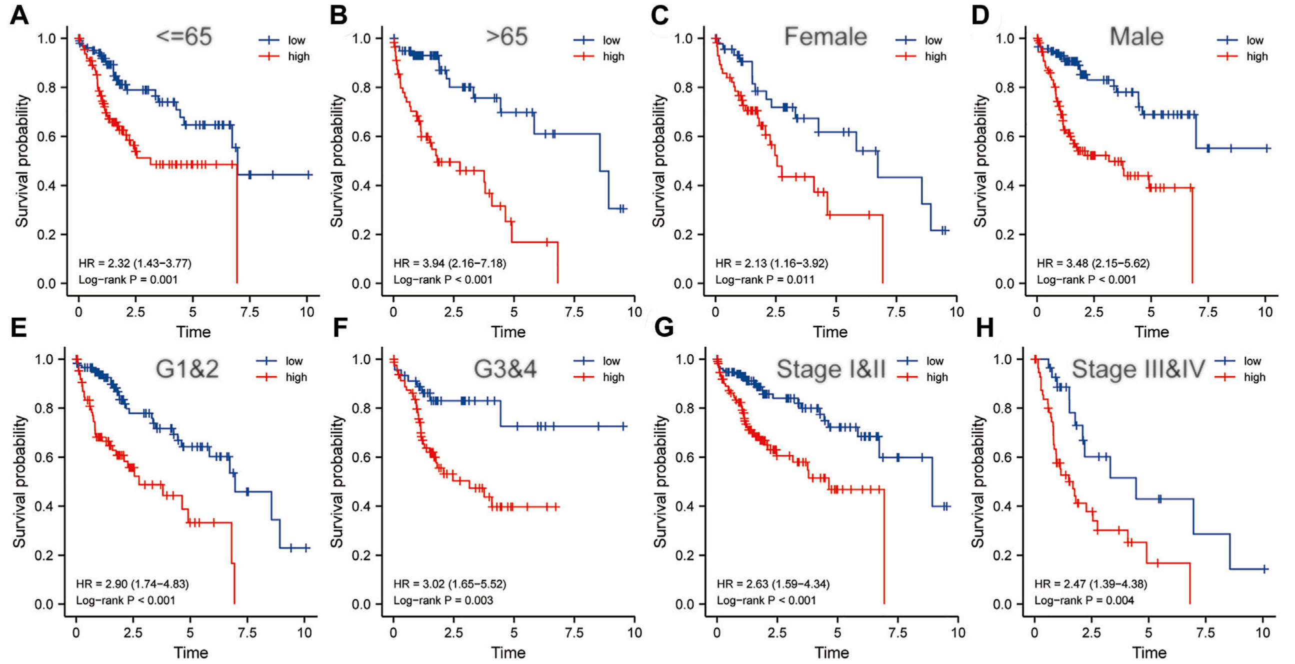


FIGURE 8 | The risk score was an important indicator of poor prognosis in each subgroup divided by clinicopathological characteristics. In various subgroups divided by clinicopathological characteristics including age **(A,B)**, gender **(C,D)**, grade **(E,F)** and pathological stage **(G,H)**, the prognosis of the high-risk groups was poorer than that of the low-risk groups. HR, risk ratio.

888, AG-014699, ATRA, AUY-922, and AZ-628, while had higher IC50s for six kinds of immunotherapy drugs including AMG-706, A-770041, AICAR, AKT inhibitor VIII, Axitinib, and AZD-0530 (**Figure 10**).

Drug Sensitivity Analysis of Five Hub Genes

To explore the potential correlations between the expressions of five key genes and drug sensitivity, we conducted drug sensitivity analysis using the CellMiner™ database. The results showed that FARSB expression was positively correlated with the drug sensitivity of Hydroxyurea (**Supplementary Figure S1A**), Vorinostat (**Supplementary Figure S1E**), Nelarabine (**Supplementary Figure S1G**), and Lomustine (**Supplementary Figure S1P**), while negatively correlated with the drug sensitivity of JNJ-42756493 (**Supplementary Figure S1O**). DHX37 expression was positively correlated with the drug sensitivity of Raltitrexed (**Supplementary Figure S1B**), Cytarabine (**Supplementary Figure S1D**), Cisplatin (**Supplementary Figure S1F**), Thiotepa (**Supplementary Figure S1H**), and Triethylenemelamine (**Supplementary Figure S1N**). YARS expression was positively correlated with the drug sensitivity of Axitinib (**Supplementary Figure S1C**), Fluphenazine (**Supplementary Figure S1K**), and Megestrol acetate (**Supplementary Figure S1M**). NOP58 expression was positively correlated with the drug sensitivity of Vorinostat (**Supplementary Figure S1I**) and 6-Thioguanine (**Supplementary Figure S1J**). The expression of CCT4 was positively correlated with the drug sensitivity of Nelarabine (**Supplementary Figure S1L**).

Constructing a Nomogram to Predict OS in HCC Patients

In order to establish a clinically applicable method for predicting the OS of HCC patients, we constructed a nomogram combining risk score and pathological stage (**Figure 11A**), and then analyzed the accuracy of the model using a calibration curve. The results showed that the 1-year, 3-year, and 5-year survival probabilities predicted by the nomogram were basically consistent with the observed survival probabilities, confirming the reliability of the nomogram (**Figure 11B**). Meanwhile, t-ROC curve suggested that the nomogram combined with pathological stage and risk score had the largest AUC. The AUCs of 1-, 3-, and 5-year survival predictions were above 0.72, which suggested that compared with the model constructed by a single prognostic factor, the nomogram combining risk scores and pathological stages was a better prognostic model for survival prediction in HCC patients (**Figure 11C**). In addition, we plotted the calculated net benefit with the threshold probabilities for HCC patients with 1-year, 3-year, and 5-year survival rates. As shown in **Figure 11D**, the net benefit of the nomogram was better than other models.

DISCUSSION

In recent years, immunotherapy has become the focus of HCC research. Immune checkpoint inhibitors, including anti-PD-1, anti-PD-L1, and anti-CTLA-4 antibodies, have shown potential therapeutic value in advanced HCC. At present, the anti-PDL1

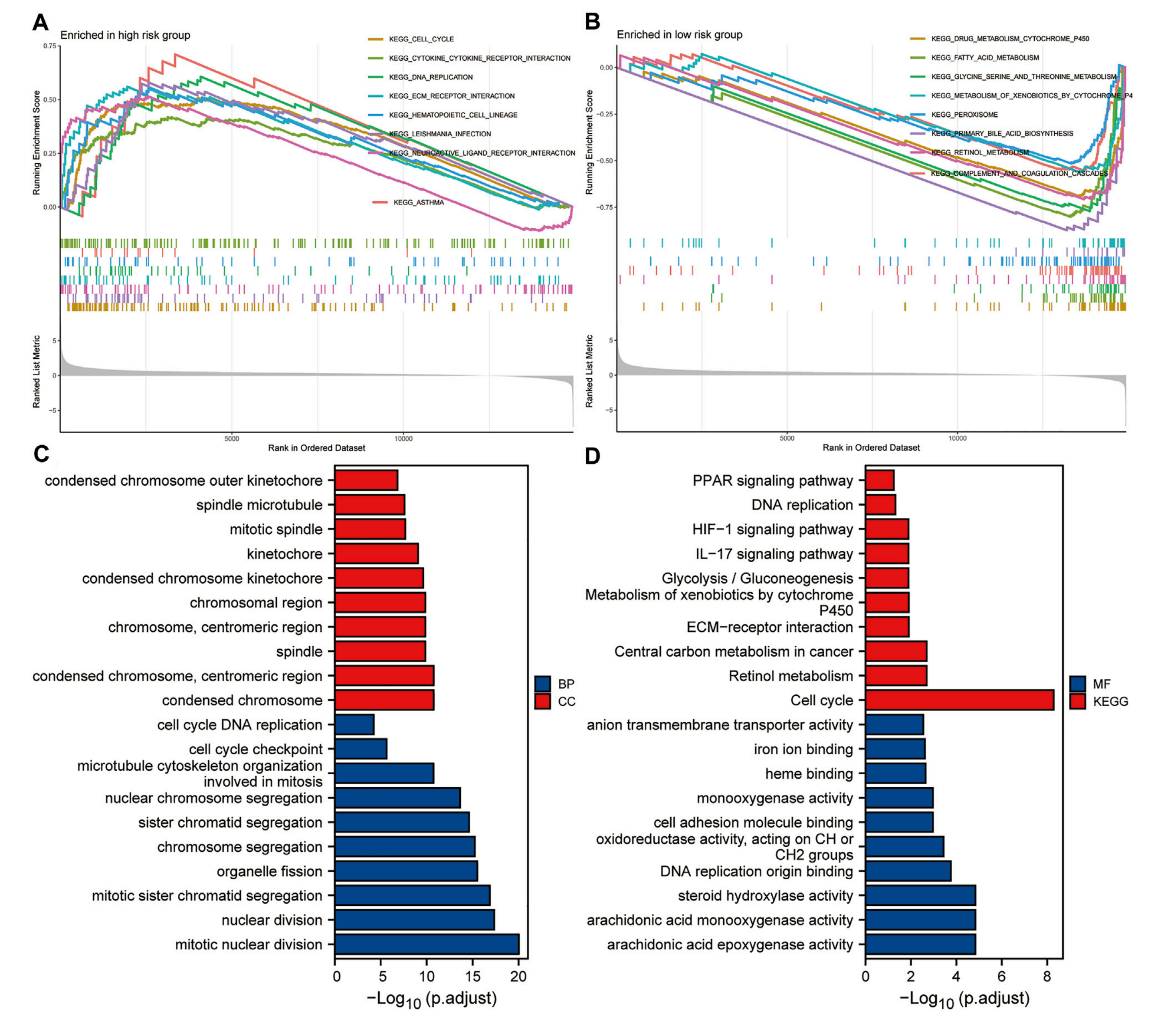


FIGURE 9 | Enrichment analysis based on the risk score. **(A,B)** GSEA analysis of high- and low-risk groups. **(C,D)** GO and KEGG analysis of DEGs between high- and low-risk groups.

antibody Atezolizumab combined with the vascular endothelial growth factor neutralizing antibody Avastin is expected to become the standard treatment for HCC. Therefore, HCC checkpoint immunotherapy combined with other systemic or local treatments is considered to be the most promising treatment option for HCC. Currently, there is an urgent need for the identification and validation of predictive biomarkers and the screening of more effective immunotherapy drugs for HCC immunotherapy.

In this study, we focused on constructing a proliferation-related gene signature for patients with HCC. Firstly, the CRISPR Library and the TCGA database were used to screen differentially expressed genes related to the proliferation of HCC cells. Then, univariate COX regression analysis, random forest algorithm and multiple combinations were used to construct a prognostic five-gene signature (FARSB, NOP58, CCT4, DHX37, and YARS). Next, the prognostic value of the five-gene signature was confirmed in both the training set and the validation set. Finally, we combined risk scores and pathological stage to construct a nomogram for clinical practice. Meanwhile, calibration curve, ROC curve and decision curve showed that the nomogram can more accurately predict the ability of OS in HCC patients. In addition, the roles of this five-gene signature in gene mutation, cancer cell stemness and immune functions were explored, respectively. Therefore, this five-gene signature is an independent prognostic predictor of HCC.

Traditional pathological staging is commonly used method for evaluating the prognosis of HCC patients. Alpha-fetoprotein

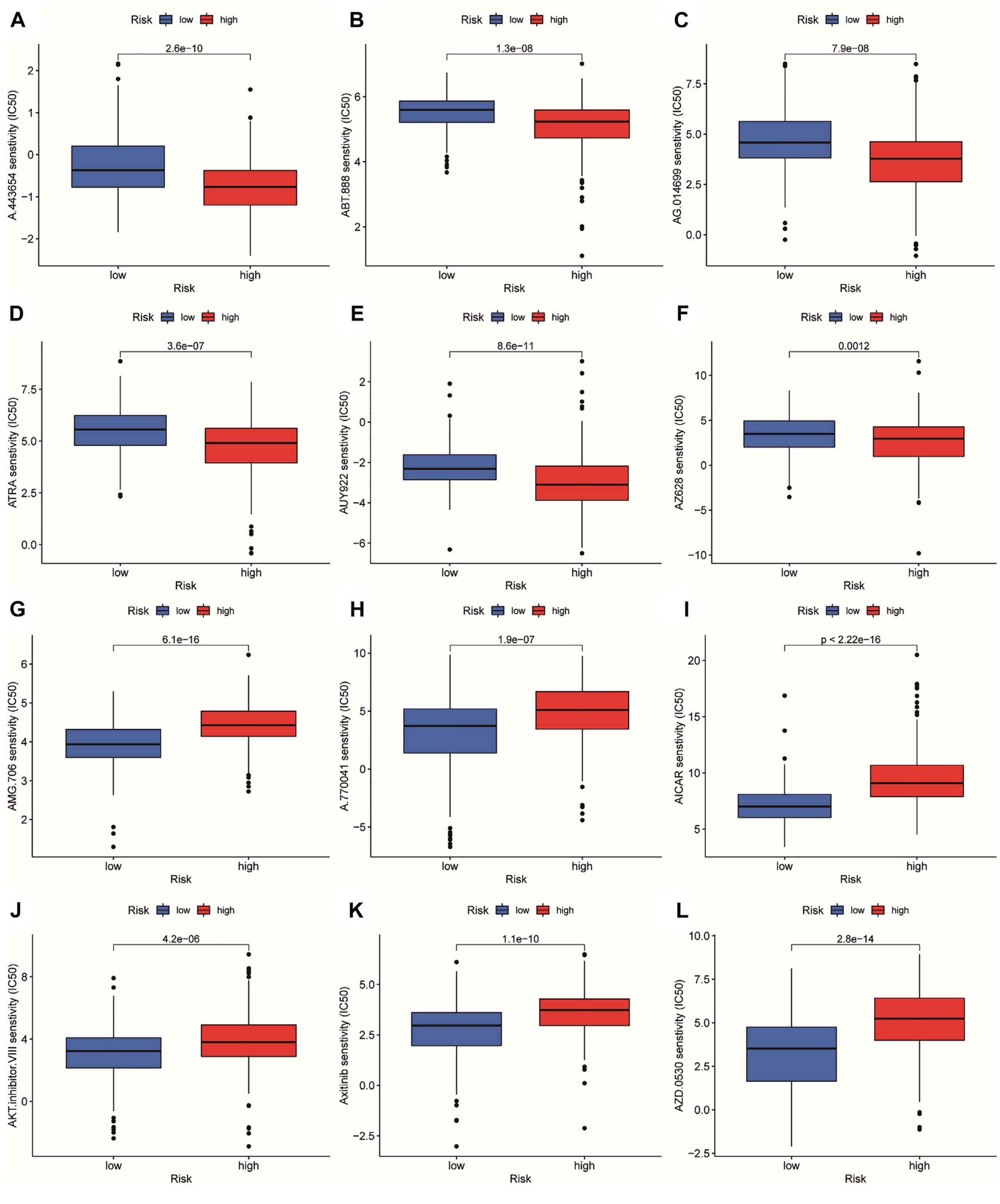


FIGURE 10 | The investigation of tumor immunotherapy. The drugs prediction of risk groups.

(AFP) is a widely used biomarker to monitor treatment response and improve prognosis. However, the high heterogeneity of HCC increases the difficulty of survival prediction. Recently, some new biomarkers have become effective tools for predicting the prognosis of HCC. For example, CCL14, CBX3/HP1, APEX1, and UBE2C are considered to be prognostic biomarkers for HCC (Wei Z. et al., 2019; Zhong X. et al., 2019; Cao et al., 2020; Gu et al., 2020). In addition, a four-gene signature including PBK,

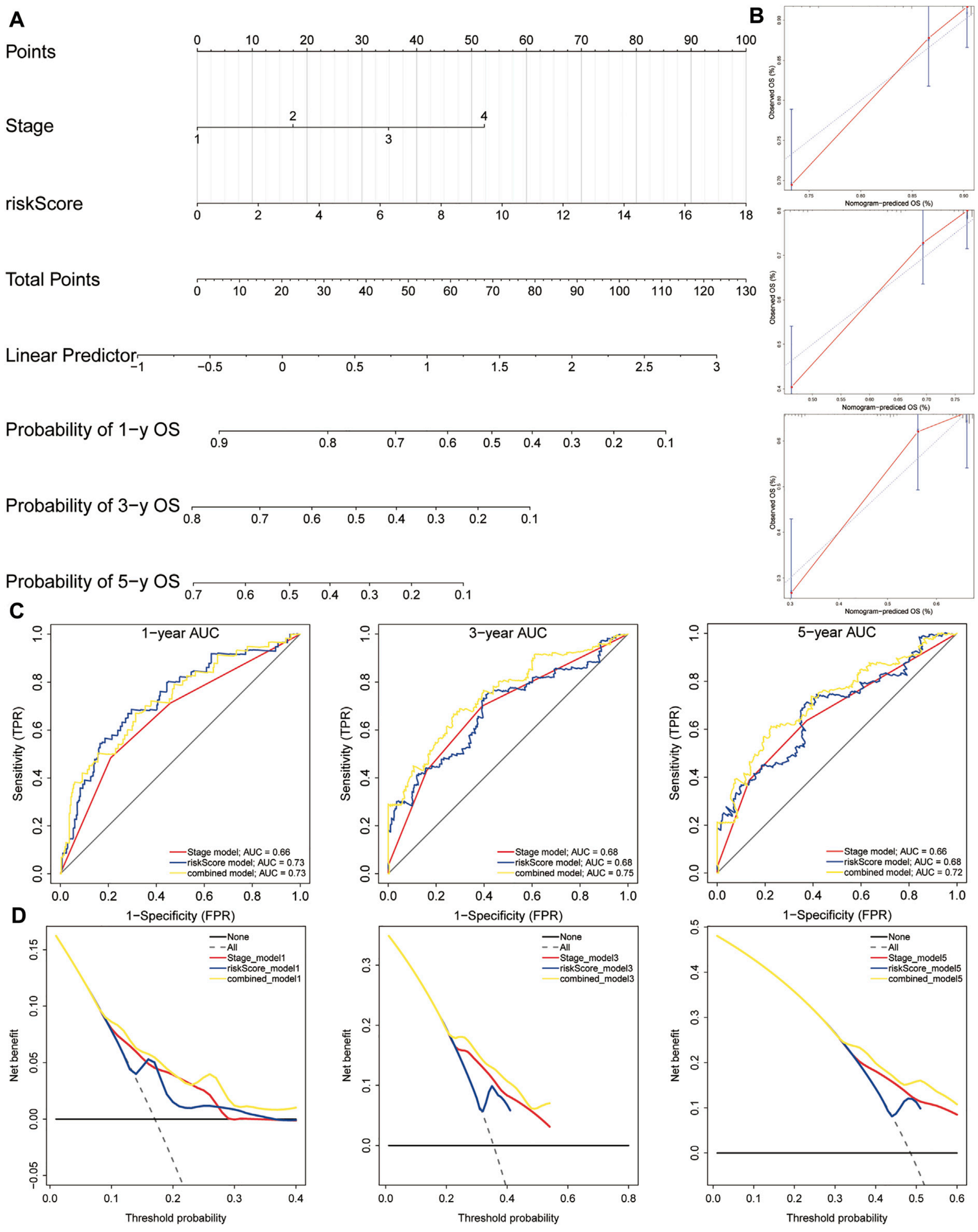


FIGURE 11 | A nomogram used for survival prediction. **(A)** A nomogram combining the five-gene signature and clinical pathological stage. **(B)** The calibration chart showed that the predicted 1-, 3-, and 5-year survival probabilities were basically consistent with actual observations. **(C)** The t-ROC analysis showed that the nomogram had good survival prediction power. **(D)** DCA curve visually evaluated the clinical benefit of the nomogram and the scope of application of the clinical benefit obtained by the model. The calculated net benefit (*Y* axis) was plotted against the threshold probabilities of patients with 1-, 3-, and 5-year survival on the *X* axis. The gray dotted line represents the hypothesis that all patients have 1-year, 3-year, and 5-year survival. The solid black line represents the hypothesis that no patient has a 1-year, 3-year, or 5-year survival period. t-ROC, time-dependent receiver operating characteristics; DCA, decision curve analysis.

CBX2, CLSPN, and CPEB3, a four-methylated mRNA signature including BRCA1, CAD, CDC20, and RBM8A, a 5-gene lncRNA signature including RP11-325L7.2, DKFZP434L187, RP11-100L22.4, DLX2-AS1, and RP11-104L21.3, as well as many other polygenic gene signatures have been shown to have prognostic value in HCC (Sun et al., 2019; Wang Y. et al., 2019; Yan et al., 2019).

Based on the important role of HCC cell proliferation in tumor progression and its impact on patient prognosis, this work constructed a prognostic gene signature associated with HCC cell proliferation. The results showed that the five-gene signature including FARSB, NOP58, CCT4, DHX37, and YARS was with good prognostic values. Meanwhile, the enrichment analysis showed that the significant enrichment pathways in the high-risk group included cell cycle, cytokine-cytokine receptor interaction, DNA replication, ECM receptor interaction and hematopoietic cell lineage. These results indicated that the five hub genes were involved in the molecular mechanism of proliferation and progression in HCC. Previous studies have shown that FARSB is involved in amino acid metabolism and tRNA aminoacylation, and plays a key role in the progression of gastric cancer (Gao et al., 2021). NOP58 is involved in the transport of mature mRNA and protein metabolism that do not depend on SLBP. NOP58 is not only negatively related to the OS of HCC patients, but may also be closely related to the recurrence of lung adenocarcinoma (Shen et al., 2021; Wang et al., 2021). CCT4 is involved in protein metabolism and is significantly related to HCC cell growth and prognosis (Li F. et al., 2021; Li W. et al., 2021). In addition, downregulation of CCT4 can significantly inhibit the migration of lung adenocarcinoma cells (Tano et al., 2010). DHX37 is an RNA helicase, which is significantly upregulated in 17 kinds of tumors (Huang et al., 2021). DHX37 could affect the prognosis of patients with HCC or lung adenocarcinoma by immune infiltration, and can be used as a prognostic biomarker for HCC and lung adenocarcinoma (Xu et al., 2020; Chen et al., 2022). Moreover, DHX37 acts as a function regulator of CD8 T cells (Dong et al., 2019). YARS1 is involved in tRNA aminoacylation and gene expression. There are no reports on the role of YARS1 in HCC for now.

It is worth mentioning that the gene signatures constructed by different methods may have different applications. For example, a four-gene metabolic signature for HCC can reflect the disorder of the metabolic microenvironment, thereby providing potential biomarkers for the metabolic treatment and treatment response prediction of HCC (Liu et al., 2020). A ferroptosis-related gene signature can be used to predict the prognosis of HCC patients (Liang et al., 2020). An immune-related lncRNA signature has the potential to measure the response to ICB immunotherapy and guide the choice of HCC immunotherapy (Zhang Y. et al., 2020). An immune-related gene signature can predict the response of HCC patients to immunotherapy (Dai et al., 2021). DNA methylation is an important regulator of gene transcription in the etiology and pathogenesis of HCC. Two HCC prognostic signatures related to DNA repair have recently been reported to help explore molecular mechanisms related to DNA repair (Li N. et al., 2019; Li G. X. et al., 2019). A gene signature related to glycolysis could help to analyze the role of glycolysis in HCC (Jiang et al., 2019). In addition, the tumor microenvironment plays an important role in the progression, recurrence and metastasis of HCC. A gene signature based on the HCC microenvironment helps to explore the role of the tumor microenvironment in HCC (Zhang F.-P. et al., 2020).

This study has some limitations. Although this study used the method of mutual verification between two independent datasets to verify the prognostic significance of the five-gene signature. However, *in vitro* experiments are still an important step to further confirm the prognostic value of this gene signature. In addition, this is a retrospective study, so it is necessary to verify the robustness of this five-gene signature in a prospective study in the future.

CONCLUSION

In summary, the study identified a new prognostic gene signature based on proliferation-related genes (FARSB, NOP58, CCT4, DHX37, and YARS). Besides, a nomogram based on the five-gene signature was constructed for clinical practice. The five-gene signature can be used for survival prediction and risk stratification for HCC patients. Moreover, potential drugs targeting this gene signature deserve further attention in the treatment of HCC.

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

YL designed the study and revised the manuscript. JuL collected and analyzed the data. JiL interpreted the data and drafted the manuscript. WL and WM revised the manuscript. All authors have read and approved the final manuscript. JuL and JiL contributed equally to this work and are co-first authors.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2022.900380/full#supplementary-material

REFERENCES

Bai, J., Yao, B., Wang, L., Sun, L., Chen, T., Liu, R., et al. (2019). lncRNA A1BG-AS1 Suppresses Proliferation and Invasion of Hepatocellular Carcinoma Cells by Targeting miR-216a-5p. *J Cell. Biochem.* 120 (6), 10310–10322. doi:10.1002/jcb.28315

Busato, D., Mossenta, M., Baboci, L., Di Cintio, F., Toffoli, G., and Dal Bo, M. (2019). Novel Immunotherapeutic Approaches for Hepatocellular Carcinoma Treatment. *Expert Rev. Clin. Pharmacol.* 12 (5), 453–470. doi:10.1080/17512433.2019.1598859

Cao, L., Cheng, H., Jiang, Q., Li, H., and Wu, Z. (2020). APEX1 is a Novel Diagnostic and Prognostic Biomarker for Hepatocellular Carcinoma. *Aging* 12 (5), 4573–4591. doi:10.18632/aging.102913

Chen, W., Ye, L., Wen, D., and Chen, F. (2019). MiR-490-5p Inhibits Hepatocellular Carcinoma Cell Proliferation, Migration and Invasion by Directly Regulating ROBO1. *Pathol. Oncol. Res.* 25 (1), 1–9. doi:10.1007/s12253-017-0305-4

Chen, H., Jiang, Z., Yang, B., Yan, G., Wang, X., and Zang, S. (2022). Exploring Prognostic Signatures of Hepatocellular Carcinoma and the Potential Implications in Tumor Immune Microenvironment. *Comb. Chem. High. Throughput Screen.* 25, 998–1004. doi:10.2174/1386207324666210309100923

Cronin, K. A., Lake, A. J., Scott, S., Sherman, R. L., Noone, A.-M., Howlader, N., et al. (2018). Annual Report to the Nation on the Status of Cancer, Part I: National Cancer Statistics. *Cancer* 124 (13), 2785–2800. doi:10.1002/cncr.31551

Dai, Y., Qiang, W., Lin, K., Gui, Y., Lan, X., and Wang, D. (2021). An Immune-Related Gene Signature for Predicting Survival and Immunotherapy Efficacy in Hepatocellular Carcinoma. *Cancer Immunol. Immunother.* 70 (4), 967–979. doi:10.1007/s00262-020-02743-0

Dong, M. B., Wang, G., Chow, R. D., Ye, L., Zhu, L., Dai, X., et al. (2019). Systematic Immunotherapy Target Discovery Using Genome-Scale *In Vivo* CRISPR Screens in CD8 T Cells. *Cell* 178 (5), 1189–1204. e23. doi:10.1016/j.cell.2019.07.044

Du, H., Xu, Q., Xiao, S., Wu, Z., Gong, J., Liu, C., et al. (2019). MicroRNA-424-5p Acts as a Potential Biomarker and Inhibits Proliferation and Invasion in Hepatocellular Carcinoma by Targeting TRIM29. *Life Sci.* 224, 1–11. doi:10.1016/j.lfs.2019.03.028

Fulgenzi, C. A. M., Talbot, T., Murray, S. M., Silletta, M., Vincenzi, B., Cortellini, A., et al. (2021). Immunotherapy in Hepatocellular Carcinoma. *Curr. Treat. Options Oncol.* 22 (10), 87. doi:10.1007/s11864-021-00886-5

Gao, W., Lu, Y. X., Wang, F., Sun, J., Bian, J. X., and Wu, H. Y. (2019). miRNA-217 Inhibits Proliferation of Hepatocellular Carcinoma Cells by Regulating KLF5. *Eur. Rev. Med. Pharmacol. Sci.* 23 (18), 7874–7883. doi:10.26355/eurrev_201909_18997

Gao, X., Guo, R., Li, Y., Kang, G., Wu, Y., Cheng, J., et al. (2021). Contribution of Upregulated Aminoacyl-tRNA Biosynthesis to Metabolic Dysregulation in Gastric Cancer. *J. Gastroenterol. Hepatol.* 36 (11), 3113–3126. doi:10.1111/jgh.15592

Geeleher, P., Cox, N. J., and Huang, R. S. (2014). Clinical Drug Response Can Be Predicted Using Baseline Gene Expression Levels and *In Vitro* Drug Sensitivity in Cell Lines. *Genome Biol.* 15 (3), R47. doi:10.1186/gb-2014-15-3-r47

Gu, Y., Li, X., Bi, Y., Zheng, Y., Wang, J., Li, X., et al. (2020). CCL14 is a Prognostic Biomarker and Correlates with Immune Infiltrates in Hepatocellular Carcinoma. *Aging* 12 (1), 784–807. doi:10.18632/aging.102656

Heagerty, P. J., Lumley, T., and Pepe, M. S. (2000). Time-Dependent ROC Curves for Censored Survival Data and a Diagnostic Marker. *Biometrics* 56 (2), 337–344. doi:10.1111/j.0006-341x.2000.00337.x

Huang, K., Pang, T., Tong, C., Chen, H., Nie, Y., Wu, J., et al. (2021). Integrative Expression and Prognosis Analysis of DHX37 in Human Cancers by Data Mining. *BioMed Res. Int.* 2021, 6576210. doi:10.1155/2021/6576210

Ji, D., Wang, Y., Sun, B., Yang, J., and Luo, X. (2019). Long Non-Coding RNA MNX1-AS1 Promotes Hepatocellular Carcinoma Proliferation and Invasion through Targeting miR-218-5p/COMMD8 axis. *Biochem. Biophys. Res. Commun.* 513 (3), 669–674. doi:10.1016/j.bbrc.2019.04.012

Ji, D., Jiang, C., Zhang, L., Liang, N., Jiang, T., Yang, B., et al. (2019). LncRNA CRNDE Promotes Hepatocellular Carcinoma Cell Proliferation, Invasion, and Migration through Regulating miR-203/BCAT1 Axis. *J. Cell. Physiol.* 234 (5), 6548–6560. doi:10.1002/jcp.27396

Jiang, L., Zhao, L., Bi, J., Guan, Q., Qi, A., Wei, Q., et al. (2019). Glycolysis Gene Expression Profilings Screen for Prognostic Risk Signature of Hepatocellular Carcinoma. *Aging* 11 (23), 10861–10882. doi:10.18632/aging.102489

Li, N., Zhao, L., Guo, C., Liu, C., and Liu, Y. (2019). Identification of a Novel DNA Repair-Related Prognostic Signature Predicting Survival of Patients with Hepatocellular Carcinoma. *Cancer Manag. Res.* 11, 7473–7484. doi:10.2147/CMAR.S204864

Li, G. X., Ding, Z. Y., Wang, Y. W., Liu, T. T., Chen, W. X., Wu, J. J., et al. (2019). Integrative Analysis of DNA Methylation and Gene Expression Identify a Six Epigenetic Driver Signature for Predicting Prognosis in Hepatocellular Carcinoma. *J. Cell. Physiol.* 234 (7), 11942–11950. doi:10.1002/jcp.27882

Li, F., Liu, C.-S., Wu, P., Ling, A.-S., Pan, Q., and Li, X.-N. (2021). CCT4 Suppression Inhibits Tumor Growth in Hepatocellular Carcinoma by Interacting with Cdc20. *Chin. Med. J. Engl.* 134 (22), 2721–2729. doi:10.1097/CM9.0000000000001851

Li, W., Liu, J., and Zhao, H. (2021). Prognostic Power of a Chaperonin Containing TCP-1 Subunit Genes Panel for Hepatocellular Carcinoma. *Front. Genet.* 12, 668871. doi:10.3389/fgene.2021.668871

Liang, J.-Y., Wang, D.-S., Lin, H.-C., Chen, X.-X., Yang, H., Zheng, Y., et al. (2020). A Novel Ferroptosis-Related Gene Signature for Overall Survival Prediction in Patients with Hepatocellular Carcinoma. *Int. J. Biol. Sci.* 16 (13), 2430–2441. doi:10.7150/ijbs.45050

Lin, M., Liu, Y., Ding, X., Ke, Q., Shi, J., Ma, Z., et al. (2019). E2F1 Transactivates IQGAP3 and Promotes Proliferation of Hepatocellular Carcinoma Cells through IQGAP3-Mediated PKC-Alpha Activation. *Am. J. Cancer Res.* 9 (2), 285–299.

Liu, Y., Wang, D., Li, Y., Yan, S., Dang, H., Yue, H., et al. (2019). Long Noncoding RNA CCAT2 Promotes Hepatocellular Carcinoma Proliferation and Metastasis through Up-Regulation of NDRG1. *Exp. Cell Res.* 379 (1), 19–29. doi:10.1016/j.yexcr.2019.03.029

Liu, G. M., Xie, W. X., Zhang, C. Y., and Xu, J. W. (2020). Identification of a Four-Gene Metabolic Signature Predicting Overall Survival for Hepatocellular Carcinoma. *J. Cell. Physiol.* 235 (2), 1624–1636. doi:10.1002/jcp.29081

Meyers, R. M., Bryan, J. G., McFarland, J. M., Weir, B. A., Sizemore, A. E., Xu, H., et al. (2017). Computational Correction of Copy Number Effect Improves Specificity of CRISPR-Cas9 Essentiality Screens in Cancer Cells. *Nat. Genet.* 49 (12), 1779–1784. doi:10.1038/ng.3984

Ming, M., Ying, M., and Ling, M. (2019). miRNA-125a-5p Inhibits Hepatocellular Carcinoma Cell Proliferation and Induces Apoptosis by Targeting TP53 Regulated Inhibitor of Apoptosis 1 and Bcl-2-Like-2 Protein. *Exp. Ther. Med.* 18 (2), 1196–1202. doi:10.3892/etm.2019.7674

Park, S. Y. (2018). Nomogram: An Analogue Tool to Deliver Digital Knowledge. *J. Thorac. Cardiovasc. Surg.* 155 (4), 1793. doi:10.1016/j.jtcvs.2017.12.107

Qu, D., Cui, F., Lu, D., Yang, Y., and Xu, Y. (2019). DEP Domain Containing 1 Predicts Prognosis of Hepatocellular Carcinoma Patients and Regulates Tumor Proliferation and Metastasis. *Cancer Sci.* 110 (1), 157–165. doi:10.1111/cas.13867

Reinhold, W. C., Sunshine, M., Liu, H., Varma, S., Kohn, K. W., Morris, J., et al. (2012). CellMiner: A Web-Based Suite of Genomic and Pharmacologic Tools to Explore Transcript and Drug Patterns in the NCI-60 Cell Line Set. *Cancer Res.* 72 (14), 3499–3511. doi:10.1158/0008-5472.CAN-12-1370

Sangro, B., Sarobe, P., Hervás-Stubbs, S., and Melero, I. (2021). Advances in Immunotherapy for Hepatocellular Carcinoma. *Nat. Rev. Gastroenterol. Hepatol.* 18 (8), 525–543. doi:10.1038/s41575-021-00438-0

Shen, Z., Liu, S., Liu, J., Liu, J., and Yao, C. (2021). Weighted Gene Co-Expression Network Analysis and Treatment Strategies of Tumor Recurrence-Associated Hub Genes in Lung Adenocarcinoma. *Front. Genet.* 12, 756235. doi:10.3389/fgene.2021.756235

Sun, Y., Zhang, F., Wang, L., Song, X., Jing, J., Zhang, F., et al. (2019). A Five lncRNA Signature for Prognosis Prediction in Hepatocellular Carcinoma. *Mol. Med. Rep.* 19 (6), 5237–5250. doi:10.3892/mmr.2019.10203

Tano, K., Mizuno, R., Okada, T., Rakwal, R., Shibato, J., Masuo, Y., et al. (2010). MALAT-1 Enhances Cell Motility of Lung Adenocarcinoma Cells by Influencing the Expression of Motility-Related Genes. *FEBS Lett.* 584 (22), 4575–4580. doi:10.1016/j.febslet.2010.10.008

Wang, X., Zhou, Z., Zhang, T., Wang, M., Xu, R., Qin, S., et al. (2019). Overexpression of miR-664 Is Associated with Poor Overall Survival and

Accelerates Cell Proliferation, Migration and Invasion in Hepatocellular Carcinoma. *Onco Targets Ther.* 12, 2373–2381. doi:10.2147/OTT.S188658

Wang, Y., Ruan, Z., Yu, S., Tian, T., Liang, X., Jing, L., et al. (2019). A Four-Methylated mRNA Signature-Based Risk Score System Predicts Survival in Patients with Hepatocellular Carcinoma. *Aging* 11 (1), 160–173. doi:10.18632/aging.101738

Wang, J., Huang, R., Huang, Y., Chen, Y., and Chen, F. (2021). Overexpression of NOP58 as a Prognostic Marker in Hepatocellular Carcinoma: A TCGA Data-Based Analysis. *Adv. Ther.* 38 (6), 3342–3361. doi:10.1007/s12325-021-01762-2

Wei, H., Hu, J., Pu, J., Tang, Q., Li, W., Ma, R., et al. (2019). Long Noncoding RNA HAGLROS Promotes Cell Proliferation, Inhibits Apoptosis and Enhances Autophagy via Regulating miR-5095/ATG12 Axis in Hepatocellular Carcinoma Cells. *Int. Immunopharmacol.* 73, 72–80. doi:10.1016/j.intimp.2019.04.049

Wei, Z., Liu, Y., Qiao, S., Li, X., Li, Q., Zhao, J., et al. (2019). Identification of the Potential Therapeutic Target Gene UBE2C in Human Hepatocellular Carcinoma: An Investigation Based on GEO and TCGA Databases. *Oncol. Lett.* 17 (6), 5409–5418. doi:10.3892/ol.2019.10232

Xiao, Z., Liu, Y., Zhao, J., Li, L., Hu, L., Lu, Q., et al. (2020). Long Noncoding RNA LINC01123 Promotes the Proliferation and Invasion of Hepatocellular Carcinoma Cells by Modulating the miR-34a-5p/TUFT1 Axis. *Int. J. Biol. Sci.* 16 (13), 2296–2305. doi:10.7150/ijbs.45457

Xu, Y., Jiang, Q., Liu, H., Xiao, X., Yang, D., Saw, P. E., et al. (2020). DHX37 Impacts Prognosis of Hepatocellular Carcinoma and Lung Adenocarcinoma through Immune Infiltration. *J. Immunol. Res.* 2020, 8835393. doi:10.1155/2020/8835393

Yan, Y., Lu, Y., Mao, K., Zhang, M., Liu, H., Zhou, Q., et al. (2019). Identification and Validation of a Prognostic Four-Genes Signature for Hepatocellular Carcinoma: Integrated ceRNA Network Analysis. *Hepatol. Int.* 13 (5), 618–630. doi:10.1007/s12072-019-09962-3

Zhang, Y., Zhang, L., Xu, Y., Wu, X., Zhou, Y., and Mo, J. (2020). Immune-Related Long Noncoding RNA Signature for Predicting Survival and Immune Checkpoint Blockade in Hepatocellular Carcinoma. *J. Cell. Physiol.* 235 (12), 9304–9316. doi:10.1002/jcp.29730

Zhang, F.-P., Huang, Y.-P., Luo, W.-X., Deng, W.-Y., Liu, C.-Q., Xu, L.-B., et al. (2020). Construction of a Risk Score Prognosis Model Based on Hepatocellular Carcinoma Microenvironment. *World J. Gastroenterol.* 26 (2), 134–153. doi:10.3748/wjg.v26.i2.134

Zhong, L., Liao, D., Zhang, M., Zeng, C., Li, X., Zhang, R., et al. (2019). YTHDF2 Suppresses Cell Proliferation and Growth via Destabilizing the EGFR mRNA in Hepatocellular Carcinoma. *Cancer Lett.* 442, 252–261. doi:10.1016/j.canlet.2018.11.006

Zhong, X., Kan, A., Zhang, W., Zhou, J., Zhang, H., Chen, J., et al. (2019). CBX3/HP1γ Promotes Tumor Proliferation and Predicts Poor Survival in Hepatocellular Carcinoma. *Aging* 11 (15), 5483–5497. doi:10.18632/aging.102132

Zongyi, Y., and Xiaowu, L. (2020). Immunotherapy for Hepatocellular Carcinoma. *Cancer Lett.* 470, 8–17. doi:10.1016/j.canlet.2019.12.002

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 Genetics

ORIGINAL RESEARCH
published: 14 July 2022
doi: 10.3389/fgene.2022.908957

Construction and Verification of a Fibroblast-Related Prognostic Signature Model for Colon Cancer

Zhe Zhao[1†], *Wenqi Li*[2†], *LiMeng Zhu*[1], *Bei Xu*[1], *Yudong Jiang*[3], *Nan Ma*[1], *LiQun Liu*[1], *Jie Qiu*[2] *and Min Zhang*[1*]

[1]Zhengzhou KingMed Center for Clinical Laboratory Co. Ltd., Zhengzhou, China, [2]Department of Newborn Infants, Children's Hospital of Nanjing Medical University, Nanjing, China, [3]Department of General Surgery, Zhongshan Hospital, Fudan University, Shanghai, China

Traditionally, cancer-associated fibroblasts (CAFs), an essential component of tumor microenvironment, were exert a crucial part in colon cancer progression. In this study, single-cell RNA-sequencing (scRNA-seq) data from 23 and bulk RNA-seq data from 452 colon cancer patients were extracted from the GEO database and TCGA-COAD and GEO databases, respectively. From single-cell analysis, 825 differentially expressed genes (DEGs) in CAFs were identified between each pair of six newly defined CAFs, named enCAF, adCAF, vaCAF, meCAF, erCAF, and cyCAF. Cell communication analysis with the iTALK package showed communication relationship between CAFs, including cell autocrine, cytokine, and growth factor subtypes, such as receptor-ligand pairs of *TNFSF14-LTBR*, *IL6-F3,* and *IL6-IL6ST*. Herein, we demonstrated the presence and prognostic value of adCAF and erCAF in colon cancer based on CIBERSORTx, combining single-cell marker genes and transcriptomics data. The prognostic significance of the enCAF and erCAF has been indirectly proved by both the correlation analysis with macrophages and CAFs, and the quantitative reverse transcription-polymerase chain reaction (qRT-PCR) experiment based on 20 paired tumor samples. A prognostic model was constructed with 10 DEGs using the LASSO Cox regression method. The model was validated using two testing datasets, indicate a significant survival accuracy ($p <$ 0.0025). Correlation analyses between clinical information, such as age, gender, tumor stage and tumor features (tumor purity and immune score), and risk scores revealed our CAF-related model's robustness and excellent performance. Cell infiltration analysis by xCell revealed that the interaction between CAFs and multiple non-specific immune cells such as macrophages and the dendritic cell was a vital factor affecting immune score and prognosis. Finally, we analyzed how common anti-cancer drugs, including camptothecin, docetaxel and bortezomib, and immunotherapy, such as anti-PD-1 treatment, could be different in low-risk and high-risk patients inferred from our CAF-related model. In conclusion, the study utilized refined colon cancer fibroblast subsets and established the prognostic effects from the interaction with nonspecific immune cell.

OPEN ACCESS

Edited by:
Tianhai Tian,
Monash University, Australia

Reviewed by:
Zhiyuan Yang,
Hangzhou Dianzi University, China
Maryam Hassan Al-Zahrani,
King Abdulaziz University, Saudi Arabia

***Correspondence:**
Min Zhang
labzm@kingmed.com.cn

†These authors have contributed equally to this work

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

Received: *04 April 2022*
Accepted: *23 June 2022*
Published: *14 July 2022*

Citation:
Zhao Z, Li W, Zhu L, Xu B, Jiang Y, Ma N, Liu L, Qiu J and Zhang M (2022) Construction and Verification of a Fibroblast-Related Prognostic Signature Model for Colon Cancer. Front. Genet. 13:908957. doi: 10.3389/fgene.2022.908957

Abbreviations: CAF, cancer-associated fibroblast; DEGs, differentially expressed genes; GEO, gene expression omnibus; GO, gene ontology; KEGG, kyoto encyclopedia of genes and genomes; LASSO, least absolute shrinkage and selection operator; NK cell, natural killer cell; OS, overall survival; scRNA-seq, single-cell RNA-sequencing; TCGA, the Cancer Genome Atlas; ME, tumor microenvironment; t-SNE, t-distributed stochastic neighbor embedding; UMAP, uniform manifold approximation and projection.

Keywords: single-cell RNA-sequencing(scRNA-seq), colon cancer, cancer-associated fibroblasts (CAFs), immune infiltration, drug sensitivity

INTRODUCTION

Colon cancer is the third most common cancer among women and men and has the second-highest cancer mortality rate worldwide (Siegel et al., 2020a; Siegel et al., 2020b). Even with intensive treatments, the 5-year overall survival (OS) rate of colon cancer is below 60% (Moghimi-Dehkordi and Safaee, 2012). What's worse, the number of colon cancer patients under 50 years old has been rising sharply in recent years, and the mortality of colon cancer in young men is the highest during 2012–2016 (Wolf et al., 2018; Kasi et al., 2019).

Colon cancer is usually caused by continuous malignant gene mutations and epigenetic changes in the colon and rectum. The tumor microenvironment (TME) plays a vital role and is one of the driving factors in many types of cancer (Kalluri, 2016). The TME is composed of various non-epithelial cells and extracellular matrices. The non-epithelial cells mainly include tumor-infiltrating immune cells, fibroblasts, and vascular endothelial cells. Tumor-infiltrating immune cells, including macrophages, T cells, B cells, NK cells, dendritic cells (DCs), myeloid-derived suppressor cells (MDSCs), and regulatory T cells (Tregs), could affect tumor development and progression through interaction with tumor cells (Quail and Joyce, 2013; Hui and Chen, 2015). Cancer-associated fibroblasts (CAFs), a type of permanently activated fibroblasts, are shown to be important in tumor development and drug resistance (Pietras and Ostman, 2010; Kobayashi et al., 2021). However, the expressions of multiple commonly used fibroblast markers, such as COL3A1 and THY1, vary greatly in different CAF subgroups (Bu et al., 2019). Therefore, in order to develop better colon cancer treatment strategies based on CAFs, new methods are required to better classify CAFs and identify how different CAFs affect tumor development differently. Single-cell sequencing can uncover the cell diversity in tumor tissues in a comprehensive and unbiased manner. In recent years, single-cell Transcriptome sequencing technology has been widely adopted in the study of TME. However, in TME of colon cancer (CC), gene markers for CAFs have not been well elucidated.

Over the past decades, technological development in omics and bioinformatics, such as bulk RNA-seq and scRNA-seq, has dramatically advanced the diagnosis and treatment of many types of cancer (Gustafson et al., 2010; Wang et al., 2010). Hae-Ock Lee et al. did scRNA-seq on two colon cancer samples and made their data publicly available (Lee et al., 2020), providing researchers with much information on the characteristics and function of different CAF subgroups. However, due to the high demand for throughput and budget, it is unrealistic to apply large-scale scRNA-seq to a large number of tumor samples. Instead, developing a strategy to explore the valuable information from these existing scRNA-seq data would be more appealing. Furthermore, databases such as TCGA provide us with rich resources, including transcriptomic profiles and clinical information. Combining bulk RNA-seq and scRNA-seq data would be a more time- and cost-efficient approach. Thus, in the current study, we explored the prognosis value of different CAF subtypes in colon cancer patients using TCGA data (i.e., bulk RNA-seq data and OS information) and scRNA-seq data. The CIBERSORTx algorithm was applied to the TCGA bulk RNA-seq data, and the Seurat package was applied to the scRNA-seq data. Finally, we established a CAF-related prognostic signature model that could predict the OS of colon cancer patients. The reliability and accuracy of our weighted model was evaluated comprehensively using the TCGA test dataset and related clinical information. Our model could be well explained by factors such as immune cell infiltration, immune scores, and specific tumorigenic pathways. We further carried out various analyses to make our model accurate for providing suggestion for clinical treatment (**Figure 1**).

RESULTS

Classification of CAFs in Colon Cancer

The high dimensional information of scRNA-seq enables the identification of CAFs out of a pool of heterogeneous cells, the clustering of CAFs into various subtypes, and the determination of DEGs in different CAF subtypes. In total, scRNA-seq data were extracted for 33 samples, including 23 tumor samples and 10 paracancerous tissue samples from the SMC cohort (GSE132465). After quality control based on the proportion of cell signatures and mitochondrial and ribosomal gene expression, all the cells were classified by the dimensionality reduction algorithms, namely, t-distributed stochastic neighbor embedding (t-SNE) and uniform manifold approximation and projection (UMAP) into seven major and 32 more detailed clusters (**Figures 2A,B**). According to instructions in the original research, we successfully repeated the stromal cell classification results. We further divided the classified stromal cells into fibroblast and non-fibroblast subgroups. Classification of different cell groups, including the fibroblast subgroup, was validated using a combination of specific gene markers. Common gene markers for CAFs are shown in **Figures 2C,D**. The CAFs did not express any other cell markers. Specific gene markers, including *SPARC*, *COL1A1*, *COL1A2*, *LUM,* and *DCN*, and common gene markers, including *COL3A1* and *THY1*, were highly expressed in a high proportion of fibroblasts. A low percentage of stromal cells expressed fibroblast gene markers at lower levels, but these stromal cells could be excluded by negative markers of fibroblasts, such as *PECAM1*. The cell types and proportions of different clusters, including stromal cells and fibroblasts, are shown in **Figure 2E**.

Overall, T cells and epithelial cells (ECs) accounted for a higher proportion, while stromal cells and B cells accounted for a lower proportion, within the tumor samples. Immature ECs and fibroblasts were the dominant cell types in stromal cells. There were significant clustering differences for fibroblasts in tumor and

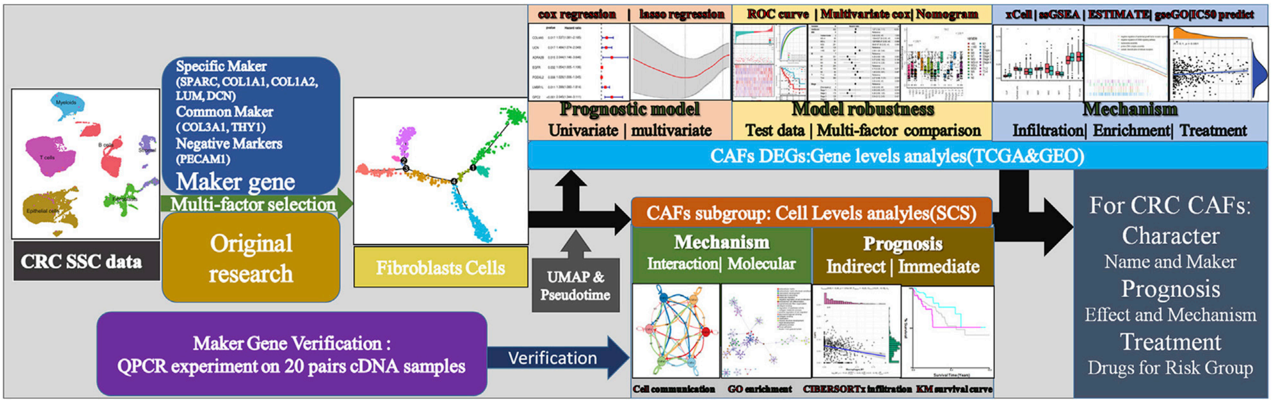


FIGURE 1 | Technical roadmap for this study. We combined experiments and database analysis to reveal the characteristics, prognostic mechanisms and treatment recommendations of CAFs from different perspectives at the gene and cell level.

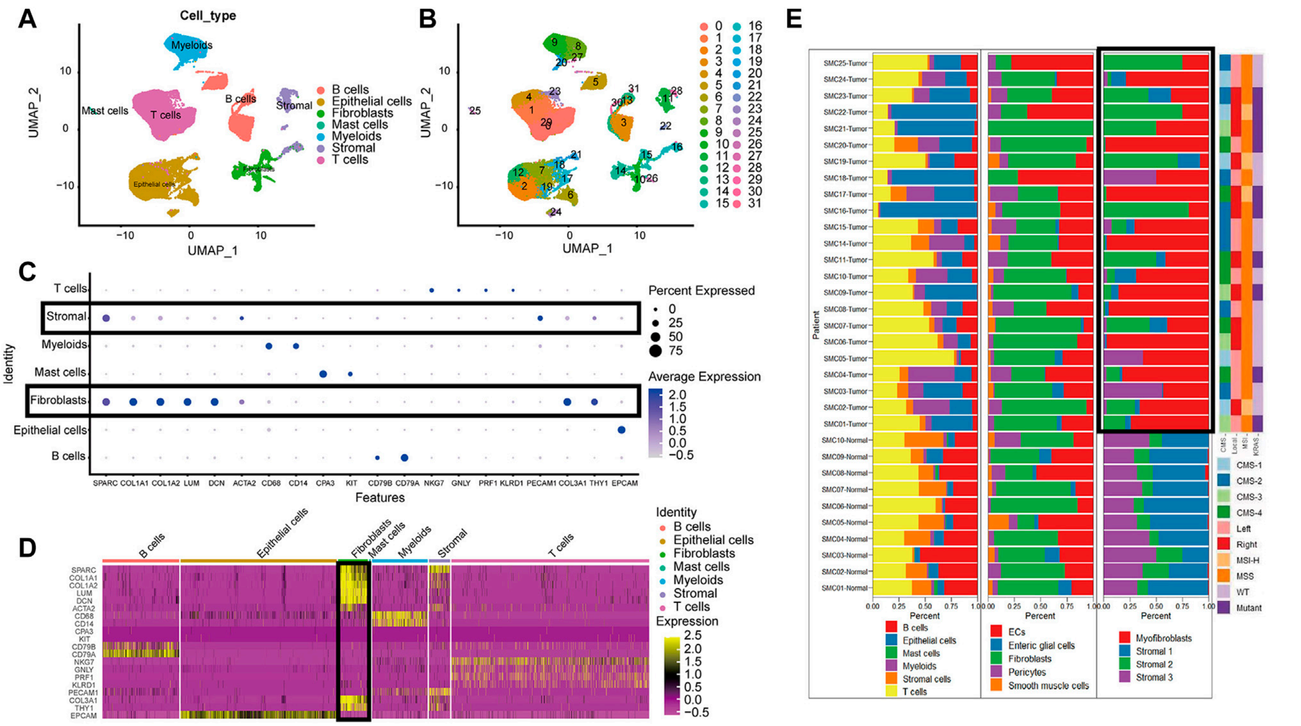


FIGURE 2 | The cell group in CRC based on scRNA-seq. **(A)**. UMAP plot of all cell from the original article **(B)**. UMAP plot of the cluster based on scRNA-seq; **(C)**. Bubble chart of maker gene within all cell type; sizes of dots show gene abundance, while shade shows gene expression level. The main difference of fibroblasts and stromal cells marked with black boxes was the negative selection maker PECAM1. **(D)**. Heatmap of maker gene expression graph for every cell. **(E)**. The proportion structure of all cells, stromal cells and fibroblasts for each patient with clinical information.

paracancerous tissues (**Figure 2E**). For instance, there were more activated fibroblasts in the tumors than predominant stromal sub I and III cells in the paracancerous tissues.

We used an unsupervised trajectory analysis to establish a novel classification for the previously classified fibroblast cells. In this approach, we divided all the fibroblast cells in the tumor samples into six subgroups, i.e., CAF1-6, according to the result of an unsupervised trajectory analysis (**Figure 3**). According to the characteristics of each subtype, we renamed CAF1-6 to enCAF (entoderm-related CAF), adCAF (adhesion-related CAF), vaCAF (vascular-related CAF), meCAF (mesenchyme-related CAF), erCAF (endoplasmic reticulum-related CAF), and cyCAF (cell cycle-related CAF), respectively. The top 10 marker genes for each subgroup are shown in **Figure 3E**.

Cell Communicational Signal Analysis and Construction of a Ligand-Receptor Interaction Atlas Among CAFs in Colon Cancer

To analyze the intercellular communication between CAF subgroups, we used iTALK, a cell communication signal analysis tool, to analyze the TCGA colon cancer samples. We explored other factors, including checkpoints, cytokines, and growth factors, which revealed mechanistic insights into the CAF subtype interactions (**Figure 3**). Many cytokines were identified within the enCAF, acting as the receptor in the intercellular communication. IL6 was most abundant in the erCAF, forming IL6-IL6ST and IL6-F3 receptor-ligand pairs with other CAF subtypes (i.e., erCAF and enCAF). In terms of immune checkpoint genes, *TNFSF14* was highly expressed in erCAF and interacted with other CAF subgroups through the TNFSF14-TNFRSF14 and TNFSF14-LTBR receptor-ligand pairs. Quantitatively, meCAF had more intensive immune checkpoint interactions with other CAF subtypes. Unexpectedly, erCAF only express the receptor, while other CAF subtypes, especially cyCAF and enCAF, could secrete growth factors, such as CTGF, and interact with other CAF subtypes through the CTGF-ITGA5 and CTGF-LRP1 ligand-receptor pairs. Overall, these results showed that the erCAF subtype interacts with the cyCAF subtype *via* the COL1A1-ITGB1 and COL1A1-CD44 signaling pathways; cyCAF subtype also interacts with the other subtypes (i.e., erCAF and enCAF) via the TIMP1-CD63 signaling pathway.

Functional Enrichment Analysis on Different CAF Subtypes and Association Between CAF Subtypes and Prognosis

To explore the CAF profiles and understand how different CAF subtypes could affect prognosis in colon cancer patients, we

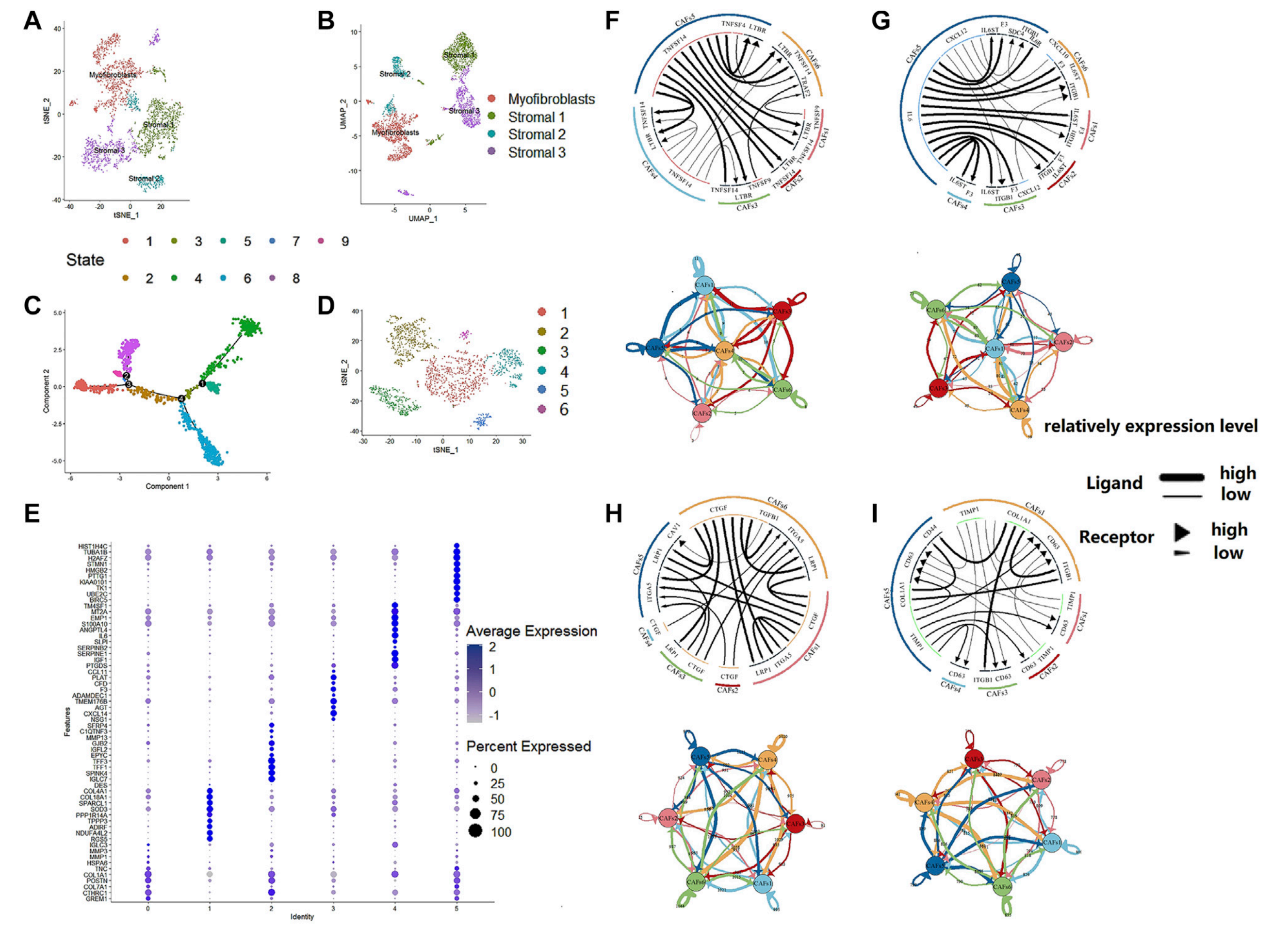


FIGURE 3 | The regroup and cell communication analysis of CAFs in CRC based on scRNA-seq. **(A)**. tSNE plot of the stromal cell from the original article **(B)**. UMAP plot of the stromal cell from the original article **(C)**. The tSNE of pseudotime trajectory analysis **(D)**. tSNE plot showed the regroup of CAFs in CRC based on the pseudotime trajectory analysis. **(E)**. Dot plot for top 10 markers of each CAF subgroup; sizes of dots show gene abundance, while shade shows gene expression level. for the subgroups of CAFs in CRC. The upper parts are the circos plots representing top20 highly expressed ligand-receptor interactions among CAF subgroups; the lower parts are the network plots showing the number of ligand-receptor interactions among CAF subgroups. **(F)**. cytokines/chemokines **(G)**. immune checkpoint genes **(H)**. growth factors **(I)**. others.

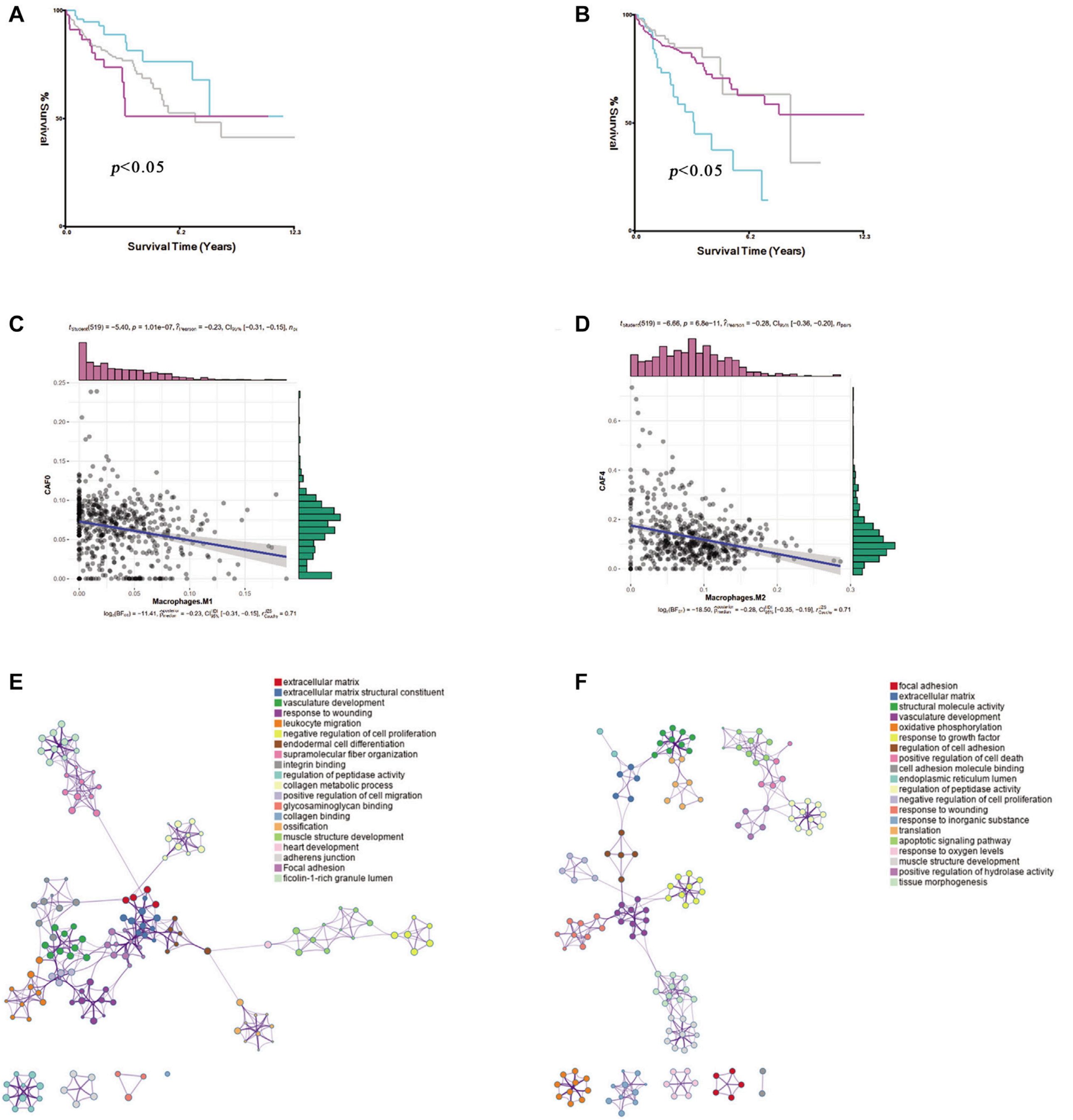


FIGURE 4 | The identification of prognostic CAF subgroup. **(A)**. The Kaplan–Meier plot of the abundance of adCAF, the red line for high abundance, the aquamarine line for low abundance **(B)**. The Kaplan–Meier plot of the abundance of cyCAF, red line for high abundance, aquamarine line for low abundance **(C)**. Figure for the Pearson correlation between enCAF and macrophage M1 **(D)**. The figure for the Pearson correlation between erCAF and macrophage M2 **(E)**. Network diagram for the GO enrichment of CAFs marker genes **(F)**. Network diagram for the KEGG enrichment of CAFs marker genes.

applied the CIBERSORTx algorithm to analyze the abundance of different CAF subtypes and immune cells in colon cancer. We found that the adCAF subtype was a risk factor (**Figure 4A**) while the cyCAF subtype was a protective factor (**Figure 4B**) regarding OS of the colon cancer patients. Moreover, we found that the abundance of the enCAF subtype was negatively associated with the abundance of the M1-type macrophages (**Figure 4C**), while the abundance of the erCAF subtype was negatively related to the abundance of the M2-type macrophages (**Figure 4D**). To explore the specific function of each CAF subtype, we carried out Gene Ontology (GO) and Kyoto Encyclopedia of Genes and Genomes (KEGG) enrichment analyses, of which the results are shown in

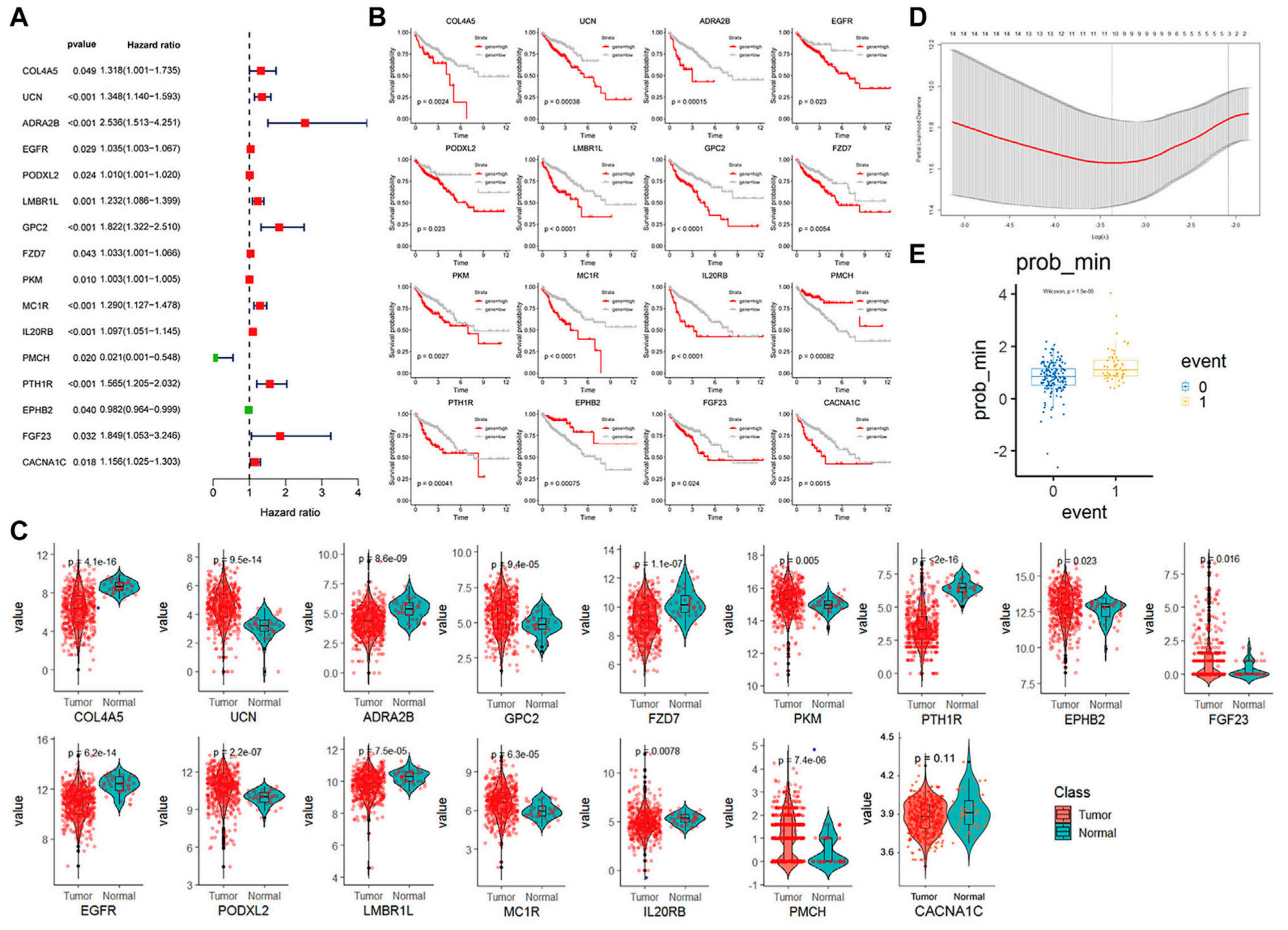


FIGURE 5 | CAF-related gene prognosis model based on TCGA data **(A)**. Forest plot of the 16 genes selected by single factor COX regression **(B)**. The Kaplan–Meier plot of the 16 genes **(C)**. Box plot showing the different expression levels of 16 genes in tumor and normal tissues. **(D)**. Box plot of survival events under lasso optimal regression parameters **(E)**. Number of lasso regression variables genes under the best regression parameters.

Figures 4E,F, and the related -log10 (*p-values*) are shown in **Supplementary Figure S1**.

Of special note, two key marker genes, *GREM1,* and *IGF1*, were significantly differently expressed in 20 pairs of colon cancer and paracancerous tissues. *GREM1* as enCAF maker gene was highly expressed in colon cancer tissues. On the contrary, *IFG1* as erCAF maker gene was highly expressed in the paracancerous tissues (**Figure 9C**).

Construction of a CAF-Related Prognostic Signature Model

We identified 825 highly expressed ligand or receptor genes in different CAF subtypes. To further explore how different CAF subtypes relate to the prognosis of colon cancer patients, we constructed a CAF-related prognostic signature model based on these 825 genes. Fifteen genes that were significantly associated with the prognosis of colon cancer patients were identified by the univariate Cox regression analysis ($p < 0.05$) (**Figure 5A**). Consistently, expression levels of these 15 genes were significantly different between colon cancer and paracancerous tissues (**Figure 5C**). The OS was significantly different in the high- and low-expression groups of each of these 16 genes (**Figure 5B**).

The TCGA colon cancer patients were divided into the training and internal testing datasets by an 8:2 ratio. The least absolute shrinkage and selection operator (LASSO) Cox regression was used to construct the CAF-related signature model. Ten genes were recovered from the LASSO regression analysis under optimal regularization parameters (**Figure 5E**). Using our model, the OS of patients with colon cancer group by different genes could be well distinguished (**Figure 5B**). Prognostic genes are weighted by lasso regression, ie the following formula is a simplified weighted model after removing expression correlations between genes. Risk score = (CACNA1C × 0.195) + (COL4A5 × 0.563) + (ADRA2B × 0.734) + (EGFR × 0.082) + (LMBR1L × 0.299) + (FZD7 × 0.119) + (PKM × 0.007) + (IL20RB × 0.384)—(PMCH × 3.74)—(EPHB2 × 0.055). Each gene here represents the transcript expression of the gene (hg38 version), and the coefficient of each gene is the

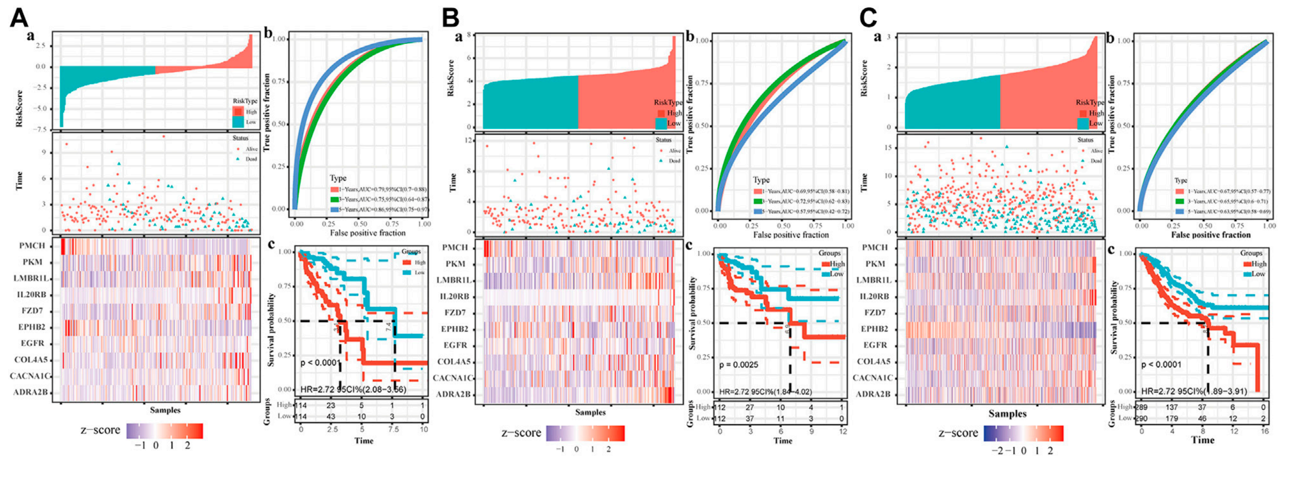


FIGURE 6 | The establishment of CAF-related prognostic signature. The upper left parts are distribution plots for the relationship between risk score and survival status **(A)**; the lower left parts are heatmaps for the expression of 10 genes in the cohort; the upper right parts are ROC curve for the CAF-related signature in the cohort **(B)**; the lower right parts are survival curves between high-risk and low-risk groups **(C)**. **(A)**. Training set **(B)**. Internal testing set **(C)**. External validation set.

weighted value. The positive and negative values represent tumor-promoting or tumor-suppressing genes, respectively.

Based on the risk model, the patients were divided into high and low scores groups, respectively. Based on this classification, the K-M plot showed a significant difference in high- and low-risk groups in the training dataset ($p < 0.0001$) and the two testing datasets ($p = 0.0025$ and $p < 0.000$, respectively). The area under the curve (AUC) values for OS prediction at 1-, 3- and 5-years of the training dataset were 0.79, 0.75, and 0.86, respectively. Consistently, the AUC values for OS prediction at 1-, 3- and 5-years of the internal and external testing datasets were 0.69, 0.72, 0.57, and 0.67, 0.65, 0.63, respectively, indicating that our signature model is robust and of great prognostic value (**Figure 6**).

Accuracy and Robustness of Our Constructed CAF-Related Prognostic Signature Model

We calculated the risk scores for all patients using our model. The risk scores were different between different subgroups when classifying using different clinical features, including old, N stage, T stage, M stage, and Tumor stage ($p < 5e\text{-}^{6}$), except for the MSI mutation feature. The distribution of scores was consistent with clinical characteristics. As shown in **Figure 7D**, patient groups with older age, M1 stage (M staging system), N2 stage (N staging system), stage 4 (tumor staging system), and T3-4 (tumor grade) had higher risk scores. The CAF-related prognostic model performed well, not interfered by multiple clinical factors (**Figure 7A**). In the multiple Cox regression analysis combining risk scores and clinical factors such as MSI mutation type, patient age, tumor grade, and TNM stage, the prognostic prediction was not affected compared with that from risk scores alone. However, age could contribute significantly to the risk model ($p = 0.004$). Overall, our risk scores correlated better with OS at 1-, 5- and 10-years compared with tumor stage and age (**Figure 7B**). The calibration curve of the model was very stable, and there were limited variations between the training and the two testing datasets (**Figure 7C**).

Possible Molecular Mechanisms Related to Our Prognostic Signature Model

To better understand the differences in immune cell infiltration status between different groups classified based on our CAF-related prognostic model, we used xCell to infer the cell infiltration ratio in each sample. Using xCell gene signatures, 11 out of 64 cell types were highly infiltrated with a ratio higher than 5%, including Th1 cells and smooth muscle cells (>25%) (**Figure 8A**). Of the top seven cell types, epithelial cells and mesenchymal stem cells (MSCs) were identified to be risk factors as these 2 cell types had a high percentage in the high-risk group. On the contrary, common lymphoid progenitors (CLPs), smooth muscle cells, classic dendritic cells (cDCs) and interstitial dendritic cells (iDCs) were identified as protective factors as these cell types had a low percentage in the low-risk group. In addition, immune scores were also significantly different between high- and low-risk groups.

To evaluate how different tumor indicators affect the accuracy of our model, we used ESTIMATE to calculate parameters, including ESTIMATE score, tumor purity, immune score, and stromal score for each of the TCGA-COAD samples and did correlation analyses between these parameters and risk scores. The analyses showed that there were significant correlations between immune scores ($R = -0.15$, $p = 0.0046$), ESTIMATE score ($R = -0.18$, $p = 0.00079$), tumor purity ($R = 0.18$, $p = 0.00079$) and stromal score ($R = -0.17$, $p = 0.0014$), and risk scores.

Checkpoint-related genes, such as *CD80*, *CD86*, *CD274* and *PDCD1*, were all highly expressed in the high-risk scores (**Figure 8D**). GO analysis showed many significantly enriched pathways, such as translational initiation, protein-DNA subunit

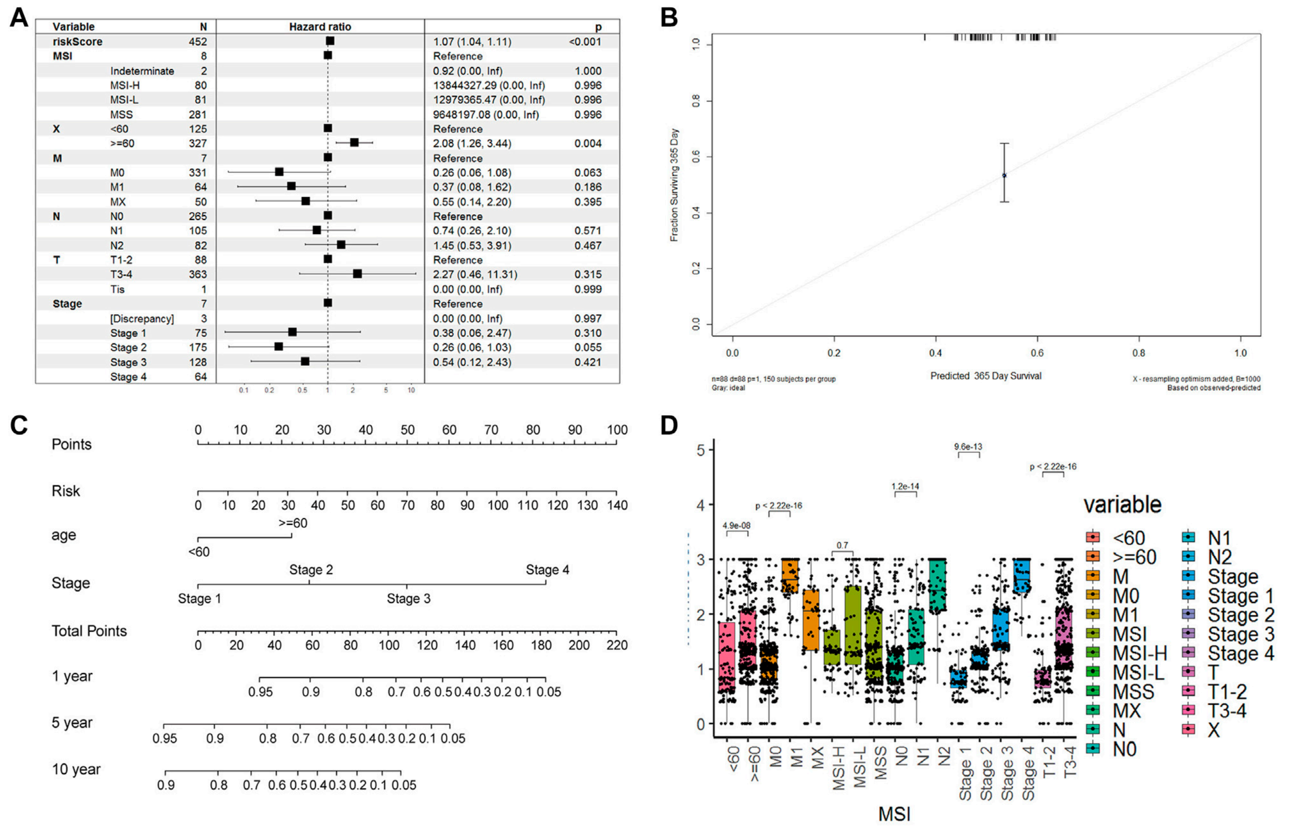


FIGURE 7 | Model robustness analysis based on TCGA clinical information. **(A)**. Forest plot of multivariate COX regression combined with multiple clinical indicators and risk scoring models. Training set **(B)**. Survival nomogram of risk model, age and stade factors **(C)**. The error fitting graph of the risk model under the training set and the test machine **(D)**. Box plot of distribution differences of risk factors in all clinical information groups.

assembly, and G2/M-related cell cycle (**Figure 8E**). The HEDGEHOG pathway, which is closely related to tumor development, was among the top 10 enriched pathways (**Figure 8B**). Our model inferred a significant linear correlation among these pathways and risk scores. In the high-risk group, six classical hallmark pathways, including the HEDGEHOG, APICAL, and NOTCH pathways, were highly activated, while two pathways, including the MYCV1 and E2F pathways, were inhibited. As shown in **Figure 8B**, most of the cancer-promoting pathways showed a strong autocorrelation in **Figure 8B**.

From the CGP database, we identified 10 drugs which was sensitive to colon tumors (**Figure 9A**). We tested these drugs and identified a total of seven drugs with relatively insignificant IC50 values. Three of the seven drugs, cisplatin, dasatinib, and BMS.536,924, showed poor drug sensitivity, while another 3, camptothecin, docetaxel, and bortezomib, showed strong drug sensitivity. For docetaxel and bortezomib, there was a significant relationship between drug sensitivity and risk scores (**Figure 9B**), indicating that docetaxel and bortezomib may be more effective in treating low-risk colon cancer patients defined using our model. It is worth noting that the drug sensitivity is more significant in the linear correlation model than in the grouping test. The discrete type of drug sensitivity data in different samples was more significant without an obvious clustering effect.

METHODS

Data Source

Bulk RNA-seq data and microarray data were downloaded from the TCGA-COAD (Network, 2012) and GSE39582 cohorts (Marisa et al., 2013), respectively. scRNA-seq data (SMC cohort) from 23 colon cancer and 10 paracancerous tissues were downloaded from the GSE132465 cohort (Lee et al., 2020). For bulk RNA-seq data from TCGA, only those with corresponding detailed clinical information were included. As a result, 452 patients from the TCGA-COAD cohort and 579 patients from the GES39582 cohort were included in our study. The TCGA-COAD cohort was used as a training dataset and an internal testing dataset with 8:2 radio. The GSE39582 cohort was used as an external testing dataset. This study followed the guidelines of the TCGA and GEO databases.

scRNA-Seq Data Preprocessing and Classification of CAF Subtypes

The quality control process was performed using the Seurat R package (version 4.0.1) (Hao et al., 2021). Low-quality cells, which were defined as cells with more than 10% mitochondrion-derived UMI counts, were removed.

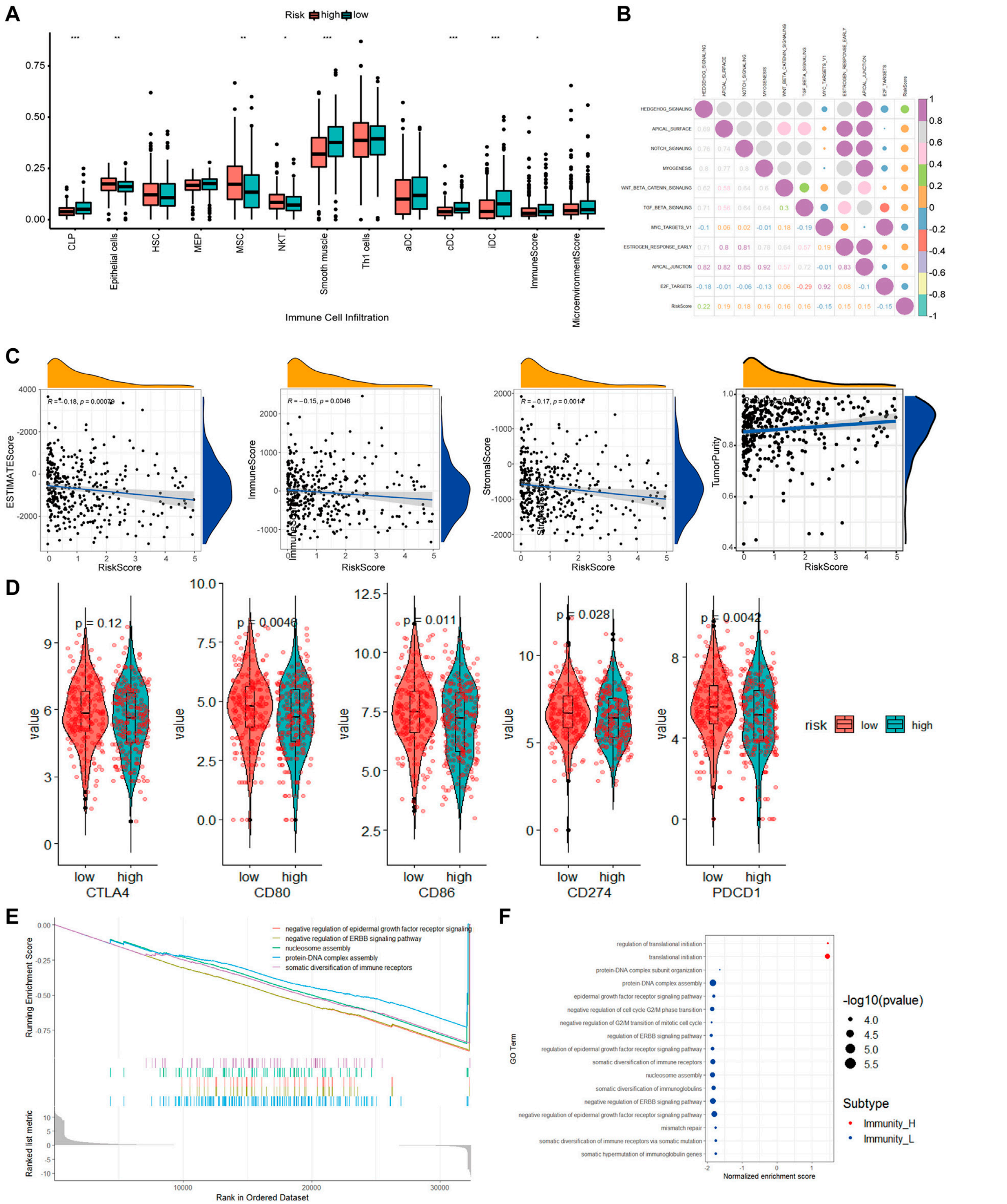


FIGURE 8 | The molecular mechanism of prognostic models based on multiple analyses. **(A)**. Differences in cell infiltration between high and low risk groups based on xCell **(B)**. Correlation heatmap of 50 classic tumor pathways and risk factors **(C)**. Violin chart of the expression levels of immune checkpoint genes in high-risk and low-risk groups **(D)**. Correlation scatter plot of risk factors and multiple infiltration scores **(E)**. The top 5 pathways of gseGO enrichment. **(F)**. Bubble chart of all significant pathways analyzed by gseGO.

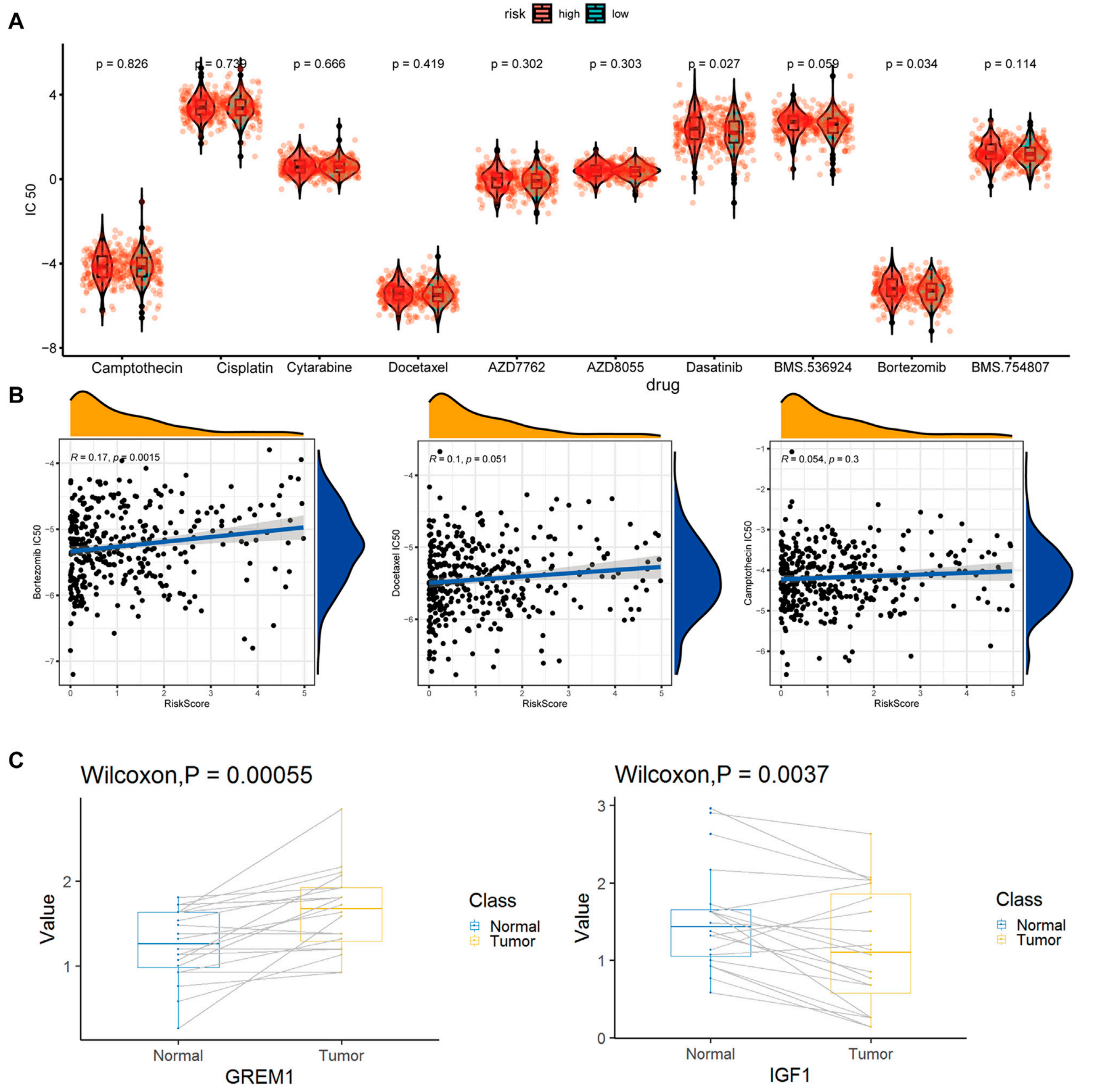


FIGURE 9 | Drug susceptibility prediction and QPCR experimental verification results. **(A)**. Statistical violin chart of IC50 predictions of 10 drugs in high and low risk groups **(B)**. Graph of linear correlation between risk factors and three drugs **(C)**. Box plot of differential expression of key maker genes for enCAF (GREM1) and erCAF (IGF1).

IntegrateData module in the Seurat was used to eliminate the batch effects among different patients. Here a relatively strict fibroblast system. 1. Preliminary classification of cells group according to the classification method in the original study; 2. Perform final verification according to specific maker and negative selection maker. 3. Perform preliminary verification and statistics according to the common maker of various types of cells (Zhou et al., 2020). The CAF definition in the current study: 1. From tumor samples; 2. Strict Fibroblasts. The CAF subtypes were firstly identified according to the definition in the original article visualized by 2D uniform manifold approximation and projection (UMAP) or t-Distributed Stochastic Neighbor Embedding (tSNE) (Becht et al., 2019).

Pseudotime Trajectory Analysis

To better classify CAFs, we tried pseudotime trajectory analysis by applying the Monocle 2 R package (Qiu et al., 2017). The "mean expression" parameter was set as > 0.125; the "num_cells_expressed" parameter was set as R10; the *p-value* was set as < 0.01 in the "differentialGeneTest" function. t-SNE plots were used for visualization of the pseudotime trajectories. The 2000 hypervariable genes were selected for analysis, and then the number of principal components (PCs) was set to 20 to obtain cell cluster clusters, and then these clusters were displayed in the form of a "tSNE" diagram. The best category was judged from the number of leaves in the quasi-sequential analysis and the clustering of tSNE.

Identification of Differently Expressed Genes (DEGs) and Enrichment Analysis

DEGs of the CAF subgroups were identified by the FindMarkers function of Seurat, with cut-offs set as fold change (FC) > 1.5 and adj. *p-value* < 0.01. GO and KEGG enrichment analyses were carried out based on the DEGs, with an adj. *p-value* < 0.05 considered significant.

Communication Analysis for CAF Subtypes

The identifying and illustrating alterations in the intercellular signaling network (iTALK) R package is a novel tool for intercellular communication analysis based on scRNA-seq (Wang et al., 2019), which could capture highly abundant downregulated or upregulated ligand-receptor gene pairs. We applied iTALK to analyze the ligand-receptor communications among the CAF subgroups and identified a total of 2,648 known ligand-receptor gene pairs. For further analysis, we further divided these gene pairs into four groups, namely, cytokines/chemokines, immune checkpoint genes, growth factors, and the rest.

Combination of Bulk-Seq and scRNA-Seq Data

CIBERSORTx, which is also known as "digital cytometry", could infer the proportion of cell types by deconvoluting bulk RNA-seq data (Steen et al., 2020a). We applied CIBERSORTx to estimate the abundance of each CAF subtype in TCGA-COAD patients. We used the software X-tile to set the optimum cut-off values. Patients in each subgroup were divided into CAF-high abundance and CAF-low abundance groups. Univariate Cox regression analysis was performed to analyze the prognostic value of different CAF subtypes in the TCGA-COAD cohort.

Quantitative Reverse Transcription-Polymerase Chain Reaction

Twenty pairs of fresh colon cancer and paracanceroush tissues from the Fudan University (Shanghai, China) were collected and snap-frozen in liquid nitrogen between October 2020 and September 2021. The samples were then stored at −80°C for later qRT-PCR analysis. In brief, total RNA was extracted using TRIzol reagent (Takara Biotechnology Co., Ltd., Dalian, China). Primers used for qRT-PCR were designed using the Primer5 software. cDNA was prepared using a reverse transcription kit (Takara Biotechnology Co., Ltd.), and qRT-PCR was carried out using the TB Green Premix ExTaq kit and the Applied Biosystems Step One Plus Real-Time PCR system. Ct values were calculated based on housekeeping genes, *ACTB* and *GAPDH*. All the primers were purchased from Takara (Dalian, China) and showed in **Supplementary Table S1**.

Construction and Validation of a CAF-Related Prognostic Signature Model for Colon Cancer

The 825 highly expressed DEGs in CAFs were used to construct a prognostic signature model. The univariate Cox regression analysis was performed in the training dataset to identify OS-related genes with $p < 0.05$. Then, LASSO regression analysis was used to optimize the model to avoid overfitting. According to the calculated coefficients from the LASSO analysis, risk score was assigned to each colon cancer patient. Finally, all these colon cancer patients were divided into high- and low-risk groups based on their risk scores, with the median risk score as the cutoff. Kaplan-Meier survival curves and scatter plots were used to visualize OS in the high- and low-risk groups. AUC was used to evaluate the time-dependent predictive accuracy of our model in the training, internal and external testing datasets.

Independence and Accuracy Test of the Prognostic Signature Model

Multiple cox regression analyses of related clinical factors and risk scores of our prognostic model were based on the "survival" R package. Survival time and survival status, combined with other clinical factors, were used to predict the prognosis by drawing a nomogram established using the "rms" R package, which was then used to illustrate the calibration curve and evaluate the prediction accuracy of the model.

Immune Infiltration Analysis

The immune score for each sample was calculated using the ESTIMATE package (Steen et al., 2020b). The proportion of different cell types within each tumor sample was calculated using the xCell package with default parameters (Aran et al., 2017). The pathway enrichment score for each sample was estimated using the GSVA package (Hnzelmann et al., 2013). Based on the 50 hallmark pathway feature gene set from MsigDB (H collection) (Liberzon et al., 2015), GO enrichment analysis was carried out by applying the gseGO function of the GSVA package and the clusterProfiler package with "c5. all.v7.1. symbols.gmt" geneset used. For analyses using clusterProfiler, specific parameters were set as follows: ont = "BP", nPerm = 1,000, minGSSize = 100, MaxGSSize = 1,000, *p*-value cutoff = 0.05. All the significant pathways were shown in the bubble chart, while only the top five significant pathways were shown in the enrichment curve.

Prediction of Drug Sensitivity

We selected candidate drugs from the CGP database and then applied the pRRophetic package (Paul et al., 2014) to the expression profile of TCGA-COAD samples to predict the drug sensitivity of these candidate drugs. The drug sensitivity of these candidate drugs was further tested in colon cancer cells from the GDSC database (Geeleher et al., 2014). The drug sensitivity is indicated by an IC50 value, which represents the drug concentration when half of the tumor cells die. Low IC50 values indicate better drug sensitivity for colon cancer in our study. It should be noted that the IC50 here is a relative estimate of drug sensitivity. Its value may be less than 0 and does not correspond exactly to the drug concentration.

Data Visualization and Correlation Analysis

The processing of single-cell data was performed using the Linux platform. Transcriptome data such as those from the TCGA and GEO databases were processed using the Windows10 platform. All the rest analyses were performed using the R 4.0.1 platform. Data cleaning, deformation, and integration were performed using the mgsub, reshape and dplyr packages. Factors such as the cell proportion, risk score, immune infiltration ratio were visualized using the ggpubr and ggplot2 packages (Wickham et al., 2016). The color matching was carried out using the RColorBrewer package (Neuwirth and Neuwirth, 2014). Correlation analysis was implemented using the cor_test and stat_compare_means modules in the R package with default parameters.

DISCUSSION

Characterization of CAF Subtypes and Potential Mechanisms in Colon Cancer at the Cellular and Molecular Levels

High-dimensional single-cell RNA-seq data are valuable resources for study at single-cell level. The abundance of fibroblasts is very different between normal and tumor tissues, indicating their importance in tumor development. Traditional bulk RNA-seq is unable to distinguish different CAF subtypes at single-cell level. However, by combining sc-RNA-seq data and bulk RNA-seq data, we were able to identify different CAF subtypes from tumor tissues, such as those from the TCGA database. Using this approach, we successfully identified six CAF subtypes in CRC, which we named enCAF, adCAF vaCAF, meCAF, erCAF, and cyCAF, respectively. We further explored the prognostic significance of these CAF subtypes and discover that two of them, adCAF and cyCAF, were significantly associated with prognosis of colon cancer patients, with the adCAF subtype as a protective factor while the cyCAF subtype as a risk factor. Another two CAFs (enCAF and erCAF) were functioning synergistically and showed an indirect link with prognosis in colon cancer patients. Enrichment analysis revealed that the prognostic significance of enCAF and erCAF related to macrophages. These two subtypes were negatively correlated with the M1 and M2 macrophage infiltration, respectively. M1 macrophages can secrete pro-inflammatory cytokines and chemokines, and present antigens, thus enhancing immune response and surveillance. On the contrary, M2 macrophages can secrete inhibitory cytokines, thus reducing the immune response (Jiawei et al., 2021). Expression levels of these modulating factors were confirmed by qRT-PCR and the key genes in the enCAF and erCAF subtypes were differentially expressed in tumor and paracanceroush tissues. Representative gene *GREM1* was highly expressed in the enCAFs of tumor tissues, which was a risk factor; Representative genes *IFG1* were highly expressed in erCAF of paracanceroush tissues.

To analyze the intercellular communications among the CAF subgroups, we applied the iTALK R package to the scRNA-seq data. Regarding immune checkpoint-related genes, the TNF superfamily member 14 (*TNFSF14*)-lymphotoxin beta receptor (*LTBR*) gene pair was most significantly differentially expressed between the erCAF and other CAF subtypes. TNFSF14 could contribute to vascular and tertiary lymphoid structure formation (Skeate et al., 2020). TNFSF14-LTBR pathway plays a vital role in immune responses in the TME of many types of cancer, but this pathway has not been reported in TME of colon cancer, suggesting it might be an important immunotherapeutic target for CRC treatment. Regarding cytokine-related genes, the *IL6-F3* and *IL6-IL6ST* gene pairs were the most widespread (**Figure 3**). In several types of cancer, such as breast cancer and hepatocellular carcinoma, CAFs can secret IL6 to promote tumor progression (Dittmer and Dittmer, 2020; Jia et al., 2020). IL6 belongs to a class of polypeptides that can bind to specific, high-affinity cell membrane receptors, regulating multiple cellular functions. The *CTGF-ITGA5* gene pair, both encoding growth factors, was differentially expressed between different CAF subgroups. Interestingly, CTGF is a known multifunctional regulator in TME that can activate CAFs, promote angiogenesis and inflammation, thus acting as an oncogene in various types of cancer (Shen et al., 2021). ITGA5 is expressed in CAFs and is responsible for the tumor-promoting effect of CAFs in colon cancer (Lu et al., 2019). Therefore, targeting the CTGF-ITGA5 pathway is promising for colon cancer treatment in patients with a erCAF. Therefore, we have characterized the prognostic significance and potential mechanisms of CAFs in colon cancer at the cellular and molecular levels. Through the detailed description of specific CAF subgroups, the underlying mechanism of CAFs function was indicated, which could be potential therapeutic targets.

In short, we performed deeper bioinformatics analyses, redefined the CAF subtypes, explored the prognostic significance of different CAF subtypes, and carried out experimental validation of key genes in CAFs, to study the role of CAFs in colon cancer development. Other than indices such as tumor size and immune cell infiltration ratio, fibroblast types and ratios may be important prognostic markers for CRC. Understanding the specific roles of different CAF subtypes would be critical for the assessment of prognosis and tumor treatment.

A Prognostic Signature Model at the Genetic Level

Although we have proven that our CAF-related prognostic signature model is accurate for prognosis assessment and promising for providing treatment recommendations, bulk RNA-seq data were impossible to be applied to this model directly. To further investigate the prognostic value of CAF-related genes, we constructed a CAF-related signature model based on the TCGA-COAD cohort and validated this model using the GSE39582 cohort. With univariate regression analysis, we identified 825 DEGs using scRNA-seq data. Among these DEGs, 16 genes were differentially expressed in colon tumor and paracancerous tissues, which showed excellent prognostic significance in TCGA-COAD patients. Moreover, through lasso regression analysis, we further removed 5 genes that were redundant and thus 10 genes were used as prognostic genes, namely, *CACNA1C*, *COL4A5*, *ADRA2B*, *EGFR*, *LMBR1L*, *FZD7*, *PKM*, *IL20RB*, *PMCH,* and *EPHB2*. A previous study represent that *EGFR* is over expressed in activated CAFs, contributing to colon cancer development (Shin et al., 2019). In addition, some types of CAF from tumors with epithelial-to-mesenchymal transition can escape tyrosine kinase inhibitors (TKIs) mediated EGFR inhibition, suggesting that these types of CAF might relate to EGFR-TKI drug resistance (Mink et al., 2010). In breast cancer, the CAF-derived exosome was able to regulate the expression of *PKM* in cancer cells (Li et al., 2020). However, the associated autocrine signaling in CAFs has not been elucidated. Autocrine signaling-related genes, such as *CACNA1C*, *COL4A5*, *ADRA2B*, *FZD7*, *IL20RB*, *PMCH*, and *EPHB2*, were implicated in some types of cancer (Ikeda et al., 2006; Merlos-Suárez et al., 2011; Kiaii et al., 2013; Phan et al., 2017; Zhang et al., 2018; Cui et al., 2019; Ye et al., 2019).

Next, we validated the established prognostic signature model from the aspects of model effect, test set deviation, and clinical feature comparison. Significant differences were observed regarding the K-M survival curve between high- and low-risk groups in the internal and external testing datasets ($p \leq 0.0025$). In machine learning on test and training sets, model bias is very limited. In the multiple cox regression analysis, risk factors were significantly correlated with all clinical features ($p < 5e^{-8}$) except for the MSI mutation feature. Moreover, the model in the multivariate regression analysis was the substitute for all factors except age, with better prediction range in the nomogram.

In short, comprehensive correlation analyses between multiple prognostic factors and risk scores from our model were performed, and the molecular mechanisms of our model were elucidated. Instead of directly acting on T cells, our prognostic model indicated that CAFs were significantly correlated with CLPs, DCs, and MSCs. CLPs are lymphatic stem cells that can differentiate into T cells, B cells, and NK cells. As a dominant cell type in the intestinal tract, the high ratio of MSCs could explain tumor cells' low proportion and low activity. DCs are professional antigen-presenting cells (APCs) in the body, where immature DCs can efficiently ingest, process, and present antigens to effectively activate naive T cells, a process important for immune response. MSCs have the tendency to promote tumor development. For example, cytokines secreted by MSCs can inhibit the function of T cells. Probably due to the complex intercellular associations, the risk scores are negatively correlated with the immune scores inferred from multiple scoring algorithms. In addition, as revealed by our model, the prognostic effects relate to 10 classical pathways, including the HEDGEHOG, APICAL, NOTCH, MYCV1, E2F pathways and the translational initiation, protein-DNA subunit assembly, and G2/M-related cell cycle pathways.

In summary, this study established a prognostic model for colon cancer based on CAF-related signature genes, which shows excellent performance compared with models using traditional clinical features. The model is based on the development pathway of cancer and the interaction with various tumor microecological cells to achieve a unified mechanism with key test indicators such as immune score and tumor purity. The model could be a powerful tool for predicting the prognosis of colon cancer patients.

CAFs on Tumor Development and Treatment

TME is a complex local ecosystem that connects tumors and other parts of the body (Qian et al., 2020). Different from T cells or macrophages that kill tumor cells directly, CAFs play roles in tumor development in an indirect way. Despite the fact that clinical treatment of colon cancer involves many complicated factors, our model could provide potential treatment recommendations based on the transcriptome profile of colon cancer patients. High expressed checkpoint-related genes indicate high activity of immunosuppressive pathways in patients of the high-risk group, who might benefit from treatment of antagonistic antibodies. Importantly, three drugs out of the 10 potential drugs in the GCP database, camptothecin, docetaxel, and bortezomib, may be potential candidates for colon cancer treatment in the low-risk group inferred from our model and may have a better therapeutic effect. In conclusion, starting from the identification of the subgroups of CAFs in colon cancer, by constructing a prediction model for the prognosis of colon cancer patients and prediction of drug sensitivity based on genomics data, the current research was expected to provide new directions and ideas for the CAF-related targeted therapy for colon 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.

AUTHOR CONTRIBUTIONS

All authors listed have made a substantial, direct, and intellectual contribution to the work and approved it for publication. ZZ and WL designed the project and carried out the bioinformatics analysis. LZ and BX performed database collection, data cleaning and mapping. YJ, NM, LL, and JQ completed the experimental work, MZ reviewed and directed the research plan and research work.

FUNDING

National Natural Science Foundation of China (81671500).

ACKNOWLEDGMENTS

We acknowledge TopEdit LLC for the linguistic editing and proofreading during the preparation of this manuscript. We thank Jianming Zeng (University of Macau), and all the members of his bioinformatics team, biotrainee, for generously sharing their experience and codes.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2022.908957/full#supplementary-material

REFERENCES

Aran, D., Hu, Z., and Butte, A. (2017). xCell: Digitally Portraying the Tissue Cellular Heterogeneity Landscape. *Genome Biol.* 18 (1), 220. doi:10.1186/s13059-017-1349-1

Becht, E., McInnes, L., Healy, J., Dutertre, C.-A., Kwok, I. W. H., Ng, L. G., et al. (2019). Dimensionality Reduction for Visualizing Single-Cell Data Using UMAP. *Nat. Biotechnol.* 37 (1), 38–44. doi:10.1038/nbt.4314

Bu, L., Baba, H., Yoshida, N., Miyake, K., Yasuda, T., Uchihara, T., et al. (2019). Biological Heterogeneity and Versatility of Cancer-Associated Fibroblasts in the Tumor Microenvironment. *Oncogene* 38 (25), 4887–4901. doi:10.1038/s41388-019-0765-y

Cui, X.-F., Cui, X.-G., and Leng, N. (2019). Overexpression of Interleukin-20 Receptor Subunit Beta (IL20RB) Correlates with Cell Proliferation, Invasion and Migration Enhancement and Poor Prognosis in Papillary Renal Cell Carcinoma. *J. Toxicol. Pathol.* 32 (4), 245–251. doi:10.1293/tox.2019-0017

Dittmer, A., and Dittmer, J. (2020). Carcinoma-Associated Fibroblasts Promote Growth of Sox2-Expressing Breast Cancer Cells. *Cancers (Basel)* 12 (11), 3435. doi:10.3390/cancers12113435

Geeleher, P., Cox, N. J., and Huang, R. S. (2014). Clinical Drug Response Can Be Predicted Using Baseline Gene Expression Levels and *In Vitro* Drug Sensitivity in Cell Lines. *Genome Biol.* 15 (3), R47–R12. doi:10.1186/gb-2014-15-3-r47

Gustafson, A. M., Soldi, R., Anderlind, C., Scholand, M. B., Qian, J., Zhang, X., et al. (2010). Airway PI3K Pathway Activation Is an Early and Reversible Event in Lung Cancer Development. *Sci. Transl. Med.* 2 (26), 26ra25. doi:10.1126/scitranslmed.3000251

Hao, Y., Hao, S., Andersen-Nissen, E., Mauck, W. M., Zheng, S., Butler, A., et al. (2021). Integrated Analysis of Multimodal Single-Cell Data. *Cell.* 184 (13), 3573–3587. e29. doi:10.1016/j.cell.2021.04.048

Hnzelmann, S., Castelo, R., and Guinney, J. (2013). GSVA: Gene Set Variation Analysis for Microarray and RNA-Seq Data. *Bmc Bioinforma.* 14 (1), 7. doi:10.1186/1471-2105-14-7

Hui, L., and Chen, Y. (2015). Tumor Microenvironment: Sanctuary of the Devil. *Cancer Lett.* 368 (1), 7–13. doi:10.1016/j.canlet.2015.07.039

Ikeda, K., Iyama, K.-i., Ishikawa, N., Egami, H., Nakao, M., Sado, Y., et al. (2006). Loss of Expression of Type IV Collagen α5 and α6 Chains in Colorectal Cancer Associated with the Hypermethylation of Their Promoter Region. *Am. J. Pathology* 168 (3), 856–865. doi:10.2353/ajpath.2006.050384

Jia, C., Wang, G., Wang, T., Fu, B., Zhang, Y., Huang, L., et al. (2020). Cancer-associated Fibroblasts Induce Epithelial-Mesenchymal Transition via the Transglutaminase 2-dependent IL-6/IL6R/STAT3 axis in Hepatocellular Carcinoma. *Int. J. Biol. Sci.* 16 (14), 2542–2558. doi:10.7150/ijbs.45446

Jiawei, Z., Xia, D. C., Dong, L. C., Yang, L., Kun, Y. J., Feng, D. H., et al. (2021). M2 Subtype Tumor Associated Macrophages (M2-TAMs) Infiltration Predicts Poor Response Rate of Immune Checkpoint Inhibitors Treatment for Prostate Cancer. *Ann. Med.* 53 (1), 730–740. doi:10.1080/07853890.2021.1924396

Kalluri, R. (2016). The Biology and Function of Fibroblasts in Cancer. *Nat. Rev. Cancer* 16 (9), 582–598. doi:10.1038/nrc.2016.73

Kasi, P. M., Shahjehan, F., Cochuyt, J. J., Li, Z., Colibaseanu, D. T., and Merchea, A. (2019). Rising Proportion of Young Individuals with Rectal and Colon Cancer. *Clin. Colorectal Cancer* 18 (1), e87–e95. doi:10.1016/j.clcc.2018.10.002

Kiaii, S., Clear, A. J., Ramsay, A. G., Davies, D., Sangaralingam, A., Lee, A., et al. (2013). Follicular Lymphoma Cells Induce Changes in T-Cell Gene Expression and Function: Potential Impact on Survival and Risk of Transformation. *Jco* 31 (21), 2654–2661. doi:10.1200/jco.2012.44.2137

Kobayashi, H., Gieniec, K. A., Wright, J. A., Wang, T., Asai, N., Mizutani, Y., et al. (2021). The Balance of Stromal BMP Signaling Mediated by GREM1 and ISLR Drives Colorectal Carcinogenesis. *Gastroenterology* 160 (4), 1224–1239. e30. doi:10.1053/j.gastro.2020.11.011

Lee, H.-O., Hong, Y., Etlioglu, H. E., Cho, Y. B., Pomella, V., Van den Bosch, B., et al. (2020). Lineage-dependent Gene Expression Programs Influence the Immune Landscape of Colorectal Cancer. *Nat. Genet.* 52 (6), 594–603. doi:10.1038/s41588-020-0636-z

Li, Y., Zhao, Z., Liu, W., and Li, X. (2020). SNHG3 Functions as miRNA Sponge to Promote Breast Cancer Cells Growth through the Metabolic Reprogramming. *Appl. Biochem. Biotechnol.* 191 (3), 1084–1099. doi:10.1007/s12010-020-03244-7

Liberzon, A., Birger, C., Thorvaldsdóttir, H., Ghandi, M., Mesirov, J. P., and Tamayo, P. (2015). The Molecular Signatures Database Hallmark Gene Set Collection. *Cell. Syst.* 1 (6), 417–425. doi:10.1016/j.cels.2015.12.004

Lu, L., Xie, R., Wei, R., Cai, C., Bi, D., Yin, D., et al. (2019). Integrin α5 Subunit Is Required for the Tumor Supportive Role of Fibroblasts in Colorectal Adenocarcinoma and Serves as a Potential Stroma Prognostic Marker. *Mol. Oncol.* 13 (12), 2697–2714. doi:10.1002/1878-0261.12583

Marisa, L., de Reyniès, A., Duval, A., Selves, J., Gaub, M. P., Vescovo, L., et al. (2013). Gene Expression Classification of Colon Cancer into Molecular Subtypes: Characterization, Validation, and Prognostic Value. *Plos Med.* 10 (5), e1001453. doi:10.1371/journal.pmed.1001453

Merlos-Suárez, A., Barriga, F. M., Jung, P., Iglesias, M., Céspedes, M. V., Rossell, D., et al. (2011). The Intestinal Stem Cell Signature Identifies Colorectal Cancer Stem Cells and Predicts Disease Relapse. *Cell. Stem Cell.* 8 (5), 511–524. doi:10.1016/j.stem.2011.02.020

Mink, S. R., Vashistha, S., Zhang, W., Hodge, A., Agus, D. B., and Jain, A. (2010). Cancer-associated Fibroblasts Derived from EGFR-TKI-Resistant Tumors Reverse EGFR Pathway Inhibition by EGFR-TKIs. *Mol. Cancer Res.* 8 (6), 809–820. doi:10.1158/1541-7786.mcr-09-0460

Moghimi-Dehkordi, B., and Safaee, A. (2012). An Overview of Colorectal Cancer Survival Rates and Prognosis in Asia. *Wjgo* 4 (4), 71–75. doi:10.4251/wjgo.v4.i4.71

Network, C. G. A. (2012). Comprehensive Molecular Characterization of Human Colon and Rectal Cancer. *Nature* 487 (7407), 330–337. doi:10.1038/nature11252

Neuwirth, E., and Neuwirth, M. E. (2014). Package 'RColorBrewer'. ColorBrewer Palettes.

Paul, G., Cox, N., and Huang, R. S. (2014). pRRophetic: An R Package for Prediction of Clinical Chemotherapeutic Response from Tumor Gene Expression Levels. *Plos One* 9 (9), e107468. doi:10.1371/journal.pone.0107468

Phan, N. N., Wang, C.-Y., Chen, C.-F., Sun, Z., Lai, M.-D., and Lin, Y.-C. (2017). Voltage-gated Calcium Channels: Novel Targets for Cancer Therapy. *Oncol. Lett.* 14 (2), 2059–2074. doi:10.3892/ol.2017.6457

Pietras, K., and Ostman, A. (2010). Hallmarks of Cancer: Interactions with the Tumor Stroma. *Exp. Cell. Res.* 316 (8), 1324–1331. doi:10.1016/j.yexcr.2010.02.045

Qian, J., Olbrecht, S., Boeckx, B., Vos, H., Laoui, D., Etlioglu, E., et al. (2020). A Pan-Cancer Blueprint of the Heterogeneous Tumor Microenvironment Revealed by Single-Cell Profiling. *Cell. Res.* 30 (9), 745–762. doi:10.1038/s41422-020-0355-0

Qiu, X., Hill, A., Packer, J., Lin, D., Ma, Y.-A., and Trapnell, C. (2017). Single-cell mRNA Quantification and Differential Analysis with Census. *Nat. Methods* 14 (3), 309–315. doi:10.1038/nmeth.4150

Quail, D. F., and Joyce, J. A. (2013). Microenvironmental Regulation of Tumor Progression and Metastasis. *Nat. Med.* 19 (11), 1423–1437. doi:10.1038/nm.3394

Shen, Y.-W., Zhou, Y.-D., Chen, H.-Z., Luan, X., and Zhang, W.-D. (2021). Targeting CTGF in Cancer: an Emerging Therapeutic Opportunity. *Trends cancer* 7 (6), 511–524. doi:10.1016/j.trecan.2020.12.001

Shin, N., Son, G. M., Shin, D.-H., Kwon, M.-S., Park, B.-S., Kim, H.-S., et al. (2019). Cancer-Associated Fibroblasts and Desmoplastic Reactions Related to Cancer Invasiveness in Patients with Colorectal Cancer. *Ann. Coloproctol.* 35 (1), 36–46. doi:10.3393/ac.2018.09.10

Siegel, R. L., Miller, K. D., Goding Sauer, A., Fedewa, S. A., Butterly, L. F., Anderson, J. C., et al. (2020b). Colorectal Cancer Statistics, 2020. *CA A Cancer J. Clin.* 70 (3), 145–164. doi:10.3322/caac.21601

Siegel, R. L., Miller, K. D., and Jemal, A. (2020a). Cancer Statistics, 2020. *CA A Cancer J. Clin.* 70 (1), 7–30. doi:10.3322/caac.21590

Skeate, J. G., Otsmaa, M. E., Prins, R., Fernandez, D. J., Da Silva, D. M., and Kast, W. M. (2020). TNFSF14: LIGHTing the Way for Effective Cancer Immunotherapy. *Front. Immunol.* 11 (922), 922. doi:10.3389/fimmu.2020.00922

Steen, C. B., Liu, C. L., Alizadeh, A. A., and Newman, A. M. (2020a). Profiling Cell Type Abundance and Expression in Bulk Tissues with CIBERSORTx. *Methods Mol. Biol.* 2117, 135–157. doi:10.1007/978-1-0716-0301-7_7

Steen, C. B., Liu, C. L., Alizadeh, A. A., and Newman, A. M. (2020b). "Profiling Cell Type Abundance and Expression in Bulk Tissues with CIBERSORTx," in *Stem Cell Transcriptional Networks* (Springer), 135–157. doi:10.1007/978-1-0716-0301-7_7

Wang, L., Hong, W., and Zhang, S. (2010). Clinical Significance of the Upregulated Osteopontin mRNA Expression in Human Colorectal Cancer. *J. Gastrointest. Surg.* 14 (1), 74–81. doi:10.1007/s11605-009-1035-z

Wang, Y., Wang, R., Zhang, S., Song, S., Jiang, C., Han, G., et al. (2019). iTALK: an R Package to Characterize and Illustrate Intercellular Communication. *Am. Soc. Hematol.*

Wickham, H. (2016). *ggplot2: Elegant Graphics for Data Analysis*. Verlag, New York: Springer. ISBN 978-3-319-24277-4 Available at: https://ggplot2.tidyverse.org

Wolf, A. M. D., Fontham, E. T. H., Church, T. R., Flowers, C. R., Guerra, C. E., LaMonte, S. J., et al. (2018). Colorectal Cancer Screening for Average-Risk Adults: 2018 Guideline Update from the American Cancer Society. *CA A Cancer J. Clin.* 68 (4), 250–281. doi:10.3322/caac.21457

Ye, C., Xu, M., Lin, M., Zhang, Y., Zheng, X., Sun, Y., et al. (2019). Overexpression of FZD7 Is Associated with Poor Survival in Patients with Colon Cancer. *Pathology - Res. Pract.* 215 (8), 152478. doi:10.1016/j.prp.2019.152478

Zhang, Y., Fang, L., Zang, Y., and Xu, Z. (2018). Identification of Core Genes and Key Pathways via Integrated Analysis of Gene Expression and DNA Methylation Profiles in Bladder Cancer. *Med. Sci. Monit.* 24, 3024–3033. doi:10.12659/msm.909514

Zhou, Y., Bian, S., Zhou, X., Cui, Y., Wang, W., Wen, L., et al. (2020). Single-cell Multiomics Sequencing Reveals Prevalent Genomic Alterations in Tumor Stromal Cells of Human Colorectal Cancer. *Cancer Cell.* 38 (6), 818–828. e5. doi:10.1016/j.ccell.2020.09.015

Conflict of Interest: Author ZZ, LZ, BX, NM, LL, and MZ were employed by the Zhengzhou KingMed Center for Clinical Laboratory Co. Ltd.

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

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: 15 July 2022
doi: 10.3389/fgene.2022.919264

BNEMDI: A Novel MicroRNA–Drug Interaction Prediction Model Based on Multi-Source Information With a Large-Scale Biological Network

Yong-Jian Guan[1], Chang-Qing Yu[1], Li-Ping Li[1,2]*, Zhu-Hong You[3], Zhong-Hao Ren[1], Jie Pan[4] and Yue-Chao Li[1]*

[1]*School of Information Engineering, Xijing University, Xi'an, China,* [2]*College of Grassland and Environment Sciences, Xinjiang Agricultural University, Urumqi, China,* [3]*School of Computer Science, Northwestern Polytechnical University, Xi'an, China,* [4]*Key Laboratory of Resources Biology and Biotechnology in Western China, Ministry of Education, College of Life Science, Northwest University, Xi'an, China*

OPEN ACCESS

Edited by:
Jia Meng,
Xi'an Jiaotong-Liverpool University,
China

Reviewed by:
Stephen S.-T. Yau,
Tsinghua University, China
Malik Yousef,
Zefat Academic College, Israel
Yuping Sun,
Guangdong University of Technology,
China

****Correspondence:***
Li-Ping Li
cs2bioinformatics@gmail.com
Chang-Qing Yu
xaycq@163.com

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

Received: *13 April 2022*
Accepted: *30 May 2022*
Published: *15 July 2022*

Citation:
Guan Y-J, Yu C-Q, Li L-P, You Z-H, Ren Z-H, Pan J and Li Y-C (2022) BNEMDI: A Novel MicroRNA–Drug Interaction Prediction Model Based on Multi-Source Information With a Large-Scale Biological Network. Front. Genet. 13:919264. doi: 10.3389/fgene.2022.919264

As a novel target in pharmacy, microRNA (miRNA) can regulate gene expression under specific disease conditions to produce specific proteins. To date, many researchers leveraged miRNA to reveal drug efficacy and pathogenesis at the molecular level. As we all know that conventional wet experiments suffer from many problems, including time-consuming, labor-intensity, and high cost. Thus, there is an urgent need to develop a novel computational model to facilitate the identification of miRNA–drug interactions (MDIs). In this work, we propose a novel bipartite network embedding-based method called BNEMDI to predict MDIs. First, the Bipartite Network Embedding (BiNE) algorithm is employed to learn the topological features from the network. Then, the inherent attributes of drugs and miRNAs are expressed as attribute features by MACCS fingerprints and *k*-mers. Finally, we feed these features into deep neural network (DNN) for training the prediction model. To validate the prediction ability of the BNEMDI model, we apply it to five different benchmark datasets under five-fold cross-validation, and the proposed model obtained excellent AUC values of 0.9568, 0.9420, 0.8489, 0.8774, and 0.9005 in ncDR, RNAInter, SM2miR1, SM2miR2, and SM2miR MDI datasets, respectively. To further verify the prediction performance of the BNEMDI model, we compare it with some existing powerful methods. We also compare the BiNE algorithm with several different network embedding methods. Furthermore, we carry out a case study on a common drug named 5-fluorouracil. Among the top 50 miRNAs predicted by the proposed model, there were 38 verified by the experimental literature. The comprehensive experiment results demonstrated that our method is effective and robust for predicting MDIs. In the future work, we hope that the BNEMDI model can be a reliable supplement method for the development of pharmacology and miRNA therapeutics.

Keywords: miRNA–drug interaction, BiNE, k-mer, MACCS fingerprint, deep neural network

INTRODUCTION

As many previous studies have shown, RNA plays a vital role in encoding, decoding, regulation, and expression of genes (Fu, 2014). Global transcriptional analyses of the human genome proved that the quantity of non-coding RNAs (ncRNA) is much larger than protein in human cells and ncRNA is involved in the regulation of stem cell pluripotency and cell division (Cawley et al., 2004; Iyer et al., 2015). In the human genome project, the newly discovered RNA genes are far more abundant than protein genes (Bentwich et al., 2005). RNA can be divided into two classes based on the length of the RNA chain, mainly including long RNA of more than 200 nucleotides and small RNA of fewer than 200 nucleotides. MicroRNAs (miRNAs) are a kind of short endogenous non-coding RNAs with 20–25 nucleotides, which may modulate the expression of genes in post-transcription (Ambros, 2001; Bartel, 2004). MiRNAs will incompletely bind to the target genes for inhibiting the transcripts, which may truncate mRNAs but does not affect the stability of mRNAs (Jiang et al., 2009).

Despite great advances in miRNA therapeutics and the theoretical knowledge between miRNAs and diseases, most of the drug targets are proteins. In human cells, less than 15% of disease-related proteins are targets of drugs (Dixon and Stockwell, 2009). This means that drug targets, which are designed through proteins, can only act on a small proportion of the human genome. In brief, most proteins are not “druggable.” As a result, ncRNAs are increasingly considered by researchers as a potential drug target. Among them, miRNA is considered a valuable drug target because it can play a key role in gene regulation when the disease occurs. Increasing number of experiments prove that there is a strong relationship between the abnormal regulation of miRNA and human diseases. For example, the expression level of miR-205 and miR-393 are potential biomarkers of mucinous colorectal cancer and colon cancers, which will be increased when cancer occurs (Eyking et al., 2016). Bommer et al. (2007) discovered that the expression level of miR-34 will be lessened in non-small cell lung cancers (Bommer et al., 2007). If miRNA can be used as drug targets, it will be conducive to the development of drug discovery and drug repositioning (Zhang et al., 2021b).

Therefore, many recent studies focus on the miRNA-based approach as a therapeutic, one of which is targeting over-expressed miRNAs (Ishida and Selaru, 2013; Bayraktar et al., 2018; Zhang et al., 2021). Kota et al. (2009) reported that miR-26a transported by adeno-associated virus (AAV) inhibits the spread of cancer cells and activates the apoptosis of cancer cells. In the previous study, Esquela-Kerscher et al. suggested that the active expression of let-7 could suppress the proliferation of tumor cells in the mice model (Fu et al., 2021). Matboli et al. (2017) demonstrated that caffeic can effectively attenuate diabetic kidney disease in rats by downregulating the expression level of miR-133b, miR-342, and miR30a (Matboli et al., 2017).

However, detecting MDI based on the experiment is a labor-intensive and time-consuming process. *In silico*, some of the prediction methods have been proposed to infer the potential interaction between miRNAs and drugs. For example, Huang et al. proposed a computational method named GCMDR, which is based on a graph convolution neural network and explores the link between miRNA and drug resistance (Ya et al., 2020). In detail, they constructed a bipartite graph integrating the fingerprint of drug compounds and miRNA functional similarity. Moreover, they learned from the idea of auto-encoder, in which they built a graph convolution-based encoder to generate the embeddings of nods and a decoder to complete the prediction task. Lv et al. (2015)constructed two homogeneous networks of miRNAs and small molecular drugs. Multiple similarity measurements (i.e., side effect, functional consistency, indication phenotype, and chemical structure) are fused to represent the node feature of miRNAs and drugs, and they implemented the improved random walk restart algorithm on the heterogeneous network, which is fused by two homogeneous networks. Thus, this method can infer the potential MDI without having to resort to the information of known MDI. But there are too many parameters required to adjust in this method. Recently, Deepthi and Jereesh, (2021) developed an ensemble approach of the convolutional neural network based on deep architecture-based classification for identifying the association between miRNAs and drugs. They treated the similarities of miRNAs and drug compounds as the biological features and reduced the dimensions of features by the PCA algorithm. Then, they constructed a convolutional deep neural network for the purpose of feature extraction. Finally, they employed SVM to predict the potential MDIs. Anyway, the aforementioned methods rely heavily on side information calculated by functional similarities such as gene functional similarity and disease phenotype similarity. Abuse of functional similarity carries the risk of label leakage. However, due to the incomplete database, a lot of side information about miRNA and drugs is missing. In most cases, researchers only have the sequence profile and phenotypic profile of biological molecules and chemical compound. Therefore, we think that an MDI prediction method based on the sequence profile rather than functional similarity should be designed.

The information on how miRNAs affect drug effects in the literature can also provide rich information (Fleuren and Alkema, 2015). Hence, Xie et al. (2017) proposed a novel text mining approach named EmDL to infer the MDIs by extracting the explicit information in the literature. They began by splitting substantial articles, which were collected from PubMed and MEDLINE, into individual sentences. For each miRNA–drug pair, the word distance between miRNA and drug appearing in the sentence was calculated to extract the representation features. Last, they leveraged the principal component analysis (PCA) algorithm to reduce the dimension of representation features and was carried out using the support vector machine (SVM) to predict whether the miRNA–drug pair was interactive (Deepthi and Jereesh, 2021). Moreover, Guo et al. (2020) creatively introduced natural language processing (NLP) to the field of biological information. For the purpose of mining the information from the chemical structure of biological entities, they regarded the miRNA sequences and drug SMILES sequences as sentences and implemented the word2vec algorithm for them. However, implementation of NLP methods required a large

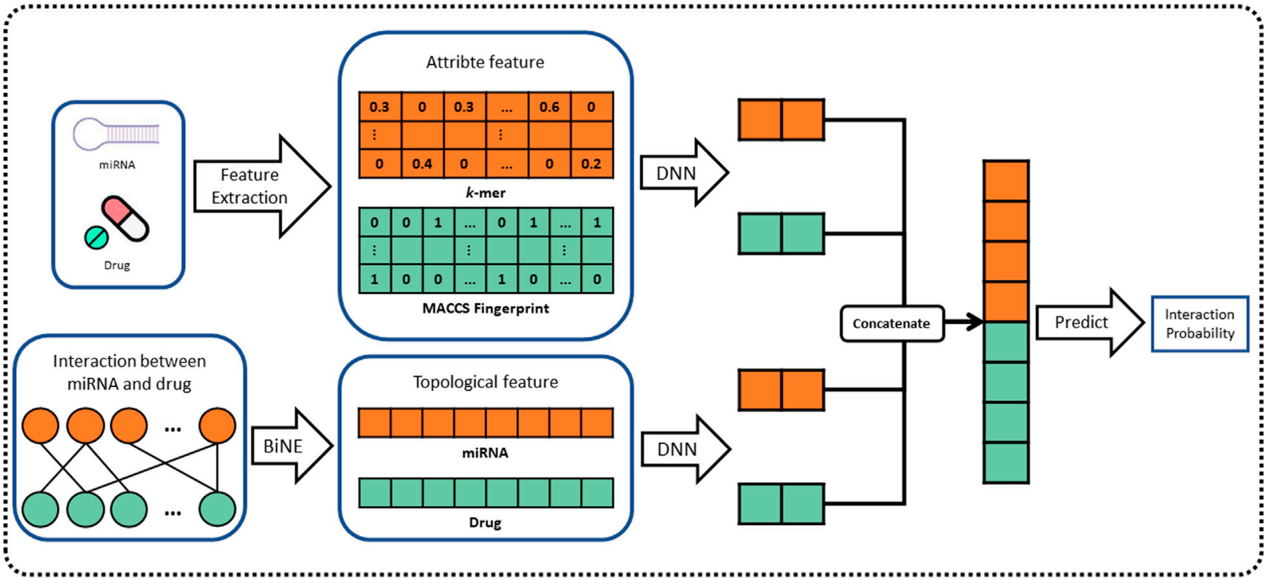


FIGURE 1 | Flowchart of the BNEMDI model for predicting potential MDIs.

corpus, and the performance of text mining-based methods will be affected by different corpora and different semantic statements.

In this work, we propose a novel computational method, named BNEMDI, which predicts miRNA–drug interactions using drug substructure fingerprint, miRNA sequence, and MDIs bipartite graph. We have collected known MDI from three databases (e.g., ncDR, RNAInter, and SM2miR) and split them into five datasets. In datasets, the MDI pairs were treated as positive samples, and the same number of unconfirmed miRNA–drug pairs was selected randomly as negative samples. The known MDIs in datasets were constructed as the bipartite graph, and the miRNAs and drug compounds are regarded as the nodes of the graph. The graph embedding methods are pervasive to reveal the complex traits of each entity (Li et al., 2021; Yue and He, 2021). Thus, a graph embedding technique called BiNE was implemented on the bipartite graph for learning the topological features of nodes (Gao et al., 2018), and BNEMDI considers not only the topological information of MDI but also the inherent attribute information of the biological entities. Specifically, the attribute features of drug compounds are denoted by MACCS substructure fingerprints, and the attribute features of miRNAs are calculated by k-mers (Kurtz et al., 2008; Cereto-Massagué et al., 2015). Finally, we constructed a neural network model based on DNN to fuse two kinds of features mentioned earlier and infer the potential miRNA–drug interaction pairs. The flowchart of BNEMDI is shown in **Figure 1**.

MATERIALS AND METHODS

Dataset

There are several databases about MDIs, for example, the RNA interaction dataset (RNAInter) (Kang et al., 2021), the database for non-coding RNAs involved in drug resistance (ncDR) (Dai et al., 2017), and the database of validated small molecules' effects on miRNA expression (SM2miR) (Liu X. et al., 2013).

TABLE 1 | Statistics of miRNAs, drugs, and miRNA–drug interactions in five datasets.

Dataset	ncDR	RNAInter	SM2miR1	SM2miR2	SM2miR3
Drug	95	281	86	113	142
miRNA	624	1,009	358	536	645
Interaction	4,457	5,739	1,110	1,697	1,940

We downloaded a total of 8,053 different experimentally verified miRNA–drug interactions from the three databases mentioned earlier. One thing is to note that the SM2miR database was created on 10 June 2012 and upgraded twice on 28 August 2013 and 27 April 2015. Thus, the SM2miR database was divided into three sub-datasets, according to three versions, named SM2miR1, SM2miR2, and SM2miR3 for convenience, respectively. Therefore, we obtained a total of five datasets and pre-processed them, such as de-redundancy and de-duplication. The details of the three databases are shown in **Table 1**. We only collected the miRNA–drug interaction pairs of *Homo sapiens* in three databases. The miRNA sequences and drug SMILES are collected from miRBase (Kozomara et al., 2019) and PubChem (Kim et al., 2021a). The drug SMILES is a specification that explicitly describes the molecular structure in ASCII strings (Weininger and sciences, 1988). The drug SMILES are transformed into MACCS fingerprints by the RDKit library.

Represent MicroRNA With *k*-mer

For obtaining genomic information on miRNA, the sequence of miRNA is represented by *k*-mer (Liu B. et al., 2013). *k*-mer is a

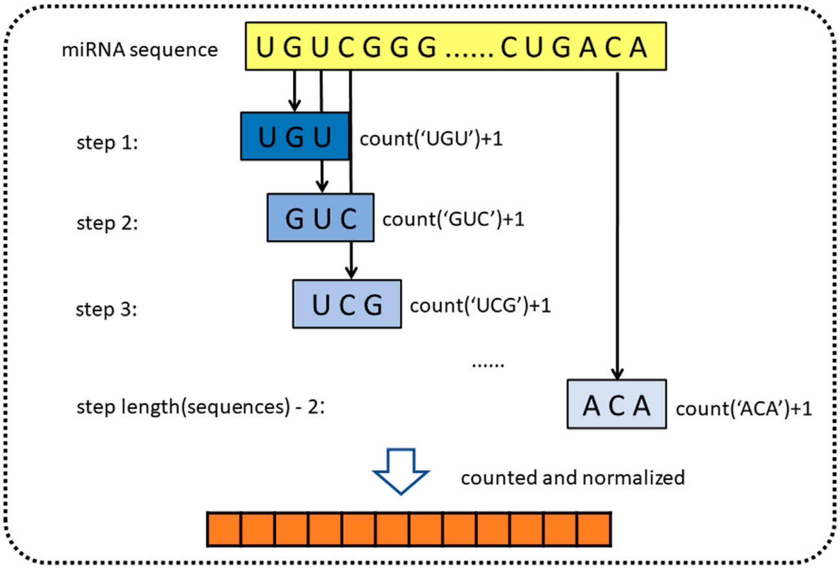


FIGURE 2 | Diagram of *k*-mer for extracting attribute sequences from miRNA sequences.

feature representation method, which is widely used in the field of bioinformatics. Yousef et al. (2017) used *k*-mer to construct simple sequence-based features to describe miRNAs for miRNA categorization (Yousef et al., 2017; Erten-Ela et al., 2018). In addition, Yi et al. (2020) also used *k*-mer to represent molecules such as lncRNA, miRNA, and protein in the molecule association network (Yi et al., 2020; Pan et al., 2022). *k*-mer is a substring of biological sequence with a length of *k*. For the miRNA sequences, we define the 3-mer of miRNA as the subsequence, such as "AGG" and "AAA." Then we sequentially extract three nucleotides from the first nucleotides, using the form of a sliding window (step length is one). Since miRNA consists of four types of bases, there are 64 (4^3) possible 3-mer patterns in a sequence. After that, we count the normalized frequencies of all 3-mer patterns. Finally, we obtained miRNA representation vectors with a length of 64 and containing miRNA sequence information. **Figure 2** shows the principle of *k*-mer.

Represent Drug Molecules With MACCS Fingerprint

In the past research, numerous kinds of descriptors have been established to portray the chemical structure of pharmaceutical compounds such as geometrical, topological constitutional, and quantum chemical properties (Cao et al., 2012). The substructure keys-based fingerprint is customarily adopted as the descriptor to represent the chemical structure. Substructure fingerprints encode molecular structure to a bit-string with a fixed length, according to the substructure of the drug instead of using 3D structural information. Plenty of previous research works have demonstrated that substructure fingerprint is effective and feasible to represent drugs. Specifically, we incorporated a dictionary that includes a list of substructure features represented as SMART strings. SMART is a system to identify substructures by the expanding rule of SMILES. After the first step of composing the dictionary, we compare each item of the dictionary to the given molecular substructures, if the SMART pattern is included in the given molecular substructure, the corresponding bit of fingerprint is set to one, and zero otherwise. An example of the substructure fingerprint determined by the given molecular substructure is displayed in **Figure 3**. Herein, we used MACCS fingerprint to compose the dictionary, which contained 166 types of general molecular substructures and covered most of the interesting chemical structures of drugs. Finally, we represented Boolean vectors of molecular drug for the length of 166.

Topological Features Extraction Based on Graph Representation Method

In this study, the graph representation learning method may encode each node by topological information and embed nodes in a low-dimension space. It is different from previous studies, in which it can extract underlying information from the network.

The challenge of MDI prediction may be formulated as a link prediction problem with a heterogeneous graph. The MDIs network is employed to construct a heterogeneous graph $G = (D, M, E)$, and there are two types of nodes, drug $\mathrm{D} = \{\mathrm{d}_1, d_2, \ldots, d_i\}$ (i is the index of drugs in the dataset) and miRNA $M = \{m_1, m_2, \ldots, m_j\}$ (j is the index of miRNA in the dataset). $E \subset D \times M$ denotes the set of edges between D and M. The edges represent the known interactions between drugs and miRNAs. If d_i and m_j have interaction, the weight of the edge is set to one, and zero otherwise. The matrix $W = [w_{ij}]$ denotes the weight of the edges between drug d^i and miRNA m^j in graph G. The graph embedding aims to look for a map function

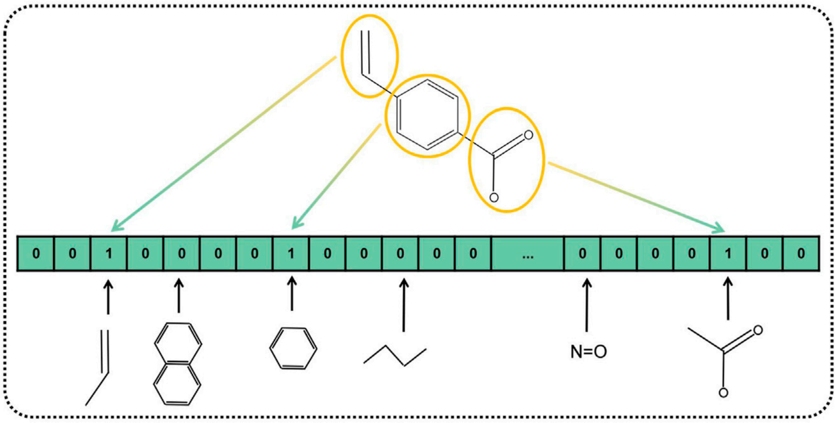


FIGURE 3 | Diagram of MACCS fingerprint-represented drug substructures.

$f: D \cup M \rightarrow R^t$, where $t < < |m| \cup |d|$. In other words, the low-dimensional presentation vectors of each node in the graph will be learned by the graph embedding method, and maintain the graph topology information and node properties (Cai et al., 2018). To achieve this aim, we utilized a graph embedding method called BiNE, which has great performance in reconstructing the original bipartite network, proposed by Gao et al. (2018). Previous research on graph embedding has raised the question of extracting explicit relations between the nodes of different types and implicit relationships among the nodes of sample types. BiNE contributed an innovative idea to solve this problem by constructing a jointly optimizing framework, consisting of three objection functions and variable weight. These three objective functions include an explicit relation and two implicit relations.

To model the explicit relations, we calculated the local proximity between two different vertices in the bipartite network, which is based on local proximity in LINE (Tang et al., 2015). We define the joint probability between two connected nodes as:

$$P(i,j) = \frac{w_{ij}}{\sum_{e_{ij}\in E} w_{ij}} \tag{1}$$

where w_{ij} is the weight of the edge between two types of nodes.

Drawing on the principle of word2vec, BiNE estimates the local proximity between two nodes by inner product (Church, 2017), and the sigmoid function is used to map the interaction value to the probability space. The joint probability of two different types of nodes in embedding space is defined as follows:

$$\hat{P}(i,j) = \frac{1}{1+\exp\left(-\mathrm{d}_i^T \mathrm{m}_j\right)}, \tag{2}$$

where $\vec{d}_i \in R^t$ and $\vec{m}_j \in R^t$ are the embedding vectors of drugs d_i and miRNAs m_j, respectively.

To get the knowledge of observed edges and learn the embedding vectors, we need to minimize the difference between empirical distribution and the reconstructed distribution. KL-divergence is used to measure the difference between the previously two joint probabilities mentioned. The first part of the joint optimizing framework can be defined as:

$$minimize\ O_1 = KL\left(P \middle\| \widehat{P}\right) = \sum_{e_{ij}\in E} P(i,j) \log\left(\frac{P(i,j)}{\widehat{P}(i,j)}\right). \tag{3}$$

Studies of recommendation systems demonstrated that implicit relations are also helpful to discover potential information in the heterogeneous graph as explicit relations (Jiang et al., 2016; Yu et al., 2018). This means that nodes of the same type are not connected in the bipartite network, but still contain a wealth of information, that is, crucial to model the implicit relationship between the nodes of the same type. BiNE constructs two homogeneous networks in accordance with the interaction profile between two types of nodes and performs the random walk on two homogeneous networks to encode the high-order proximity of the origin network.

To reveal the 2nd proximity of the heterogeneous graph, BiNE utilizes co-HITS (Deng et al., 2009) to generate two weighted homogeneous networks (drug–drug network and miRNA–miRNA network). In accordance with co-HITS, the correlation coefficient between two nodes can be defined as:

$$w_{ij}^M = \sum_{k\in D} w_{ik} w_{jk};\ w_{ij}^D = \sum_{k\in M} w_{ki} w_{kj} \tag{4}$$

where w_{ij} is the weight of the edge e_{ij}. Intuitively, suppose an $i \times j$ MDI bipartite matrix G_b, the drug–drug network can be denoted by a $i \times i$ matrix $G_b G_b^T$, and the miRNA–miRNA network can be represented by a $j \times j$ matrix $G_b^T G_b$.

Truncated random walks are employed on two homogeneous networks previously generated to obtain the corpus of node sequences. Therefore, the biased and self-adaptive random walk generator, which may maintain the vertex distribution, is introduced to produce the corpus of node sequences with true validity and effectiveness. Its core design can be described as "richer get richer." Specifically, the greater centrality of a node, the more likely that node will be the starting point for the random walk to begin. The centrality of nodes in the homogeneous network is measured by HITS (Kleinberg, 1999). Compared with other random walk-based measures, a probability is specified to stop the random walk in each step. Therefore, the node sequence generated by our method does not have a fixed length because the variable-length sequences are more simulated to natural language.

The skip-gram model is carried out to process the samples of two corpora obtained from truncated random walk. If two nodes frequently appear in the same context of a node sequence, the skip-gram model will assign them similar embedding vectors.

In order to learn the implicit relations, two objection functions are defined to maintain the high-order proximity by maximizing the conditional probability. The symbols of $C_S(d_i)$ and $C_S(m_j)$ represent the context of node d_i and m_j in a sequence S, respectively. For the corpus of the drug homogeneous network P^D, the objection function is as follows:

$$maximize\ O_2 = \prod_{d_i \in S \wedge P^D} \prod_{d_c \in C_S(d_i)} P(d_c|d_i) \tag{5}$$

Then, the corpus of the miRNA homogeneous network P^M is treated in the same way, and the objection function is expressed as:

$$maximize\ O_3 = \prod_{m_j \in S \wedge P^M} \prod_{m_c \in C_S(m_j)} P(m_c|m_i) \tag{6}$$

Similar to the LINE (Tang et al., 2015), the conditional probability $P(d_c|d_i)$ and $P(m_c|m_i)$ are defined using the inter product kernel and softmax function:

$$P(d_c|d_i) = \frac{\exp\left(\vec{d}_i^T \vec{\theta}_c\right)}{\sum_{k=1}^{|D|}\left(\vec{d}_i^T \vec{\theta}_k\right)}, P(m_c|m_i) = \frac{\exp\left(\vec{m}_j^T \vec{\vartheta}_c\right)}{\sum_{k=1}^{|M|}\left(\vec{m}_j^T \vec{\vartheta}_k\right)} \tag{7}$$

where $|D|$ and $|M|$ represent the number of drug compounds and miRNAs, respectively. The context vectors corresponding to two types of nodes are denoted as $\vec{\theta}_c$ and $\vec{\vartheta}_c$.

Finally, the three components of the objective function are combined into the joint optimization framework for learning the low-dimension embedding vectors of the bipartite network. The overall jointly optimizing function is defined as follows:

$$maximize\ L = \alpha \log O_2 + \beta \log O_3 - \gamma O_1 \tag{8}$$

where α, β, and γ are parameters of explicit relation and implicit relation.

To improve computational efficiency, a negative sampling method is adopted to approach the complicated denominator of the sigmoid function. In particular, nodes are divided into different buckets by locality-sensitive hashing (LSH) (Wang et al., 2013) and randomly selected as the negative samples. Finally, the joint framework is optimized by the stochastic gradient ascent (SGA) algorithm. The first part of the optimizing framework $L_1 = -\gamma O_1$ is maximized to update embedding vectors $\vec{d}_i$ and $\vec{m}_j$, and the updated rules of embedding vectors $\vec{d}_i$ and $\vec{m}_j$ are expressed as follows:

$$\vec{d}_i = \vec{d}_i + \lambda\left\{\gamma w_{ij}\left[1 - \sigma\left(\vec{d}_i^T \vec{m}_j\right)\right] \cdot \vec{m}_j\right\} \tag{9}$$

$$\vec{m}_j = \vec{m}_j + \lambda\left\{\gamma w_{ij}\left[1 - \sigma\left(\vec{d}_i^T \vec{m}_j\right)\right] \cdot \vec{d}_i\right\} \tag{10}$$

where λ represents the learning rate, and σ represents the sigmoid function. Then, the part of $\alpha \log O_2$ and $\beta \log O_3$ are also maximized to update embedding vectors $\vec{d}_i$ and $\vec{m}_j$ to follow the rules:

$$\vec{d}_i = \vec{d}_i + \lambda\left\{\sum_{z \in \{d_c\} \cup N_S^{ns}(d_i)} \alpha\left[I(z, d_i) - \sigma\left(\vec{d}_i^T \vec{\theta}_z\right)\right] \bullet \vec{\theta}_z\right\}, \tag{11}$$

$$\vec{m}_j = \vec{m}_j + \lambda\left\{\sum_{z \in \{m_c\} \cup N_S^{ns}(m_j)} \beta\left[I\left(z, m_j\right) - \sigma\left(\vec{m}_j^T \vec{\vartheta}_z\right)\right] \bullet \vec{\vartheta}_z\right\}, \tag{12}$$

where $I(z, d_i)$ and $I(z, m_j)$ is an indicator function that confirms whether the node z belongs to the context of d_i and m_j, respectively. The context of nodes is updated as:

$$\vec{\theta}z = \vec{\theta}z + \lambda\left\{\alpha\left[I(z, d_i) - \sigma\left(\vec{d}_i^T \vec{\theta}_z\right)\right] \bullet \vec{d}_i\right\} \tag{13}$$

$$\vec{\vartheta}z = \vec{\vartheta}z + \lambda\left\{\beta\left[I\left(z, m_j\right) - \sigma\left(\vec{m}_j^T \vec{\vartheta}_z\right)\right] \bullet \vec{m}_j\right\} \tag{14}$$

Building Predictor

In this section, we will introduce how to predict whether the miRNA–drug pairs have underlying interaction. After feature extraction, the attribute and topological features of miRNAs and drugs were concatenated and fed into the DNN model for fusing as unified dimension representation vectors. Finally, a dense layer with 256 neurons is used to complete the classification task. Specifically, suppose that the nodes miRNA and nodes drug are d_i and m_j, and the representation features of them are f_i and f_j, respectively. The possibility of interaction between d_i and m_j can be defined as:

$$P^{ij} = \sigma\left(f_i^T \oplus f_j\right) \tag{15}$$

where σ means the sigmoid function and $\oplus$ means the concatenation. P^{ij} represent the prediction score between d_i and m_j, if the P^{ij} is greater than 0.5 means, d_i is to interact with m_j, and vice versa. The binary cross-entropy was used as the loss function, and the “Adam” algorithm was used to optimize the model.

RESULTS AND DISCUSSION

Evaluation Criteria

As MDI prediction is a binary classification problem for each pair of miRNA and drugs, we used some evaluation criteria to measure the performance of the proposed model, including accuracy (Acc.), sensitivity (Sen), specificity (Spec.), also precision (Prec.), and Matthews correlation coefficient (MCC). They are defined as:

TABLE 2 | Performance of the proposed method on five datasets.

Fold	AUC	AUPR (%)	Acc (%)	Sen (%)	Spec (%)	Prec (%)	MCC (%)
ncDR	0.9568 ± 0.0010	95.65 ± 0.13	88.75 ± 0.10	89.13 ± 0.19	88.39 ± 0.13	88.47 ± 0.11	77.51 ± 0.20
RNAInter	0.9420 ± 0.0016	93.88 ± 0.16	87.23 ± 0.34	88.99 ± 1.05	85.47 ± 1.30	85.98 ± 0.94	74.52 ± 0.65
SM2miR1	0.8489 ± 0.0021	84.61 ± 0.25	77.24 ± 0.39	80.82 ± 0.33	73.66 ± 0.68	75.42 ± 0.49	54.62 ± 0.78
SM2miR2	0.8774 ± 0.0023	87.04 ± 0.17	79.92 ± 0.21	81.12 ± 0.33	78.73 ± 0.23	79.22 ± 0.19	59.86 ± 0.42
SM2miR3	0.9005 ± 0.0026	89.34 ± 0.20	81.86 ± 0.65	79.47 ± 1.42	84.24 ± 2.05	83.49 ± 1.62	63.81 ± 1.37

$$Acc. = \frac{TN + TP}{TN + TP + FN + FP} \tag{16}$$

$$Sen. = \frac{TP}{FP + FN}, \tag{17}$$

$$Spec. = \frac{TN}{TN + FP}, \tag{18}$$

$$Prec. = \frac{TP}{TP + FP} \tag{19}$$

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \tag{20}$$

Here, TP and TN are signs of the number of correct positive samples and correct negative samples predicted by the model, respectively. Correspondingly, FP and FN are signs of the number of false positive samples and false negative samples predicted by the model, respectively (Pan et al., 2020). Following previous studies, the receiver-operating characteristic (ROC) and precision-recall (PR) are implemented to visually display the result of the experiment, and the area under ROC (AUC) and PR (AUPR) are used to assess the comprehensive performance of the proposed model.

Prediction Performance on Different Datasets

To systematically evaluate the performance of the BNEMDI model, our proposed model is implemented to predict potential MDI pairs on five different datasets, and five-fold cross-validations are implemented for obtaining a more accurate assessment. In detail, the dataset will be divided into five parts, each part will serve as the testing set in turn, and the rest as training sets. Afterward, **Table 2** lists various evaluation values to illustrate the prediction performance of BNEMDI. As can be seen in **Table 2**, we get the result of the experiment with the accuracy of 88.75% (ncDR), 87.23% (RNAInter), 77.24% (SM2miR1), 79.92% (SM2miR2), and 81.86% (SM2miR3). The standard deviations of accuracy are 0.1, 0.34, 0.39, 0.21, and 0.65%, respectively. To directly illustrate the prediction performance of BNEMDI on each dataset, **Figure 4** presents the ROC and PR curves of the result of five-fold cross-validations on five datasets. The proposed model BNEMDI achieves average AUCs of 0.9568 (ncDR), 0.9420 (RNAInter), 0.8489 (SM2miR1), 0.8774 (SM2miR2), and 0.9005 (SM2miR3). The standard deviations of five-fold cross-validations are 0.001, 0.0016, 0.0021, 0.0023, and 0.0026, respectively. It is apparent from these criteria values that our proposed model BNEMDI is stable and effective.

Previously, some studies have conducted MDI prediction experiments on the ncDR dataset (Huang et al., 2018; Huang et al., 2020). Herein, we compared our proposed model with these models and some classical methods like collaborative filtering (CF) and matrix factorization (MF) (Boutsidis and Gallopoulos, 2008; Su and Khoshgoftaar, 2009). The evaluation criteria are AUC and the results are shown in **Table 3**. In GCMDR and HMDPI, the attribute features of miRNAs and drugs were constructed using miRNA expression profile, drug substructure fingerprints, gene ontology, and disease ontology. Huang et al. constructed the GCMDR model by combining graph convolution and auto-encoder to learn deep features. In the GCMDR model, the dimensional latent factor, units in hidden layer, maximum Chebyshev polynomial degree, and training epochs are set to 25, 100, 3, and 200, respectively. In EPLMI, they implemented a two-way diffusion method on the weighted network to generate resource vectors which can be defined as:

$$R_{\ln cRNA} = \sum\nolimits_{m=1}^{nm} \frac{A^{w}_{a,m} \cdot A_{*,m}}{\sum\nolimits_{i=1}^{nl} A^{w}_{i,m}} \tag{21}$$

$$R_{miRNA} = \sum\nolimits_{l=1}^{nl} \frac{A^{w}_{l,b} \cdot A_{l,*}}{\sum\nolimits_{i=1}^{nm} A^{w}_{l,i}} \tag{22}$$

where A^w is the weighted adjacency matrixes constructed by similarity, and A is the adjacency matrixes. Other experimental parameters are set to default.

In the methods based on CF, the self-similarities of miRNA and drug are calculated by the Pearson correlation coefficient (PCC), which is defined as:

$$P_{*}(a,b) = \frac{\sum\nolimits_{i=1}^{N}\left(f_{ai} - \overline{f_a}\right)\left(f_{bi} - \overline{f_b}\right)}{\sqrt{\sum\nolimits_{i=1}^{N}\left(f_{ai} - \overline{f_a}\right)^2 \sum\nolimits_{i=1}^{N}\left(f_{bi} - \overline{f_b}\right)^2}}, \tag{23}$$

where f_a and f_b represent the features of two same types of elements (miRNA or drug). Based on the PCC, self-similarity matrixes and adjacency matrixes M for miRNA and drug, the predicted score matrix of drug-based CF can be defined as:

$$M^{'}_{drug}\left(d_i, m_j\right) = \frac{\sum\nolimits_{k=1}^{n_d} P_{drug}\left(d_i, d_k\right) \cdot M_{k,j}}{n_d} \tag{24}$$

where M' is the predicted matrix and n_d is the number of drugs in the dataset.

Correspondingly, the predicted score matrix of miRNA-based CF can be defined as:

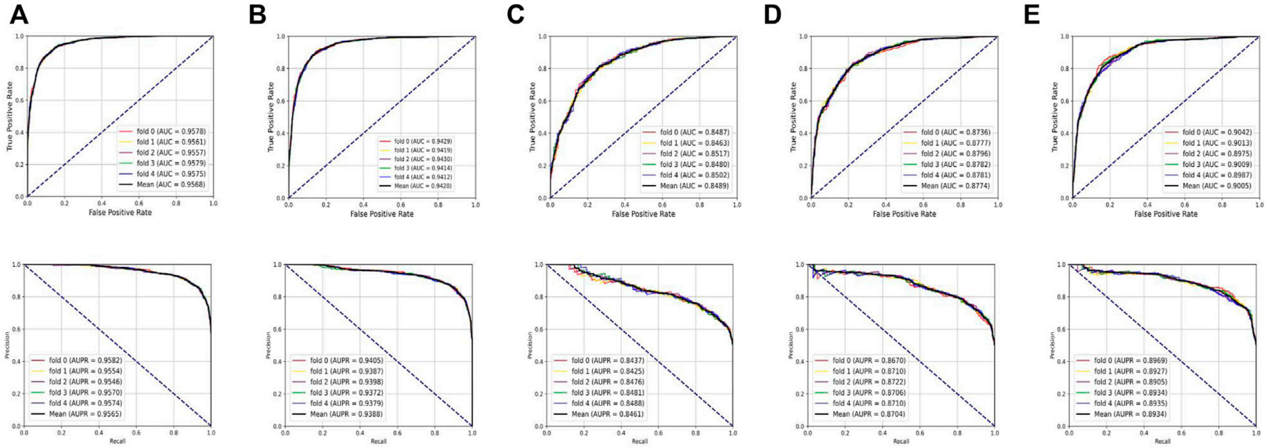


FIGURE 4 | Prediction performance of BNEMDI based on ROC and PR curves. **(A)** Five-fold cross-validation ROC and PR curves on the ncDR dataset. **(B)** Five-fold cross-validation ROC and PR curves on the RNAInter dataset. **(C)** Five-fold cross-validation ROC and PR curves on the SM2miR1 dataset. **(D)** Five-fold cross-validation ROC and PR curves on the SM2miR2 dataset. **(E)** Five-fold cross-validation ROC and PR curves on the SM2miR3 dataset.

TABLE 3 | Comparison of the prediction performance based on the ncDR dataset (N/A means not available).

Method	ncDR	RNAInter	SM2miR1	SM2miR2	SM2miR3
GCMDR	0.9359 ± 0.0006	N/A	N/A	N/A	N/A
EPLMI	0.8971 ± 0.0009	N/A	N/A	N/A	N/A
Neighbor-based CF	0.8644 ± 0.0009	0.8532 ± 0.0007	0.6289 ± 0.0017	0.7346 ± 0.0027	0.8654 ± 0.0015
Drug-based CF	0.7313 ± 0.0008	0.7120 ± 0.0010	0.6982 ± 0.0026	0.6993 ± 0.0013	0.7030 ± 0.0016
miRNA-based CF	0.8235 ± 0.0015	0.8364 ± 0.0022	0.6325 ± 0.0019	0.6534 ± 0.0014	0.7644 ± 0.0009
SVD-based MF	0.6007 ± 0.0052	0.6189 ± 0.0044	0.5978 ± 0.0050	0.6039 ± 0.0051	0.6045 ± 0.0045
BNEMDI	0.9568 ± 0.0010	0.9420 ± 0.0016	0.8489 ± 0.0021	0.8774 ± 0.0023	0.9005 ± 0.0026

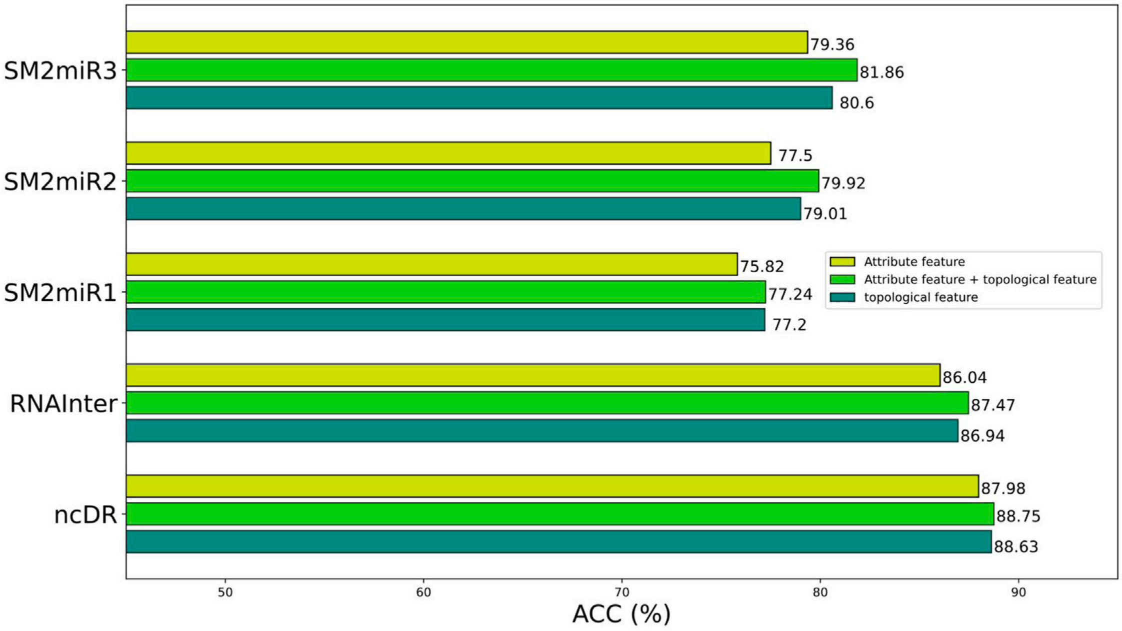


FIGURE 5 | Prediction performance of different features on different datasets.

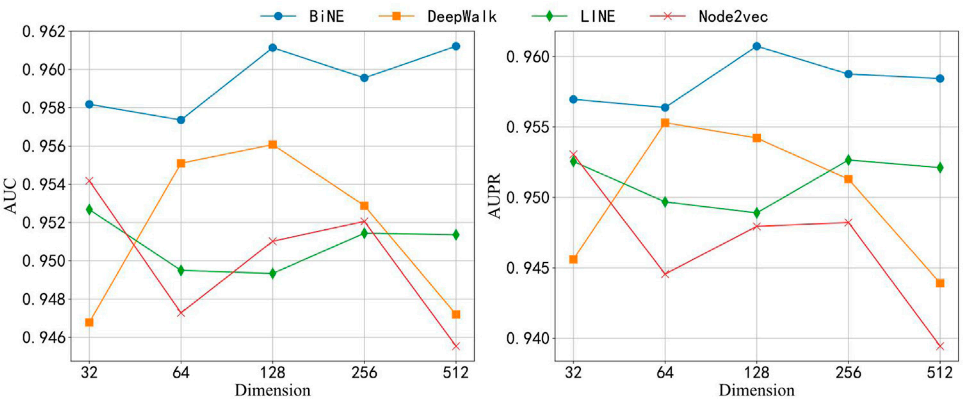


FIGURE 6 | AUC and AUPR of four network embedding methods in different dimensions.

TABLE 4 | Average performance of the different classifiers on ncDR datasets.

Classifier	AUC	AUPR (%)	Acc (%)	Sen (%)	Spec (%)	Prec (%)	MCC (%)
NB	0.9166 ± 0.0035	90.16 ± 0.53	86.49 ± 0.54	81.69 ± 1.22	91.30 ± 0.87	90.38 ± 0.81	73.34 ± 1.05
SVM	0.9415 ± 0.0033	92.93 ± 0.57	86.84 ± 0.63	85.91 ± 0.38	87.77 ± 1.21	87.55 ± 1.08	73.70 ± 1.28
LR	0.9473 ± 0.0029	94.27 ± 0.48	87.56 ± 0.77	86.56 ± 0.32	88.56 ± 1.45	88.34 ± 1.32	75.14 ± 1.56
RF	0.9502 ± 0.0036	94.42 ± 0.56	88.38 ± 0.87	88.18 ± 0.88	88.58 ± 1.79	88.56 ± 1.58	76.77 ± 1.76
BNEMDI	0.9573 ± 0.0009	95.65 ± 0.13	88.75 ± 0.10	89.13 ± 0.19	88.39 ± 0.13	88.47 ± 1.11	77.51 ± 0.20

$$M'_{miRNA}\left(d_i, m_j\right) = \frac{\sum_{k=1}^{n_m} P_{miRNA}\left(m_i, m_k\right) \cdot M_{i,k}}{n_m} s \qquad (25)$$

Neighbor-based CF takes into account both drug-based CF and miRNA-based CF and is defined as:

$$M'_{neighbor}\left(d_i, m_j\right) = \frac{M'_{drug}\left(d_i, m_j\right) + M'_{miRNA}\left(d_i, m_j\right)}{2} \qquad (26)$$

Several studies on drug target interaction prediction or drug repositioning have used similarity-related information to construct the prediction models. Although they gain optimistic results on datasets, it seems difficult for the model to work in real-world scenarios. However, the similarity itself is related to interactions of biological entities, and the abuse of similarity will potentially lead to label leakage. The prediction ability of label leaking models is easily overestimated when it implements on a known dataset. In this experiment, after dividing the dataset into training sets and test sets, only the training set was extracted topological features and used to construct the prediction model for avoiding label leakage. For instance, there are 4,457 MDI pairs in the ncDR dataset, of which only 3,565 MDI pairs will be extracted as features and used to construct the prediction model. But there are no such issues in the attribute features.

Ablation Experiment

To better construct representation vectors, we considered attribute features and topological features of nodes in the miRNA–drug bipartite network. In this section, we are going to discuss the impact of different features on the performance of BNEMDI. We consider three kinds of features: attribute feature, topological feature, and the combination of them to separately construct the representation vectors and the corresponding prediction model. The accuracy is used as the standard to compare the influence of various features on the model.

Figure 5 shows the prediction performance of models based on different features. In general, the topological features are more effective than the attribute features. Therefore, we concluded that the topological features make a great contribution to the proposed model. Although attribute features do not perform as well as the topological feature, the production of attribute features only requires sequence information like SMILES and miRNA sequences. Thus, the attribute features are suitable as the representation vectors for the new samples.

The attribute features are constructed by the sequence profile information of nodes in the relationship network and contain chemical structure information of the miRNAs and drugs. The topological features consider high-order implicit transition relationships and explicit relations, which provide distinct similarity information of homologous nodes. This makes it easier for miRNA and drug relationship pairs with similar structures to known MDI to be considered interacting, and vice versa. In principle, the combination of topological features and attribute features will make the effect more pronounced.

Compare With Other Embedding Methods and Classifiers

The topological feature extracted by BiNE is important for building the BNEMDI model. To highlight the advantages of

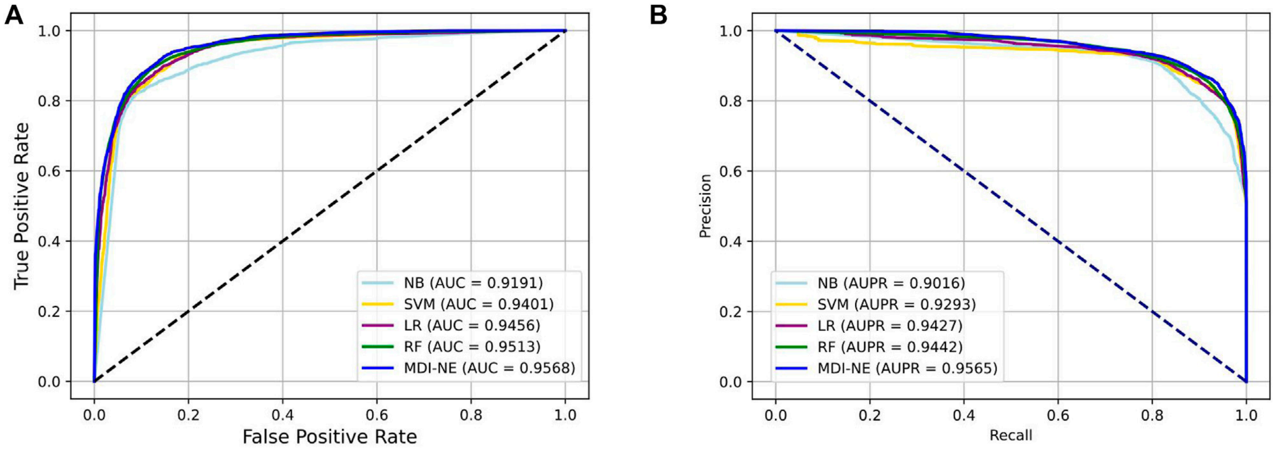


FIGURE 7 | Comparison of BNEMDI with different classifiers under five-fold cross-validation. **(A)** ROC curve on the ncDR MDI dataset. **(B)** PR curve on the ncDR MDI dataset.

TABLE 5 | Top 30 potential MDIs predicted by BNEMDI.

Drug (CID)	miRNA	Evidence	Drug (CID)	miRNA	Evidence
60750	hsa-miR-24-3p	Unconfirmed	60750	hsa-miR-29c-3p	29807360
60750	hsa-miR-205-5p	31602229	2767	hsa-miR-1236-3p	30805558
2767	hsa-miR-193b-3p	27918099	5310940	hsa-miR-660-5p	Unconfirmed
3385	hsa-miR-10a-5p	Unconfirmed	60750	hsa-miR-532-5p	Unconfirmed
3385	hsa-miR-33b-5p	Unconfirmed	31703	hsa-miR-18a-5p	Unconfirmed
3385	hsa-miR-376a-3p	Unconfirmed	3385	hsa-miR-431-5p	Unconfirmed
2520	hsa-miR-126-3p	Unconfirmed	5310940	hsa-miR-196a-5p	Unconfirmed
60750	hsa-miR-124-3p	35127724	5310940	hsa-miR-101-3p	31934027
3385	hsa-miR-93-5p	30573973	60750	hsa-miR-1908-5p	Unconfirmed
31703	hsa-miR-19b-3p	30343695	6857599	hsa-miR-200c-3p	25757925
3385	hsa-miR-32-5p	29530052	36314	hsa-miR-141-3p	26025631
2767	hsa-miR-363-3p	25416050	119307	hsa-miR-181d-5p	Unconfirmed
5310940	hsa-miR-373-3p	Unconfirmed	3385	hsa-miR-620	Unconfirmed
3385	hsa-miR-576-5p	Unconfirmed	3385	hsa-miR-9-3p	Unconfirmed
36462	hsa-miR-21-5p	23834154	5310940	hsa-miR-128-3p	30890168
60750	hsa-miR-24-3p	Unconfirmed	60750	hsa-miR-29c-3p	29807360
60750	hsa-miR-205-5p	31602229	2767	hsa-miR-1236-3p	30805558
2767	hsa-miR-193b-3p	27918099	5310940	hsa-miR-660-5p	Unconfirmed
3385	hsa-miR-10a-5p	Unconfirmed	60750	hsa-miR-532-5p	Unconfirmed
3385	hsa-miR-33b-5p	Unconfirmed	31703	hsa-miR-18a-5p	Unconfirmed
3385	hsa-miR-376a-3p	Unconfirmed	3385	hsa-miR-431-5p	Unconfirmed
2520	hsa-miR-126-3p	Unconfirmed	5310940	hsa-miR-196a-5p	Unconfirmed

The CID of PubChem is used to indicate known MDIs in the RNAInter dataset. The first column records the top 1–25 MDIs. The second column records the top 26–50 MDIs. The evidence is indicated by the PubMed ID of the experimental literature.

BiNE, we compare BiNE to three state-of-the-art graph representation methods and discuss their performance in different dimensions. In a similar way to BiNE, several state-of-the-art network embedding methods (i.e., DeepWalk (Perozzi et al., 2014), LINE (Tang et al., 2015), and node2vec (Grover and Leskovec, 2016)) are used to learn the embedding vectors of each node and compare to BiNE. DeepWalk carries out the random walk on the graph to generate node sequences, and the node sequences are regarded as sentences to learn embedding vectors by word2vec (Church, 2017). LINE combines the first-order and second-order proximities and optimizes them using the asynchronous stochastic gradient algorithm (ASGD) (Recht et al., 2011). Node2vec is an extension of DeepWalk. It introduces depth-first search (DFS) and breadth-first search (DFS) to the process of the random walk. BFS may explore the structural properties of the graph, and DFS may reflect the homogeneity between similar nodes. Based on the experiment, the best result may be obtained when hyper-parameters p and q are set to 0.5 (Gao et al., 2018). Moreover, the parameters of other embedding methods are set to their default settings except for the dimension of the node embedding vector.

Here, we analyze the performance of models in different dimensions of the node embedding vector. We have carried on the experiment to each embedding method separately in

TABLE 6 | Top 50 associated miRNA of drug 5-FU predicted by BNEMDI.

Drug (CID)	miRNA	Evidence	Drug (CID)	miRNA	Evidence
3385	hsa-miR-21-5p	31918721	3385	hsa-miR-181b-5p	Unconfirmed
3385	hsa-miR-221-3p	27726102	3385	hsa-miR-26b-5p	30662808
3385	hsa-miR-126-3p	Unconfirmed	3385	hsa-miR-194-5p	30451820
3385	hsa-miR-200c-3p	28411308	3385	hsa-miR-103a-3p	27247088
3385	hsa-miR-222-3p	19956872	3385	hsa-miR-208a-3p	Unconfirmed
3385	hsa-let-7c-5p	33051247	3385	hsa-miR-18a-5p	32884453
3385	hsa-miR-214-3p	Unconfirmed	3385	hsa-miR-20b-5p	27878272
3385	hsa-miR-155-5p	30741544	3385	hsa-miR-663a	confirmed
3385	hsa-miR-93-5p	32426273	3385	hsa-miR-145-5p	32801865
3385	hsa-miR-18b-5p	25990502	3385	hsa-miR-24-3p	31646794
3385	hsa-miR-143-3p	19843160	3385	hsa-miR-19a-3p	24460313
3385	hsa-miR-181a-3p	29795190	3385	hsa-let-7a-5p	35071455
3385	hsa-miR-16-5p	18449891	3385	hsa-miR-4661-3p	Unconfirmed
3385	hsa-miR-27b-3p	24401318	3385	hsa-miR-27a-3p	24401318
3385	hsa-miR-107	26636340	3385	hsa-miR-200b-3p	32714549
3385	hsa-miR-34c-5p	Unconfirmed	3385	hsa-miR-9-5p	Unconfirmed
3385	hsa-miR-17-5p	32426273	3385	hsa-miR-101-3p	34086111
3385	hsa-miR-34a-5p	31802650	3385	hsa-miR-196a-5p	Unconfirmed
3385	hsa-miR-125b-5p	28176874	3385	hsa-miR-200a-3p	28496200
3385	hsa-miR-497-5p	26673620	3385	hsa-miR-802	Unconfirmed
3385	hsa-miR-29b-3p	34155879	3385	hsa-miR-197-3p	26055341
3385	hsa-miR-20a-5p	31760170	3385	hsa-miR-30b-5p	miR-30b
3385	hsa-miR-1915-3p	Unconfirmed	3385	hsa-miR-181b-2-3p	Unconfirmed
3385	hsa-miR-210-3p	31468617	3385	hsa-miR-100-5p	Unconfirmed
3385	hsa-miR-25-3p	35014676	3385	hsa-miR-153-3p	Unconfirmed

The CID of PubChem is used to indicate known MDIs in the RNAInter dataset. The first column records the top 1–25 MDIs. The second column records the top 26–50 MDIs. The evidence is indicated by the PubMed ID of the experimental literature.

five different dimensions, 32, 64, 128, 256, and 512. We also employed these embedding approaches to learn the topological features from the bipartite network and combine the attribute features of drug compounds and miRNAs to construct this prediction model. **Figure 6** shows the results of each model that was applied to the ncDR dataset. The y axis of **Figure 6** depicts the AUC and AUPR of each prediction model, and the x-axis depicts five kinds of node-embedding dimensions. According to **Figure 6**, we can draw a conclusion that the model with the BiNE embedding method gets the best result among these methods. The main reason for the outstanding performance of BiNE is that it considers unique information of drug and miRNA nodes while processing the relations in the miRNA–drug bipartite network. BiNE calculates the second-order proximity of nodes in the miRNA–drug bipartite network to learn the implicit relation between drugs and miRNAs, which can get more efficient similarity compared with the similarities based on domain knowledge (Yue and He, 2021).

When the dimension of embedding vectors is 64, the BiNE model achieves the lowest AUC and PR values of 0.957 and 0.956, respectively. To avoid overfitting, the dimension of embedding vectors generated by the BiNE model was set to 64 in the subsequent experiments.

We further evaluate the impact of the classifier on the overall model by comparing it with several popular machine learning classifiers, including random forest (RF), naive Bayes (NB), logistics regress (LR), and SVM classifiers (Gui et al., 2015). The features extracted by the same method of the proposed model were used as the input of the aforementioned classifiers for five-fold cross-validations on the ncDR dataset.

Table 4 exhibits the average performance of the five-fold cross-validations of each classifier on the ncDR dataset. As shown in **Table 4**, NB, SVM, LR, and RF obtained an average accuracy of 86.49, 86.84, 87.56, and 88.38%, respectively. The BNEMDI achieved the highest accuracy of 88.75%. We gained an average AUC score of 0.9167, 0.9416, 0.9473, 0.9505, and 0.9573, and an average PR score of 94.27, 90.16, 94.40, 92.93, and 95.65% for NB, LR, RF, and BNEMDI. For a more intuitive comparison, **Figure 7** depicts the corresponding ROC and PR curves. The proposed model leads in most evaluation metrics with the highest AUC of 0.9573 and the highest AURR of 0.9565 and has a relatively low standard deviation. Synthetically, BNEMDI not only has an excellent performance in various evaluation criteria but also is more stable than other classifiers.

CASE STUDY

In this subsection, we carried out a case study on the RNAInter dataset. All of the known MDIs were used to construct representation vectors to predict all candidate miRNA–drug pairs in the dataset. Then, we ranked these candidate miRNA–drug pairs according to the predicted scores in the descending order. The top 30 predicted relationships are shown in **Table 5**. Among the top 10, 20, and 30 predicted relationships, 7, 12, and 18 relationships are verified by the previous literature in PubMed, respectively.

Furthermore, to demonstrate the prediction ability for new drugs, we selected 5-fluorouracil (5-FU, CID:3385) as the investigated drug of the case study, which is a chemotherapy drug widely used in digestive system cancer and breast cancer (Wigmore et al., 2010). The MDIs related to 5-FU were removed from the dataset and the rest of MDIs were used to train the prediction model. Then we implemented the BNEMDI model to identify potential miRNAs that may interact with 5-FU. The top 50 predicted miRNAs are shown in **Table 6**. Among the top 10, 20, and 50 predicted miRNAs, there were 9, 17, and 37 miRNAs, which confirmed that they may interact with 5-FU by the previous literature.

For instance, Valeri et al. discovered that miRNA-21-5p, which ranks first in the top 50 predicted miRNAs can downregulate the expression level of human DNA MutS homolog 2 leading to 5-FU resistance in colon cancer patients (Liang et al., 2020). Moreover, the study proposed by Zhao et al. (2016) confirmed the overexpression of hsa-miR-221-3p will reduce the sensitivity of 5-FU and proved it can be a potential drug target for pancreatic cancer (Zhao et al., 2016). Moreover, through functional analysis, Jilek et al. (2020) demonstrated that has-let-7c-5p can elevate the exposure of 5-FU. They suggested that has-let-7c-5p and 5-FU can attenuate thymidylate synthase, which indicates that 5-FU can cooperate with has-let-7c-5p against hepatocellular carcinoma (Jilek et al., 2020). As stated before, this case study shows that BNEMDI can effectively find out the miRNAs interacting with given drugs.

CONCLUSION

MDI prediction plays an important role in new drug target research. In this article, we proposed a novel computational model to predict unknown MDIs, namely, BNEMDI. We adopted a bipartite network embedding method BiNE to extract the topological feature from the MDI network. The chemical structure of drugs and the base sequence information of miRNAs are represented as the attribute feature by MACCS fingerprints and k-mer. When performed on five datasets (ncDR, RNAInter, SM2miR1, SM2miR2, and SM2miR3), BNEMDI gained average AUC values of 88.75, 87.23, 77.24, 79.92, and 81.86% under five-fold cross-validation, respectively. In addition, we experimented with other popular network embedding methods in different dimensions. Moreover, the case study on a common drug for cancer and all of the candidate miRNA–drug pairs demonstrated that the proposed model could be an effective tool for predicting MDI in real scenarios. The comprehensive results indicated that BNEMDI is a reliable and stable MDI predictor, economizing time and labor for drug target studies. Even so, the BNEMDI model possesses drawbacks. For new drugs and miRNAs, they are independent nodes in the bipartite network. The network embedding methods cannot learn any information from these independent nodes. Only attribute features can represent these nodes, and then the new interaction network can be updated according to the wet experimental results. In the future, we expect to seek more efficient network embedding methods and feature descriptors for mining the relationship between drugs and miRNAs.

DATA AVAILABILITY STATEMENT

Publicly available datasets were analyzed in this study. These data can be found at: http://www.jianglab.cn/ncDR/index.jsp (ncDR) http://www.jianglab.cn/SM2miR/ (SM2miR) http://www.rnainter.org/ (RNAInter).

AUTHOR CONTRIBUTIONS

Conceptualization, methodology, and software: Y-JG; validation and formal analysis: L-PL; investigation: Y-CL; resources: Z-HR; data curation and visualization: JP; writing—original draft preparation: JP; writing—review and editing: C-QY; supervision: Y-JG; project administration: Y-CL; funding acquisition: Z-HY All authors have read and agreed to the published version of the manuscript.

FUNDING

This work was supported by the Science and Technology Innovation 2030-New Generation Artificial Intelligence Major Project (No.2018AAA0100103) and in part by the NSFC Program, under Grant Nos. 61873212, 62072378, and 62002297.

ACKNOWLEDGMENTS

The authors thank the reviewers for their valuable suggestions.

REFERENCES

Ambros, V. (2001). microRNAs. *Cell.* 107, 823–826. doi:10.1016/s0092-8674(01)00616-x

Bartel, D. P. (2004). MicroRNAs. *Cell.* 116, 281–297. doi:10.1016/s0092-8674(04)00045-5

Bayraktar, R., Van Roosbroeck, K., and Reviews, M. (2018). miR-155 in Cancer Drug Resistance and as Target for miRNA-Based Therapeutics. *Cancer Metastasis Rev.* 37, 33–44. doi:10.1007/s10555-017-9724-7

Bentwich, I., Avniel, A., Karov, Y., Aharonov, R., Gilad, S., Barad, O., et al. (2005). Identification of Hundreds of Conserved and Nonconserved Human microRNAs. *Nat. Genet.* 37, 766–770. doi:10.1038/ng1590

Bommer, G. T., Gerin, I., Feng, Y., Kaczorowski, A. J., Kuick, R., Love, R. E., et al. (2007). p53-mediated Activation of miRNA34 Candidate Tumor-Suppressor Genes. *Curr. Biol.* 17, 1298–1307. doi:10.1016/j.cub.2007.06.068

Boutsidis, C., and Gallopoulos, E. (2008). SVD Based Initialization: A Head Start for Nonnegative Matrix Factorization. *Pattern Recognit.* 41, 1350–1362. doi:10.1016/j.patcog.2007.09.010

Cai, H., Zheng, V. W., Chang, K. C.-C., and Engineering, D. (2018). A Comprehensive Survey of Graph Embedding: Problems, Techniques, and Applications. *IEEE Trans. Knowl. Data Eng.* 30, 1616–1637. doi:10.1109/tkde.2018.2807452

Cao, D.-S., Liu, S., Xu, Q.-S., Lu, H.-M., Huang, J.-H., Hu, Q.-N., et al. (2012). Large-scale Prediction of Drug-Target Interactions Using Protein Sequences and Drug Topological Structures. *Anal. Chim. Acta* 752, 1–10. doi:10.1016/j.aca.2012.09.021

Cawley, S., Bekiranov, S., Ng, H. H., Kapranov, P., Sekinger, E. A., Kampa, D., et al. (2004). Unbiased Mapping of Transcription Factor Binding Sites along Human Chromosomes 21 and 22 Points to Widespread Regulation of Noncoding RNAs. *Cell.* 116, 499–509. doi:10.1016/s0092-8674(04)00127-8

Cereto-Massagué, A., Ojeda, M. J., Valls, C., Mulero, M., Garcia-Vallvé, S., and Pujadas, G. (2015). Molecular Fingerprint Similarity Search in Virtual Screening. *Methods* 71, 58–63. doi:10.1016/j.ymeth.2014.08.005

Church, K. W. (2017). Word2Vec. *Nat. Lang. Eng.* 23, 155–162. doi:10.1017/s1351324916000334

Dai, E., Yang, F., Wang, J., Song, Q., Weiwei, A., Lihong, W., et al. (2017). ncDR: a Comprehensive Resource of Non-coding RNAs Involved in Drug Resistance. 33**,** 4010–4011. doi:10.1093/bioinformatics/btx523

Deepthi, K., and Jereesh, A. S. (2021). An Ensemble Approach Based on Multi-Source Information to Predict Drug-MiRNA Associations *via* Convolutional Neural Networks. *IEEE Access* 9, 38331–38341. doi:10.1109/access.2021.3063885

Deng, H., Lyu, M. R., and King, I. (2009). "A Generalized Co-hits Algorithm and its Application to Bipartite Graphs," in Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining, 239–248.

Dixon, S. J., and Stockwell, B. R. (2009). Identifying Druggable Disease-Modifying Gene Products. *Curr. Opin. Chem. Biol.* 13, 549–555. doi:10.1016/j.cbpa.2009.08.003

Erten-Ela, S., Kiristi, M., Varlec, A., Bozduman, F., Remskar, M., Oksuz, L., et al. (2018). Platinum-free Counter Electrodes of Plasma-Modified Hybrid Nanomaterials for Dye-Sensitised Solar Cells. *Int. J. Sustain. Energy* 37, 640–653. doi:10.1080/14786451.2017.1333995

Eyking, A., Reis, H., Frank, M., Gerken, G., Schmid, K. W., and Cario, E. (2016). MiR-205 and MiR-373 Are Associated with Aggressive Human Mucinous Colorectal Cancer. *PLoS ONE* 11, e0156871. doi:10.1371/journal.pone.0156871

Fleuren, W. W. M., and Alkema, W. (2015). Application of Text Mining in the Biomedical Domain. *Methods* 74, 97–106. doi:10.1016/j.ymeth.2015.01.015

Fu, X.-D. (2014). Non-coding RNA: a New Frontier in Regulatory Biology. *Non-coding RNA a new Front. Regul. Biol.* 1, 190–204. doi:10.1093/nsr/nwu008

Fu, Z., Wang, L., Li, S., Chen, F., Au-Yeung, K. K.-W., and Shi, C. J. F. I. P. (2021). MicroRNA as an Important Target for Anticancer Drug Development. *Front. Pharmacol.* 12, 2212. doi:10.3389/fphar.2021.736323

Gao, M., Chen, L., He, X., and Zhou, A. (2018). "Bine: Bipartite Network Embedding," in The 41st international ACM SIGIR conference on research & development in information retrieval, 715–724.

Grover, A., and Leskovec, J. J. A. (2016). node2vec: Scalable Feature Learning for Networks. *KDD* 2016, 855–864. doi:10.1145/2939672.2939754

Gui, J., Liu, T., Tao, D., Sun, Z., and Tan, T. (2015). Representative Vector Machines: a Unified Framework for Classical Classifiers. *IEEE Trans. Cybern.* 46, 1877–1888. doi:10.1109/TCYB.2015.2457234

Guo, Z.-H., You, Z.-H., Li, L.-P., Chen, Z.-H., Yi, H.-C., and Wang, Y.-B. (2020). "Inferring Drug-miRNA Associations by Integrating Drug SMILES and MiRNA Sequence Information," in International Conference on Intelligent Computing. Springer, 279–289. doi:10.1007/978-3-030-60802-6_25

Huang, Y.-A., Chan, K. C. C., and You, Z.-H. (2018). Constructing Prediction Models from Expression Profiles for Large Scale lncRNA-miRNA Interaction Profiling. 34**,** 812–819.doi:10.1093/bioinformatics/btx672

Huang, Y. A., Hu, P., Chan, K. C. C., and You, Z. H. (2020). Graph Convolution for Predicting Associations between miRNA and Drug Resistance. *Bioinformatics* 36, 851–858. doi:10.1093/bioinformatics/btz621

Ishida, M., and Selaru, F. M. (2013). miRNA-based Therapeutic Strategies. *Curr. Pathobiol. Rep.* 1, 63–70. doi:10.1007/s40139-012-0004-5

Iyer, M. K., Niknafs, Y. S., Malik, R., Singhal, U., Sahu, A., Hosono, Y., et al. (2015). The Landscape of Long Noncoding RNAs in the Human Transcriptome. *Nat. Genet.* 47, 199–208. doi:10.1038/ng.3192

Jiang, M., Cui, P., Yuan, N. J., Xie, X., and Yang, S. (2016). "Little Is Much: Bridging Cross-Platform Behaviors through Overlapped Crowds," in Thirtieth AAAI Conference on Artificial Intelligence.

Jiang, Q., Wang, Y., Hao, Y., Juan, L., Teng, M., Zhang, X., et al. (2009). miR2Disease: a Manually Curated Database for microRNA Deregulation in Human Disease. *Nucleic Acids Res.* 37, D98–D104. doi:10.1093/nar/gkn714

Jilek, J. L., Tu, M.-J., Zhang, C., Yu, A.-M., and Disposition (2020). Pharmacokinetic and Pharmacodynamic Factors Contribute to Synergism between Let-7c-5p and 5-fluorouracil in Inhibiting Hepatocellular Carcinoma Cell Viability. *Drug Metab. Dispos.* 48, 1257–1263. doi:10.1124/dmd.120.000207

Kang, J., Tang, Q., He, J., Li, L., Yang, N., Yu, S., et al. (2021). RNAInter v4.0: RNA Interactome Repository with Redefined Confidence Scoring System and Improved Accessibility. *Nucleic Acids Res.* 50 (D1), D326–D332. doi:10.1093/nar/gkab997

Kim, S., Chen, J., Cheng, T., Gindulyte, A., He, J., He, S., et al. (2021a). PubChem in 2021: New Data Content and Improved Web Interfaces. *Nucleic Acids Res.* 49, D1388–D1395. doi:10.1093/nar/gkaa971

Kleinberg, J. M. (1999). Authoritative Sources in a Hyperlinked Environment. *J. ACM* 46 (5), 604–632.

Kota, J., Chivukula, R. R., O'donnell, K. A., Wentzel, E. A., Montgomery, C. L., Hwang, H.-W., et al. (2009). Therapeutic microRNA Delivery Suppresses Tumorigenesis in a Murine Liver Cancer Model. *Cell.* 137, 1005–1017. doi:10.1016/j.cell.2009.04.021

Kozomara, A., Birgaoanu, M., and Griffiths-Jones, S. (2019). miRBase: from microRNA Sequences to Function. 47**,** D155–D162. doi:10.1093/nar/gky1141

Kurtz, S., Narechania, A., Stein, J. C., and Ware, D. (2008). A New Method to Compute K-Mer Frequencies and its Application to Annotate Large Repetitive Plant Genomes. *BMC Genomics* 9, 517–518. doi:10.1186/1471-2164-9-517

Li, M. M., Huang, K., and Zitnik, M. (2021). Representation Learning for Networks in Biology and Medicine: Advancements, Challenges, and Opportunities.

Liu, B., Shi, Y., Yuan, J., Hu, X., Zhang, H., Li, N., et al. (2013). Estimation of Genomic Characteristics by Analyzing K-Mer Frequency in De Novo Genome Projects.

Liang, G., Zhu, Y., Ali, D. J., Tian, T., Xu, H., Si, K., et al. (2020). Engineered Exosomes for Targeted Co-delivery of miR-21 Inhibitor and Chemotherapeutics to Reverse Drug Resistance in Colon Cancer. *J. Nanobiotechnology* 18, 10–15. doi:10.1186/s12951-019-0563-2

Liu, X., Wang, S., Meng, F., Zhang, Y., Wang, J., Dai, E., et al. (2013). SM2miR: a Database of the Experimentally Validated Small Molecules' Effects on microRNA Expression. *Bioinformatics* 29, 409–411. doi:10.1093/bioinformatics/bts698

Lv, Y., Wang, S., Meng, F., Yang, L., Wang, Z., Wang, J., et al. (2015). Identifying Novel Associations between Small Molecules and miRNAs Based on Integrated Molecular Networks. *Bioinformatics* 31, 3638–3644. doi:10.1093/bioinformatics/btv417

Matboli, M., Eissa, S., Ibrahim, D., Hegazy, M. G. A., Imam, S. S., and Habib, E. K. (2017). Caffeic Acid Attenuates Diabetic Kidney Disease *via* Modulation of Autophagy in a High-Fat Diet/Streptozotocin- Induced Diabetic Rat. *Sci. Rep.* 7, 2263–2312. doi:10.1038/s41598-017-02320-z

Pan, J., You, Z.-H., Li, L.-P., Huang, W.-Z., Guo, J.-X., Yu, C.-Q., et al. (2022). DWPPI: A Deep Learning Approach for Predicting Protein–Protein Interactions in Plants Based on Multi-Source Information with a Large-Scale Biological Network. *Front. Bioeng. Biotechnol.* 10. doi:10.3389/fbioe.2022.807522

Pan, J., You, Z.-H., Yu, C.-Q., Li, L.-P., and Zhan, X.-K. (2020). "Predicting Protein-Protein Interactions from Protein Sequence Information Using Dual-Tree Complex Wavelet Transform," in International Conference on Intelligent Computing. Springer, 132–142. doi:10.1007/978-3-030-60802-6_13

Perozzi, B., Al-Rfou, R., and Skiena, S. (2014). "Deepwalk: Online Learning of Social Representations," in Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining, 701–710.

Recht, B., Re, C., Wright, S., and Niu, F. (2011). Hogwild!: A Lock-free Approach to Parallelizing Stochastic Gradient Descent *Adv. Neural. Inf. Process Syst.* 24.

Su, X., and Khoshgoftaar, T. M. (20092009). A Survey of Collaborative Filtering Techniques. *Adv. Artif. Intell.* 2009, 1–19. doi:10.1155/2009/421425

Tang, J., Qu, M., Wang, M., Zhang, M., Yan, J., and Mei, Q. (2015). "Line: Large-Scale Information Network Embedding," in Proceedings of the 24th international conference on world wide web, 1067–1077.

Wang, H., Cao, J., Shu, L., and Rafiei, D. (2013). "Locality Sensitive Hashing Revisited: Filling the Gap between Theory and Algorithm Analysis," in

Proceedings of the 22nd ACM international conference on Information & Knowledge Management, 1969–1978.

Weininger, D., and Sciences, C. (1988). SMILES, a Chemical Language and Information System. 1. Introduction to Methodology and Encoding Rules. *J. Chem. Inf. Model.* 28, 31–36. doi:10.1021/ci00057a005

Wigmore, P. M., Mustafa, S., El-Beltagy, M., Lyons, L., Umka, J., and Bennett, G. (2010). "Effects of 5-FU," in *Chemo Fog* (Springer), 157–164. doi:10.1007/978-1-4419-6306-2_20

Xie, W. B., Yan, H., and Zhao, X. M. (2017). EmDL: Extracting miRNA-Drug Interactions from Literature. *IEEE/ACM Trans. Comput. Biol. Bioinform* 16, 1722–1728. doi:10.1109/TCBB.2017.2723394

Ya, H., P, H., Kcc, C., and You, Z-H. (2020). Graph Convolution for Predicting Associations between miRNA and Drug Resistance. *Bioinformatics* 36, 851–858. doi:10.1093/bioinformatics/btz621

Yi, H.-C., You, Z.-H., Guo, Z.-H., Huang, D.-S., and Chan, C. C. (2020). Learning Representation of Molecules in Association Network for Predicting Intermolecular Associations. *IEEE/ACM Trans. Comput. Biol. Bioinforma.* 18 (6), 2546–2554. doi:10.1109/tcbb.2020.2973091

Yousef, M., Khalifa, W., Acar, İ. E., and Allmer, J. (2017). MicroRNA Categorization Using Sequence Motifs and K-Mers. *BMC Bioinforma.* 18, 170–179. doi:10.1186/s12859-017-1584-1

Yu, L., Zhang, C., Pei, S., Sun, G., and Zhang, X. (2018). "Walkranker: A Unified Pairwise Ranking Model with Multiple Relations for Item Recommendation," in Proceedings of the AAAI Conference on Artificial Intelligence.

Yue, Y., and He, S. J. B. B. (2021). DTI-HeNE: a Novel Method for Drug-Target Interaction Prediction Based on Heterogeneous Network Embedding. 22**,** 1–20.doi:10.1186/s12859-021-04327-w

Zhang, S., Cheng, Z., Wang, Y., and Han, T. (2021b). The Risks of miRNA Therapeutics: In a Drug Target Perspective. *Drug Des. Dev. Ther.* 15, 721–733. doi:10.2147/dddt.s288859

Zhang, S., Cheng, Z., Wang, Y., and Han, T. (2021). The Risks of miRNA Therapeutics: In a Drug Target Perspective. *Drug Des. Devel Ther.* Vol. 15, 721–733. doi:10.2147/dddt.s288859

Zhao, L., Zou, D., Wei, X., Wang, L., Zhang, Y., Liu, S., et al. (2016). MiRNA-221-3p Desensitizes Pancreatic Cancer Cells to 5-fluorouracil by Targeting RB1. *Tumor Biol.* 37, 16053–16063. doi:10.1007/s13277-016-5445-8

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

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 Genetics

TYPE Original Research
PUBLISHED 29 August 2022
DOI 10.3389/fgene.2022.817919

OPEN ACCESS

EDITED BY
Bin Lu,
University of South China, China

REVIEWED BY
Srikanta Goswami,
National Institute of Biomedical Genomics (NIBMG), India
Guo Wen Zhi,
Zhengzhou University, China

*CORRESPONDENCE
Xiaoyan Wang,
wangxiaoyan@csu.edu.cn
Yajun Shou,
2204120125@csu.edu.cn

†These authors have contributed equally to this work

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

RECEIVED 13 December 2021
ACCEPTED 20 July 2022
PUBLISHED 29 August 2022

CITATION
Tao S, Tian L, Wang X and Shou Y (2022), A pyroptosis-related gene signature for prognosis and immune microenvironment of pancreatic cancer.
Front. Genet. 13:817919.
doi: 10.3389/fgene.2022.817919

A pyroptosis-related gene signature for prognosis and immune microenvironment of pancreatic cancer

Sifan Tao[1,2], Li Tian[2], Xiaoyan Wang[2*†] and Yajun Shou[1,3*†]

[1]Department of Gastroenterology, The Second Xiangya Hospital, Central South University, Changsha, Hunan, China, [2]Department of Gastroenterology, The Third Xiangya Hospital, Central South University, Changsha, Hunan, China, [3]Research Center of Digestive Disease, Central South University, Changsha, Hunan, China

Pancreatic cancer is one of the most lethal tumors owing to its unspecific symptoms during the early stage and multiple treatment resistances. Pyroptosis, a newly discovered gasdermin-mediated cell death, facilitates anti- or pro-tumor effects in a variety of cancers, whereas the impact of pyroptosis in pancreatic cancer remains unclear. Therefore, we downloaded RNA expression and clinic data from the TCGA-PAAD cohort and were surprised to find that most pyroptosis-related genes (PRGs) are not only overexpressed in tumor tissue but also strongly associated with overall survival. For their remarkable prognostic value, cox regression analysis and lasso regression were used to establish a five-gene signature. All patients were divided into low- and high-risk groups based on the media value of the risk score, and we discovered that low-risk patients had better outcomes in both the testing and validation cohorts using time receiver operating characteristic (ROC), nomograms, survival, and decision analysis. More importantly, a higher somatic mutation burden and less immune cell infiltration were found in the high-risk group. Following that, we predicted tumor response to chemotherapy and immunotherapy in both low- and high-risk groups, which suggests patients with low risk were more likely to respond to both immunotherapy and chemotherapy. To summarize, our study established an effective model that can help clinicians better predict patients' drug responses and outcomes, and we also present basic evidence for future pyroptosis related studies in pancreatic cancer.

KEYWORDS

pyroptosis, pancreatic cancer, immune microenvironment, prognostic model, therapeutic response prediction

Introduction

Pancreatic cancer (PAAD), which is primarily composed of pancreatic ductal adenocarcinoma, is one of the most fatal malignancies in the United States, with a survival rate of about 10% (Siegel et al., 2021). The poor prognosis and stable incidence rates of PAAD cases were not only associated with increased exposure to risk factors such

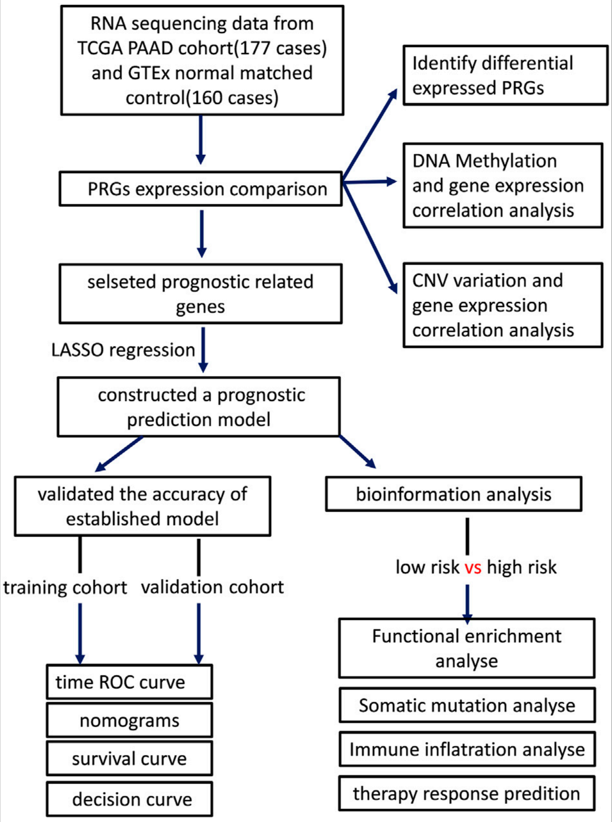

RNA sequencing data from TCGA PAAD cohort(177 cases) and GTEx normal matched control(160 cases)
PRGs expression comparison
Identify differential expressed PRGs
DNA Methylation and gene expression correlation analysis
CNV variation and gene expression correlation analysis
selseted prognostic related genes
LASSO regression
constructed a prognostic prediction model
validated the accuracy of established model
bioinformation analysis
training cohort
validation cohort
low risk vs high risk
time ROC curve
nomograms
survival curve
decision curve
Functional enrichment analyse
Somatic mutation analyse
Immune inflatration analyse
therapy response predition

FIGURE 1
The workflow of our study.

as obesity, diabetes, tobacco use, and alcohol consumption, but also with nonspecific symptoms at the early stage (Stolzenberg-Solomon et al., 2013; Rebours et al., 2015; Walter et al., 2016). Worse still, only modest progress has been achieved in reducing the mortality rate of PAAD. Though immunotherapy has proved to be a promising treatment in many other malignancies, few PAAD patients benefited from ICIs (Torphy et al., 2018; Galluzzi et al., 2020). The "cold" tumor microenvironment is one of the primary reasons for its immunotherapy resistance (O'Donnell et al., 2019). The tumor microenvironment of PAAD is mainly composed of immunosuppressive cells, such as tumor-associated macrophages, myeloid-derived suppressor cells, and regular T-cells (Clark et al., 2007). Additionally, it is believed that an unusually intense desmoplastic reaction surrounding PAAD contributes to the formation of a barrier that prevents immune infiltration and chemotherapy exposure (Provenzano et al., 2012; Ho et al., 2020). Therefore, it is critical to investigate the molecular pathways related to PAAD microenvironment.

Pyroptosis is defined as the caspase (CASP) family-driven programmed necrotic cell death mediated by gasdermin (GSDM) (Shi et al., 2015). When triggered by bacterial, viral, toxin, or chemotherapy, pyroptosis can release pro-inflammatory cytokines and immunogenic material, promoting the activation and infiltration of immune cells (Loveless et al., 2021; Yu et al., 2021). Pyroptotic cell death is characterized by cellular swelling and bubble-like protrusions forming on the cell membrane surface, as well as the release of IL1 and IL18 (Loveless et al., 2021; Yu et al., 2021). Cancers of all forms are closely related to pyroptosis (Yu et al., 2021). On one hand, inducing pyroptosis was originally considered a promising therapeutic strategy for increasing anti-tumor immune response. On the other hand, the activation of multiple signaling pathways and the release of cytokines can lead to tumorigenesis and drug resistance (Xia et al., 2019). The connection between PAAD and pyroptosis is still unclear. Recent work demonstrated that STE20-like kinase 1 slowed PAAD progression by triggering ROS-mediated pyroptosis, implying that pyroptosis may be a potential therapeutic target for PAAD (Cui et al., 2019).

One possible reason for the depressing outcomes of immunotherapy is that PAAD cells can avoid cell death induction (Chen et al., 2021). Thus, we sought to advance our understanding of the pyroptotic pathway in PAAD and construct a pyroptosis-related gene (PRG) prognostic signature. Our study provided an effective prognostic model as well as basic evidence for subsequent pyroptosis-related studies in PAAD.

Materials and methods

Data extraction

The workflow of our study is revealed in Figure 1. The UCSC Xena (Goldman et al., 2020) (Xean, http://xena.ucsc.edu/) was used to obtain the RNA sequencing profile and clinical following data of the TCGA-PAAD cohort and GTEx cohort. Xena was also implemented to integrate normalized counts from TCGA-PAAD and GTex cohort due to limited matched controls in the TCGA-PAAD cohort. All PAAD patients without survival following were excluded in this study. In this cohort, there are 177 PAAD patients and 167 normal pancreatic tissue. The GISTIC copy number dataset and DNA methylation data for all selected patients were obtained from cBioportal (https://www.cbioportal.org/), while the somatic mutation data of patients was downloaded from TCGA (https://portal.gdc.cancer.gov/). Additionally, we downloaded two extra GEO datasets (GSE28735 and GSE62452, https://www.ncbi.nlm.nih.gov/geo/) and ICGC sequencing profiles from ICGC (https://daco.icgc.org/) as independent validation cohorts (Zhang et al., 2012; Yang et al., 2016).

Identify differential expressed genes and perform functional analysis

The 33 PRGs were selected from a previously published study and are listed in Supplementary Table S1 (Ye et al., 2021). The

"DESeq2" package was used to identify differentially expressed genes (DEGs) (Love et al., 2014). Additionally, we conducted correlation analyses of gene expression and methylation using the cBioportal (http://cbioportal.org) (Cerami et al., 2012). The Mann-Whitney or unpaired *t*-test was used to investigate gene expression differences across distinct copy number variations (CNV). The function of DEGs was analyzed using KEGG enrichment analysis and gene set enrichment analysis (GSEA) *via* the "clusterProfiler" R package (Yu et al., 2012). *p*-values < 0. 05 were defined as statistically significant.

The construction of prognostic prediction models

To begin, univariate cox regressions were utilized to examine the relationships between individual 33 PRGs and overall survival (OS) in the TCGA cohort. *p*-value < 0.05 was set as the threshold to identify prognostic-related PRGs. LASSO regression analysis was then used to select significant PRGs and minimize the likelihood of overfitting. Based on these selected PRGs, the prognostic model was constructed using multivariate cox regression analysis. The risk score for OS was constructed as the following formula:

$$risk\ score = \sum_{i}^{5} Xi*\beta i$$

Where X represents the gene expression level and β represents the regression coefficient calculated by multivariate Cox regression. All patients were separated into high- and low-risk groups based on the media value of the risk score.

Validation of the prognostic prediction model

To evaluate the accuracy of the prediction model, time receiver operating characteristic (ROC) curve, nomograms, Kaplan-Meier survival curve, and decision curve were established in the TCGA cohort and validation ICGC cohort. The ROC curves at 1-, 3-, and 5- years were generated using the R package "timeROC" (Blanche et al., 2013). The Kaplan-Meier survival curve was generated by using the R package "survival" (Grambsch, 2000). The decision curve and the following clinic impact curve were finished by the R package "rmda" (Brown, 2018). And the R package "regplot" (Marshall, 2020) was used to perform the nomogram analysis.

Molecular variation analysis and tumor mutation burden between subgroups

After combining the copy number dataset with the somatic mutation dataset of TCGA, we visualized the top 15 genes with the highest mutational frequencies and compared their somatic mutation status across subgroups using the R package "maftool" (Mayakonda et al., 2018). The TMB value of each patient was also calculated through "maftool", and the Mann-Whitney or unpaired *t*-test was used to compare TMB values across subgroups (Mayakonda et al., 2018). *p*-values < 0.05 were considered statistically significant.

Comprehensive immune characteristics analysis between subgroups

By relating gene expression data to cell purity data, the "ESITMATE" R package was utilized to determine the activities of tumor cells, immune cells, and stromal cells inside the tumor environment (Yoshihara et al., 2013). We next used single-sample GSEA through the "GSVA" R package to determine the relative proportions of 28 different types of tumor-infiltrating immune cells (Hanzelmann et al., 2013). Supplementary Table S2 contains all the gene sets for targeted immune cells. Apart from that, the relative expression levels of the ICIs-targeted genes were determined using FPKM values and compared using Mann-Whitney or unpaired *t*-test.

Immunotherapy and chemotherapeutic response prediction

The TIDE (Tumor Immune Dysfunction and Exclusion) web tool (http://tide.dfci.harvard.edu/) was used to predict immunotherapy responses (Jiang et al., 2018). Patients with a lower TIDE score were considered to have a better response to immunotherapy. Besides, based on the GDSC (Genomics of Drug Sensitivity in Cancer) database, the R package "oncoPredict" was used to perform ridge regression analysis on each sample to predict IC50 values for targeted drugs (Maeser et al., 2021). A Mann-Whitney or unpaired *t*-test was used to compare TIDE scores and IC50 values across subgroups. *p*-values<0.05 were considered statistically significant.

Results

Alterations of pyroptosis-related genes RNA expression in pancreatic cancer

To begin, we identified differentially expressed PRGs between PAAD tissue and normal pancreatic tissue from the TCGA-GTEx integrated cohort. The heatmap of PRGs revealed that nearly all PRGs are significantly overexpressed within PAAD tissue (Figure 2A). More specifically, the expression of AIM2, CASP1, CASP3, CASP5, GSDMA, GSDMC, IL1B, IL6, IL18, NLRP1, NLRP2, NLRP3, NLRP7, NOD2, TNF, GPX4, and PYCARD increased more than twofold, whereas CASP9 expression decreased (Figure 2B). Following that, we

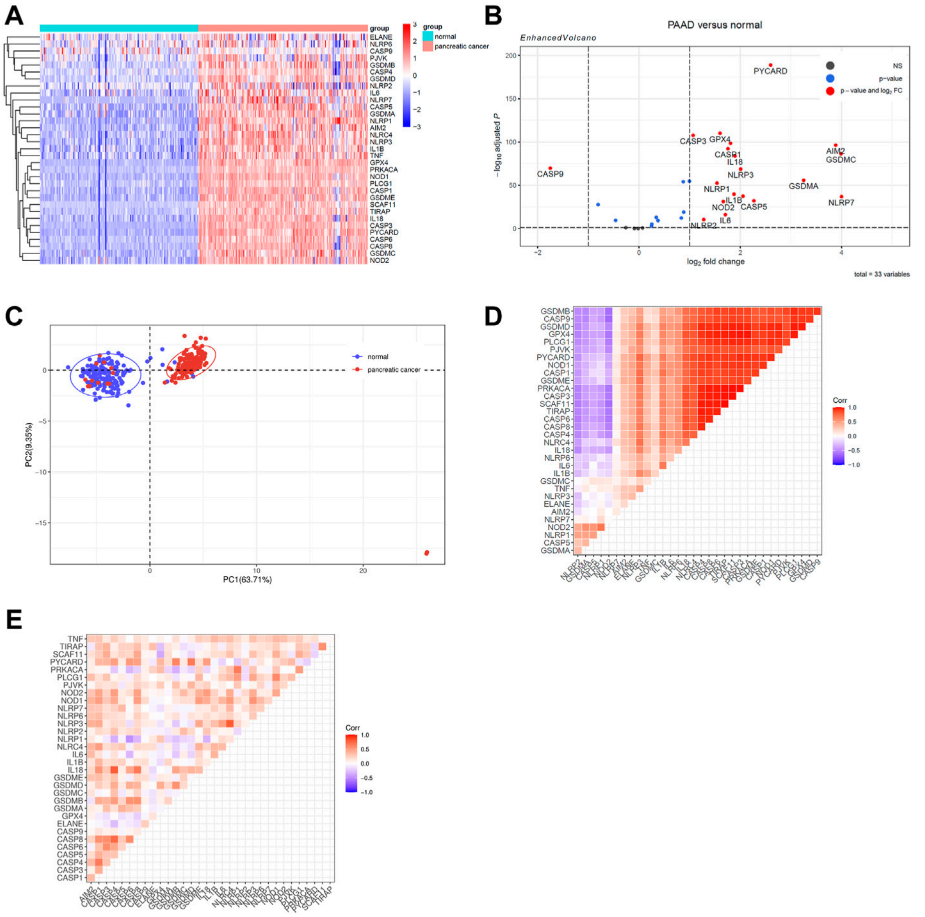


FIGURE 2
Identify differentially expressed PRGs between PAAD and normal pancreatic tissue. Genes with |log2 fold change (log2FC) | > 1 and adjusted *p* value < 0.05 were considered as differentially expressed genes. **(A)**A heatmap to show PRGs expression within normal tissue (FPKM data from GTEx cohort) and PAAD tissue (FPKM data from TCGA cohort). **(B)** The Volcano plot created using the "Enhanced Volcano" R package to show differently expressed PRGs. **(C)** Principal component analysis was processed to identify PRGs expression characters between the normal pancreatic tissue and the PAAD tissue. **(D)** A heatmap of correlation matrix of the PRGs within normal tissue from GTEx cohort. **(E)** A heatmap of correlation matrix of the PRGs within PAAD tissue from TCGA cohort.

analyzed two additional GEO datasets (GSE28735 and GSE62452) to see whether this differential expression is widespread, which showed a significantly less trend of increase (Supplementary Figures S1A,B) (Zhang et al., 2012; Yang et al., 2016). Considering that the samples of GSE28735 and GSE62452 were taken from tumor and paired adjacent normal tissue, while control samples for the TCGA cohort were derived from healthy pancreas samples from a different cohort, the batch effects may partially account for the difference. Nevertheless, all three cohorts revealed unequivocally that PRGs were activated in PAAD and 18 of these PRGs were overexpressed in all of the datasets when setting $p < 0.05$ as threshold. We next enriched these 18 PRGs into pyroptosis signaling pathways and discovered that caspase-1, 3, and 8-dependent pyroptosis, as well as gasdmin B-mediated pyroptosis, were all closely related with pancreatic cancer (Supplementary Figure S1C). In general, multiple pyroptosis mechanisms are commonly activated in pancreatic cancer.

Then, principal component analysis was processed to identify PRGs expression characteristics between normal pancreatic tissue and PAAD, which revealed a clear distinction among samples (Figure 2C). To achieve a better understanding of the relationship among PRGs, the correlation matrix was constructed by calculating the Pearson correlation coefficient between each two genes within either normal samples from the GTEx cohort or PAAD samples from the TCGA cohort. In normal pancreatic tissue, the majority of PRGs were found to be remarkably positively linked with each others while only five genes were shown to be adversely connected to other PRGs, including NLRP2, GSDMA, CASP5, NLRP1, and NOD2 (Figure 2D). Among the PAAD samples, the expression of PRGs was likewise

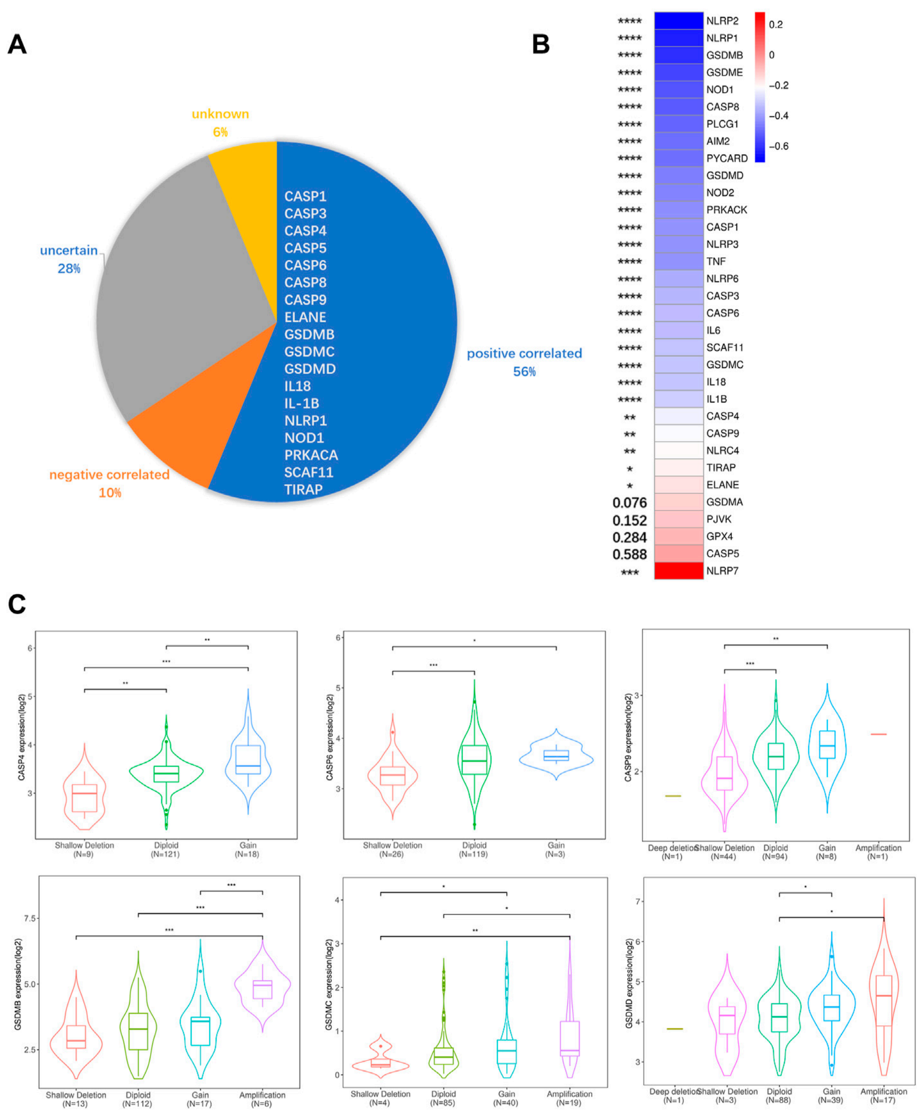


FIGURE 3
DNA methylation, CNV, and gene expression correlation analysis. **(A)** Correlations between CNV and PRGs expression. Positive correlation was defined as certain PRGs expression increased while copy number augmented. Negative correlation was defined as certain PRGs expression decreased while copy number augmented. Uncertain was defined as both expression increasement and decrement can be observed while copy number augmented. Unknown was defined as no significant differences between different CNV groups. **(B)** Pearson correlative value between methylation (HM450) versus mRNA expression z-scores relative to all samples of each PRGs. **(C)** Violin plots of example positive correlated PRGs. The rest PRGs are presented in Supplementary Figure S2. Significance was determined using the Mann-Whitney or unpaired t-test. Data shown are means ± SD, *$p < 0.05$, **$p < 0.01$, ***$p < 0.001$, ****$p < 0.0001$.

positively correlated, which suggested that the co-interaction of PRGs may have a role in PAAD development (Figure 2E).

DNA methylation and copy number variation affect the pyroptosis-related genes expression

To elucidate possible explanations for the increased expression of PRGs in the TCGA cohort, we analyzed DNA methylation and CNV. Both DNA methylation and CNV have been implicated in the regulation of gene expression in a variety of cancers (Stranger et al., 2007; Daniel et al., 2011). To ascertain if CNV influences PRGs expression, we divided the TCGA cohort into five or fewer groups based on their copy number for each gene, which included deletion, shallow deletion, diploid, gain, and amplification. We discovered that copy number is positively correlated with gene expression in more than half of the PRGs, suggesting a significant role for CNV in gene regulation. Besides that, copy number is negatively correlated with gene expression in 10% of PRGs and has no correlation in the

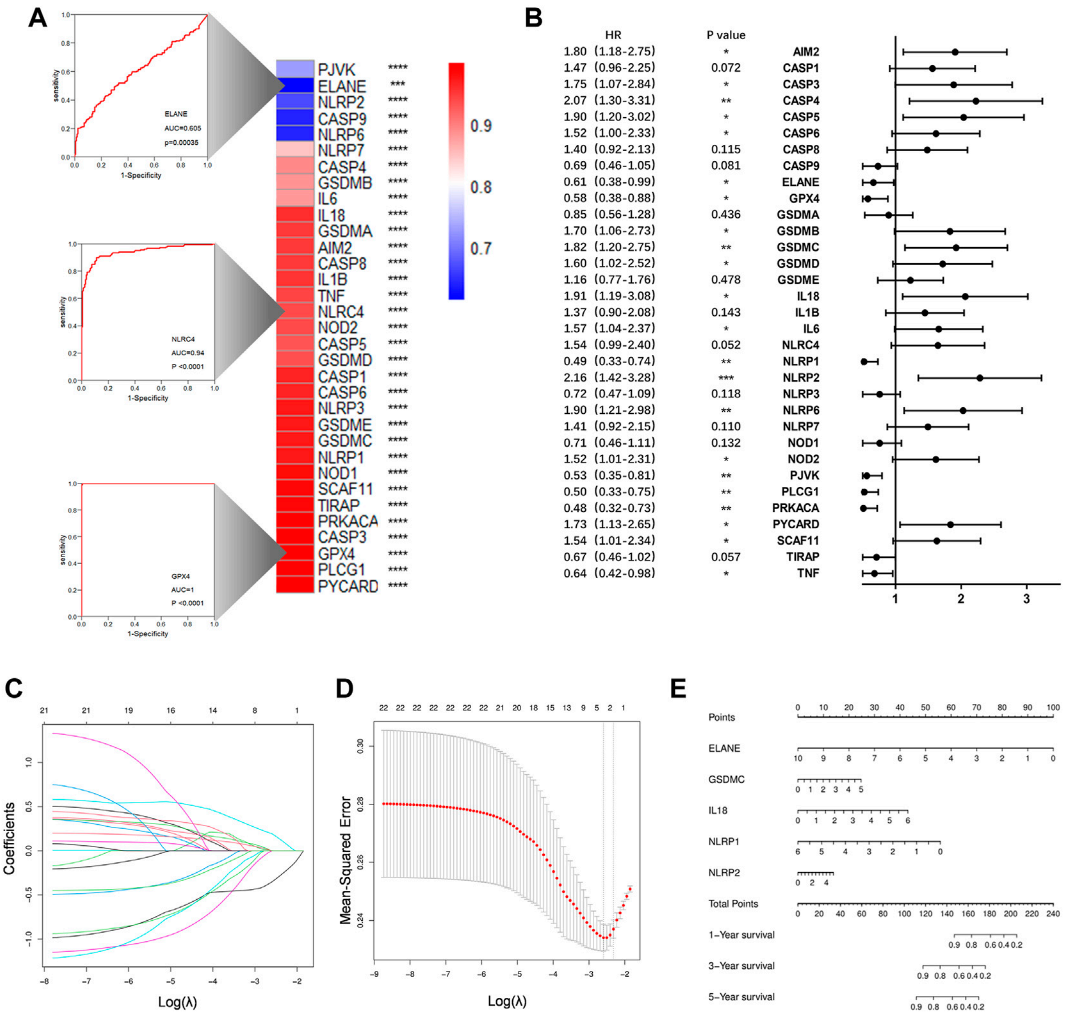


FIGURE 4
Construction of a prognostic prediction model. **(A)**Heatmap to show AUC values for each PRGs. Three example ROC curves are displayed on the left. **(B)** Hazard ratios analyzed *via* univariate cox regression to evaluate the prognostic ability for each PRGs. **(C)** LASSO coefficient profile of PRGs. **(D)** Ten times cross-validation for parameter selections in the LASSO cox regression. **(E)**The nomogram incorporating 5 selected PRGs. *$p < 0.05$, **$p < 0.01$, ***$p < 0.001$, ****$p < 0.0001$.

remaining PRGs. (Figures 3A,C; Supplementary Figure S2). Since the CNV alone could not fully account for the increased PRGs expression, we performed a correlation analysis between DNA methylation and PRGs expression, revealing that the expression of 28/33 PRGs is negatively correlated with DNA methylation (Figure 3B). This indicates both DNA demethylation and copy number increasement contribute to the overexpression of PRGs in PAAD.

Construction of a prognostic gene signature

The ROC curves for each PRGs revealed that the majority of PRGs had a high predictive value for diagnosis, implying that they may contribute to PAAD tumorigenesis (Figure 4A). To further assess their prognostic potential, we performed a univariate cox analysis between each PRG and OS, and 22 genes were screened out (with $p < 0.05$) (Figure 4B). Lasso regression analysis was then used to identify the most prognostic genes, and 5 genes were chosen by the vertical grey line in Figure 4D (Figures 4C,D). Finally, the model was determined by multivariate cox regression within selected PRGs. Among them, GSDMC, IL18, and NLRP2 are all associated with an increased risk, while the other two confer a protective effect (Figure 4E). The formula of the risk score was: risk score = (GSDMC*0.2302) -(ELANE*0.4664)+ (IL18*0.3341)—(NLRP1*0.4324)+ (NLRP2*0.1297). Taking the median risk score as the cut-off value, we classified all TCGA patients into low- and high-risk groups. Detailed clinical

TABLE 1 Clinical characteristics between risk score related subgroups.

	Total (n = 177)	Risk level		p-value
		low (n = 88)	High (n = 89)	
Age(year)				0.8259
<65	81 (45.76%)	41 (46.59%)	40 (44.94%)	
≥65	96 (54.24%)	47 (53.41%)	49 (55.06%)	
Gender				0.4021
Male	97 (54.80%)	51 (57.95%)	46 (51.69%)	
Female	80 (45.20%)	37 (42.05%)	43 (48.31%)	
TNM stage				0.2225
Stage I	21 (11.86%)	14 (15.91%)	7 (7.87%)	
Stage II	145 (81.92)	70 (79.55.22)	75 (84.27%)	
Stage III-IV	8 (4.52%)	3 (3.41%)	5 (5.62%)	
Unknown	3 (1.69%)	1 (1.14%)	2 (2.25%)	
Histologic grade				0.0085
G1-G2	125 (70.62%)	70 (79.55%)	55 (61.80%)	
G3-G4	50 (28.25%)	17 (19.32%)	33 (37.08%)	
unknown	2 (1.13%)	1 (1.14%)	1 (1.12%)	
Disease type				0.0553
Adenomas and adenocarcinomas	30 (16.95%)	21 (23.86%)	9 (10.11%)	
Cystic, mucinous, and serous neoplasms	5 (2.82%)	3 (3.41%)	2 (2.25%)	
Ductal and lobular neoplasms	141 (79.66%)	63 (71.59%)	78 (87.64%)	
Epithelial neoplasms, NOS	1 (0.56%)	1 (1.14%)	0	
Family history of cancer				0.5449
YES	62 (35.03%)	32 (36.36%)	30 (33.71%)	
NO	47 (26.55%)	27 (30.68%)	20 (22.47%)	
Unknown	68 (38.42%)	29 (32.95%)	39 (43.82%)	
Family history of pancreatitis				0.7299
Yes	13 (7.34%)	6 (6.82%)	7 (7.87%)	
No	127 (71.75)	65 (73.86%)	62 (69.66%)	
Unknown	37 (20.90%)	17 (19.32%)	20 (22.47%)	
Overall survive				<0.0001
Alive	85 (48.02%)	59 (67.05%)	26 (29.21%)	
Dead	92 (51.98%)	29 (32.95%)	63 (70.79%)	

information is presented in Table 1 Regardless of histologic stage, disease type or OS, the majority of clinicopathological characteristics are evenly distributed among two groups. An increased risk score, on the other hand, may indicate a higher histological grade and a greater likelihood of ductal and lobular origins.

Prognostic value of pyroptosis-related genes signature in TCGA and validation cohort

To assess the prognostic efficacy of this signature, we calculated the probability of 3-years OS in the TCGA cohort (Figure 5A) and a validation cohort, ICGC (Supplementary Figure S3A). The results indicated that the model had a high predictive capacity in both cohorts. Additionally, time dependent ROC analysis was used to assess the sensitivity and specificity of this model. As for the TCGA cohort, beside 1-year, both 3-years and 5-years corresponding areas under the curve (AUC) are over 0.75 (Figure 5B), whereas the ICGC cohort's accuracy is lower, with a 1-year AUC of 0.661 and a 3-years AUC of 0.528 (Supplementary Figure S3B). However, its poor performance for predicting longer time survival status may be explained by the fact that only 10% of patients in the ICGC cohort survive till the third year. Following that, similar to the TCGA cohort, all 89 patients in the ICGC cohort were equally divided into low- and high-risk groups based on their risk score, and we observed

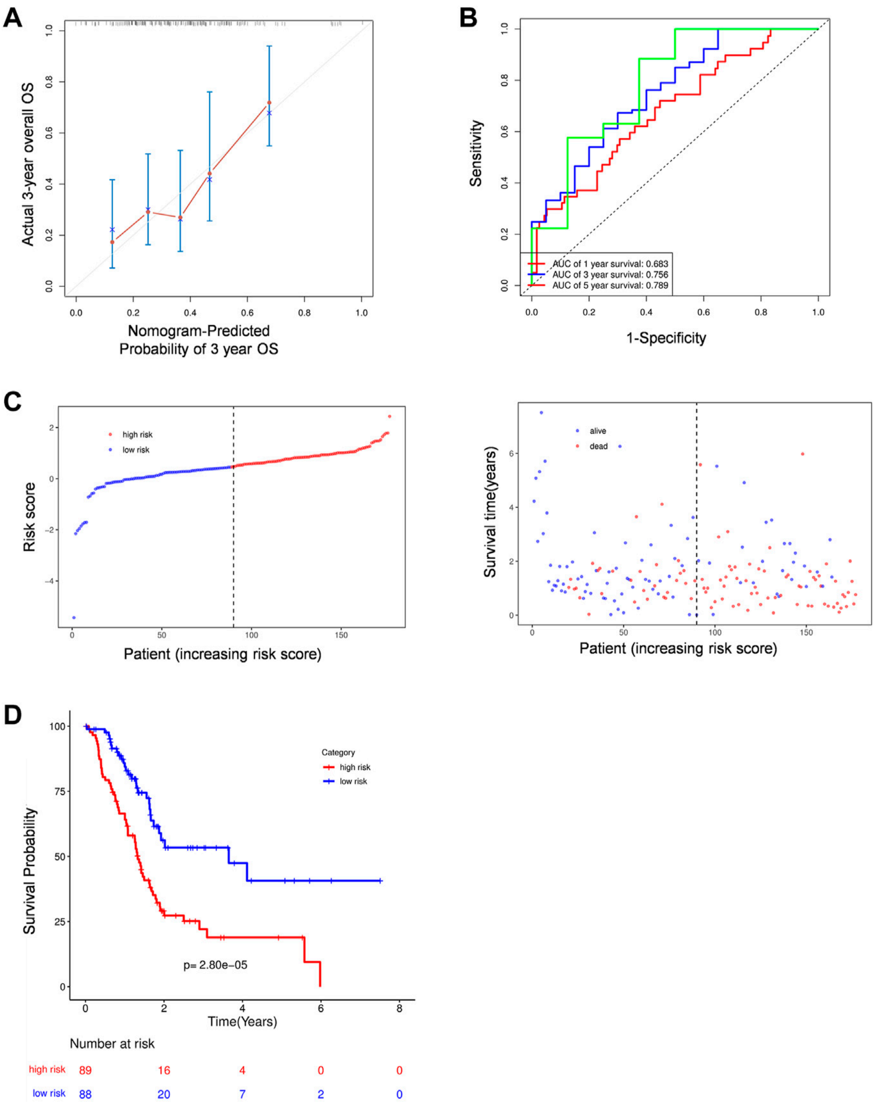


FIGURE 5
The prognostic analysis of PRGs signature. **(A)** Calibration plots of the nomogram for predicting OS within 3 years basing on PRGs signature in the TCGA cohorts. **(B)** Time dependent ROC analysis in the TCGA cohort. **(C–D)** The plots of risk score and alive status**(C)** as well as Kaplan-Meier survival analysis **(D)** in the TCGA cohorts.

an obvious difference in OS between the two groups. Higher risk patients were associated with more deaths and tended to have shorter survival time in both cohorts (Figures 5C,D; Supplementary Figures S3C,D).

pyroptosis-related genes model outperforms clinical characteristics in prognosis

Following that, we compare the predictive accuracy of our model to that of clinicopathological characteristics. Both univariate and multivariate analyses indicated that risk score is an independent predictor, moreover, age and disease type also demonstrated their independent predictive ability with $p < 0.1$ as the threshold value (Table 2). After combining these three variables, a nomogram model was built to evaluate its clinical utility (Figure 6A). Then, we processed decision curve and ROC analysis to compare the clinical benefit of the composite nomogram to that of a risk score or clinical characteristics alone. While the composite model performed better than the basic clinical factors in terms of prognosis accuracy, it demonstrated limited clinical net benefit compared to the risk score (Figure 6B). Additionally, the time-related AUCs of the risk

TABLE 2 Univariate and multivariate cox regression analysis for prognostic model and clinical characteristics.

Variable	Univariate analysiss		Multivariate analysiss	
	Hazard ratio (95% Cl)	*p*-value	Hazard ratio (95% Cl)	*p*-value
Risk score	2.72 (1.88–3.93)	<0.0001	2.52 (1.69–3.76)	<0.0001
Age	1.03 (1.01–1.05)	0.0076	1.02 (1.00–1.04)	0.0559
Gender				
Female	References			
Male	0.81 (0.54–1.22)	0.3111		
Tumor stage				
Stage I	References			
Stage II	2.33 (1.07–5.09)	0.0334		
Stage III	1.25 (0.15–10.28)	0.8323		
Stage IV	1.56 (0.32–7.61)	0.5824		
Histology grade				
G1	References			
G2	1.95 (1.00–3.79)	0.0487		
G3	2.62 (1.30–5.27)	0.0071		
G4	1.65 (0.21–12.85)	0.6346		
Disease type				
AA[a]	References		References	
CMS[a]	4.80 (1.27–18.21)	0.0210	3.24 (0.82–12.86)	0.0948
DL[a]	3.16 (1.52–6.57)	0.0020	1.52 (0.71–3.25)	0.2786
History of chronic pancreatitis				
No	References			
Yes	1.18 (0.56–2.47)	0.6649		
Family history of cancers				
No	References			
Yes	1.12 (0.65–1.92)	0.6858		

[a]AA, is short for Adenomas and Adenocarcinomas.
CMS, is short for Cystic, Mucinous and Serous Neoplasms.
DL, is short for Ductal and Lobular Neoplasms.

score model were consistently greater than those of the composed model at each time point, suggesting that the risk score possessed the greatest clinical utility (Figure 6C).

Bioinformation analysis based on the pyroptosis-related genes model

We identified 365 genes with increased expression and 1,514 genes with decreased expression in the high-risk group as compared to the low-risk group (Figures 7A,B). These DEGs were then used to conduct KEGG enrichment and GSEA analysis to further investigate the biological pathway correlated with risk score. Interestingly, DEGs were predominantly enriched in organismal systems such as endocrine, nervous, and circulatory systems (Figure 7C). Meanwhile, the GSEA results demonstrate that several pathways, including calcium signaling, cAMP signaling, cGMP-PKG signaling pathways and so on, are down-regulated in the high-risk group (Figure 7D). Apart from functional analysis, we then looked at the somatic mutation status of TCGA patients. As expected, high-risk individuals have a considerably higher somatic mutation burden, typically for the genes KARS and TP53, which are known to be the primary drivers of PAAD (Kleeff et al., 2016) (Figure 7E; Supplementary Figure S4A). Consistently, the tumor mutation burden (TMB) was also found to be considerably greater in the high-risk group than in the low-risk group (Supplementary Figure S4B).

Given that KRAS and TP54 have been linked to other cell death processes such apoptosis and ferroptosis, we attempted to identity the specific correlation between oncogenes and pyroptosis by comparing the expression of PRGs between KRAS or TP53 mutated and unmutated individuals (Chen et al., 2021). Despite the fact that GSDMC, NOD2, and

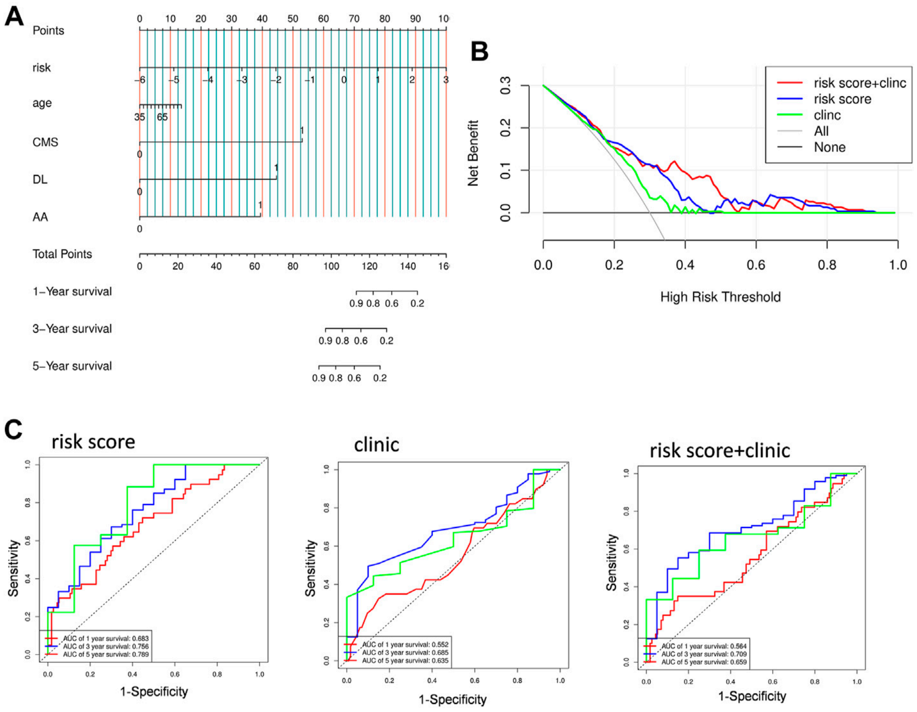


FIGURE 6
Validation prognostic efficiency of PRGs signature. **(A)** Nomogram predicted 1- ,3-, and 5-years OS based on prognostic model combined with clinical characteristic in the TCGA cohort. CMS: Cystic, Mucinous and Serous Neoplasms; DL: Ductal and Lobular Neoplasms; AA: Adenomas and Adenocarcinomas. **(B)** The decision curve of the risk score, clinical characteristic and their combination. **(C)** time-dependent ROC curves for the risk score, clinical characteristic, or their combination.

IL18 were modestly elevated while NLRP1 and NLRP6 were lowered, the majority of PRGs between the mutant and non-mutant groups were not significantly different (data not shown). The link between pyroptosis and gene mutation is not evident based on the existing findings, and more research is needed to understand the particular interaction between the two.

Immunity features underlying the pyroptosis-related genes model

We further characterize their immune environment heterogeneity by elucidating the association between risk score and immune state. The ESTIMATE web tool was first used to determine cell distribution, and it revealed that high-risk group had significantly less stromal cell and immune cell infiltration. Meanwhile, the testing group ICGC cohort presented a similar trend, though without a statistically significant difference (Figure 8A; Supplementary Figure S4C). Additionally, the compositions of specific cell types were determined through ssGSEA, showing that the infiltration of a considerable number of immune cell types were reduced in high risk group, including effector memory CD4+T-cells, effector memory CD8+T-cells, and type I helper cells, which are known to have anti-tumor effects. Apart from these, eosinophils, macrophages, mast cells, monocytes, myeloid derived suppressor cells, and plasmacytoid dendritic cells were found to be adversely associated with risk score (Figures 8B,C).

Therapy response features underlying the pyroptosis-related gene model

We suspected that a higher risk score would be correlated with a weaker response to immunotherapy and other bio-agents, given that patients in the high-risk group exhibited reduced immune cell infiltration. Then, the TIDE analysis corroborated our hypothesis, demonstrating that individuals at low-risk are more likely to respond to ICI treatment but without statistical significance (Figure 9A). Moreover, patients in the high-risk group have higher exclusion score but a lower dysfunction score, suggesting that immunological exclusion was the primary cause of their poor outcomes (Figure 9A). Notably, while both increased and decreased expression of the ICI target gene can be observed, the link

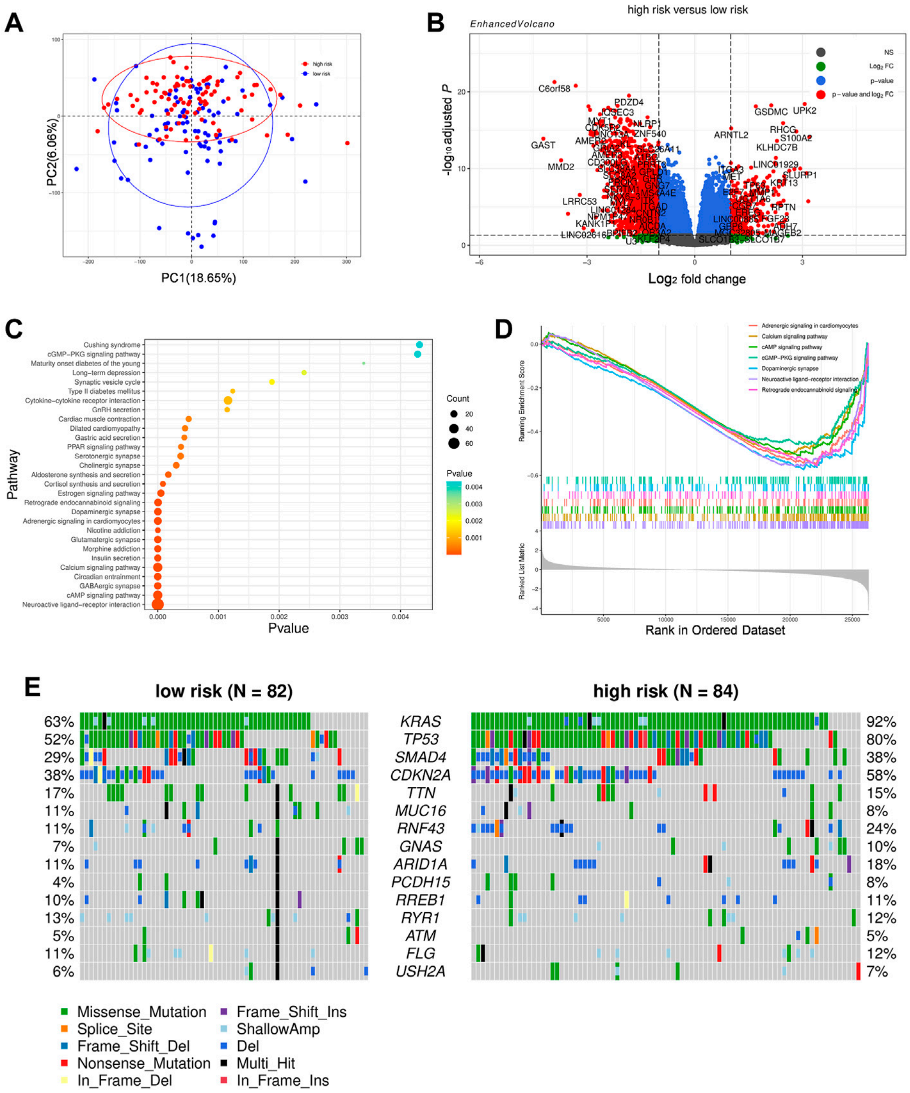


FIGURE 7
Comparison of the subgroups of TCGA cohort. **(A)** Principal component analysis of the TCGA cohort grouped by high and low risk. **(B)** A volcano plot represented DEGs between the high- and low-risk groups of TCGA cohort. **(C)** Function enrichment analysis of DEGs based on the KEGG signaling pathway. **(D)**GSEA result of DEGs based upon KEGG signaling pathway. **(E)** Distribution of frequently mutated genes in different TCGA subgroups.

between specific ICI and risk score requires further investigation (Supplementary Figure S4D). Apart from that, we used onco predict to predict the IC50 values for FDA-approved drugs in high- and low- risk patients. Among the six most commonly used drugs, the low-risk group had considerably lower projected IC50 values for olaparib, irirntecan, and gemcitabine, implying that lower risk is associated with better outcomes from these chemotherapeutic drugs (Figure 9B). Overall, patients in the high-risk group were less sensitive to both immunotherapy and chemotherapy in general, which may have contributed to their poor prognosis.

Discussion

PAAD is always diagnosed at an advanced stage because of the lack of identifiable symptoms, and only a minority of patients can benefit from conventional surgical treatment or cytotoxic chemotherapy (Von Hoff et al., 2013; Walter et al., 2016). As a result, PAAD is currently one of the top 10 most lethal tumors (Rahib et al., 2014). The immunosuppressive and desmoplastic milieu of PAAD is a substantial impediment to optimizing therapeutic efficacy, including difficulties in drug transport and limited responses to ICI-based immunotherapy (Li et al., 2020). Stimulating the immunogenic cell death of tumor cells is

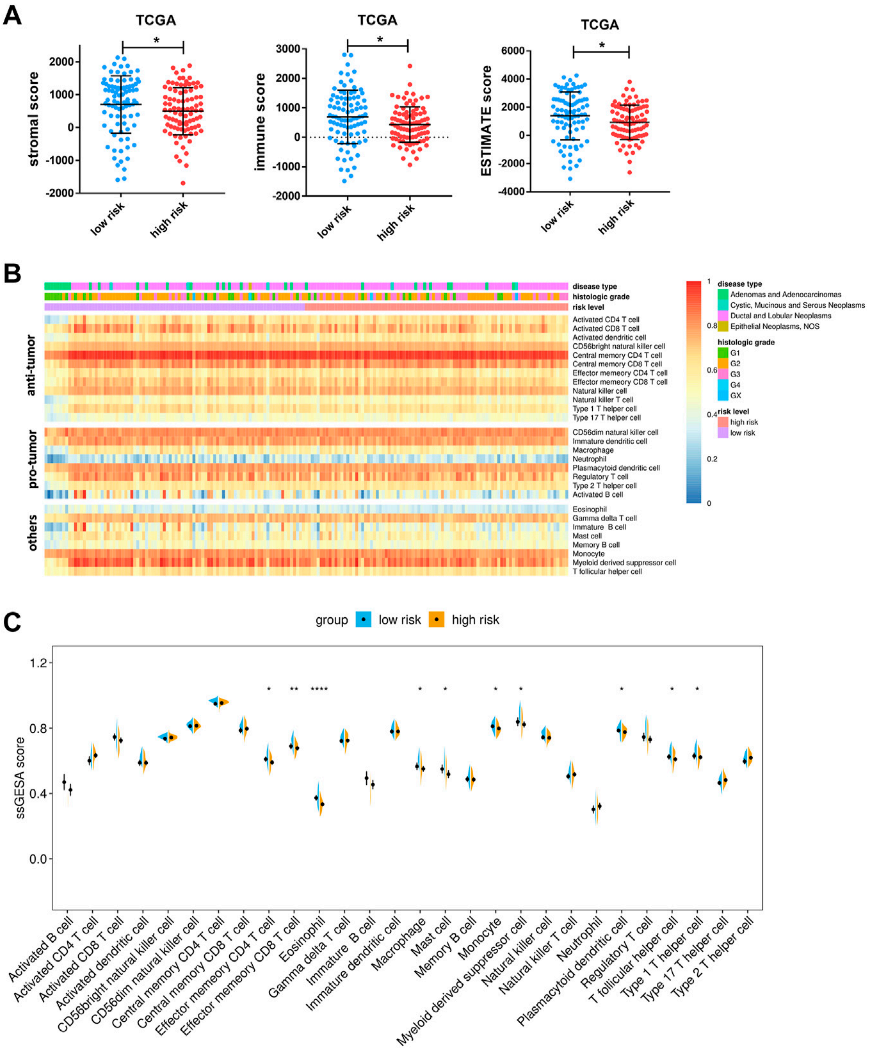


FIGURE 8
Associations between risk score and tumor microenvironment. **(A)** Comparison of stromal scores, immune scores, and ESTIMATE scores between the high- and low-risk groups of TCGA cohorts. **(B)** Heatmap of ssGSEA enrichment scores of 28 immune cell types in the TCGA cohort. Notably, the cells are grouped according to their widely accepted role in cancer, including anti-tumor, pro-tumor, and others. **(C)** Comparison of ssGSEA enrichment scores of 28 types of immune cells between the high- and low-risk groups in the TCGA cohort. Data are presented as means ± SD. Significant was determined using Mann-Whitney or unpaired *t*-test. *$p < 0.05$ **$p < 0.005$, and ****$p < 0.00005$.

regarded to be an efficient method of converting the “cool” tumor microenvironment to a “hot” environment (Kroemer et al., 2013). Given that tumor cells show intrinsic resistance to apoptosis, targeting pyroptosis might be a more efficient strategy for boosting immunotherapy (Huang et al., 2018). Our study investigated the combined effects of various PRGs in PAAD and developed a prognostic model capable of reliably predicting patient survival status and response to prospective targeted therapy.

In this study, we were surprised to find that the majority of PRGs expressed significantly differently between normal pancreatic tissue and PAAD, reflecting a fundamental change in pyroptosis activity. Gene overexpression can occur for a variety of reasons, including gene amplification, activating mutation, or epigenetic modification (Stranger et al., 2007; Daniel et al., 2011). In our case, most of these upregulations occur in part as a result of increased copy number or demethylation. Additionally, the majority of overexpressed PRGs are strongly associated with poor prognosis, indicating that they may contribute to survival state prediction. Thus, using univariate cox and lasso regression to avoid overfitting, five prognostic PRGs were chosen. Following that, we generated a signature comprised of five PRGs (ELANE, GSDMC, IL18, NLRP1, and NLRP2) by multivariate cox, which named risk

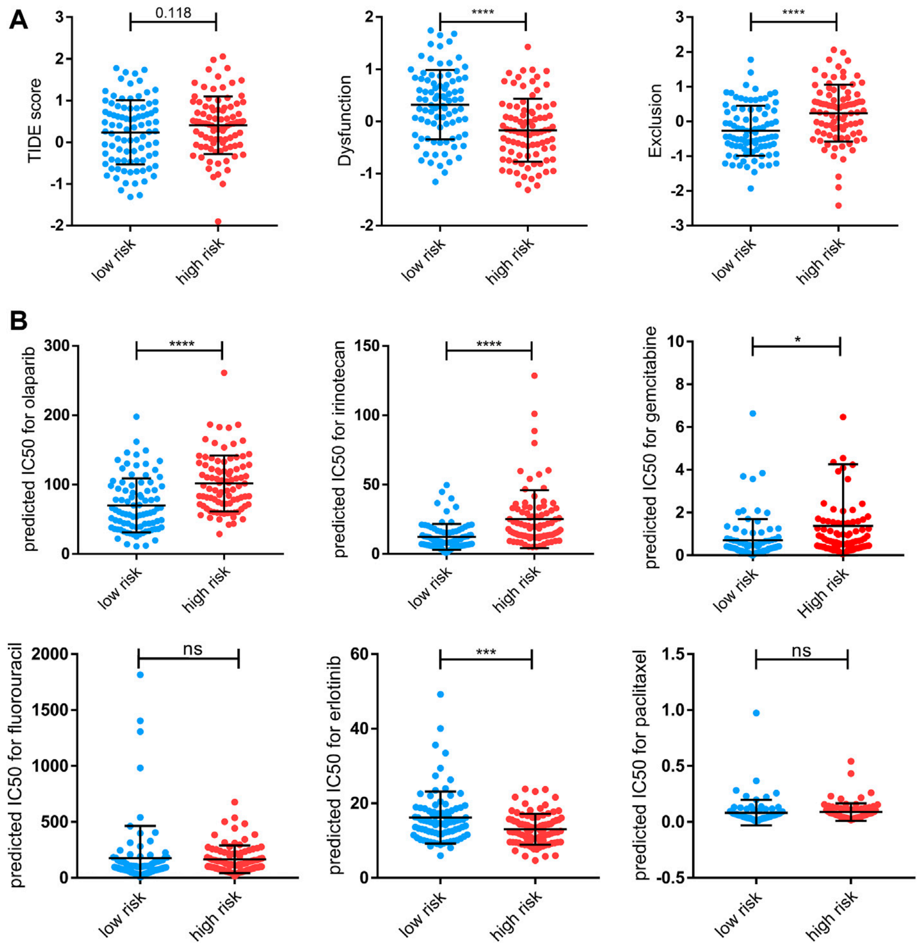


FIGURE 9
Therapy response features underlying the PRGs model. **(A)** Comparison of TIDE score, T-cell dysfunction ("Dysfunction") score, and T-cell exclusion ("Exclusion") scores between the high- and low-risk groups of the TCGA cohort. **(B)** Predicted IC50 for olaparib, irinotecan, gemcitabine, fluorouracil, erlotinib, and paclitaxel for low-risk and high-risk groups. Data shown are means ± SD. Symbols represent the individual patients. Significant was determined using the Mann-Whitney or unpaired *t*-test. **p* < 0.05 ***p* < 0.005, ****p* < 0.0005, and *****p* < 0.00005.

score, and validated its accuracy in both the training and validation cohorts. Among these core genes, higher ELANE and NLRP1 expression suggested a favorable prognosis for the patients. Consistently, Cui et al. (2021) recently demonstrated that neutrophil-derived active neutrophil elastase (ELANE) not only kills numerous types of cancer cells while sparing proximal non-cancer cells by liberating the CD95 death domain that interacts with histone H1 isoforms, but also inhibits metastasis *via* CD8+T mediated abscopal effect. Furthermore, it has been discovered that NLRP1 downregulation promotes tumorigenesis, including lung adenocarcinoma and colorectal cancer (Chen et al., 2015; Shen et al., 2021). On the other hand, overexpression of GSDMC, IL18, and NLRP2 were associated with a poor prognosis in patients with PAAD. Hou et al. (2020) showed GSDMC mediated non-canonical pyroptosis upon caspase-8 activation and that high GSDMC expression correlated with poor survival. It is difficult to thoroughly elucidate the role of IL18 in cancer. A high level of IL18 in pancreatic tumor tissue was associated with a shorter survival time, increased invasion, and metastasis, whereas a high IL18 level in plasma was correlated with a longer survival time (Guo et al., 2016). By combining our signature with previous studies, we were able to confirm and truly illustrate the predictive usefulness of these core PRGs.

Additionally, the singnature revealed differences in several pathways between the two groups. Due to the fact that the number of downregulated genes was much more than the number of upregulated genes, the majority of pathways, such as GABAergic synapase and insulin secretion, were enriched by downregulated genes, and these pathways may have a correlation with PAAD progression and prognosis. For example, gaba suppresses PAAD by inhibiting the β-adrenergic cascade and nicotine-induced cell proliferation (Al-Wadei et al., 2011; Al-Wadei et al., 2013; Al-Wadei et al., 2016). cAMP has both pro- and anti-tumor effects in malignancies (Tagliaferri et al., 1988; Ligumsky et al., 2012; Almahariq et al., 2015); To our surprise, the calcium signaling pathway and the neuroactive ligand-receptor interaction pathway, both of which are associated with a poor prognosis (Bettaieb et al., 2021; Qian et al., 2021), were downregulated in the high-risk group. However, the link between pyroptosis and these pathways is currently unknown and needs further investigation.

The pro- or anti- tumor effects of proptosis are somehow determined by the surrounding microenvironment (Hou et al., 2021). Several investigators reported the pyroptosis of tumor cells can induce inflammatory response in microenvironment and attracting CD4^{+} and CD8+T-cell populations (Wang et al., 2020). In our case, though multiple PRGs are robustly overexpressed within PAAD, it is evident pancreatic tumor microenvironment exhibits an immunosuppressive condition (Zhu et al., 2014; Jiang et al., 2016; Kumar et al., 2022). One possible explanation for this is that, unlike acute pyroptosis induction, chronic induction of pyroptosis in some tumors can result chronic inflammation, which leads to a tumor-promoting microenvironment (Tsuchiya, 2021). Besides, extracellular ATP released from pyroptotic cells can be rapidly broken down into adenosine, an immunosuppressive substance, the gradual release of modest amounts of ATP from pyroptotic tumor cells may impact antitumor immunity (Vultaggio-Poma et al., 2020; Tsuchiya, 2021). Apart from that, the pytoptosis that happened in the center region of the tumor could result in chronic tumor necrosis, which suppressed the anti-tumor immunity and accelerated tumor progression (Hou et al., 2020). In our model, patients with lower risk scores were infiltrated with more immune cells, including several anti-tumor immune cells. So that if therapy-induced pyroptosis is expected to improve the pancreatic tumor microenvironment it may be important to determine the appropriate extent of pyroptosis induction, which should be neither too strong nor too weak (Tsuchiya, 2021).

Apart from the immune cell landscape, this signature also showed a significant correlation with somatic mutation status and therapeutic response. The patients with higher risk scores carried more mutation burden, with more mutations in KARS, TP53, ADAMTS12, SMAD4 FAT4, DCHS1, and CDKN2A mutations. Among these genes, KARS, CDKN2A, TP53, and SMAD4 are four major genes involved in the progression of PAAD (Kleeff et al., 2016). However, it is unclear whether these oncogenes are involved in pyroptosis. Moreover, TIDE analysis revealed that PAAD patients with lower risk scores had a higher likelihood of achieving durable benefits from immunotherapy. PAAD is also characterized by a remarkable tolerance to chemotherapy (Kleeff et al., 2016). Thus, to test the PRGs signature's predictive utility in clinical practice, we next predicted the sensitivity to FDA-proved PAAD chemotherapeutic drugs based on gene expression profiles. Similar to immunotherapy, a low-risk score was associated with a better response to olaparib, irinotecan, and gemcitabine. In general, our findings demonstrated that patients with low-risk scores were more likely to be have a reduced mutation burden and benefit from both immunotherapy and chemotherapy.

In this study, we created a valuable PRGs signature and thoroughly explored its correlations with prognosis, immune infiltration, somatic gene mutation, and treatment response. Our model performs well in predicting patient prognosis and treatment response. Moreover, we laid the groundwork for a more complete understanding of pyroptosis's role in PAAD. However, our work is still in its early stage and the limitations of this study are clear. Further clinical trials need to be conducted to fully verify the accuracy of this model. The true involvement of pyroptosis in cancer remains a mystery, and additional researches are required.

Data availability statement

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

Author contributions

Study concept and design: ST, YS, and XW. Data analysis and interpretation: ST, YS, and LT. Manuscript writing: ST and YS. Final approval of manuscript: All authors.

Conflict of interest

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

Publisher's note

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

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2022.817919/full#supplementary-material

References

Al-Wadei, H. A., Ullah, M. F., and Al-Wadei, M. (2011). GABA (gamma-aminobutyric acid), a non-protein amino acid counters the beta-adrenergic cascade-activated oncogenic signaling in pancreatic cancer: A review of experimental evidence. *Mol. Nutr. Food Res.* 55 (12), 1745–1758. doi:10.1002/mnfr.201100229

Al-Wadei, M. H., Al-Wadei, H. A., and Schuller, H. M. (2013). Gamma-amino butyric acid (GABA) prevents the induction of nicotinic receptor-regulated signaling by chronic ethanol in pancreatic cancer cells and normal duct epithelia. *Cancer Prev. Res.* 6 (2), 139–148. doi:10.1158/1940-6207.CAPR-12-0388

Al-Wadei, M. H., Banerjee, J., Al-Wadei, H. A., and Schuller, H. M. (2016). Nicotine induces self-renewal of pancreatic cancer stem cells *via* neurotransmitter-driven activation of sonic hedgehog signalling. *Eur. J. Cancer* 52, 188–196. doi:10.1016/j.ejca.2015.10.003

Almahariq, M., Chao, C., Mei, F. C., Hellmich, M. R., Patrikeev, I., Motamedi, M., et al. (2015). Pharmacological inhibition and genetic knockdown of exchange protein directly activated by cAMP 1 reduce pancreatic cancer metastasis *in vivo*. *Mol. Pharmacol.* 87 (2), 142–149. doi:10.1124/mol.114.095158

Bettaieb, L., Brule, M., Chomy, A., Diedro, M., Fruit, M., Happernegg, E., et al. (2021). Ca(2+) signaling and its potential targeting in pancreatic ductal carcinoma. *Cancers (Basel)* 13 (12), 3085. doi:10.3390/cancers13123085

Blanche, P., Dartigues, J. F., and Jacqmin-Gadda, H. (2013). Estimating and comparing time-dependent areas under receiver operating characteristic curves for censored event times with competing risks. *Stat. Med.* 32 (30), 5381–5397. doi:10.1002/sim.5958

Brown, M. (2018). rmda: Risk model decision analysis. Available: https://CRAN.R-project.org/package=rmda.

Cerami, E., Gao, J., Dogrusoz, U., Gross, B. E., Sumer, S. O., Aksoy, B. A., et al. (2012). The cBio cancer genomics portal: An open platform for exploring multidimensional cancer genomics data. *Cancer Discov.* 2 (5), 401–404. doi:10.1158/2159-8290.CD-12-0095

Chen, C., Wang, B., Sun, J., Na, H., Chen, Z., Zhu, Z., et al. (2015). DAC can restore expression of NALP1 to suppress tumor growth in colon cancer. *Cell Death Dis.* 6, e1602. doi:10.1038/cddis.2014.532

Chen, X., Zeh, H. J., Kang, R., Kroemer, G., and Tang, D. (2021). Cell death in pancreatic cancer: From pathogenesis to therapy. *Nat. Rev. Gastroenterol. Hepatol.* 18 (11), 804–823. doi:10.1038/s41575-021-00486-6

Clark, C. E., Hingorani, S. R., Mick, R., Combs, C., Tuveson, D. A., and Vonderheide, R. H. (2007). Dynamics of the immune reaction to pancreatic cancer from inception to invasion. *Cancer Res.* 67 (19), 9518–9527. doi:10.1158/0008-5472.CAN-07-0175

Cui, C., Chakraborty, K., Tang, X. A., Zhou, G., Schoenfelt, K. Q., Becker, K. M., et al. (2021). Neutrophil elastase selectively kills cancer cells and attenuates tumorigenesis. *Cell* 184 (12), 3163–3177. e3121. doi:10.1016/j.cell.2021.04.016

Cui, J., Zhou, Z., Yang, H., Jiao, F., Li, N., Gao, Y., et al. (2019). MST1 suppresses pancreatic cancer progression via ROS-induced pyroptosis. *Mol. Cancer Res.* 17 (6), 1316–1325. doi:10.1158/1541-7786.MCR-18-0910

Daniel, F. I., Cherubini, K., Yurgel, L. S., de Figueiredo, M. A., and Salum, F. G. (2011). The role of epigenetic transcription repression and DNA methyltransferases in cancer. *Cancer* 117 (4), 677–687. doi:10.1002/cncr.25482

Galluzzi, L., Humeau, J., Buque, A., Zitvogel, L., and Kroemer, G. (2020). Immunostimulation with chemotherapy in the era of immune checkpoint inhibitors. *Nat. Rev. Clin. Oncol.* 17 (12), 725–741. doi:10.1038/s41571-020-0413-z

Goldman, M. J., Craft, B., Hastie, M., Repecka, K., McDade, F., Kamath, A., et al. (2020). Visualizing and interpreting cancer genomics data via the Xena platform. *Nat. Biotechnol.* 38 (6), 675–678. doi:10.1038/s41587-020-0546-8

Grambsch, T. M. T. A. P. M. (2000). *Modeling survival data: Extending the cox model.* New York: Springer.

Guo, X., Zheng, L., Jiang, J., Zhao, Y., Wang, X., Shen, M., et al. (2016). Blocking NF-κB is essential for the immunotherapeutic effect of recombinant IL18 in pancreatic cancer. *Clin. Cancer Res.* 22 (23), 5939–5950. doi:10.1158/1078-0432.CCR-15-1144

Hanzelmann, S., Castelo, R., and Guinney, J. (2013). Gsva: Gene set variation analysis for microarray and RNA-seq data. *BMC Bioinforma.* 14, 7. doi:10.1186/1471-2105-14-7

Ho, W. J., Jaffee, E. M., and Zheng, L. (2020). The tumour microenvironment in pancreatic cancer - clinical challenges and opportunities. *Nat. Rev. Clin. Oncol.* 17 (9), 527–540. doi:10.1038/s41571-020-0363-5

Hou, J., Hsu, J. M., and Hung, M. C. (2021). Molecular mechanisms and functions of pyroptosis in inflammation and antitumor immunity. *Mol. Cell* 81, 4579–4590. doi:10.1016/j.molcel.2021.09.003

Hou, J., Zhao, R., Xia, W., Chang, C. W., You, Y., Hsu, J. M., et al. (2020). PD-L1-mediated gasdermin C expression switches apoptosis to pyroptosis in cancer cells and facilitates tumour necrosis. *Nat. Cell Biol.* 22 (10), 1264–1275. doi:10.1038/s41556-020-0575-z

Huang, X., Xiao, F., Li, Y., Qian, W., Ding, W., and Ye, X. (2018). Bypassing drug resistance by triggering necroptosis: Recent advances in mechanisms and its therapeutic exploitation in leukemia. *J. Exp. Clin. Cancer Res.* 37 (1), 310. doi:10.1186/s13046-018-0976-z

Jiang, H., Hegde, S., Knolhoff, B. L., Zhu, Y., Herndon, J. M., Meyer, M. A., et al. (2016). Targeting focal adhesion kinase renders pancreatic cancers responsive to checkpoint immunotherapy. *Nat. Med.* 22 (8), 851–860. doi:10.1038/nm.4123

Jiang, P., Gu, S., Pan, D., Fu, J., Sahu, A., Hu, X., et al. (2018). Signatures of T cell dysfunction and exclusion predict cancer immunotherapy response. *Nat. Med.* 24 (10), 1550–1558. doi:10.1038/s41591-018-0136-1

Kleeff, J., Korc, M., Apte, M., La Vecchia, C., Johnson, C. D., Biankin, A. V., et al. (2016). Pancreatic cancer. *Nat. Rev. Dis. Prim.* 2, 16022. doi:10.1038/nrdp.2016.22

Kroemer, G., Galluzzi, L., Kepp, O., and Zitvogel, L. (2013). Immunogenic cell death in cancer therapy. *Annu. Rev. Immunol.* 31, 51–72. doi:10.1146/annurev-immunol-032712-100008

Kumar, S., Schoonderwoerd, M. J. A., Kroonen, J. S., de Graaf, I. J., Sluijter, M., Ruano, D., et al. (2022). Targeting pancreatic cancer by TAK-981: A SUMOylation inhibitor that activates the immune system and blocks cancer cell cycle progression in a preclinical model. *Gut* 2021, 324834. doi:10.1136/gutjnl-2021-324834

Li, K. Y., Yuan, J. L., Trafton, D., Wang, J. X., Niu, N., Yuan, C. H., et al. (2020). Pancreatic ductal adenocarcinoma immune microenvironment and immunotherapy prospects. *Chronic Dis. Transl. Med.* 6 (1), 6–17. doi:10.1016/j.cdtm.2020.01.002

Ligumsky, H., Wolf, I., Israeli, S., Haimsohn, M., Ferber, S., Karasik, A., et al. (2012). The peptide-hormone glucagon-like peptide-1 activates cAMP and inhibits growth of breast cancer cells. *Breast Cancer Res. Treat.* 132 (2), 449–461. doi:10.1007/s10549-011-1585-0

Love, M. I., Huber, W., and Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* 15 (12), 550. doi:10.1186/s13059-014-0550-8

Loveless, R., Bloomquist, R., and Teng, Y. (2021). Pyroptosis at the forefront of anticancer immunity. *J. Exp. Clin. Cancer Res.* 40 (1), 264. doi:10.1186/s13046-021-02065-8

Maeser, D., Gruener, R. F., and Huang, R. S. (2021). oncoPredict: an R package for predicting *in vivo* or cancer patient drug response and biomarkers from cell line screening data. *Brief. Bioinform.* 22, bbab260. doi:10.1093/bib/bbab260

Marshall, R. (2020). regplot: Enhanced regression nomogram plot. Available: https://CRAN.R-project.org/package=regplot.

Mayakonda, A., Lin, D. C., Assenov, Y., Plass, C., and Koeffler, H. P. (2018). Maftools: Efficient and comprehensive analysis of somatic variants in cancer. *Genome Res.* 28 (11), 1747–1756. doi:10.1101/gr.239244.118

O'Donnell, J. S., Teng, M. W. L., and Smyth, M. J. (2019). Cancer immunoediting and resistance to T cell-based immunotherapy. *Nat. Rev. Clin. Oncol.* 16 (3), 151–167. doi:10.1038/s41571-018-0142-8

Provenzano, P. P., Cuevas, C., Chang, A. E., Goel, V. K., Von Hoff, D. D., and Hingorani, S. R. (2012). Enzymatic targeting of the stroma ablates physical barriers to treatment of pancreatic ductal adenocarcinoma. *Cancer Cell* 21 (3), 418–429. doi:10.1016/j.ccr.2012.01.007

Qian, X., Jiang, C., Shen, S., and Zou, X. (2021). GPRC5A: An emerging prognostic biomarker for predicting malignancy of Pancreatic Cancer based on bioinformatics analysis. *J. Cancer* 12 (7), 2010–2022. doi:10.7150/jca.52578

Rahib, L., Smith, B. D., Aizenberg, R., Rosenzweig, A. B., Fleshman, J. M., and Matrisian, L. M. (2014). Projecting cancer incidence and deaths to 2030: The unexpected burden of thyroid, liver, and pancreas cancers in the United States. *Cancer Res.* 74 (11), 2913–2921. doi:10.1158/0008-5472.CAN-14-0155

Rebours, V., Gaujoux, S., d'Assignies, G., Sauvanet, A., Ruszniewski, P., Levy, P., et al. (2015). Obesity and fatty pancreatic infiltration are risk factors for pancreatic precancerous lesions (PanIN). *Clin. Cancer Res.* 21 (15), 3522–3528. doi:10.1158/1078-0432.CCR-14-2385

Shen, E., Han, Y., Cai, C., Liu, P., Chen, Y., Gao, L., et al. (2021). Low expression of NLRP1 is associated with a poor prognosis and immune infiltration in lung adenocarcinoma patients. *Aging (Albany NY)* 13 (5), 7570–7588. doi:10.18632/aging.202620

Shi, J., Zhao, Y., Wang, K., Shi, X., Wang, Y., Huang, H., et al. (2015). Cleavage of GSDMD by inflammatory caspases determines pyroptotic cell death. *Nature* 526 (7575), 660–665. doi:10.1038/nature15514

Siegel, R. L., Miller, K. D., Fuchs, H. E., and Jemal, A. (2021). Cancer statistics, 2021. *Ca. Cancer J. Clin.* 71 (1), 7–33. doi:10.3322/caac.21654

Stolzenberg-Solomon, R. Z., Schairer, C., Moore, S., Hollenbeck, A., and Silverman, D. T. (2013). Lifetime adiposity and risk of pancreatic cancer in the NIH-AARP Diet and Health Study cohort. *Am. J. Clin. Nutr.* 98 (4), 1057–1065. doi:10.3945/ajcn.113.058123

Stranger, B. E., Forrest, M. S., Dunning, M., Ingle, C. E., Beazley, C., Thorne, N., et al. (2007). Relative impact of nucleotide and copy number variation on gene expression phenotypes. *Science* 315 (5813), 848–853. doi:10.1126/science.1136678

Tagliaferri, P., Katsaros, D., Clair, T., Ally, S., Tortora, G., Neckers, L., et al. (1988). Synergistic inhibition of growth of breast and colon human cancer cell lines by site-selective cyclic AMP analogues. *Cancer Res.* 48 (6), 1642–1650.

Torphy, R. J., Zhu, Y., and Schulick, R. D. (2018). Immunotherapy for pancreatic cancer: Barriers and breakthroughs. *Ann. Gastroenterol. Surg.* 2 (4), 274–281. doi:10.1002/ags3.12176

Tsuchiya, K. (2021). Switching from apoptosis to pyroptosis: Gasdermin-elicited inflammation and antitumor immunity. *Int. J. Mol. Sci.* 22 (1), E426. doi:10.3390/ijms22010426

Von Hoff, D. D., Ervin, T., Arena, F. P., Chiorean, E. G., Infante, J., Moore, M., et al. (2013). Increased survival in pancreatic cancer with nab-paclitaxel plus gemcitabine. *N. Engl. J. Med.* 369 (18), 1691–1703. doi:10.1056/NEJMoa1304369

Vultaggio-Poma, V., Sarti, A. C., and Di Virgilio, F. (2020). Extracellular ATP: A feasible target for cancer therapy. *Cells* 9 (11), E2496. doi:10.3390/cells9112496

Walter, F. M., Mills, K., Mendonca, S. C., Abel, G. A., Basu, B., Carroll, N., et al. (2016). Symptoms and patient factors associated with diagnostic intervals for pancreatic cancer (SYMPTOM pancreatic study): A prospective cohort study. *Lancet. Gastroenterol. Hepatol.* 1 (4), 298–306. doi:10.1016/S2468-1253(16)30079-6

Wang, Q., Wang, Y., Ding, J., Wang, C., Zhou, X., Gao, W., et al. (2020). A bioorthogonal system reveals antitumour immune function of pyroptosis. *Nature* 579 (7799), 421–426. doi:10.1038/s41586-020-2079-1

Xia, X., Wang, X., Cheng, Z., Qin, W., Lei, L., Jiang, J., et al. (2019). The role of pyroptosis in cancer: Pro-cancer or pro-"host. *Cell Death Dis.* 10 (9), 650. doi:10.1038/s41419-019-1883-8

Yang, S., He, P., Wang, J., Schetter, A., Tang, W., Funamizu, N., et al. (2016). A novel MIF signaling pathway drives the malignant character of pancreatic cancer by targeting NR3C2. *Cancer Res.* 76 (13), 3838–3850. doi:10.1158/0008-5472.CAN-15-2841

Ye, Y., Dai, Q., and Qi, H. (2021). A novel defined pyroptosis-related gene signature for predicting the prognosis of ovarian cancer. *Cell Death Discov.* 7 (1), 71. doi:10.1038/s41420-021-00451-x

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

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, P., Zhang, X., Liu, N., Tang, L., Peng, C., and Chen, X. (2021). Pyroptosis: Mechanisms and diseases. *Signal Transduct. Target. Ther.* 6 (1), 128. doi:10.1038/s41392-021-00507-5

Zhang, G., Schetter, A., He, P., Funamizu, N., Gaedcke, J., Ghadimi, B. M., et al. (2012). DPEP1 inhibits tumor cell invasiveness, enhances chemosensitivity and predicts clinical outcome in pancreatic ductal adenocarcinoma. *PLoS One* 7 (2), e31507. doi:10.1371/journal.pone.0031507

Zhu, Y., Knolhoff, B. L., Meyer, M. A., Nywening, T. M., West, B. L., Luo, J., et al. (2014). CSF1/CSF1R blockade reprograms tumor-infiltrating macrophages and improves response to T-cell checkpoint immunotherapy in pancreatic cancer models. *Cancer Res.* 74 (18), 5057–5069. doi:10.1158/0008-5472.CAN-13-3723

TYPE Original Research
PUBLISHED 06 September 2022
DOI 10.3389/fgene.2022.959109

Systematic analysis of prognostic and immunologic characteristics associated with coronavirus disease 2019 regulators in acute myeloid leukemia

Mingjie Shi[1,2], Lidan Chen[3], Yue Wei[4], Riling Chen[1,2,5], Runmin Guo[1,2]* and Fei Luo[1,2,5]*

[1]Key Laboratory of Research in Maternal and Child Medicine and Birth Defects, Guangdong Medical University, Foshan, China, [2]Matenal and Child Research Institute, Shunde Women and Children's Hospital (Maternity andChild Healthcare Hospital of Shunde Foshan), Guangdong Medical University, Foshan, China, [3]First College of Clinical Medicine, Guangdong Medical University, Zhanjiang, China, [4]Department of Ultrasound, Shunde Women and Children's Hospital (Maternity and Child Healthcare Hospital of Shunde Foshan), Guangdong Medical University, Foshan, China, [5]Department of Hematology-Oncology, Shunde Women and Children's Hospital (Maternity and Child Healthcare Hospital of Shunde Foshan), Guangdong Medical University, Foshan, China

OPEN ACCESS

EDITED BY
Christoph Schaniel,
Icahn School of Medicine at Mount Sinai, United States

REVIEWED BY
Panagiotis Karagiannis,
King's College London, United Kingdom
Afsane Vaezi,
Iran University of Medical Sciences, Iran

*CORRESPONDENCE
Fei Luo,
luofei_00@163.com
Runmin Guo,
grm2021@126.com

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

RECEIVED 09 June 2022
ACCEPTED 12 August 2022
PUBLISHED 06 September 2022

CITATION
Shi M, Chen L, Wei Y, Chen R, Guo R and Luo F (2022), Systematic analysis of prognostic and immunologic characteristics associated with coronavirus disease 2019 regulators in acute myeloid leukemia.
Front. Genet. 13:959109.
doi: 10.3389/fgene.2022.959109

The coronavirus disease 2019 (COVID-19) pandemic has so far damaged the health of millions and has made the treatment of cancer patients more complicated, and so did acute myeloid leukemia (AML). The current problem is the lack of understanding of their interactions and suggestions of evidence-based guidelines or historical experience for the treatment of such patients. Here, we first identified the COVID-19-related differentially expressed genes (C-DEGs) in AML patients by analyzing RNA-seq from public databases and explored their enrichment pathways and candidate drugs. A total of 76 C-DEGs associated with the progress of AML and COVID-19 infection were ultimately identified, and the functional analysis suggested that there are some shared links between them. Their protein–protein interactions (PPIs) and protein–drug interactions were then recognized by multiple bioinformatics algorithms. Moreover, a COVID-19 gene-associated prognostic model (C-GPM) with riskScore was constructed, patients with a high riskScore had poor survival and apparently immune-activated phenotypes, such as stronger monocyte and neutrophil cell infiltrations and higher immunosuppressants targeting expressions, meaning which may be one of the common denominators between COVID-19 and AML and the reason what complicates the treatment of the latter. Among the study's drawbacks is that these results relied heavily on publicly available datasets rather than being clinically confirmed. Yet, these findings visualized those C-DEGs' enrichment pathways and inner associations, and the C-GPM based on them could accurately predict survival outcomes in AML patients, which will be helpful for further optimizing therapies for AML patients with COVID-19 infections.

KEYWORDS

acute myeloid leukemia, COVID-19, differentially expressed genes, prognosis, drug molecule, protein–protein interaction

Introduction

Acute myeloid leukemia (AML) is a common malignancy in adults and is characterized by abnormal proliferation of primitive and naive myeloid cells in the bone marrow and peripheral blood, which has the lowest 5-year survival rate in all leukemia types (Döhner et al., 2015; Westermann and Bullinger 2021). Coronavirus disease 2019 (COVID-19) is an infectious disease caused by the SARS-CoV-2 virus and is mainly manifested by fever, dry cough, fatigue, etc. (Cui et al., 2019; V'Kovski et al., 2021). AML patients have a high risk of getting infected by SARS-CoV-2 owing to their poor resistance and immunity, and the difficulty of treating is undoubtedly greatly increased when AML patients are infected by SARS-CoV-2. Reports on improving the treatment and care of AML patients infected with COVID-19 are also being published (Ferrara et al., 2020; Khan et al., 2020). Farah et al. (2020) established minimal residual disease monitoring in the treatment of NPM1-mutant AML for someone who used updated chemotherapy that had fewer myelosuppressive regimens. Patel et al. (2021) demonstrated that AML patients could activate the immune responses to SARS-CoV-2 even facing immune suppression by chemotherapy. In addition, many studies have previously found that certain gene sets (such as autophagy and immunity) are important in the progression of AML, while the role of COVID-19-related genes in its process is still unknown (Yan et al., 2019; Fu et al., 2021). Thus, identifying the regulatory molecules between them may facilitate providing novel and effective therapeutics for AML patients with COVID-19. In this study, we attempt to identify COVID-19-related differentially expressed genes (C-DEGs), explore their interactions with one another, and discover their candidate drug molecules by multiple bioinformatics. Another prognostic model was constructed through those identified DEGs, and the prognosis performance was validated in the GSE37642 database, and its relationship to the immune microenvironment was also subsequently assessed (Figure 1).

Materials and methods

Download and processing of the AML dataset

First, we have comprehensively analyzed the RNA-seq datasets of AML subjects, which were downloaded from The Cancer Genome Atlas (TCGA, https://portal.gdc.cancer.gov/), Genotype-Tissue Expression (GTEx, https://gtexportal.org/home/), and Gene Expression Omnibus (GEO, https://www.ncbi.nlm.nih.gov/geo/) databases. A total of 176 patients with AML in TCGA,

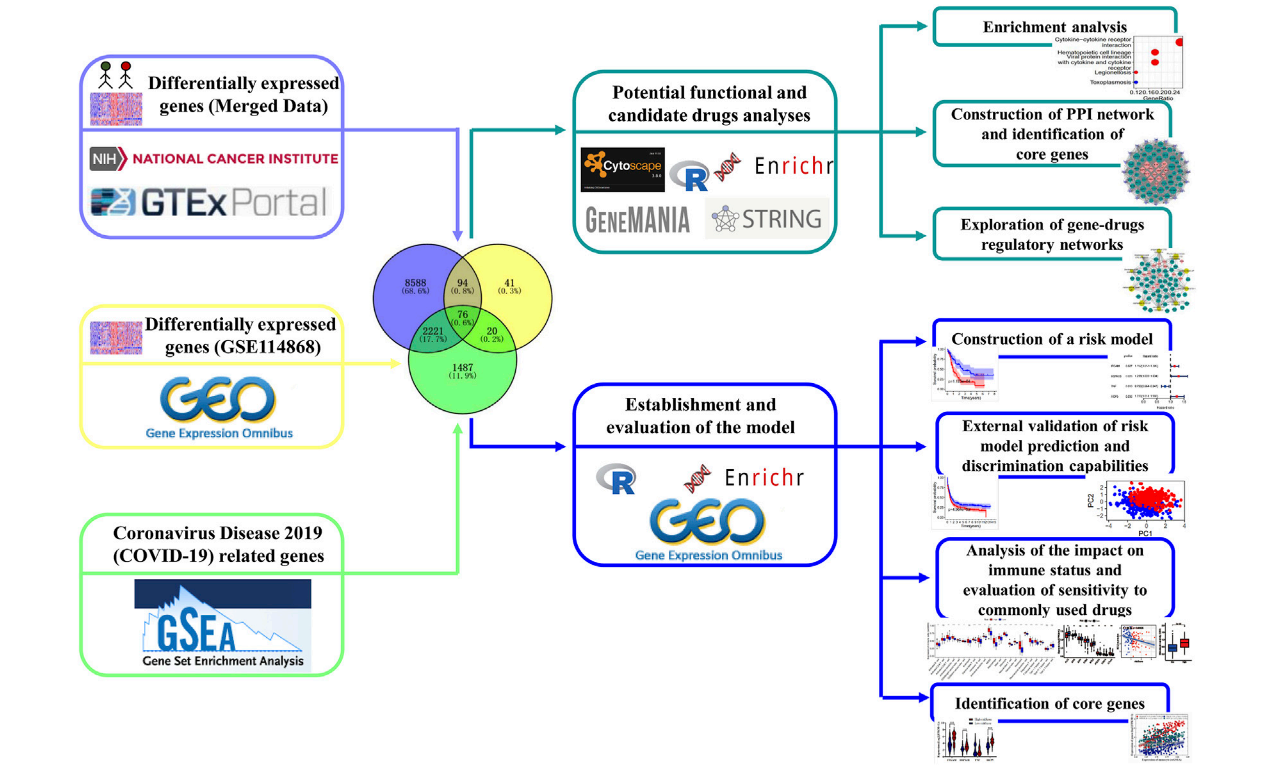

Differentially expressed genes (Merged Data)
NIH NATIONAL CANCER INSTITUTE
GTExPortal
Differentially expressed genes (GSE114868)
Gene Expression Omnibus
Coronavirus Disease 2019 (COVID-19) related genes
GSEA
Gene Set Enrichment Analysis
8588 (68.6%)
94 (0.8%)
41 (0.3%)
76 (0.6%)
2221 (17.7%)
20 (0.2%)
1487 (11.9%)
Potential functional and candidate drugs analyses
Cytoscape
R
Enrichr
GENEMANIA
STRING
Establishment and evaluation of the model
R
Enrichr
Gene Expression Omnibus
Enrichment analysis
Construction of PPI network and identification of core genes
Exploration of gene-drugs regulatory networks
Construction of a risk model
External validation of risk model prediction and discrimination capabilities
Analysis of the impact on immune status and evaluation of sensitivity to commonly used drugs
Identification of core genes

FIGURE 1
Flow diagram exhibiting the process analysis of the study.

TABLE 1 Basic information on datasets in the study.

Series accession number	Platform used	No. of normal samples	No. of tumorous samples	FAB morphology code (%)	Gender (%)	Mean age [min, max]	Vital status (%)	Survive time ($\bar{x}\pm s$)
Merge (GTEx and TCGA-LAML)	Illumina RNAseq	70	179 (153 samples with complete clinical data)	M0: 13 (8.5); M1: 34 (22.2); M2: 35 (22.9) M3: 14 (9.2); M4: 34 (22.2); M5: 17 (11.1); M6: 2 (1.3); M7: 3 (2.0) Unknown: 1 (0.7)	Female: 73 (47.8) Male: 80 (52.2)	54.2 [18, 88]	Alive: 59 (38.6) Dead: 94 (61.4)	620.8 ± 585.2
GSE114868	Affymetrix Human Transcriptome Array 2.0 (GPL17586)	20	194	NA	NA	NA	NA	NA
GSE37642	Affymetrix Human Genome U133A Array (GPL96) and Affymetrix Human Genome U133 Plus 2.0 Array (GPL570)	0	553	M0: 22 (4.0); M1: 113 (20.4); M2: 164 (29.7) M3: 26 (4.7); M4: 121 (21.9); M5: 66 (11.9); M6: 22 (4.0); M7: 3 (0.5) Unknown: 16 (2.9)	NA	54.9 [18, 85]	Alive: 147 (26.6) Dead: 406 (73.4)	997.0 ± 1,292.5

Abbreviation: FAB, French American British.

70 normal samples in GTEx, and 194 AML patients and 20 healthy subjects in GSE114868 were used to identify differentially expressed genes. Limma packages with normalizeBetweenArrays function were utilized to merge and emend the data from TCGA and GTEx databases. Additionally, we extracted samples from the clinic in TCGA, according to the following criteria: (a) removing duplicated subjects referred to as formalin-fixed and paraffin-embedded; (b) dislodging subjects with insufficient clinical data; and (c) taking the average of duplicated genes or the same ensemble ID. In total, 152 patients in TCGA and 553 patients in GSE37642 were ultimately incorporated into our study to construct a prognosis model and evaluate its predictive performance (Table 1).

Identification of COVID-19-related differentially expressed genes and functional enrichment analysis

The COVID-19-related gene sets comprising 3,804 genes were downloaded from the Gene Set Enrichment Analysis (GSEA, http://www.gsea-msigdb.org/gsea/index.jsp) database. The Limma package (http://bioconductor.org/packages/release/bioc/html/limma.html) with Benjamini–Hochberg correction and the DESEq2 package (http://bioconductor.org/packages/release/bioc/html/DESeq2.html) were applied to identify differentially expressed genes using Padj <0.001 as all screening criteria. C-DEGs were identified by Venn analysis, and their annotations and functional enrichment analysis on Gene Ontology (GO) and Kyoto Encyclopedia of Genes and Genomes (KEGG) were executed by "clusterProfiler" and "enrichplot" packages, respectively. A *p*-value < 0.05 was deemed as a threshold.

Protein–protein interaction network analysis and hub genes' identification

The protein–protein interaction network is composed of proteins that interact with each other to participate in all aspects of biological processes such as signal transmission, gene expression regulation, energy and material metabolism, and cell cycle regulation. Information about the roles of multiple proteins in cells can be integrated into databases and visualized through protein network diagrams. GeneMANIA (http://genemania.org/) (Franz et al., 2018) and STRING (version 11.5, https://string-db.org/) (Szklarczyk et al., 2021) were utilized to explore the PPI networks and hub genes of those identified C-DEGs for further understanding of the physical and functional interactions between AML and COVID-19. All results were visualized by Cytoscape (v.3.7.1, https://cytoscape.org/) (Shannon et al., 2003), which is an open-source network visualization tool to produce an improved performance for different interactions.

Exploration of candidate drugs

In addition, we also explored the protein–drug interactions or candidate drug molecules based on these identified C-DEGs by the Drug Signatures Database (DSigDB) *via* Enrichr (https://maayanlab.cloud/Enrichr/), and the latter is a web-based comprehensive gene set enrichment analysis tool and can utilize the DSigDB resource to explore related drugs and small molecules.

Construction of the risk model and analysis of its effect on tumor-infiltrating immune cells and expression of common or emerging immune checkpoints

Samples with completed clinical data in TCGA and GSE37642 databases were applied to construct and validate a prognosis model related to C-DEGs, respectively. Univariate Cox and LASSO regression analyses identified the potential prognosis-associated C-DEGs, and then, multivariate Cox regression analysis was executed to build COVID-19 gene-related prognostic models (C-GPMs) and to assure it was not overfitted. The risk score (riskScore) of each individual was estimated by the following formula: riskScore = [Coefficient 1] * [Expression1] + [Coefficient 2] * [Expression 2] + [Coefficient 3] * [Expression 3] + [Coefficient n] * [Expression n], and the coefficient of each factor was calculated by the LASSO-Cox model. The Kaplan–Meier curve and the receiver operating characteristic curves (ROCs) were utilized to measure the discriminative ability of the C-GPM. In addition, to explore the relationship of this model on the immune microenvironment, the ssGSEA and ESTIMATE algorithms were utilized to calculate the abundance of tumor-infiltrating immune cells (TIICs) and the scores of the tumor microenvironment in each sample, and their correlation and differentiation were separately analyzed by Spearman analysis and Wilcoxon signed-rank test, as they were common or emerging immune checkpoints.

Cell culture and treatment

The AML cell line KG-1 was provided by Shanghai Yihe Applied Biotechnology Co., Ltd. and was cultured in RPMI-1640 (Gibco, Life Technologies, Carlsbad, CA, USA) that contained 10% fetal bovine serum and 1% penicillin/streptomycin. Cells were grown in a humidified atmosphere with 5% CO2 at 37°C. KG-1 cells were seeded in complete RPMI-1640 medium at appropriate cell numbers and then incubated in the presence of ATRA or RAD001 for the indicated times. RAD001 (everolimus) and all-trans retinoic acid (ATRA) were purchased from APExBIO (Houston, USA) and Sigma Chemical Co. Ltd. (St. Louis, MO), respectively. RNA isolation and real-time PCR were performed based on the corresponding kit instructions.

Cell counting Kit-8

The CCK8 assay was conducted in accordance with the manufacturer's instructions (GK10001, GLPBIO). For the assay, 2000 cells/well in 96-well plates containing 100 μL of the culture medium were seeded. A measure of 10 μL of the CCK8 reagent was added to each well at the indicated time, and the plates were given shock for 20 s and then incubated at 37 °C for 2 h. Lastly, we measured the OD value of each hole at 450 nm. These experiments were performed with three replicates, and five parallel samples were measured each time.

Statistical analysis

Statistical analyses were performed by R software (version: 3.5.2) with multiple packages (including Limma, ggplot2, glmnet, rms, preprocessCore, survminer, and ConsensusClusterPlus) and GraphPad Prism (version 8.4.3, La Jolla, CA, United States) software. Student's t-test was used to test for significant differences between any two groups of data, and one-way ANOVA was used when evaluating multiple groups of data. All hypothetical tests were two-tailed, and a p-value < 0.05 was considered statistically significant.

Result

Identification of common DEGs associated with COVID-19 in acute myeloid leukemia and enrichment analysis

The COVID-19-related gene sets comprising 3,804 genes were downloaded from the GSEA database (Supplementary Table S1). A total of 76 C-DEGs were identified with DESeq2 and Limma packages using the adjusted p-value < 0.001 as screening criteria (Figure 2A, Supplementary Tables S2 and S3). The top 15 enriched GO terms strikingly exhibited in the bubble chart were intimately concerned with immune inflammation and tumor progressions, such as the positive regulation of cytokine production, the cytokine-mediated signaling pathway, myeloid leukocyte activation, leukocyte chemotaxis, and migration in biological processes (BP); immune, cytokine, and pattern recognition receptors' activity, heat shock protein binding, and protein folding chaperone in molecular function (MF); and secretory and tertiary granule lumen, cytoplasmic vesicle lumen, and tertiary granule in cellular component (CC) (Figure 2B) (Supplementary Table S4). The KEGG pathways were mainly involved in hematopoietic cell lineage, viral protein interaction with cytokines and cytokine receptors, cytokine–cytokine receptor interaction, and other immune or viral infection-related pathways (Figure 2B) (Supplementary Table S5).

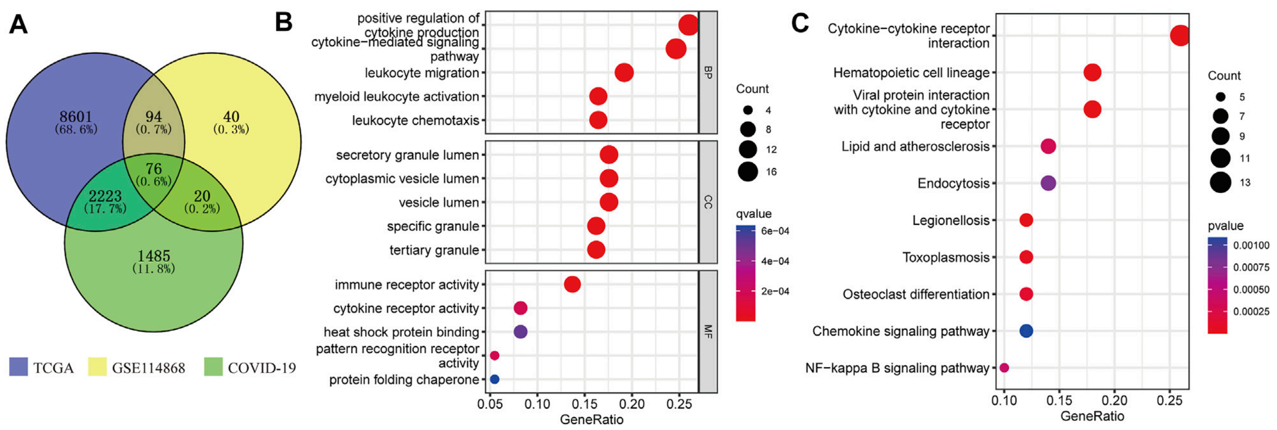


FIGURE 2
Identification and enrichment analysis of C-DEGs. **(A)** C-DEGs were identified by Venn analysis. GO **(B)** and KEGG **(C)** analyses of those C-DEGs.

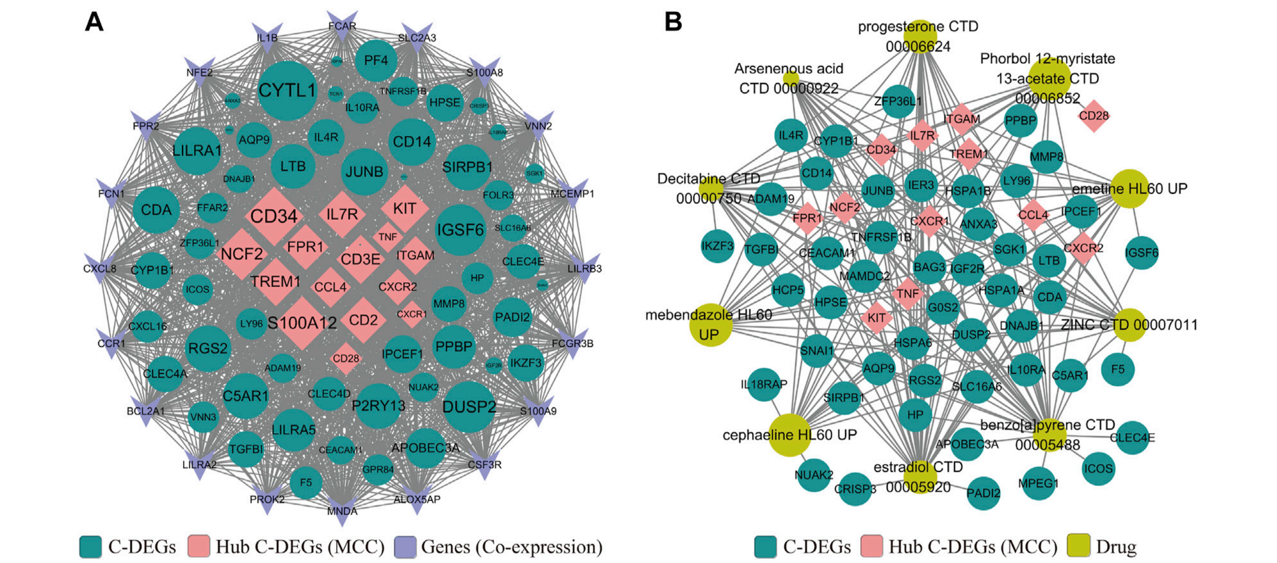


FIGURE 3
Construction of PPI **(A)** and protein–drug interaction **(B)** networks on those identified C-DEGs by GeneMANIA, STRING, Enrichr, and Cytoscape. Pink represents core C-DEGs that were identified by the MCC algorithm of the Cytoscape plugin (cytoHubba), green represents other C-DEGs, sky blue represents co-expressed genes identified by GeneMANIA, and ginger represents candidate drugs obtained in DSigDB *via* the Enrichr database.

Construction of the protein–protein interaction network and identification of their candidate drugs

Next, a PPI network was built to systematically analyze the interaction of those C-DEGs in biological systems and understand the response mechanism of biological signals and energy metabolism in special physiological states in-depth, as well as the functional connections among proteins. In our study, PPI networks associated with C-DEGs were constructed through GeneMANIA and STRING tools, and 15 hub signatures (*TNF*, *ITGAM*, *CCL4*, *IL7R*, *CD28*, *CXCR1*, *S100A12*, *CD2*, *TREM1*, *FPR1*, *CD3E*, *CD34*, *NCF2*, *KIT*, and *CXCR2*) were identified based on the network maximal clique centrality (MCC) algorithm of Cytoscape plugin (cytoHubba) (Figure 3A). NetworkAnalyst 3.0 and DrugBank were then employed to

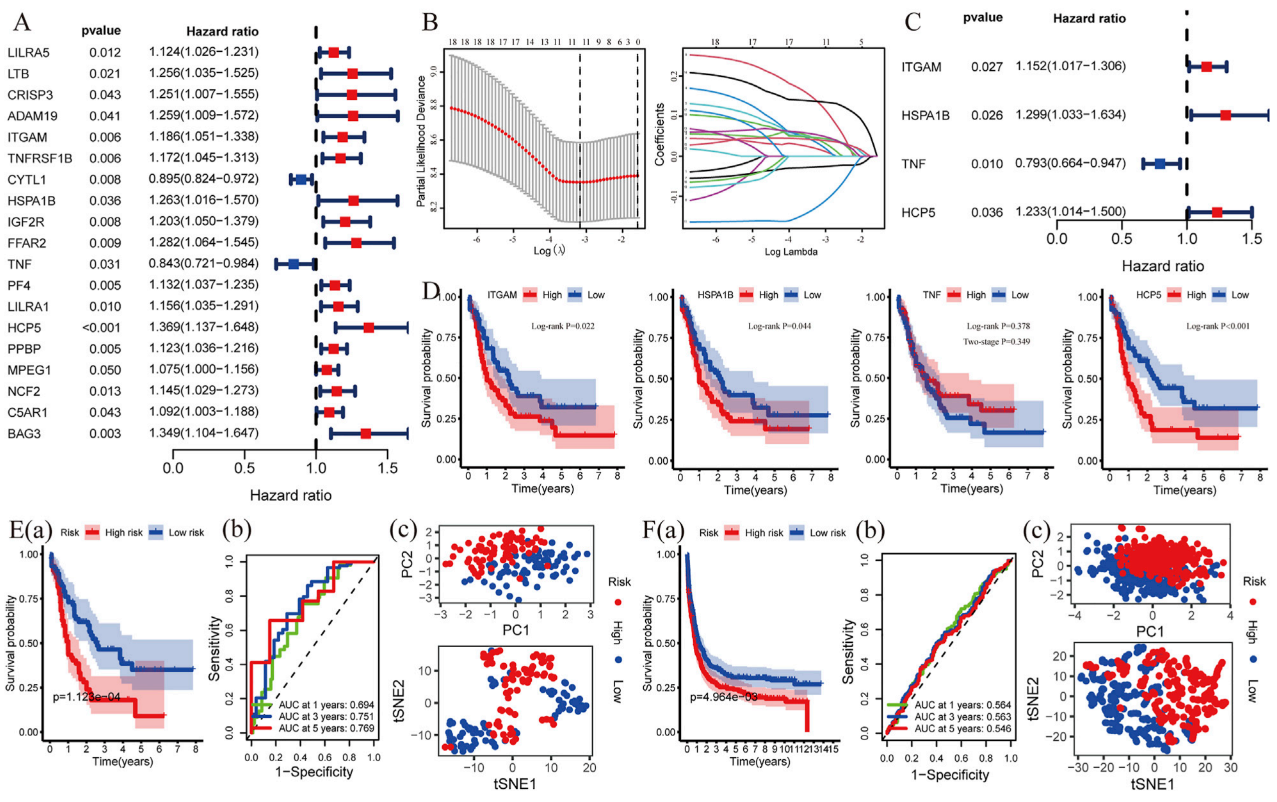


FIGURE 4
Construction and validation of a risk model associated with identifying C-DEGs. **(A)** Nineteen C-DEGs with prognosis in AML were identified by univariate Cox regression analysis ($p < 0.05$). **(B,C)** LASSO- and multivariate Cox regression analyses were applied to build a risk model (C-GPM), and the forest map exhibited four C-DEGs with their p values and hazard ratios (HR) with confidence intervals (CI). **(D)** Kaplan–Meier method was used to calculate their differences in overall survival. **(E)** Predictive and discriminative abilities of the C-GPM in AML were evaluated by multiple methods, including survival- **(a)**, ROC- **(b)**, PCA-, and t-SNE **(c)** analyses. **(F)** Validation of the predictive and discriminative abilities of the C-GPM in the GSE37642 dataset.

explore potential and available drugs targeting these C-DEGs, and a total of 101 drugs were separated using an adjusted p-value <0.001 as the threshold (Supplementary Table S6). Here, we visualized 11 of them that targeted more genes, including estradiol, benzo [a]pyrene, decitabine, progesterone, ZINC, cephaeline, arsenenous acid, emetine, mebendazole, and phorbol 12-myristate 13-acetate (Figure 3B).

Construction and validation of a risk model with four C-DEGs for AML

To explore whether these C-DEGs are associated with patients' overall survival, a total of 19 genes were integrated into the Lasso regression analyses after univariate Cox regression (p-value < 0.05, Figures 4A,B; Supplementary Table S7). A multivariate Cox proportional hazards regression model was subsequently utilized to construct the C-GPM with riskScore; patients were divided into high- and low-risk groups using the median riskScore as cutoff (Supplementary Table S8). A C-GPM consisting of four genes (*TNF*, *ITGAM*, *HSPA1B*, and *HCP5*) was identified, and they could all serve as independent indices for predicting the patients' overall survival (Figure 4C). *ITGAM*, *HSPA1B*, and *HCP5* were then confirmed to negatively correlate with the prognosis of AML patients using the Kaplan–Meier method (Figure 4D), and patients with high riskScore had significantly shorter overall survival than those with low riskScore and a favorite prognostic predictive value in determining the survival rates of AML patients (1-year AUC = 0.694, 3-year AUC = 0.751, and 5-year AUC = 0.772; Figure 4E (a) and (b)). Furthermore, the principal component analysis (PCA) and t-distributed random neighbor embedding (t-SNE) analysis showed that the C-GPM could well differentiate patients into two different risk groups (Figure 4Ec). These findings were subsequently validated in the GSE37642 database (Figure 4F).

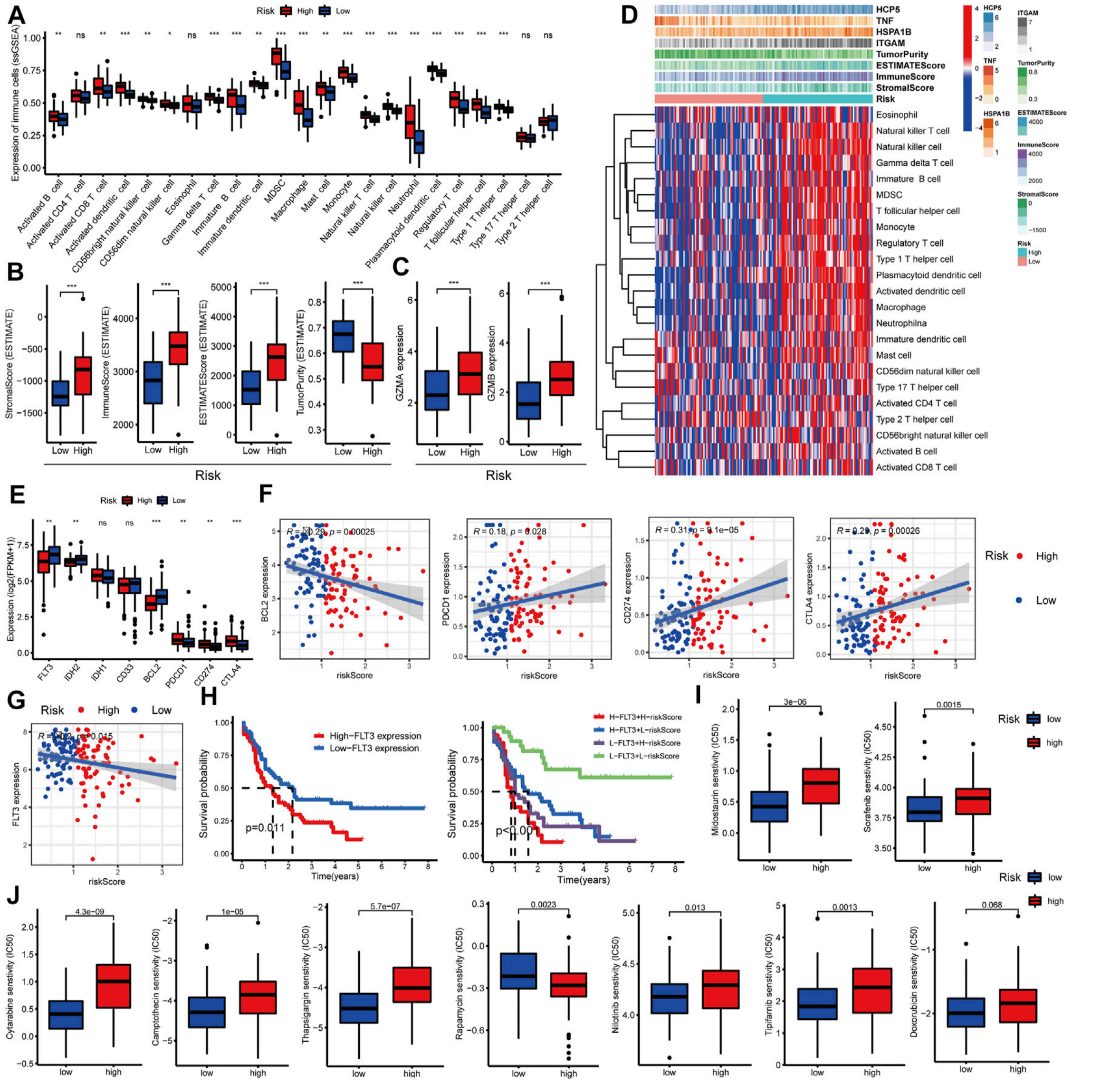


FIGURE 5
Evaluation of the relationship between the C-GPM and immune microenvironment. **(A,B)** Differences in common TIICs and the tumor microenvironment in the C-GPM were assessed, and the results indicated that patients in the high-risk group had a more pronounced immune or inflammatory activation phenotype. **(C)** Exploration of the difference between GZMA and GZMB that represents immune infiltration and immune cytolytic activity in the C-GPM. **(D)** Heatmap was used to directly show the correlation between the C-GPM with four C-DEGs and the immune microenvironment. **(E,F)** Exploration of whether emerging therapeutic targets are differentially expressed in the two distinct groups of the C-GPM and analysis of their correlation with riskScore. **(G)** Analysis of the correlation between the novel therapeutic target (FLT3) and riskScore. **(H)** Survival analysis of FLT3 with or without riskScore. **(I)** Analysis of the effect of the C-GPM on the sensitivity of midostaurin and sorafenib that modulates the receptor tyrosine kinase FLT3. **(J)** Analysis of the effect of the C-GPM on the sensitivity of other common AML drugs.

Effects of the C-GPM on the immune status and tumor microenvironment

In addition, several recent studies have indicated that the abundance of TIICs within the tumor microenvironment (TME) could predict phases of tumor inflammation and were related to the poor prognosis of AML patients. Thus, we explored the impact of the C-GPM on them based on ssGSEA and ESTIMATE algorithms. The cohort was stratified into high riskScore (N = 76) and low riskScore (N = 77) groups according to their medians; most TIICs were more abundant in high-risk groups (Wilcoxon signed-rank test, p-value < 0.05, Figure 5A), and stromal (StromalScore), immune (ImmuneScore), and ESTIMATE

(ESTIMATEScore) scores were all increased with statistically significant differences in the high-riskScore group, while the tumor purity (TumorPurity) was contrary to their trend (Wilcoxon signed-rank test, p-value < 0.05, Figure 5B). Also, we explored the expression levels of granzyme A (GZMA) and granzyme B (GZMB) representing immune infiltration and immune cytolytic activity (Arias et al., 2017). They all showed higher expression in the high-riskScore group, as was expected (Wilcoxon signed-rank test, p-value < 0.05, Figure 5C). All these findings suggested that patients with high riskScore had more pronounced immune and inflammatory responses along with more risks and worse prognoses (Figure 5D).

Analysis of the impact on immunotherapy and evaluation of sensitivity to commonly used drugs

Many explorations have shown that immune checkpoint testing is a reliable way to assess the patient's response to immunotherapy, which is blossoming into the backbone of cancer treatment, while AML patients with high expression of conventional immune checkpoints [such as programmed cell death 1 (PDCD1, best known as PD1)] did not well benefit from immunotherapy based on most clinical trials, and these are closely related to immune complications in AML patients. In our study, patients with high riskScore had higher expressions of the common immune checkpoints such as PD1, programmed cell death ligand 1 (PDL1/CD274), and cytotoxic T-lymphocyte antigen 4 (CTLA4) and had lower levels of the emerging checkpoints including fms-like tyrosine kinase-3 (FLT3), isocitrate dehydrogenase (NADP (+)) 2 (IDH2), and B-cell leukemia/lymphoma 2 (BCL2) (Figures 5E,F). Additionally, studies have shown that FLT3 is highly expressed in more than 70% of AML patients, and for this, it was considered an important target for the treatment of AML. Here, the FLT3 expression was negatively correlated with riskScore and prognosis of patients with AML; patients with low FLT3 combined with low riskScore had significantly better overall survival than others. Also, those with low riskScore had more sensitivity to drugs, such as midostaurin and sorafenib, that modulate the receptor tyrosine kinase FLT3; the former has been approved as a new treatment option for relapsed or refractory FLT3-mutated AML (Figures 5G–I). In addition, we also explored the effect of C-GPM on the sensitivity of other common AML drugs, and patients in the high-riskScore group were more sensitive and beneficial to cytarabine, camptothecin, thapsigargin, nilotinib, and tipifarnib but were less sensitive to rapamycin. The sensitivity to doxorubicin had no significant difference in both groups (Figure 5J).

HCP5 might be a novel prognostic immune-related biomarker of AML

Also, we delved into the differential expression of these four genes between AML and healthy patients, high-, and low-risk groups, respectively (Figure 6A). HCP5 and ITGAM were both highly expressed in AML patients and high-risk groups, and the latter was considered a marker for monocytes, and a very strong correlation between them was confirmed (Figure 6B). HSPA1B, a receptor that assisted virus entry into cells, is highly expressed in patients in the high-risk group, while TNF has no significant difference in both risk groups. Among them, HCP5 has been shown to have a prognostic role in multiple external datasets, and its high expression is closely associated with poor prognosis in AML (Figure 6C). We subsequently found *in vitro* that silencing of HCP5 significantly affected the proliferation of KG-1 cells derived from the human acute myelogenous leukemia cell line (Figures 6D,E). We identified 413 genes significantly associated with HCP5 through the LinkedOmics database and found that they were mainly associated with immunity in AML patients, including regulation of T-cell activation, lymphocyte-mediated immunity, and Th17 cell differentiation (Supplementary Figure S1). Additionally, all-trans retinoic acid (ATRA) is a traditional drug for the treatment of AML, and everolimus (RAD001) is a new type of immunosuppressant. ATRA (1 μM) in combination with RAD001 (10 nM) strikingly downregulated the expression of HCP5. Notably, RAD001 (10 nM) significantly enhanced the ATRA-inhibited cell growth, and these were more pronounced in HCP5-silenced cells (Figure 6F).

Discussion

The COVID-19 pandemic has been going on worldwide for over two years; although posing a huge threat to the health of normal people, it also seriously affects the treatment of cancer patients (Henderson et al., 2021). AML is the most common leukemia in adults and accounts for about 80% of all cases, and its treatment and care have been further complicated by the COVID-19 pandemic (Ferrara et al., 2020; Wilde et al., 2020). In this study, we attempted to analyze the potential connections between COVID-19 infections and AML and to explore its impact on prognosis and candidate medications' susceptibility of AML patients with COVID-19, providing a new insight into their clinical diagnosis and treatment.

Here, we identified 76 C-DEGs in AML, and they were mainly involved in hematopoietic cell lineage, viral protein interaction with cytokines and cytokine receptors, cytokine–cytokine receptor interaction, and other viral infection or tumor progression pathways. In addition, the PPI network constructed by STRING and GeneMANIA has shown that most of them have obvious interactions, and 15 hub C-DEGs (TNF, ITGAM, CCL4, IL7R, CD28, CXCR1, S100A12, CD2,

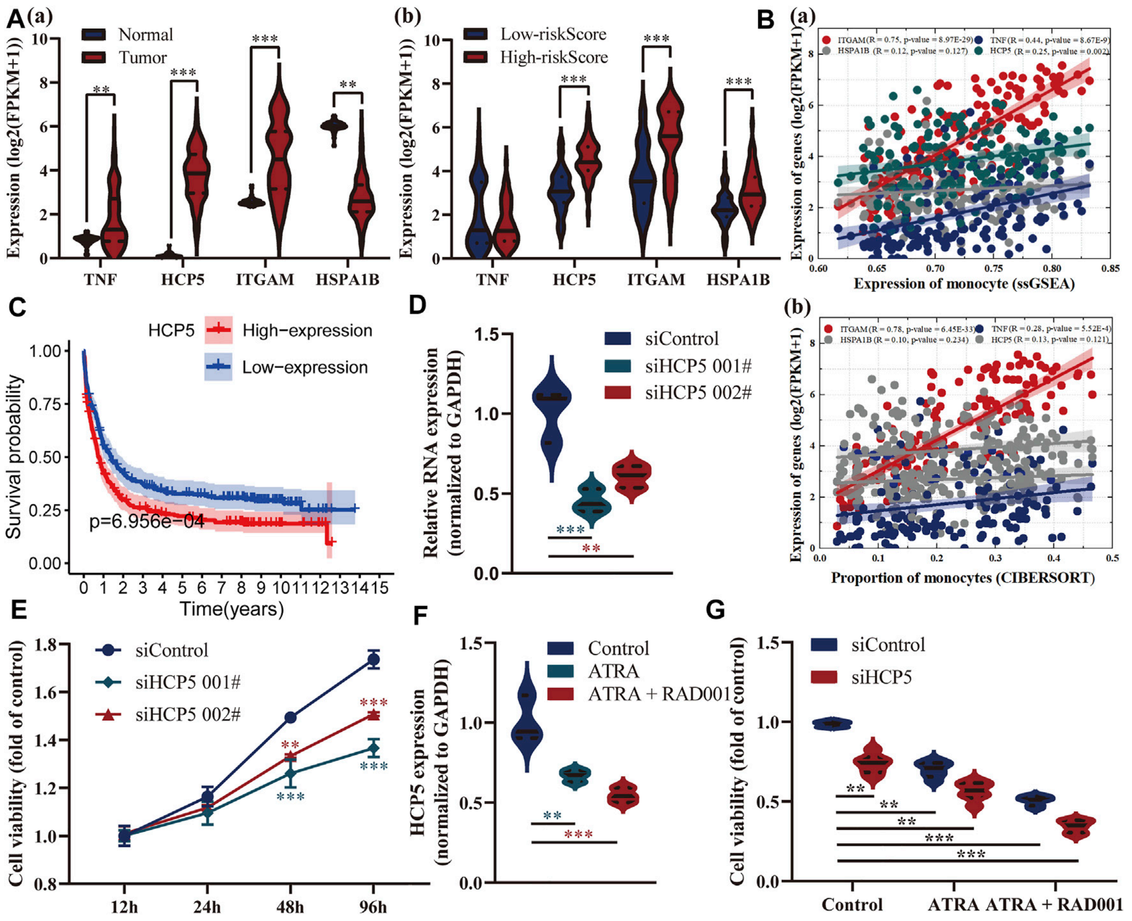


FIGURE 6
Exploration of the relationship between these genes in the model and AML. **(A)** Analysis of differential expressions of these four genes between AML and healthy patients **(a)**, high-, and low-risk groups **(b)**, respectively. **(B)** Assessment of their correlation with monocytes. **(C)** Survival analysis of HCP5 in the GSE37642 dataset showed that its expression was associated with prognosis. **(D,E)** CCK8 results showed that silencing of HCP5 significantly affected the proliferation of KG-1 cells. **(F,G)** ATRA (1 μM) in combination with RAD001 (10 nM) strikingly downregulated the expression of HCP5 and inhibited cell growth, especially in HCP5-silenced cells.

TREM1, FPR1, CD3E, CD34, NCF2, KIT, and CXCR2) were identified and were involved in the maintenance of human homeostasis. Meanwhile, their candidate drugs were also explored, multiple drugs (estradiol, benzo [a]pyrene, decitabine, progesterone, ZINC, cephaeline, arsenenous acid, emetine, mebendazole, and phorbol 12-myristate 13-acetate) were identified, and most have been reported to have antitumor effects. For example, mebendazole has been reported to exhibit potent antileukemic activity against AML, and it is considered to have the potential to bind to SARS-CoV-2 B.1.1.7 (alpha) and P.1 (gamma) variants (He et al., 2018; Yele et al., 2022).

Subsequently, a C-GPM consisting of four genes (*TNF*, *ITGAM*, *HSPA1B*, and *HCP5*) was constructed by multiple analyses, and these genes could serve as independent indices for predicting the patients' overall survival. TNF, generally known as TNF-α, is mainly secreted by mononuclear macrophages and is a cytokine involved in systemic inflammation. Some studies have reported that TNF can inhibit the replication of different viruses to exert antiviral effects, as well as regulate the function of immunocytes (Liu et al., 1998; Bruunsgaard et al., 2003). Integrin subunit alpha M (ITGAM) is implicated in mediating the uptake of pathogens and in various adhesive interactions of macrophages, monocytes, and granulocytes, and it is also required for CD177-PRTN3-mediated activation of TNF-sensitized neutrophils (Boguslawska et al., 2016; Lyu et al., 2020). Heat shock protein family A (Hsp70) member 1B (HSPA1B) is one of three protein-encoding genes belonging to the HSP70 family and is involved in the human immune response after infection with Epstein–Barr virus, *Legionella*, and influenza A (Sistonen et al., 1994; Soncin et al., 1997; Kiang and Tsokos, 1998). HLA complex P5

(HCP5) is a long non-coding RNA, and emerging studies have recently indicated that it plays an important role in the progression of AML. Research has shown that HCP5 promoted lung adenocarcinoma metastasis *via* the miR-203/SNAI axis and tumor growth and upregulated the expression of PD-L1/CD274 *via* a competing endogenous RNA mechanism of sponging miR-150–5p, and these were also consistent with our findings (Jiang et al., 2019; Xu et al., 2020). The C-GPM could well stratify patients into high- and low-risk groups based on their median riskScores, and patients in the former had worse overall survival, which had also been demonstrated in external cohorts.

Furthermore, we explored the impact of the C-GPM on the immune microenvironment, which was considered a vital criterion in tumor progression and metastasis. The results suggested that patients with high riskScore had more TIICs, higher scores of the tumor microenvironment, and lower tumor purity, implying their immune and inflammatory responses were in a more active state, which increased the difficulty of treatment and the risk of life for AML patients. Notably, patients with high riskScore were shown to have a poor prognosis; this phenomenon may be associated with their immune active status, including immune checkpoints that are highly expressed. Many explorations have shown that immune checkpoint testing is a reliable way to assess the patient's response to immunotherapy, which is blossoming into the backbone of cancer therapy. Studies show that AML patients with high expression of conventional immune checkpoints (such as *PD1*, *CD274*, and *CTLA4*) did not benefit from immunotherapy, and these are closely related to immune complications (Berger et al., 2008; Chen et al., 2020). In this study, patients with high riskScore had significantly poor prognoses and had obviously high expressions of common immune checkpoints, including *PD1*, *CD274*, and *CTLA4*. With the rapid development of targeted therapy, more and more targets have been discovered with the potential for anticancer, which means that more patients are expected to be covered by targeted therapy drugs. FLT3 is a type III receptor tyrosine kinase (RTK) and plays an important role in the proliferation, differentiation, and survival of hematopoietic stem cells and precursor B cells (Smith et al., 2012; Wu et al., 2018). FLT3 can lead to abnormal cell proliferation and induce tumorigenesis, especially those closely related to the occurrence and development of AML. Studies have shown that FLT3 is highly expressed in more than 70% of AML patients, and for this, FLT3 is considered an emerging important target for the treatment of AML (Gebru and Wang, 2020). Here, the FLT3 expression was negatively correlated with riskScore and prognosis of patients with AML, patients with low FLT3 combined with low riskScore had significantly better overall survival than others, and those with low riskScore had more sensitivity to its inhibitors, such as drugs like sorafenib and the recently approved midostaurin for relapsed or refractory FLT3-mutant AML (Antar et al., 2017; Brinton et al., 2020; Döhner et al., 2020). In addition, we also explored the effect of the C-GPM on the sensitivity of other common AML drugs, and patients in the high-riskScore group were more sensitive and beneficial to cytarabine, camptothecin, thapsigargin, nilotinib, and tipifarnib but were less sensitive to rapamycin. Among them, nilotinib, a second-generation tyrosine kinase inhibitor, is primarily utilized in the treatment of chronic myeloid leukemia and has limited results in the treatment of AML. According to studies, nilotinib has a considerable suppressing effect on $CD8^+$ T-lymphocyte activity, which could be one of the reasons why it is more sensitive to patients with a high riskScore (Chen et al., 2008). Tipifarnib is a chemical being explored for the treatment of AML and other types of cancer, and it exhibits substantial immunosuppression. These features may be used as a targeting approach in AML therapy in high-riskScore patients with strong immune activation (Bai et al., 2012; Guo et al., 2020). Additionally, the sensitivity of doxorubicin had no significant difference in both groups; this may be because doxorubicin promotes tumor cell metastasis by releasing inflammatory chemicals, which in turn aggregates monocytes and macrophages and worsens the underlying illness (Keklikoglou et al., 2019). Thapsigargin has antiviral properties and is thought to help curb the spread of epidemics including COVID-19 (Al-Beltagi et al., 2021a; Al-Beltagi et al., 2021b). These results suggest that there are established links between COVID-19 infection and AML progression, and they are related to the overall immune status of patients.

In addition, we did not retrieve reports using public datasets to explore the role of COVID-19-related gene sets in AML patients, although there have been many reports in other tumors. For example, Huang et al. identified a novel prognostic signature and nomogram based on SARS-CoV-2-related genes as reliable prognostic predictors for KIRC patients and provided potential therapeutic targets for KIRC with COVID-19 infection, and Liang et al. revealed commonality in specific gene expression by patients with COVID-19 and LUAD (Huang et al., 2021; Liang et al., 2022). Both COVID-19 and cancers provide complicated challenges that require ongoing research and development in medicine. It is challenging to confirm the results of most research excursions in the clinic because they rely on publicly available datasets, which is one of the primary reasons why these methods have, up to this point, been unable to produce convincing findings. Additionally, the methods by which these putative differential genes contribute to AML development or COVID-19 infection have not been exhaustively investigated.

Conclusion

AML patients have a high risk of getting infected by SARS-CoV-2 owing to their poor resistance and immunity, so we have analyzed the potential interactions between AML and COVID-19 infection by multiple bioinformatics, such as identifying C-DEGs, exploring their interactions with one another, and discovering candidate drug molecules by the DSigDB database. In addition, a prognostic model with satisfactory prediction performance was constructed through these identified C-DEGs, and patients were divided into high- and low-risk groups with distinct overall survival. We found that patients in the former had poor prognoses and had apparently immune-activated phenotypes, such as more immune cell infiltrations and higher expression of immunosuppressive points. Instead, patients in the latter had more sensitivity to emerging targeted inhibitors, such as midostaurin and sorafenib that modulate the receptor tyrosine kinase FLT3. At present, the number of people infected with SARS-CoV-2 is still increasing sharply; more research studies on the commonality of COVID-19 and other diseases are necessary to provide more treatments for patients.

Data availability statement

Publicly available datasets were analyzed in this study. The publicly available dataset can be downloaded from the UCSC Xena browser (https://gdc.xenahubs.net) and the GEO database (https://www.ncbi.nlm.nih.gov/gds/) under the accession numbers GSE114868 and GSE37642.

Author contributions

MS and FL conceived and designed this work, and the former wrote the manuscript and made corresponding corrections. LC carried out the accurate arrangement and modification of words in the manuscript. RC, RG, and YW contributed essential tools and interpreted the results. All authors who contributed to this manuscript approved the submitted version.

Funding

This work was supported by the grants from the Medical Research Project of Foshan Municipal Health Bureau (20210190) and the National Natural Science Foundation of China (81873649).

Conflict of interest

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

Publisher’s note

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

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2022.959109/full#supplementary-material

References

Al-Beltagi, S., Goulding, L., Chang, D., Mellits, K., Hayes, C., Gershkovich, P., et al. (2021a). Emergent SARS-CoV-2 variants: Comparative replication dynamics and high sensitivity to thapsigargin. *Virulence* 12, 2946–2956. doi:10.1080/21505594.2021.2006960

Al-Beltagi, S., Preda, C., Goulding, L., James, J., Pu, J., Skinner, P., et al. (2021b). Thapsigargin is a broad-spectrum inhibitor of major human respiratory viruses: Coronavirus, respiratory syncytial virus and influenza A virus. *Viruses* 13, 284. doi:10.3390/v13020234

Antar, A., Otrock, Z. K., El-Cheikh, J., Kharfan-Dabaja, M. A., Battipaglia, G., Mahfouz, R., et al. (2017). Inhibition of FLT3 in AML: A focus on sorafenib. *Bone Marrow Transpl.* 52, 344–351. doi:10.1038/bmt.2016.251

Arias, M., Martínez-Lostao, L., Santiago, L., Ferrandez, A., Granville, D., and Pardo, J. (2017). The untold story of granzymes in oncoimmunology: Novel opportunities with old acquaintances. *Trends Cancer* 3, 407–422. doi:10.1016/j.trecan.2017.04.001

Bai, F., Villagra, A., Zou, J., Painter, J., Connolly, K., Blaskovich, M., et al. (2012). Tipifarnib-mediated suppression of T-bet-dependent signaling pathways. *Cancer Immunol. Immunother.* 61, 523–533. immunotherapy: CII. doi:10.1007/s00262-011-1109-0

Berger, R., Rotem-Yehudar, R., Slama, G., Landes, S., Kneller, A., Leiba, M., et al. (2008). Phase I safety and pharmacokinetic study of CT-011, a humanized antibody interacting with PD-1, in patients with advanced hematologic malignancies. *Clin. Cancer Res.* 14, 3044–3051. doi:10.1158/1078-0432.ccr-07-4079

Boguslawska, J., Kedzierska, H., Poplawski, P., Rybicka, B., Tanski, Z., and Piekielko-Witkowska, A. (2016). Expression of genes involved in cellular adhesion and extracellular matrix remodeling correlates with poor survival of patients with renal cancer. *J. Urol.* 195, 1892–1902. doi:10.1016/j.juro.2015.11.050

Brinton, L., Zhang, P., Williams, K., Canfield, D., Orwick, S., Sher, S., et al. (2020). Synergistic effect of BCL2 and FLT3 co-inhibition in acute myeloid leukemia. *J. Hematol. Oncol.* 13, 139. doi:10.1186/s13045-020-00973-4

Bruunsgaard, H., Andersen-Ranberg, K., Hjelmborg, J., Pedersen, B. K., and Jeune, B. (2003). Elevated levels of tumor necrosis factor alpha and mortality in centenarians. *Am. J. Med.* 115, 278–283. doi:10.1016/s0002-9343(03)00329-2

Chen, C., Liang, C., Wang, S., Chio, C., Zhang, Y., Zeng, C., et al. (2020). Expression patterns of immune checkpoints in acute myeloid leukemia. *J. Hematol. Oncol.* 13, 28. doi:10.1186/s13045-020-00853-x

Chen, J., Schmitt, A., Chen, B., Rojewski, M., Rübeler, V., Fei, F., et al. (2008). Nilotinib hampers the proliferation and function of CD8+ T lymphocytes through inhibition of T cell receptor signalling. *J. Cell. Mol. Med.* 12, 2107–2118. doi:10.1111/j.1582-4934.2008.00234.x

Cui, J., Li, F., and Shi, Z. L. (2019). Origin and evolution of pathogenic coronaviruses. *Nat. Rev. Microbiol.* 17, 181–192. doi:10.1038/s41579-018-0118-9

Döhner, H., Weisdorf, D., and Bloomfield, C. (2015). Acute myeloid leukemia. *N. Engl. J. Med.* 373, 1136–1152. doi:10.1056/NEJMra1406184

Döhner, K., Thiede, C., Jahn, N., Panina, E., Gambietz, A., Larson, R., et al. (2020). Impact of NPM1/FLT3-ITD genotypes defined by the 2017 European LeukemiaNet in patients with acute myeloid leukemia. *Blood* 135, 371–380. doi:10.1182/blood.2019002697

Farah, N., Burt, R., Ibrahim, A., Baker, R., and Kottaridis, P. (2020). Concerns about how to use established minimal residual disease monitoring in the treatment of NPM1-mutant acute myeloid leukemia (AML) following reduced intensity chemotherapy protocols for AML given as a result of the COVID-19 pandemic. *Br. J. Haematol.* 190, e208–e210. doi:10.1111/bjh.16985

Ferrara, F., Zappasodi, P., Roncoroni, E., Borlenghi, E., and Rossi, G. (2020). Impact of Covid-19 on the treatment of acute myeloid leukemia. *Leukemia* 34, 2254–2256. doi:10.1038/s41375-020-0925-7

Franz, M., Rodriguez, H., Lopes, C., Zuberi, K., Montojo, J., Bader, G., et al. (2018). GeneMANIA update 2018. *Nucleic Acids Res.* 46, W60–W64. doi:10.1093/nar/gky311

Fu, D., Zhang, B., Wu, S., Zhang, Y., Xie, J., Ning, W., et al. (2021). Prognosis and characterization of immune microenvironment in acute myeloid leukemia through identification of an autophagy-related signature. *Front. Immunol.* 12, 695865. doi:10.3389/fimmu.2021.695865

Gebru, M., and Wang, H. (2020). Therapeutic targeting of FLT3 and associated drug resistance in acute myeloid leukemia. *J. Hematol. Oncol.* 13, 155. doi:10.1186/s13045-020-00992-1

Guo, J., Shirozu, K., Akahoshi, T., Mizuta, Y., Murata, M., and Yamaura, K. (2020). The farnesyltransferase inhibitor tipifarnib protects against autoimmune hepatitis induced by Concanavalin A. *Int. Immunopharmacol.* 83, 106462. doi:10.1016/j.intimp.2020.106462

He, L., Shi, L., Du, Z., Huang, H., Gong, R., Ma, L., et al. (2018). Mebendazole exhibits potent anti-leukemia activity on acute myeloid leukemia. *Exp. Cell Res.* 369, 61–68. doi:10.1016/j.yexcr.2018.05.006

Henderson, L. A., Canna, S. W., Friedman, K. G., Gorelik, M., and Lapidus, S. K. (2021). American College of rheumatology clinical guidance for multisystem inflammatory syndrome in children associated with SARS-CoV-2 and hyperinflammation in pediatric COVID-19. *Version 2* 73, e13–e29. doi:10.1002/art.41616

Huang, Y., Chen, S., Xiao, L., Qin, W., Li, L., Wang, Y., et al. (2021). A novel prognostic signature for survival prediction and immune implication based on SARS-CoV-2-related genes in kidney renal clear cell carcinoma. *Front. Bioeng. Biotechnol.* 9, 744659. doi:10.3389/fbioe.2021.744659

Jiang, L., Wang, R., Fang, L., Ge, X., Chen, L., Zhou, M., et al. (2019). HCP5 is a SMAD3-responsive long non-coding RNA that promotes lung adenocarcinoma metastasis via miR-203/SNAI axis. *Theranostics* 9, 2460–2474. doi:10.7150/thno.31097

Keklikoglou, I., Cianciaruso, C., Güç, E., Squadrito, M. L., Spring, L. M., Tazzyman, S., et al. (2019). Chemotherapy elicits pro-metastatic extracellular vesicles in breast cancer models. *Nat. Cell Biol.* 21, 190–202. doi:10.1038/s41556-018-0256-3

Khan, A. M., Ajmal, Z., Raval, M., and Tobin, E. (2020). Concurrent diagnosis of acute myeloid leukemia and COVID-19: A management challenge. *Cureus* 12, e9629. doi:10.7759/cureus.9629

Kiang, J. G., and Tsokos, G. C. (1998). Heat shock protein 70 kDa: Molecular biology, biochemistry, and physiology. *Pharmacol. Ther.* 80, 183–201. doi:10.1016/s0163-7258(98)00028-x

Liang, X., Chen, Y., and Fan, Y. (2022). Bioinformatics approach to identify common gene signatures of patients with coronavirus 2019 and lung adenocarcinoma. *Environ. Sci. Pollut. Res. Int.* 29, 22012–22030. doi:10.1007/s11356-021-17321-9

Liu, J., Marino, M. W., Wong, G., Grail, D., Dunn, A., Bettadapura, J., et al. (1998). TNF is a potent anti-inflammatory cytokine in autoimmune-mediated demyelination. *Nat. Med.* 4, 78–83. doi:10.1038/nm0198-078

Lyu, T., Jiang, Y., Jia, N., Che, X., Li, Q., Yu, Y., et al. (2020). SMYD3 promotes implant metastasis of ovarian cancer via H3K4 trimethylation of integrin promoters. *Int. J. Cancer* 146, 1553–1567. doi:10.1002/ijc.32673

Patel, P., Lapp, S., Grubbs, G., Edara, V., Rostad, C., Stokes, C., et al. (2021). Immune responses and therapeutic challenges in paediatric patients with new-onset acute myeloid leukaemia and concomitant COVID-19. *Br. J. Haematol.* 194, 549–553. doi:10.1111/bjh.17517

Shannon, P., Markiel, A., Ozier, O., Baliga, N., Wang, J., 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

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. Cell. Biol.* 14, 2087–2099. doi:10.1128/mcb.14.3.2087-2099.1994

Smith, C. C., Wang, Q., Chin, C. S., Salerno, S., Damon, L. E., Levis, M. J., et al. (2012). Validation of ITD mutations in FLT3 as a therapeutic target in human acute myeloid leukaemia. *Nature* 485, 260–263. doi:10.1038/nature11016

Soncin, F., Prevelige, R., and Calderwood, S. K. (1997). Expression and purification of human heat-shock transcription factor 1. *Protein Expr. Purif.* 9, 27–32. doi:10.1006/prep.1996.0672

Szklarczyk, D., Gable, A., Nastou, K., Lyon, D., Kirsch, R., Pyysalo, S., et al. (2021). The STRING database in 2021: Customizable protein-protein networks, and functional characterization of user-uploaded gene/measurement sets. *Nucleic Acids Res.* 49, D605–D612. doi:10.1093/nar/gkaa1074

V'Kovski, P., Kratzel, A., Steiner, S., and Stalder, H. (2021). Coronavirus biology and replication: Implications for SARS-CoV-2. *Nat. Rev. Microbiol.* 19, 155–170. doi:10.1038/s41579-020-00468-6

Westermann, J., and Bullinger, L. (2021). Precision medicine in myeloid malignancies. *Semin. Cancer Biol.* 84, 153–169. doi:10.1016/j.semcancer.2021.03.034

Wilde, L., Isidori, A., Keiffer, G., Palmisiano, N., and Kasner, M. (2020). Caring for AML patients during the COVID-19 crisis: An American and Italian experience. *Front. Oncol.* 10, 1689. doi:10.3389/fonc.2020.01689

Wu, M., Li, C., and Zhu, X. (2018). FLT3 inhibitors in acute myeloid leukemia. *J. Hematol. Oncol.* 11, 133. doi:10.1186/s13045-018-0675-4

Xu, S., Wang, Q., Kang, Y., Liu, J., Yin, Y., Liu, L., et al. (2020). Long noncoding RNAs control the modulation of immune checkpoint molecules in cancer. *Cancer Immunol. Res.* 8, 937–951. doi:10.1158/2326-6066.cir-19-0696

Yan, H., Qu, J., Cao, W., Liu, Y., Zheng, G., Zhang, E., et al. (2019). Identification of prognostic genes in the acute myeloid leukemia immune microenvironment based on TCGA data analysis. *Cancer Immunol. Immunother.* 68, 1971–1978. immunotherapy : CII. doi:10.1007/s00262-019-02408-7

Yele, V., Sanapalli, B., and Mohammed, A. (2022). Imidazoles and benzimidazoles as putative inhibitors of SARS-CoV-2 B.1.1.7 (alpha) and P.1 (gamma) variant spike glycoproteins: A computational approach. *Chem. Zvesti* 76, 1107–1117. doi:10.1007/s11696-021-01900-8

TYPE Original Research
PUBLISHED 29 September 2022
DOI 10.3389/fgene.2022.997970

OPEN ACCESS

EDITED BY
Youri I. Pavlov,
University of Nebraska Medical Center,
United States

REVIEWED BY
Olga Rechkoblit,
Icahn School of Medicine at Mount Sinai, United States
Arindam Bose,
Dana-Farber Cancer Institute,
United States

*CORRESPONDENCE
Ying Yuan,
yuanying1999@zju.edu.cn

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

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

RECEIVED 19 July 2022
ACCEPTED 19 August 2022
PUBLISHED 29 September 2022

CITATION
Zhu N, Zhao Y, Mi M, Lu Y, Tan Y, Fang X, Weng S and Yuan Y (2022), REV1: A novel biomarker and potential therapeutic target for various cancers. *Front. Genet.* 13:997970.
doi: 10.3389/fgene.2022.997970

REV1: A novel biomarker and potential therapeutic target for various cancers

Ning Zhu[1,2†], Yingxin Zhao[1,2†], Mi Mi[1,2], Yier Lu[1,2], Yinuo Tan[1], Xuefeng Fang[1], Shanshan Weng[1] and Ying Yuan[1,2,3*]

[1]Department of Medical Oncology, The Second Affiliated Hospital, Zhejiang University School of Medicine, Hangzhou, Zhejiang, China, [2]Cancer Institute, Key Laboratory of Cancer Prevention and Intervention, Ministry of Education, The Second Affiliated Hospital, Zhejiang University School of Medicine, Hangzhou, Zhejiang, China, [3]Cancer Center, Zhejiang University, Hangzhou, Zhejiang, China

Background: REV1 is a member of the translesion synthesis DNA polymerase Y family. It is an essential player in a variety of DNA replication activities, and perform major roles in the production of both spontaneous and DNA damage-induced mutations. This study aimed to explore the role of REV1 as a prognostic biomarker and its potential function regulating the sensitivity of anti-tumor drugs in various cancers.

Methods: We analyzed the impact of REV1 gene alterations on patient prognosis and the impact of different REV1 single nucleotide polymorphisms (SNP) on protein structure and function using multiple online prediction servers. REV1 expression was assessed using data from Oncomine, TCGA, and TIMER database. The correlation between REV1 expression and patient prognosis was performed using the PrognoScan and Kaplan-Meier plotter databases. The IC50 values of anti-cancer drugs were downloaded from the Genomics of Drug Sensitivity in Cancer database and the correlation analyses between REV1 expression and each drug pathway's IC50 value in different tumor types were conducted.

Results: Progression free survival was longer in REV1 gene altered group comparing to unaltered group [Median progression free survival (PFS), 107.80 vs. 60.89 months, *p* value = 7.062e-3]. REV1 SNP rs183737771 (F427L) was predicted to be deleterious SNP. REV1 expression differs in different tumour types. Low REV1 expression is associated with better prognosis in colorectal disease specific survival (DSS), disease-free survival (DFS), gastric overall survival (OS), post progression survival (PPS) and ovarian (OS, PPS) cancer while high REV1 expression is associated with better prognosis in lung [OS, relapse free

Abbreviations: TLS, Translesion synthesis; SNPs, single nucleotide polymorphisms; IBS, Illustrator for biological sequences; CNA, Copy number amplification; nsSNPs, nonsynonymous SNPs; GDSC, Genomics of Drug Sensitivity in Cancer; UBM, ubiquitin-binding motif; CTD, C-terminal domain; DSS, disease specific survival; DFS, disease-free survival; CHOL, cholangiocarcinoma; ESCA, esophageal carcinoma; HNSC, head and neck cancer; LIHC, liver hepatocellular carcinoma; LUSC, lung squamous cell carcinoma; STAD, stomach adenocarcinoma; BLCA, bladder urothelial carcinoma; BRCA, breast invasive carcinoma; KICH, kidney chromophobe; KIRC, kidney renal clear cell carcinoma; KIRP, kidney renal papillary cell carcinoma; PRAD, prostate adenocarcinoma; THCA, thyroid carcinoma; UCEC, uterine corpus endometrial carcinoma; PPS, post progression survival; FP, first progression; DMFS, distant metastasis-free survival; HR, hazard radio

survival (RFS), first progession (FP), PPS] and breast (DSS, RFS) cancer. In colon adenocarcinoma and rectum adenocarcinoma and lung adenocarcinoma, low expression of REV1 may suggest resistance to drugs in certain pathways. Conversely, high expression of REV1 in acute myeloid leukemia, brain lower grade glioma, small cell lung cancer and thyroid carcinoma may indicate resistance to drugs in certain pathways.

Conclusion: REV1 plays different roles in different tumor types, drug susceptibility, and related biological events. REV1 expression is significantly correlated with different prognosis in colorectal, ovarian, lung, breast, and gastric cancer. REV1 expression can be used as predictive marker for various drugs of various pathways in different tumors.

KEYWORDS

REV1, translesion synthesis (TLS), prognostic biomarker, drug sensitivity, single nucleotide polymorphism (SNP)

1 Introduction

Translesion synthesis (TLS) is a DNA damage bypass process involving a group of DNA polymerases including REV1, Pol η, ι, κ, and ζ, which together tolerate DNA damage and lead to mutations (Yamanaka et al., 2017). REV1 is a member of the TLS DNA polymerase Y family with dCMP transferase activity that helps bypass certain lesions and functions as a scaffolding protein associated with several TLS DNA polymerases (Gan et al., 2008; Washington et al., 2010). REV1 is an essential player in a variety of DNA replication activities, and perform major roles in the production of both spontaneous and DNA damage-induced mutations (Lawrence, 2002). A number of studies indicate that REV1 has been linked to the development of some cancers (Sakiyama et al., 2005; He et al., 2008; Dumstorf et al., 2009; Xu et al., 2013; Goricar et al., 2014). Studying REV1 will increase our understanding of the origin of cancer, as mutations are an important feature of cancer development.

In our study, we analyzed the impact of REV1 gene alterations on patient outcomes and the impact of different REV1 single nucleotide polymorphisms (SNPs) on protein structure and function using multiple online prediction servers. We also explored the link between the expression of REV1 and cancer patient outcome as well as the correlation between REV1 expression and IC50 of various drugs in different tumor types.

Our study conducted a comprehensive assessment of REV1, explored the role of REV1 in various cancers as a prognostic biomarker and further highlight a potential function whereby REV1 may regulate the sensitivity of tumor cells to specific drugs.

2 Materials and methods

2.1 Molecular structure of REV1

The summary of molecular structure and function of REV1 were based on the comprehensive literature results. Illustrator for biological sequences (IBS) (http://ibs.biocuckoo.org/) was used to create a schematic of the protein domain (Liu et al., 2015).

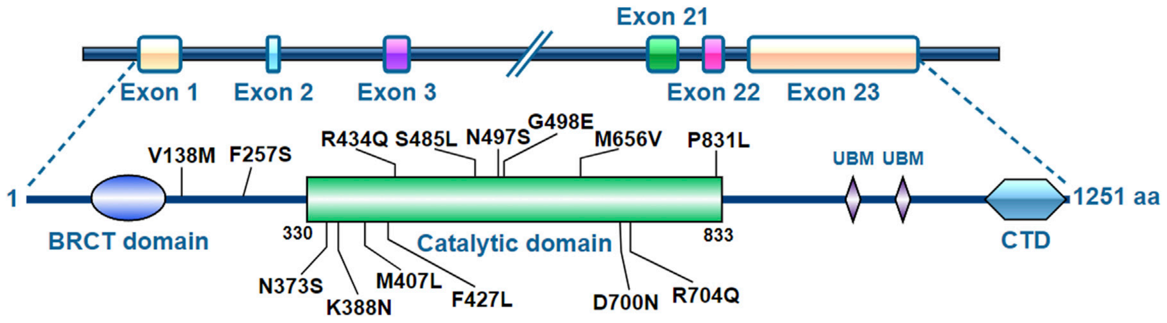


FIGURE 1
Molecular structure of REV1.

2.2 REV1 mutation analysis

Copy number amplification (CNA) and mutation status of REV1 in pan-cancer samples were obtained from the cBioPortal database (https://www.cbioportal.org/). We also analyzed the impact of REV1 gene alterations on patient outcomes using cBioPortal database. In addition, REV1 SNPs were searched in NCBI dbSNP database (https://www.ncbi.nlm.nih.gov/snp/), we selected 14 nonsynonymous SNPs (nsSNPs) which have been reported for further analyses. Eight different online servers (SIFT, PolyPhen-2, PANTHER, SNPs&GO, PROVEAN, PredictSNP, Mutation Taster2, and Mutation assessor) based on multiple algorithms (Supplementary Table S1), were used for structural and functional annotation. Stability analysis was carried out using I-Mutant2.0, MUpro, and iStable (Supplementary Table S1). We performed 3D modeling analysis of the wild-type and mutant UmuC domains of REV1 protein.

2.3 REV1 expression analysis

We assessed the expression of REV1 in multiple tumour and normal tissue types using the Oncomine database. We also analyzed immunohistochemistry staining images from the Human Protein Atlas (HPA) website (https://www.proteinatlas.org/) to study the protein expression of REV1 in multiple tumour and normal tissue types. We further used the TCGA and TIMER databases to assess how REV1 expression differs in particular tumour types. We also explored the link between the expression of REV1 and cancer patient outcome using the PrognoScan and Kaplan-Meier plotter databases (Supplementary Table S1).

2.4 Drug sensitivity analysis

The IC50 values of drugs, each drug corresponding signaling pathway and gene expression profiles in the relative cell lines (E-MTAB3610) were downloaded from the Genomics of Drug Sensitivity in Cancer (GDSC) database (https://www.cancerrxgene.org/) and ArrayExpress database (https://www.ebi.ac.uk/arrayexpress/). The correlation analyses between REV1 expression and each drug pathway's IC50 value in different tumor types were conducted.

2.5 Statistical analysis

Statistical analyses were conducted using R software (version 3.6.2) and attached packages. Survival analyses were performed using the log-rank test, the Kaplan-Meier method, and the Cox regression model. Correlation analyses were conducted using Pearson's test. Two-tailed p value < 0.05 was considered as statistically significant. p-value Significant Codes: $0 \le$ *** $< 0.001 \le$ ** $< 0.01 \le$ * $< 0.05 \le \ < 0.1$.

3 Results

3.1 Molecular structure and function of REV1

The human REV1 gene locates at chromosome 2q11.2 and has 23 exons encoding the REV1 protein consisting of 1,251 amino acids (Figure 1). The middle part of the REV1 protein (amino acids 330–833) is the catalytic domain, while the C- and N-terminal regions contain protein-protein interaction domains [e.g., ubiquitin-binding motif (UBM), C-terminal domain (CTD) and BRCT] (Swan et al., 2009; Sale, 2013).

REV1 is a member of the eukaryotic Y-polymerase family of TLS DNA polymerases (Murakumo et al., 2001; Guo et al., 2003; Ohashi et al., 2004; Tissier et al., 2004). It plays a central role in TLS by participating in protein-protein interactions through two distinct interfaces at its CTD. It recruits the TLS polymerases POL κ, POL ι, and POL η using one interface, and recruits POL ζ through interaction with REV7 components by another interface (Pozhidaeva et al., 2012; Wojtaszek et al., 2012; Yamanaka et al., 2017). In addition to this non-catalytic role, REV1, unique among polymerases, has its own catalytic activity as a deoxycytidyl transferase, utilizing a protein-template directed mechanism (Nair et al., 2005; Swan et al., 2009).

3.2 The role of REV1 in carcinogenesis and prognosis

3.2.1 REV1 single nucleotide polymorphisms in various cancers

Figure 2A showed CNA and mutation status of REV1 in pan-cancer samples from cBioPortal database. Survival analysis of REV1 gene alterations suggested that progression free survival (PFS) was longer in REV1 gene altered group comparing to unaltered group (Median PFS, 107.80 vs. 60.89 months, p value = 7.062e-3) (Figure 2B). However, the disease specific survival (DSS) and disease-free survival (DFS) of the two groups were not statistically different based on data from cBioPortal database.

There were 23,494 REV1 SNPs in NCBI dbSNP database, 19,933 SNPs in intronic region, 1,183 SNPs in 5′UTR region, 432 SNPs in 3′UTR region, and 926 SNPs were nsSNPs and 422 SNPs were synonymous SNPs in coding sequence. We selected 14 nsSNPs which have been reported for further analyses. Out of 14 nsSNPs, only rs183737771 (F427L) was predicted to be deleterious SNPs in all computational algorithms (Table 1). Nine nsSNPs were predicted to decrease

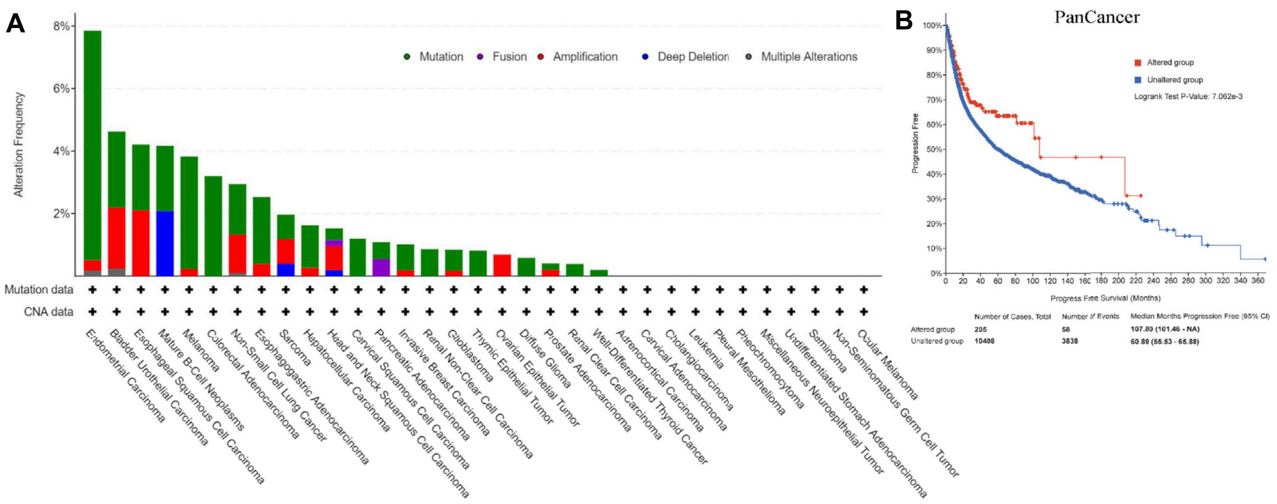


FIGURE 2
Copy number amplification and mutation status of REV1 **(A)** and corresponding progression free survival **(B)** in pan-cancer samples.

protein stability, meanwhile, one nsSNP was predicted to increase (Table 1).

To further analyze why F427L is a deleterious SNP, we performed 3D model predictions. We selected 3gqc.1. A as template in 3D modeling analysis (Figure 3). We found that compared with wild-type domain, F427L has no significant effect on protein structure. However, the wild-type 3D structure (3GQC) showed that position 427 is involved in the contact of Mg^{2+} and dCTP (Figure 4A). Project HOPE analysis revealed that after the phenylalanine at position 427 is mutated to leucine, the mutated residue is smaller than that of the wild-type residue, then creates a blank space in the protein core (Figures 4B,C), which probably interferes with the interaction with Mg^{2+} and leads to the loss of interaction with dCTP, thereby affecting protein function (Venselaar et al., 2010).

3.2.2 REV1 expression in different cancer and normal tissues

The expression of REV1 in multiple tumour and normal tissue types from the Oncomine database revealed that expression of this gene was elevated relative to normal tissue controls for brain, renal, prostate cancers, melanoma, myeloma and sarcoma. We also found that relative to normal tissue controls, REV1 expression was lower in breast, ovarian cancer and lymphoma tissues (Figure 5A). Detailed findings in particular tumour types are showed in Supplementary Table S2.

REV1 expression in tumor tissue samples and normal tissue samples from TCGA and TIMER databases were assessed to figure out how REV1 expression differs in particular tumour types. The results suggested that the expression of REV1 was significantly elevated relative to normal controls in cholangiocarcinoma (CHOL), esophageal carcinoma (ESCA), head and neck cancer (HNSC), liver hepatocellular carcinoma (LIHC), lung squamous cell carcinoma (LUSC) and stomach adenocarcinoma (STAD). In contrast, low REV1 expression was observed in eight cancer types, namely, bladder urothelial carcinoma (BLCA), breast invasive carcinoma (BRCA), kidney chromophobe (KICH), kidney renal clear cell carcinoma (KIRC), kidney renal papillary cell carcinoma (KIRP), prostate adenocarcinoma (PRAD), thyroid carcinoma (THCA) and uterine corpus endometrial carcinoma (UCEC). Differences between the expression of REV1 in tumors and normal adjacent tissue samples in the TCGA data set are shown in Figure 5B.

Immunohistochemistry staining analysis from HPA yielded similar results (Figures 6, 7). REV1 proteins were not expressed or medium expressed in normal lung tissues, while high protein expressions of REV1 were expressed in lung cancer tissues (Figure 6A). Medium protein expressions of REV1 were observed in urinary bladder, breast, prostate and endometrium normal tissues, while not detected (breast cancer, endometrial carcinoma) or low (urothelial carcinoma, prostate adenocarcinoma) protein expressions of REV1 were observed in corresponding cancer tissues (Figures 7A,B,D,F). REV1 proteins were high expressed in normal kidney tissues, while low protein expressions of REV1 were expressed in kidney adenocarcinoma tissues (Figure 7C).

3.2.3 The association between REV1 expression and patient prognosis

The association between the REV1 expression and patient prognosis were analyzed using the PrognoScan database (Supplementary Tables S3–S7). We found that high REV1 expression is associated with poorer prognosis in

TABLE 1 High risk nsSNPs identified in silico programs and effect of nsSNPs on protein stability predicted by silico programs.

SNP ID	AA change	Structural and functional annotation								Protein stability		
		SIFT	Polyphen2	PANTHER	SNPs&Go	PROVEAN	PredictSNP	Mutation Taster2	Mutation assessor	i-Mutant2.0	Mupro	iStable
rs3087399	N373S	Tolerated	Benign	Possibly damaging	Neutral	Deleterious	Neutral	Benign	Medium	Decrease	Increase	Increase
rs3087386	F257S	Tolerated	Benign	Possibly damaging	Neutral	Neutral	Neutral	Benign	Neutral	Decrease	Decrease	Decrease
rs3087403	V138M	Tolerated	Benign	Possibly damaging	Neutral	Neutral	Neutral	Benign	Low	Decrease	Decrease	Decrease
rs72550807	K388N	Deleterious	Probably Damaging	Possibly damaging	Neutral	Deleterious	Neutral	Benign	Medium	Decrease	Decrease	Decrease
rs138841507	S485L	Tolerated	Benign	Possibly damaging	Neutral	Neutral	Neutral	Benign	Neutral	Increase	Increase	Increase
rs144046214	G498E	Deleterious	Probably Damaging	Possibly damaging	Neutral	Deleterious	Deleterious	Benign	Medium	Decrease	Increase	Increase
rs183737771	F427L	Deleterious	Probably Damaging	Possibly damaging	Disease	Deleterious	Deleterious	Deleterious	High	Decrease	Decrease	Decrease
rs148052685	R434Q	Deleterious	Probably Damaging	Possibly damaging	Neutral	Deleterious	Deleterious	Deleterious	Low	Decrease	Decrease	Decrease
rs3087394	M656V	Deleterious	Benign	Possibly damaging	Neutral	Neutral	Neutral	Benign	Medium	Decrease	Decrease	Decrease
rs28382941	D700N	Deleterious	Benign	Possibly damaging	Neutral	Deleterious	Neutral	Benign	Medium	Decrease	Decrease	Decrease
rs28382942	R704Q	Tolerated	Benign	Possibly damaging	Neutral	Neutral	Neutral	Benign	Neutral	Decrease	Decrease	Decrease
rs139685542	P831L	Tolerated	Benign	Possibly damaging	Neutral	Neutral	Neutral	Benign	Neutral	Decrease	Increase	Increase
rs200184935	M407L	Tolerated	Benign	Possibly damaging	Neutral	Neutral	Neutral	Benign	Low	Decrease	Increase	Increase
rs180712764	N497S	Tolerated	Probably Damaging	Possibly damaging	Neutral	Neutral	Neutral	Benign	Neutral	Decrease	Increase	Increase

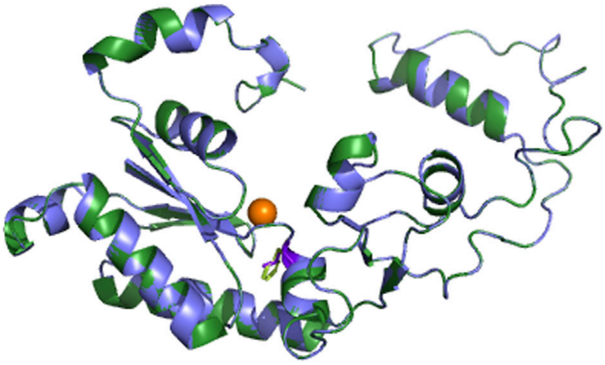


FIGURE 3
3D models of UmuC domain.

colorectal (DSS: HR, 3.15, 95% CI, 1.05–9.39, *p* = 0.039999; DFS: HR, 2.72, 95% CI, 1.04–7.16, *p* = 0.041971) and ovarian cancer (overall survival (OS) 1: HR, 1.72, 95% CI, 1.10–2.67, *p* = 0.016573; OS 2: HR, 1.50, 95% CI, 1.06–2.12, *p* = 0.022128) (Figures 8A–D). However, as shown in Figures 8E–H, survival curves identified that high expression of REV1 indicated better prognosis in lung [OS 1: HR, 0.54, 95% CI, 0.37–0.79, *p* = 0.001760; OS 2: HR, 0.37, 95% CI, 0.20–0.70, *p* = 0.002163; relapse free survival (RFS): HR, 0.09, 95% CI, 0.03–0.28, *p* = 0.000033] and breast cancer (DSS: HR, 0.56, 95% CI, 0.38–0.81, *p* = 0.002398).

The Kaplan-Meier plotter database was also employed to assess the relationship between REV1 expression and patient prognosis in a range of cancer types. As shown in Figures 9A–C, the elevation expression of REV1 to be significantly linked with a poorer prognosis in gastric cancer [OS: HR, 1.4, 95% CI, 1.11–1.76, *p* = 0.0038; post progression survival (PPS): HR, 1.45, 95% CI, 1.11–1.9, *p* = 0.0069] and ovarian cancer (PPS: HR, 1.35, 95% CI, 1.04–1.75, *p* = 0.024). However, we found reduced REV1 expression to be correlated with poorer patient prognosis in lung cancer (first progession (FP): HR, 0.42, 95% CI, 0.28–0.61, *p* = 2.3e-06; OS: HR, 0.53, 95% CI, 0.43–0.65, *p* = 6.2e-10; PPS: HR, 0.54, 95% CI, 0.34–0.84, *p* = 0.0055) (Figures 9D–F), Rectum adenocarcinoma (OS: HR, 0.24, 95% CI, 0.08–0.74, *p* = 0.0084; RFS: HR, 0.08, 95% CI, 0.01–0.99, *p* = 0.022) (Figures 9G,H) and breast cancer (RFS: HR, 0.77, 95% CI, 0.66–0.89, *p* = 0.00073) (Figure 9I). There was no statistically significant relationship between the expression of REV1 and the prognosis of breast cancer patients [distant metastasis-free survival (DMFS), OS and PPS], gastric cancer (FP) and ovarian cancer (OS and PFS) (Supplementary Figure S1).

3.3 The role of REV1 in sensitivity of cancer treatment

3.3.1 The association between REV1 expression and cancer treatment sensitivity

As shown in Figure 10 and Table S8, REV1 expression is negatively correlated with IC50 of drugs of DNA replication, hormone-related, JNK and p38 signaling, kinases, protein stability and degradation, RTK signaling, and WNT signaling pathways in colon adenocarcinoma and rectum adenocarcinoma and drugs of kinases, RTK signaling, and WNT signaling pathways in lung adenocarcinoma. In contrast, in acute myeloid leukemia, brain lower grade glioma, small cell lung cancer and thyroid carcinoma, REV1 expression is positively

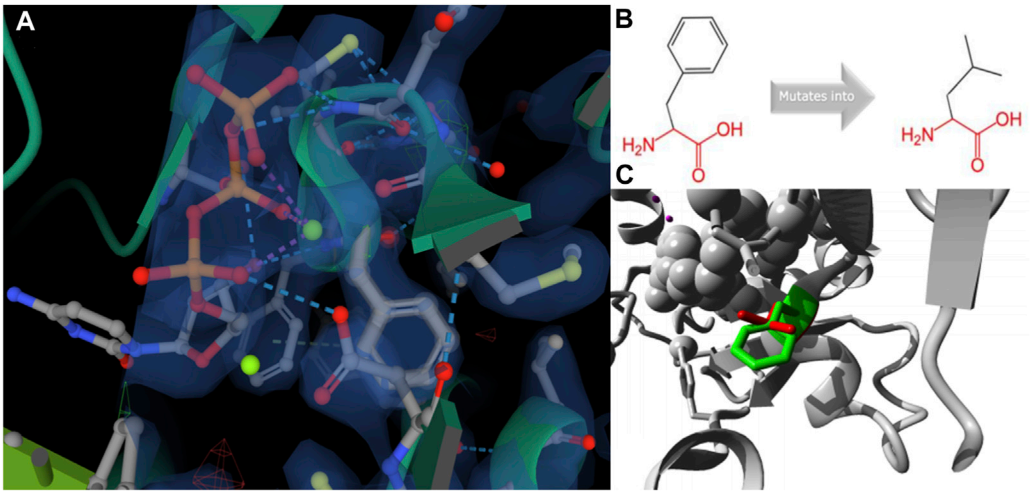


FIGURE 4
3D structure **(A, C)** and chemical structure changes **(B)** of position 427 on REV1 protein.

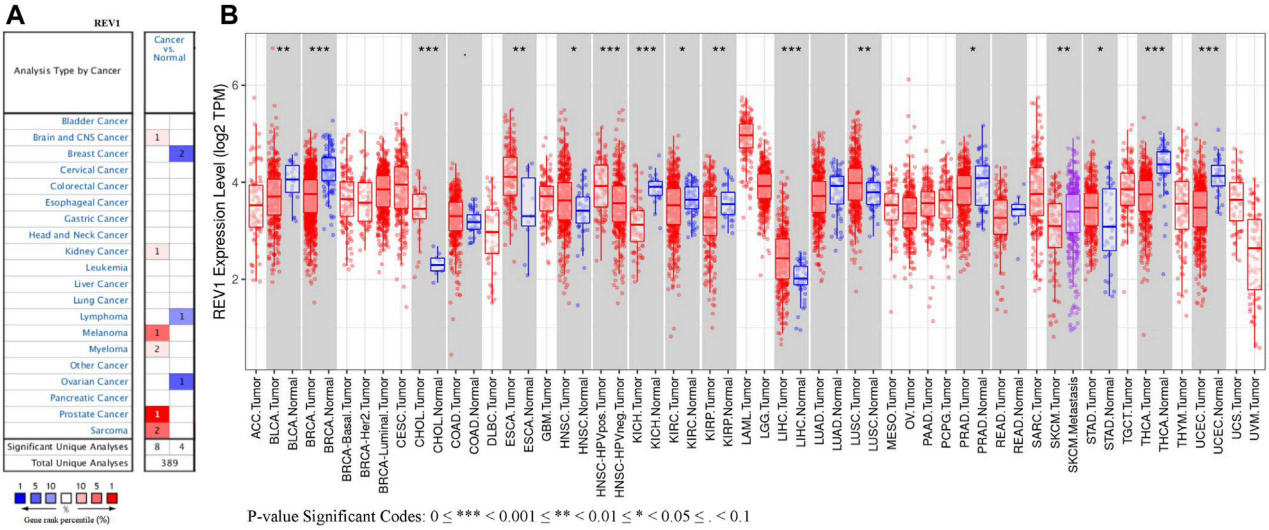


FIGURE 5
The expression level of REV1 in different types of tumor tissues and normal tissues. **(A)** Oncomine database. **(B)** TIMER database.

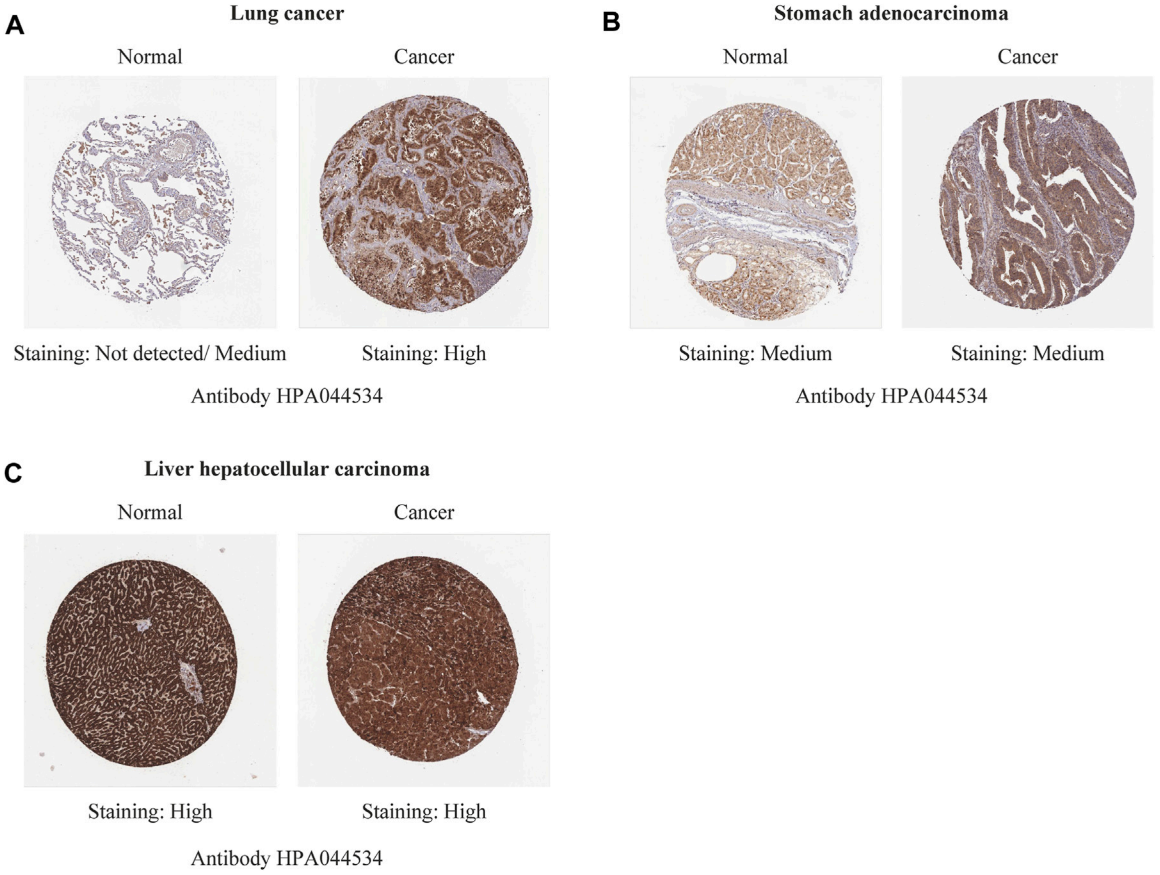


FIGURE 6
The immunohistochemistry staining images of REV1 in different types of tumor tissues and normal tissues [lung **(A)**, stomach **(B)** and liver **(C)**].

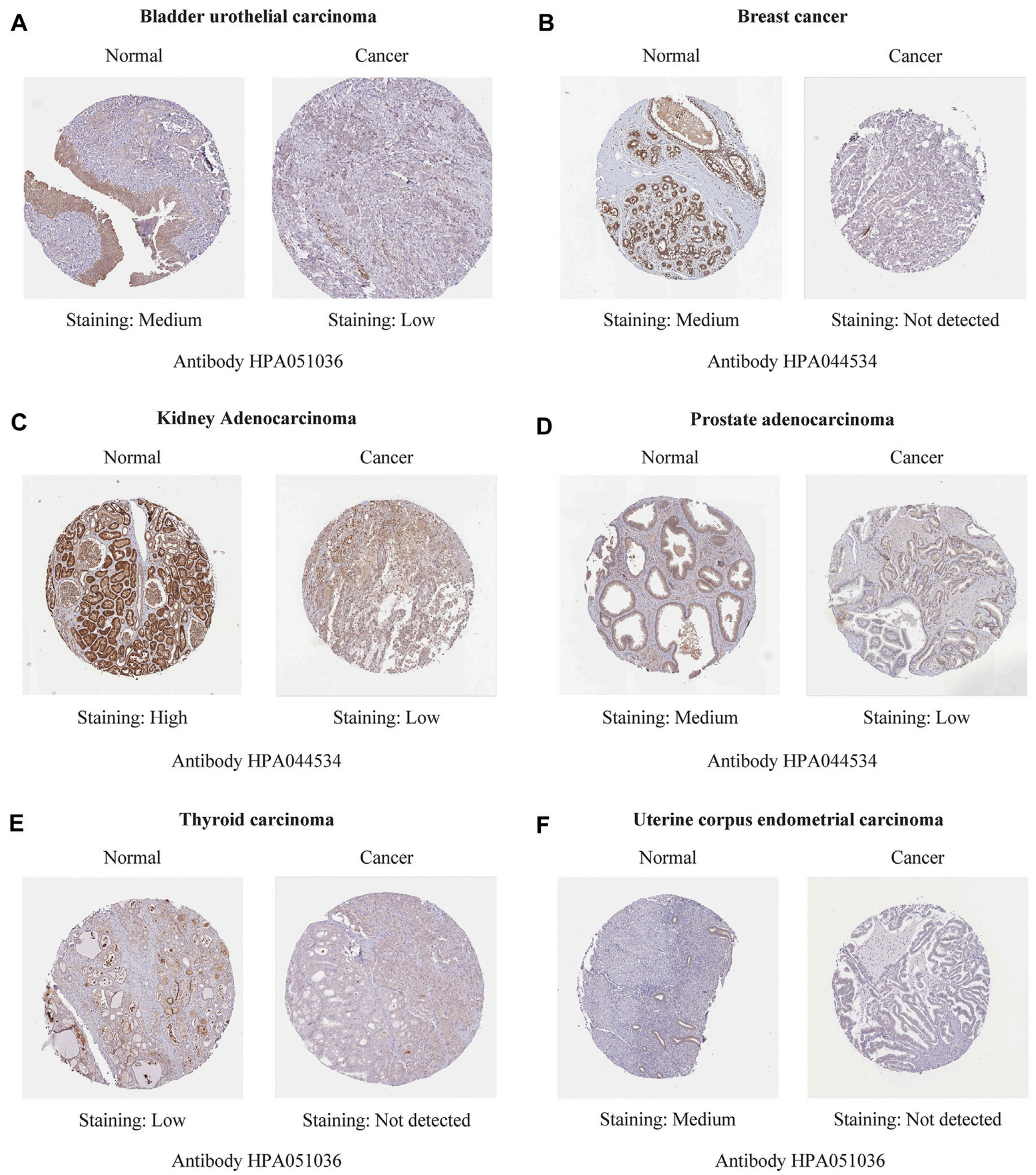


FIGURE 7
The immunohistochemistry staining images of REV1 in different types of tumor tissues and normal tissues [urinary bladder **(A)**, breast **(B)**, kidney **(C)**, prostate **(D)**, thyroid **(E)**, and endometrium **(F)**].

correlated with IC50 of drugs of various pathways (acute myeloid leukemia: Cell cycle, Genome integrity, Hormone-related, kinases, RTK signaling, and WNT signaling pathways; brain lower grade glioma: Cell cycle, Chromatin histone methylation, DNA replication, JNK and p38 signaling, kinases, PI3K/MTOR signaling, and RTK signaling pathways; small cell lung cancer: Cell cycle, Chromatin histone acetylation, Cytoskeleton, DNA replication, ERK MAPK signaling, Hormone-related, IGF1R signaling, JNK and p38 signaling, Mitosis, kinases, PI3K/MTOR signaling, RTK signaling, and WNT signaling pathways; thyroid carcinoma: Cell cycle, ERK MAPK signaling, Hormone-related, JNK and p38 signaling, kinases, PI3K/MTOR signaling, and RTK signaling pathways). In acute lymphoblastic leukemia, REV1 expression is negatively

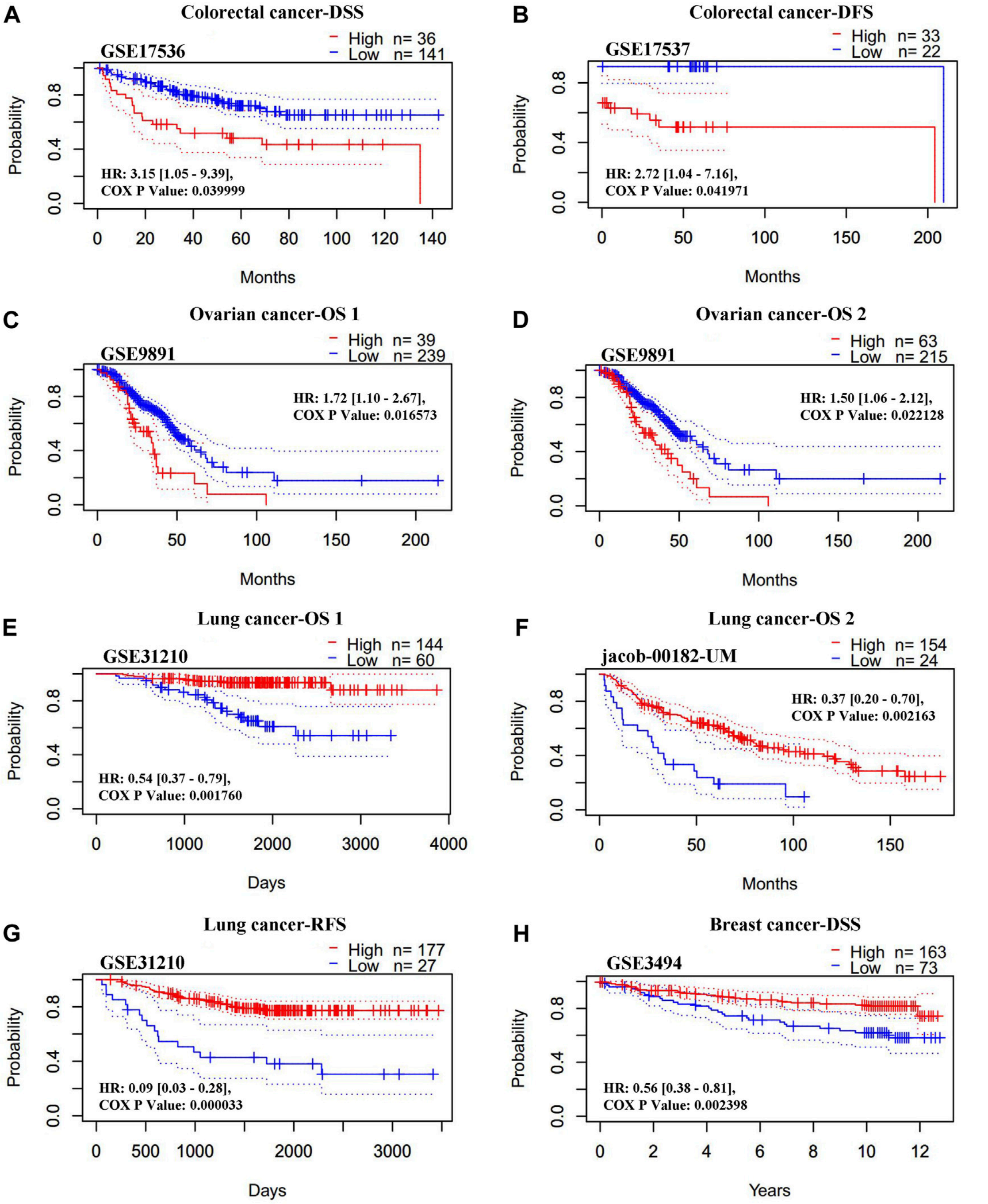


FIGURE 8
Correlation between REV1 expression and prognosis of various types of cancer [colorectal cancer **(A, B)**, ovarian cancer **(C, D)**, lung cancer **(E–G)**, and breast cancer **(H)**] (PrognoScan database).

correlated with IC50 of drugs of Genome integrity and metabolism pathways while positively correlated with IC50 of drugs of RTK signaling pathway. Thus, REV1 expression can be used as predictive marker for various drugs of various pathways in different tumors.

4 Discussion

REV1 is one of the key proteins in TLS. TLS is a DNA damage tolerance process, which contributes to cell survival by bypass of the unrepaired DNA lesions. TLS functions in an error-prone

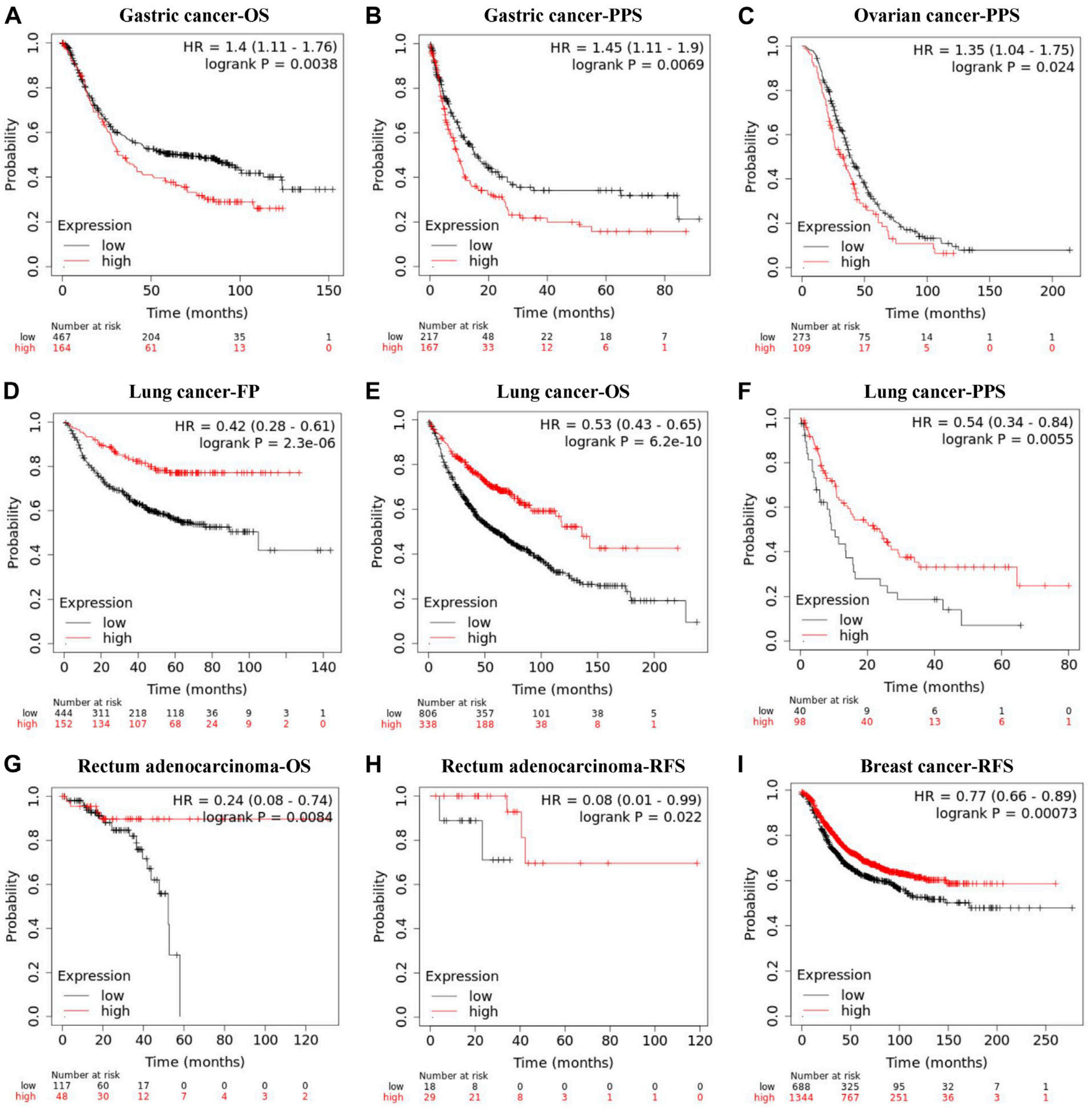
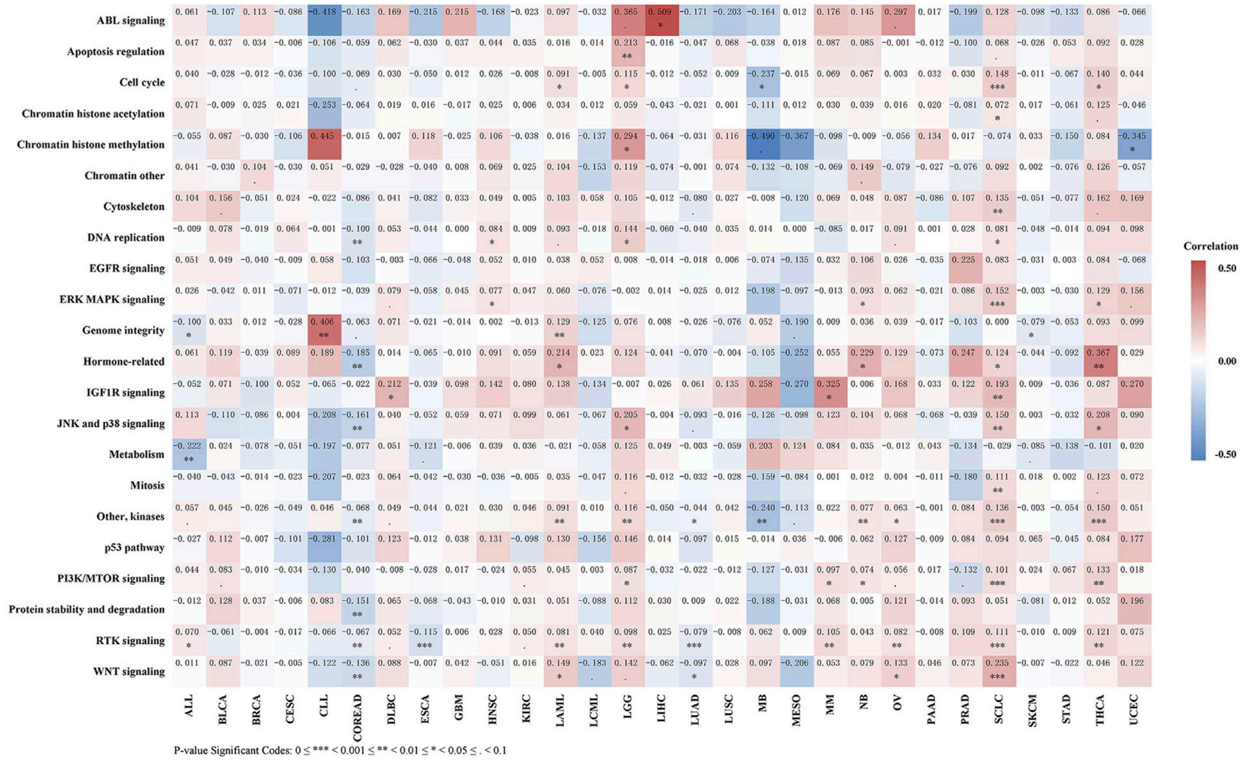


FIGURE 9
Correlation between REV1 expression and prognosis of various types of cancer [gastric cancer **(A, B)**, ovarian cancer **(C)**, lung cancer **(D–F)**, rectum adenocarcinoma **(G, H)**, and breast cancer **(I)**] (Kaplan-Meier Plotter database).

manner and sometimes can actively promote the generation of mutations (Waters et al., 2009). These accumulated errors in the DNA may play a key role in the initiation of various types of cancers (Schoket, 2004). REV1 also plays important role in replication stress response (RSR). The RSR is activated in response to DNA lesions or intrinsic replication fork barriers. Replication forks stalled at DNA lesions can restart replication by firing dormant origins, repriming replication, reversing the stalled fork or activating the DNA damage tolerance pathways (Zeman and Cimprich, 2014). ssDNA gaps are frequent structures that accumulate on newly synthesized DNA under conditions of replication stress (Quinet et al., 2014). Tirman et al. (2021) showed that REV1-POLζ-mediated trans-DNA damage synthesis promotes ssDNA gap filling in G2 and S, affecting cell survival and genome stability. Similar findings were found by Nayak et al. (2020), Moreover, Yang et al. (2015) found that



	ALL	BLCA	BRCA	CESC	CLL	COREAD	DLBC	ESCA	GBM	HNSC	KIRC	LAML	LCML	LGG	LIHC	LUAD	LUSC	MB	MESO	MM	NB	OV	PAAD	PRAD	SCLC	SKCM	STAD	THCA	UCEC
ABL signaling	0.061	-0.107	0.113	-0.086	-0.418	-0.163	0.169	-0.215	0.215	-0.168	-0.023	0.097	-0.032	0.365 .	0.509 *	-0.171	-0.203	-0.164	0.012	0.176	0.145	0.297 .	0.017	-0.199	0.128	-0.098	-0.133	0.086	-0.066
Apoptosis regulation	0.047	0.037	0.034	-0.006	-0.106	-0.059	0.062	-0.030	0.037	0.044	0.035	0.016	0.014	0.213 **	-0.016	-0.047	0.068	-0.038	0.018	0.087	0.085	-0.001	-0.012	-0.100	0.068 .	-0.026	0.053	0.092	0.028
Cell cycle	0.040	-0.028	-0.012	-0.036	-0.100	-0.069 .	0.030	-0.050	0.012	0.026	-0.008	0.091 *	-0.005	0.115 *	-0.012	-0.052	0.009	-0.237 *	-0.015	0.069	0.067	0.003	0.032	0.030	0.148 ***	-0.011	-0.067	0.140 *	0.044
Chromatin histone acetylation	0.071	-0.009	0.025	0.021	-0.253	-0.064	0.019	0.016	-0.017	0.025	0.006	0.034	0.012	0.059	-0.043	-0.021	0.001	-0.111	0.012	0.030	0.039	0.016	0.020	-0.081	0.072 *	0.017	-0.061	0.125 .	-0.046
Chromatin histone methylation	-0.055	0.087	-0.030	-0.106	0.445	-0.015	0.007	0.118	-0.025	0.106	-0.038	0.016	-0.137	0.294 *	-0.064	-0.031	0.116	-0.490 .	-0.367	-0.098	-0.009	-0.056	0.134	0.017	-0.074	0.033	-0.150	0.084	-0.345 *
Chromatin other	0.041	-0.030	0.104 .	-0.030	0.051	-0.029	-0.028	-0.040	0.008	0.069	0.025	0.104	-0.153	0.119	-0.074	-0.001	0.074	-0.132	-0.108	-0.069	0.149 .	-0.079	-0.027	-0.076	0.092	0.002	-0.076	0.126	-0.057
Cytoskeleton	0.104	0.156 .	-0.051	0.024	-0.022	-0.086	0.041	-0.082	0.033	0.049	0.005	0.103	0.058	0.105	-0.012	-0.080 .	0.027	-0.008	-0.120	0.069	0.048	0.087	-0.086	0.107	0.135 **	-0.051	-0.077	0.162 .	0.169
DNA replication	-0.009	0.078	-0.019	0.064	-0.001	-0.100 **	0.053	-0.044	0.000	0.084 *	0.009	0.093 .	-0.018	0.144 *	-0.060	-0.040	0.035	0.014	0.000	-0.085	0.017	0.091 .	0.001	0.028	0.081 *	-0.048	-0.014	0.094	0.098
EGFR signaling	0.051	0.049	-0.040	-0.009	0.058	-0.103	-0.003	-0.066	-0.048	0.052	0.010	0.038	0.052	0.008	-0.014	-0.018	0.006	-0.074	-0.135	0.032	0.106	0.026	-0.035	0.225	0.083	-0.031	0.003	0.084	-0.068
ERK MAPK signaling	0.026	-0.042	0.011	-0.071	-0.012	-0.039	0.079 .	-0.058	0.045	0.077 *	0.047	0.060	-0.076	-0.002	0.014	-0.025	0.012	-0.198	-0.097	-0.013	0.093 *	0.062	-0.021	0.086	0.152 ***	-0.003	-0.030	0.129 *	0.156 .
Genome integrity	-0.100 *	0.033	0.012	-0.028	0.406 **	-0.063 .	0.071	-0.021	-0.014	0.002	-0.013	0.129 **	-0.125	0.076	0.008	-0.026	-0.076	0.052	-0.190 .	0.009	0.036	0.039	-0.017	-0.103	0.000	-0.079 *	-0.053	0.093	0.099
Hormone-related	0.061	0.119	-0.039	0.089	0.189	-0.185 **	0.014	-0.065	-0.010	0.091	0.059	0.214 *	0.023	0.124	-0.041	-0.070	-0.004	-0.105	-0.252	0.055	0.229 *	0.129	-0.073	0.247	0.124 *	-0.044	-0.092	0.367 **	0.029
IGF1R signaling	-0.052	0.071	-0.100	0.052	-0.065	-0.022	0.212 *	-0.039	0.098	0.142	0.080	0.138	-0.134	-0.007	0.026	0.061	0.135	0.258	-0.270	0.325 *	0.006	0.168	0.033	0.122	0.193 **	0.009	-0.036	0.087	0.270
JNK and p38 signaling	0.113	-0.110	-0.086	0.004	-0.208	-0.161 **	0.040	-0.052	0.059	0.071	0.099	0.061	-0.067	0.205 *	-0.004	-0.093 .	-0.016	-0.126	-0.098	0.123	0.104	0.068	-0.068	-0.039	0.150 **	0.003	-0.032	0.208 *	0.090
Metabolism	-0.222 **	0.024	-0.078	-0.051	-0.197	-0.077	0.051	-0.121 .	-0.006	0.039	0.036	-0.021	-0.058	0.125	0.049	-0.003	-0.059	0.203	0.124	0.084	0.035	-0.012	0.043	-0.134	-0.029	-0.085 .	-0.138	-0.101	0.020
Mitosis	-0.040	-0.043	-0.014	-0.023	-0.207	-0.023	0.064	-0.042	-0.030	-0.036	-0.005	0.035	-0.047	0.116 .	-0.012	-0.032	-0.028	-0.159	-0.084	0.001	0.012	0.004	-0.011	-0.180	0.111 **	0.018	0.002	0.123 .	0.072
Other, kinases	0.057 .	0.045	-0.026	-0.049	0.046	-0.068 **	0.049 .	-0.044	0.021	0.030	0.046	0.091 **	0.010	0.116 **	-0.050	-0.044 *	0.042	-0.240 **	-0.113 .	0.022	0.077 **	0.063 *	-0.001	0.084	0.136 ***	-0.003	-0.054	0.150 ***	0.051
p53 pathway	-0.027	0.112	-0.007	-0.101	-0.281	-0.101	0.123	-0.012	0.038	0.131	-0.098	0.130	-0.156	0.146	0.014	-0.097	0.015	-0.014	0.036	-0.006	0.062	0.127	-0.009	0.084	0.094	0.065	-0.045	0.084	0.177
PI3K/MTOR signaling	0.044	0.083 .	-0.010	-0.034	-0.130	-0.040	-0.008	-0.028	0.017	-0.024	0.055 .	0.045	0.003	0.087 *	-0.032	-0.022	-0.012	-0.127	-0.031	0.097 *	0.074 *	0.056 .	0.017	-0.132 .	0.101 ***	0.024	0.067	0.133 **	0.018
Protein stability and degradation	-0.012	0.128	0.037	-0.006	0.083	-0.151 **	0.065	-0.068	-0.043	-0.010	0.031	0.051	-0.088	0.112	0.030	0.009	0.022	-0.188	-0.031	0.068	0.005	0.121	-0.014	0.093	0.051	-0.081	0.012	0.052	0.196
RTK signaling	0.070 *	-0.061	-0.004	-0.017	-0.066	-0.067 **	0.052 .	-0.115 ***	0.006	0.028	0.050 .	0.081 **	0.040	0.098 **	0.025	-0.079 ***	-0.008	0.062	0.009	0.105 **	0.043	0.082 **	-0.008	0.109	0.111 ***	-0.010	0.009	0.121 **	0.075
WNT signaling	0.011	0.087	-0.021	-0.005	-0.122	-0.136 **	0.088	-0.007	0.042	-0.051	0.016	0.149 *	-0.183 .	0.142 .	-0.062	-0.097 *	0.028	0.097	-0.206	0.053	0.079	0.133 *	0.046	0.073	0.235 ***	-0.007	-0.022	0.046	0.122

P-value Significant Codes: 0 ≤ *** < 0.001 ≤ ** < 0.01 ≤ * < 0.05 ≤ . < 0.1

FIGURE 10
Correlation heatmap of REV1 expression and drug pathway's IC50 in different tumors.

REV1 protects nascent replication tracts intact by stabilizing RAD51 filaments, to block nascent replication tracts from degradation in response to replication stress.

Studies have shown that REV1 may play an oncogenic role in lung and intestinal tumor. Overexpression of REV1 promotes the development of carcinogen-induced intestinal adenomas *via* accumulation of point mutation and suppression of apoptosis proportionally to the REV1 expression level (Sasatani et al., 2017). And an animal experiment showed that in 27% of the carcinogen-exposed mice, REV1 inhibition completely abolished tumor formation (Dumstorf et al., 2009). We assessed REV1 expression using tumor tissue data and normal tissue data from TCGA and TIMER databases. We found that REV1 was overexpressed in six cancer types, namely, CHOL, ESCA, HNSC, LIHC, LUSC, and STAD. In contrast, low REV1 expression was observed in only eight cancer types, namely, BLCA, BRCA, KICH, KIRC, KIRP, PRAD, THCA, and UCEC. It can be seen that REV1 has different effects on the occurrence and development of tumors in different tumor types.

In our studies, we also focused on the role of REV1 in various cancers as a prognostic biomarker and the potential function of REV1 regulating the sensitivity of tumor cells to specific drugs. These findings may provide a certain reference for the development of REV1-targeted therapeutics.

REV1 as a potential prognostic biomarker, our research found REV1 gene alterations are related to PFS prognosis in pan-cancer samples while nsSNP F427L is predicted to be deleterious SNP. That means different REV1 SNPs play different roles in different tumor types, drug susceptibility, and related biological events. According to current literature, Yeom et al. (2016) divided the 12 germline variants of REV1 into three types, the "wild-type-like" variants (K388N, S485L, and G498E), the hypoactive variants (F427L, R434Q, M656V, D700N, R704Q, and P831L) and the hyperactive variants (N373S, M407L, and N497S). These mutations may lead to different mutant phenotype and susceptibility to certain chemical and viral carcinogens in affected individuals (Yeom et al., 2016). REV1 N373S SNP was associated with an increased risk of cervical cancer, while F257S SNP is associated with a reduced risk of cervical cancer and an increased risk of lung cancer in people with heavy smoking (Sakiyama et al., 2005; He et al., 2008; Dumstorf et al., 2009; Xu et al., 2013). For REV1 SNP in non-coding region, rs6761390, rs3792142, and rs3792136 have been reported in literature. Specifically, rs6761390 locates at promoter, a putative transcription factor binding site;

rs3792142 is a tag SNP located at intron 5 and rs3792136 was located in the intron region (Xu et al., 2013). Minor allele carriers of two REV1 SNPs (rs6761390 and rs3792142) had significantly more often large tumours and tumours with high histological grade and stage than the common homozygotes (Varadi et al., 2011). The heterozygote of REV1 rs3087386 (F257S) and rs3792136 were independent prognostic factors for lung cancer survival with hazard radio (HR) 1.54 (95% CI: 1.12–2.12) and 1.44 (95% CI: 1.06–1.97) respectively (Xu et al., 2013).

As for the correlation between REV1 expression and prognosis, our study suggests that low REV1 expression is associated with better prognosis in colorectal (DSS, DFS), gastric (OS, PPS) and ovarian (OS, PPS) cancer while high REV1 expression is associated with better prognosis in lung (OS, RFS, FP, PPS) and breast (DSS, RFS) cancer. These indicated that the functions of REV1 are different or even opposite in different tumors.

REV1 as a potential predictor of drug sensitivity, our study shows that in different tumor types, the expression of REV1 is correlated with the sensitivity of drugs in different mechanisms of pathways. In colon adenocarcinoma and rectum adenocarcinoma and lung adenocarcinoma, low expression of REV1 may suggest resistance to drugs in certain pathways. Conversely, high expression of REV1 in acute myeloid leukemia, brain lower grade glioma, small cell lung cancer and thyroid carcinoma may indicate resistance to drugs in certain pathways.

Currently, there is some evidence that inhibition of TLS polymerase including REV1 not only sensitizes tumor cells to chemotherapeutic drugs, but also reduces the acquisition of drug-induced mutations associated with tumor resistance (Xie et al., 2010). Therefore, TLS inhibition may have dual anticancer effects, and inhibition of TLS polymerase is a promising new approach to improve cancer therapy (Yamanaka et al., 2017).

In recent years, many scientists devoted to develop the small molecule inhibitors targeting REV1 and achieved certain results. In 2017, Sail et al. (2017) identified the first small molecules that exhibit anti-TLS activity in human cancer cells through disruption of the protein-protein interactions between the C-terminal domain of REV1 and the REV1-interacting region. In 2018, Vanarotti et al. (2018) identified the first small-molecule compound, MLAF50, that inhibited the interaction of REV1 UBM2 with ubiquitin through directly binding to REV1 UBM2. In 2019, Wojtaszek et al. (2019) discovered a small molecule inhibitor, JH-RE-06, targeting a nearly featureless surface of REV1 that interacts with the REV7 subunit of POL ζ. Binding of JH-RE-06 induces REV1 dimerization, which blocks the REV1-REV7 interaction and POL ζ recruitment. Recently, Pernicone et al. (2022) discovered two small molecules (ZINC97017995 and ZINC25496030) from the ZINC12 subset Library that disrupt the assembly of MAD2L2-Rev1 and the formation of an active TLS complex. The above studies all show that these small molecules combined with cisplatin could enhance the sensitivity of human cancer cells to cisplatin while have minimal toxicity on their own. However, the current researches on these small molecule inhibitors are limited to *in vitro* cell experiments and a small number of mouse subcutaneous tumor drug experiments, and clinical trials have not yet been carried out.

In addition, there are many literatures suggesting that REV1 SNP or high level of the REV1 expression is associated with resistance to certain drugs. Goricar et al. (2014), Goricar et al. (2015) reported that the V138M SNP of REV1 gene is associated with poor response to cisplatin chemotherapy in patients with malignant mesothelioma and osteosarcoma. The study also revealed that REV1 gene SNPs were associated with the hematological toxicity of cisplatin in malignant mesothelioma. The REV1 allele V138M SNP (rs3087403) is associated with an increased risk of leukopenia and neutropenia. In contrast, patients with at least one REV1 allele of the F257S SNP (rs3087386) had a reduced risk of neutropenia compared with patients with two wild-type alleles (Goricar et al., 2014). Other researchers have demonstrated *in vivo* experiments that REV1 plays a key role in the development of acquired cyclophosphamide resistance (Xie et al., 2010). In addition, studies have shown that reducing the expression level of REV1 can promote the sensitivity of cells to cisplatin and PARP inhibition (Wang et al., 2012). And high levels of REV1 expression led to resistance of ovarian cancer cells to cisplatin. This may be related to the fact that REV1 can enhance cell survival and the generation of drug-resistant variants in surviving populations (Okuda et al., 2005; Lin et al., 2006). However, recently, studies by Kanayo E. Ikeh et al. (2021) suggested that REV1 inhibition may provide cytoprotection by inducing autophagy during radiation therapy. Thus, REV1 may be an important biomarker in tumor treatment. On the one hand, its decreased level during chemotherapy such as cisplatin and cyclophosphamide will be an indicator of a good response of patients to treatment. On the other hand, patients receiving radiation therapy need to increase REV1 levels.

The results of the TLS process can be error-free or error-prone, depending on the type of DNA damage and the specific polymerase. Therefore, REV1 may be protective or mutagenic for each specific lesion during DNA damage-induced cellular mutagenesis (Yeom et al., 2016). That also explains that the effects of REV1 on the occurrence, development, prognosis and drug sensitivity of tumors vary greatly among different tumor types and different SNPs, which reminds us that REV1, as a novel biomarker and potential therapeutic target, requires specific analysis of specific problems. Our study conducted a comprehensive assessment of REV1, to a certain extent, which may provide some hints in this regard.

5 Conclusion

In conclusion, REV1 plays different roles in different tumor types, drug susceptibility, and related biological events. REV1 gene alterations are related to PFS prognosis and nsSNP F427L is predicted to be deleterious SNP. REV1 expression is significantly correlated with different prognosis in colorectal, ovarian, lung, breast, and gastric cancer. REV1 expression can be used as predictive marker for various drugs of various pathways in different tumors.

Data availability statement

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

Ethics statement

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

Author contributions

NZ and YZ contributed to the study conception and design. Literature searching and data analysis were performed by NZ, YZ, MM, YL, and YT. The first draft of the manuscript was written by NZ and YZ and the paper was revised by XF, SW, and YY critically. All authors commented on previous versions of the manuscript. All authors read and approved the final manuscript.

Funding

This study was supported by the Key R&D Program of Zhejiang Province (2021C03125 to YY), the National Natural Science Foundation of China (81872481 to YY; 82102708 to YT), and Leading Innovative and Entrepreneur Team Introduction Program of Zhejiang (2019R01007 to YY).

Conflict of interest

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

Publisher's note

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

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2022.997970/full#supplementary-material

References

Dumstorf, C. A., Mukhopadhyay, S., Krishnan, E., Haribabu, B., and McGregor, W. G. (2009). REV1 is implicated in the development of carcinogen-induced lung cancer. *Mol. Cancer Res.* 7 (2), 247–254. doi:10.1158/1541-7786.MCR-08-0399

Gan, G. N., Wittschieben, J. P., Wittschieben, B. O., and Wood, R. D. (2008). DNA polymerase zeta (pol zeta) in higher eukaryotes. *Cell Res.* 18 (1), 174–183. doi:10.1038/cr.2007.117

Goricar, K., Kovac, V., and Dolzan, V. (2014). Polymorphisms in translesion polymerase genes influence treatment outcome in malignant mesothelioma. *Pharmacogenomics* 15 (7), 941–950. doi:10.2217/pgs.14.14

Goricar, K., Kovac, V., Jazbec, J., Zakotnik, B., Lamovec, J., and Dolzan, V. (2015). Translesion polymerase genes polymorphisms and haplotypes influence survival of osteosarcoma patients. *Omics a J. Integr. Biol.* 19 (3), 180–185. doi:10.1089/omi.2014.0159

Guo, C., Fischhaber, P. L., Luk-Paszyc, M. J., Masuda, Y., Zhou, J., Kamiya, K., et al. (2003). Mouse Rev1 protein interacts with multiple DNA polymerases involved in translesion DNA synthesis. *EMBO J.* 22 (24), 6621–6630. doi:10.1093/emboj/cdg626

He, X., Ye, F., Zhang, J., Cheng, Q., Shen, J., and Chen, H. (2008). REV1 genetic variants associated with the risk of cervical carcinoma. *Eur. J. Epidemiol.* 23 (6), 403–409. doi:10.1007/s10654-008-9251-5

Ikeh, K. E., Lamkin, E. N., Crompton, A., Deutsch, J., Fisher, K. J., Gray, M., et al. (2021). REV1 inhibition enhances radioresistance and autophagy. *Cancers (Basel)* 13 (21), 5290. doi:10.3390/cancers13215290

Lawrence, C. W. (2002). Cellular roles of DNA polymerase zeta and Rev1 protein. *DNA Repair* 1 (6), 425–435. doi:10.1016/s1568-7864(02)00038-1

Lin, X., Okuda, T., Trang, J., and Howell, S. B. (2006). Human REV1 modulates the cytotoxicity and mutagenicity of cisplatin in human ovarian carcinoma cells. *Mol. Pharmacol.* 69 (5), 1748–1754. doi:10.1124/mol.105.020446

Liu, W., Xie, Y., Ma, J., Luo, X., Nie, P., Zuo, Z., et al. (2015). IBS: an illustrator for the presentation and visualization of biological sequences. *Bioinformatics* 31 (20), 3359–3361. doi:10.1093/bioinformatics/btv362

Murakumo, Y., Ogura, Y., Ishii, H., Numata, S., Ichihara, M., Croce, C. M., et al. (2001). Interactions in the error-prone postreplication repair proteins hREV1, hREV3, and hREV7. *J. Biol. Chem.* 276 (38), 35644–35651. doi:10.1074/jbc.M102051200

Nair, D. T., Johnson, R. E., Prakash, L., Prakash, S., and Aggarwal, A. K. (2005). Rev1 employs a novel mechanism of DNA synthesis using a protein template. *Science* 309 (5744), 2219–2222. doi:10.1126/science.1116336

Nayak, S., Calvo, J. A., Cong, K., Peng, M., Berthiaume, E., Jackson, J., et al. (2020). Inhibition of the translesion synthesis polymerase REV1 exploits replication gaps as a cancer vulnerability. *Sci. Adv.* 6 (24), eaaz7808. doi:10.1126/sciadv.aaz7808

Ohashi, E., Murakumo, Y., Kanjo, N., Akagi, J., Masutani, C., Hanaoka, F., et al. (2004). Interaction of hREV1 with three human Y-family DNA polymerases. *Genes* 9 (6), 523–531. doi:10.1111/j.1356-9597.2004.00747.x

Okuda, T., Lin, X., Trang, J., and Howell, S. B. (2005). Suppression of hREV1 expression reduces the rate at which human ovarian carcinoma cells acquire resistance to cisplatin. *Mol. Pharmacol.* 67 (6), 1852–1860. doi:10.1124/mol.104.010579

Pernicone, N., Elias, M., Onn, I., Tobi, D., and Listovsky, T. (2022). Disrupting the MAD2L2-rev1 complex enhances cell death upon DNA damage. *Mol. (Basel, Switz).* 27 (3), 636. doi:10.3390/molecules27030636

Pozhidaeva, A., Pustovalova, Y., D'Souza, S., Bezsonova, I., Walker, G. C., and Korzhnev, D. M. (2012). NMR structure and dynamics of the C-terminal domain from human Rev1 and its complex with Rev1 interacting region of DNA polymerase η. *Biochemistry* 51 (27), 5506–5520. doi:10.1021/bi300566z

Quinet, A., Vessoni, A. T., Rocha, C. R., Gottifredi, V., Biard, D., Sarasin, A., et al. (2014). Gap-filling and bypass at the replication fork are both active mechanisms for tolerance of low-dose ultraviolet-induced DNA damage in the human genome. *DNA Repair* 14, 27–38. doi:10.1016/j.dnarep.2013.12.005

Sail, V., Rizzo, A. A., Chatterjee, N., Dash, R. C., Ozen, Z., Walker, G. C., et al. (2017). Identification of small molecule translesion synthesis inhibitors that target the rev1-CT/RIR protein-protein interaction. *ACS Chem. Biol.* 12 (7), 1903–1912. doi:10.1021/acschembio.6b01144

Sakiyama, T., Kohno, T., Mimaki, S., Ohta, T., Yanagitani, N., Sobue, T., et al. (2005). Association of amino acid substitution polymorphisms in DNA repair genes TP53, POLI, REV1 and LIG4 with lung cancer risk. *Int. J. Cancer* 114 (5), 730–737. doi:10.1002/ijc.20790

Sale, J. E. (2013). Translesion DNA synthesis and mutagenesis in eukaryotes. *Cold Spring Harb. Perspect. Biol.* 5 (3), a012708. doi:10.1101/cshperspect.a012708

Sasatani, M., Xi, Y., Kajimura, J., Kawamura, T., Piao, J., Masuda, Y., et al. (2017). Overexpression of Rev1 promotes the development of carcinogen-induced intestinal adenomas via accumulation of point mutation and suppression of apoptosis proportionally to the Rev1 expression level. *Carcinogenesis* 38 (5), 570–578. doi:10.1093/carcin/bgw208

Schoket, B. (2004). The role of DNA adducts in smoking-related carcinogenesis. *Magy. Onkol.* 48 (3), 201–205.

Swan, M. K., Johnson, R. E., Prakash, L., Prakash, S., and Aggarwal, A. K. (2009). Structure of the human Rev1-DNA-dNTP ternary complex. *J. Mol. Biol.* 390 (4), 699–709. doi:10.1016/j.jmb.2009.05.026

Tirman, S., Quinet, A., Wood, M., Meroni, A., Cybulla, E., Jackson, J., et al. (2021). Temporally distinct post-replicative repair mechanisms fill PRIMPOL-dependent ssDNA gaps in human cells. *Mol. Cell* 81 (19), 4026–4040. doi:10.1016/j.molcel.2021.09.013

Tissier, A., Kannouche, P., Reck, M. P., Lehmann, A. R., Fuchs, R. P., and Cordonnier, A. (2004). Co-localization in replication foci and interaction of human Y-family members, DNA polymerase pol eta and REVl protein. *DNA Repair* 3 (11), 1503–1514. doi:10.1016/j.dnarep.2004.06.015

Vanarotti, M., Grace, C. R., Miller, D. J., Actis, M. L., Inoue, A., Evison, B. J., et al. (2018). Structures of REV1 UBM2 domain complex with ubiquitin and with a small-molecule that inhibits the REV1 UBM2-ubiquitin interaction. *J. Mol. Biol.* 430 (17), 2857–2872. doi:10.1016/j.jmb.2018.05.042

Varadi, V., Bevier, M., Grzybowska, E., Johansson, R., Enquist, K., Henriksson, R., et al. (2011). Genetic variation in genes encoding for polymerase zeta subunits associates with breast cancer risk, tumour characteristics and survival. *Breast Cancer Res. Treat.* 129 (1), 235–245. doi:10.1007/s10549-011-1460-z

Venselaar, H., Te Beek, T. A., Kuipers, R. K., Hekkelman, M. L., and Vriend, G. (2010). Protein structure analysis of mutations causing inheritable diseases. An e-Science approach with life scientist friendly interfaces. *BMC Bioinforma.* 11, 548. doi:10.1186/1471-2105-11-548

Wang, Y., Huang, J. W., Calses, P., Kemp, C. J., and Taniguchi, T. (2012). MiR-96 downregulates REV1 and RAD51 to promote cellular sensitivity to cisplatin and PARP inhibition. *Cancer Res.* 72 (16), 4037–4046. doi:10.1158/0008-5472.CAN-12-0103

Washington, M. T., Carlson, K. D., Freudenthal, B. D., and Pryor, J. M. (2010). Variations on a theme: eukaryotic Y-family DNA polymerases. *Biochim. Biophys. Acta* 1804 (5), 1113–1123. doi:10.1016/j.bbapap.2009.07.004

Waters, L. S., Minesinger, B. K., Wiltrout, M. E., D'Souza, S., Woodruff, R. V., and Walker, G. C. (2009). Eukaryotic translesion polymerases and their roles and regulation in DNA damage tolerance. *Microbiol. Mol. Biol. Rev.* 73 (1), 134–154. doi:10.1128/MMBR.00034-08

Wojtaszek, J., Liu, J., D'Souza, S., Wang, S., Xue, Y., Walker, G. C., et al. (2012). Multifaceted recognition of vertebrate Rev1 by translesion polymerases ζ and κ. *J. Biol. Chem.* 287 (31), 26400–26408. doi:10.1074/jbc.M112.380998

Wojtaszek, J. L., Chatterjee, N., Najeeb, J., Ramos, A., Lee, M., Bian, K., et al. (2019). A small molecule targeting mutagenic translesion synthesis improves chemotherapy. *Cell* 178 (1), 152–159. doi:10.1016/j.cell.2019.05.028

Xie, K., Doles, J., Hemann, M. T., and Walker, G. C. (2010). Error-prone translesion synthesis mediates acquired chemoresistance. *Proc. Natl. Acad. Sci. U. S. A.* 107 (48), 20792–20797. doi:10.1073/pnas.1011412107

Xu, H. L., Gao, X. R., Zhang, W., Cheng, J. R., Tan, Y. T., Zheng, W., et al. (2013). Effects of polymorphisms in translesion DNA synthesis genes on lung cancer risk and prognosis in Chinese men. *Cancer Epidemiol.* 37 (6), 917–922. doi:10.1016/j.canep.2013.08.003

Yamanaka, K., Chatterjee, N., Hemann, M. T., and Walker, G. C. (2017). Inhibition of mutagenic translesion synthesis: A possible strategy for improving chemotherapy? *PLoS Genet.* 13 (8), e1006842. doi:10.1371/journal.pgen.1006842

Yang, Y., Liu, Z., Wang, F., Temviriyanukul, P., Ma, X., Tu, Y., et al. (2015). FANCD2 and REV1 cooperate in the protection of nascent DNA strands in response to replication stress. *Nucleic Acids Res.* 43 (17), 8325–8339. doi:10.1093/nar/gkv737

Yeom, M., Kim, I. H., Kim, J. K., Kang, K., Eoff, R. L., Guengerich, F. P., et al. (2016). Effects of twelve germline missense variations on DNA lesion and G-quadruplex bypass activities of human DNA polymerase REV1. *Chem. Res. Toxicol.* 29 (3), 367–379. doi:10.1021/acs.chemrestox.5b00513

Zeman, M. K., and Cimprich, K. A. (2014). Causes and consequences of replication stress. *Nat. Cell Biol.* 16 (1), 2–9. doi:10.1038/ncb2897

frontiers | Frontiers in Genetics

TYPE Original Research
PUBLISHED 14 October 2022
DOI 10.3389/fgene.2022.1032779

Drug-protein interaction prediction *via* variational autoencoders and attention mechanisms

Yue Zhang*, Yuqing Hu, Huihui Li and Xiaoyong Liu

School of Computer Science, Guangdong Polytechnic Normal University, Guangzhou, China

OPEN ACCESS

EDITED BY
Leyi Wei,
Shandong University, China

REVIEWED BY
Jiazhou Chen,
South China University of Technology, China
Yushan Qiu,
Shenzhen University, China

*CORRESPONDENCE
Yue Zhang,
zhangyue@gpnu.edu.cn

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

RECEIVED 31 August 2022
ACCEPTED 30 September 2022
PUBLISHED 14 October 2022

CITATION
Zhang Y, Hu Y, Li H and Liu X (2022), Drug-protein interaction prediction via variational autoencoders and attention mechanisms.
Front. Genet. 13:1032779.
doi: 10.3389/fgene.2022.1032779

During the process of drug discovery, exploring drug-protein interactions (DPIs) is a key step. With the rapid development of biological data, computer-aided methods are much faster than biological experiments. Deep learning methods have become popular and are mainly used to extract the characteristics of drugs and proteins for further DPIs prediction. Since the prediction of DPIs through machine learning cannot fully extract effective features, in our work, we propose a deep learning framework that uses variational autoencoders and attention mechanisms; it utilizes convolutional neural networks (CNNs) to obtain local features and attention mechanisms to obtain important information about drugs and proteins, which is very important for predicting DPIs. Compared with some machine learning methods on the C.elegans and human datasets, our approach provides a better effect. On the BindingDB dataset, its accuracy (ACC) and area under the curve (AUC) reach 0.862 and 0.913, respectively. To verify the robustness of the model, multiclass classification tasks are performed on Davis and KIBA datasets, and the ACC values reach 0.850 and 0.841, respectively, thus further demonstrating the effectiveness of the model.

KEYWORDS

drug-protein interactions (DPIs), variational autoencoder (VAE), attention mechanism, convolutional neural network (CNN), deep learning - artificial neural network

Introduction

Finding gene-drug relationships is important not only for understanding a certain mechanism of drug molecules, but also for developing treatments for patients. The gene-drug relationship is many-to-many, which is much more complex than a gene-to-drug or a drug-to-gene, and also explains the complex relationship between gene-drug. The gene-drug relationship has similarities to the drug-protein relationship (Chen et al., 2019; Huang et al., 2021).

In the prediction of RNA-binding proteins, limited by the huge cost of biological experiments, it is difficult to fully understand the underlying mechanisms of alternative splicing (AS) and related RNA-binding proteins (RBPS) in regulating the epithelial-mesenchymal transition (EMT) process. This needs to be achieved by means of computational methods (Qiu et al., 2021b) proposed an inductive matrix-based model to study the relationship between RBP and AS during EMT. The main purpose of the model is to compensate for missing and unknown RBP-AS relationships (Qiu et al., 2021a) proposed a method based on weighted data fusion with sparse matrix tri-factorization to conduct experiments. The AS-RBP relationship is explored by assigning different weights to the source data. Both methods achieve good results. At the same time, this has parallels with the drug-protein relationship. It achieves the desired effect by looking for a drug to inhibit an binding site of a protein.

Drug-protein interactions (DPIs) exploration is a critical step in the drug discovery process. With the discovery of new drugs, the field of drug development continues to expand, and awareness regarding the repositioning of existing drugs and new interactions involving approved drugs is of increasing concern (Oprea and Mestres, 2012). Based on biological experiments, it usually takes 10–20 years and much money (US$ 0.5–260 million) to develop a new drug (Avorn, 2015), so it is important to explore the interactions between drugs and proteins. In recent years, computer-aided methods have achieved good results and contributed significantly to the prediction of DPIs. The application of artificial intelligence in chemical research can accelerate the development of high-precision DPIs prediction methods.

In the past decade, the problem of predicting the interactions between drugs and proteins has been solved using traditional machine learning methods, which solve binary classification problems (Yamanishi et al., 2010; Liu et al., 2016; Nascimento et al., 2016; Keum and Nam, 2017). Due to the rise and popularity of deep learning, it has become a popular choice for solving DPIs predictions (Unterthiner and Mayr, 2014; Tian et al., 2016) used a deep neural network (DNN) to explore the interactions between drugs and proteins instead of traditional machine learning methods, which directed the subsequent research on drug and protein interactions toward deep learning approaches, such as convolutional neural networks (CNNs), recurrent neural networks (RNNs) (Gao et al., 2018; Mayr et al., 2018) and stacked autoencoders (Wang et al., 2018).

In general, DPIs approaches can be divided into three categories: docking-based methods, machine learning-based methods, and deep learning-based methods. Docking-based methods require the best site and protein structure to be found and combined, but such a technique usually time-consuming, and many datasets lack three-dimensional protein structures (Gschwend et al., 1996). Machine learning-based methods (Faulon et al., 2008; Bleakley and Yamanishi, 2009; Ballester and Mitchell, 2010) usually require manual features, and the features passed to the model before modeling occurs require manual participation, which demands considerable feature extraction experience and expertise. Deep learning-based methods have been applied to many fields in biology (Min et al., 2016; Zeng et al., 2019; Zhang et al., 2019, 2021; Wu et al., 2022b); DPIs prediction performance has been improved through the framework structure and network parameters of deep learning. For example, the DeepDTA approach of (Öztürk et al., 2018) learns internal high-level features by extracting the features of drugs and proteins as the network inputs and then predicts the relationships between drugs and proteins. The WideDTA method proposed by (Öztürk et al., 2019) is similar to DeepDTA, and the network framework is roughly unchanged; the main difference is that when inputting features, WideDTA extracts the features of drug proteins from multiple aspects as model inputs. Notably, a graph-based network architecture called GraphDTA (Nguyen et al., 2019), which treats drugs as a graph structure to predict DPIs, has also been developed. A Novel Graph Neural Network for Predicting Drug-Protein Interactions called BridgeDTA (Wu et al., 2022a), which introduces a class of nodes named hyper-nodes, which bridge different proteins/drugs to work as the protein-protein and drug-drug associations. HOGMMNC is a higher order graph matching with multiple network constraints model. It mainly obtains the fixed structural relationship in multi-source data through hypergraph matching, so as to identify the relationship between genes and drugs, and improve the accuracy and reliability of the identification relationship (Chen et al., 2019). These deep learning methods all have three similarities. *1*) They encode drugs and proteins. *2*) They extract the high-level features of drugs and proteins through their network structures. *3*) They predict the features obtained in *2*) through a fully connected (FC) layer. The advantage of these methods is that the process is not too cumbersome (it is simple). Furthermore, we exploit the strengths of these network frameworks for the prediction of DPIs.

A variational autoencoder (VAE) is a machine learning model that can reconstruct a variable x based on a latent feature Z. Unlike a simple autoencoder, it can learn the distribution of latent variables and then sample from this distribution to generate new samples. The model has been shaped and used in various fields, such as image processing (Walker et al., 2017; Liu et al., 2018) and text processing (Mayr et al., 2018). We use this model to predict DPIs. Experimental results show that better results are achieved on some datasets. Our contributions are as follows.

1) A variational autoencoder is designed to provide a probabilistic way of describing the latent representation of drugs and proteins, denoted *via* mean and variance of the hidden state distribution. Such generative way effectively reduce the redundant information in the raw samples to ease for leaning drug-protein interactions.
2) Discriminative local features on drugs and proteins are extracted *via* deep CNN. A specially designed attention mechanism is incorporated to focus on the key interactive information on both drugs and proteins, thus obtaining strong drug-to-drug and protein-to-protein relationships.

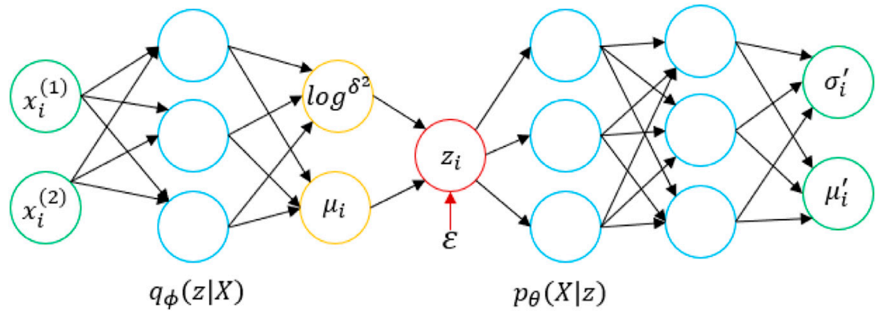


FIGURE 1
Structural diagram of the utilized VAE.

3) Extensive experiments on C.elegans and Human dataset, BindingDB dataset, Davis dataset and KIBA dataset. Datasets demonstrate that the proposed method can robustly identify the drug-protein interactions.

Methods

A VAE network to identify drug and protein interactions

We input a set of drug molecules D and a protein sequence T, and a VAE (Kingma and Welling, 2014) learns the distribution of a multidimensional variable x based on an independent and identically distributed latent variable ($X = \{x^i\}_1^N$, where N is the number of samples). The framework of this strategy is shown in Figure 1, where $x_i^{(j)}$represents the jth feature of the ith sample.

First, a data point x_i is input into the encoder. Through the neural network, we obtain the parameters of the approximate posterior distribution $q_\varnothing(z|x_i)$ obeyed by the hidden variable z. The posterior distribution is a Gaussian distribution, and the output of the encoder includes the parameters σ_i^2 and μ_i of the Gaussian distribution that $z|x_i$ obeys. With the parameters σ_i^2 and μ_i of the $z|x_i$ distribution, we sample one ϵ_i from $\mathcal{N}(0, I)$ and use the reparameterization trick to set $z_i = \mu_i + \sigma_i \square \epsilon_i$ where z_i represents the value of a similar sample x_i and □ represents the elementwise multiplication operation. The decoder needs to fit the likelihood distribution $p_\varnothing(X|z_i)$ and feed a z_i to the decoder, which returns the parameters of the distribution that $X|z_i$ obeys; the likelihood also obeys a Gaussian distribution. After obtaining the parameters of the distribution of $|z_i$, we sample from the distribution to generate a sample x_i. In the last step, we do not sample and directly regard the μ_i' output by the model as the sample x_i generated by z_i. Then, we can obtain the objective function of the VAE, and we only need to maximize $\mathcal{L}$.

$$\mathcal{L}(p_\theta, q_\varnothing) = -D_{KL}(q_\varnothing, p) + \mathbb{E}_{q_\varnothing}[log p_\theta(X|z)]$$

$-D_{KL}(q_\varnothing, p)$is the Kullback-Leibler (KL) divergence of the two distributions$p$ and q and is also a regular term. $\mathbb{E}_{q_\varnothing}[log p_\theta(X|z)]$ is often referred to as the reconstruction loss.

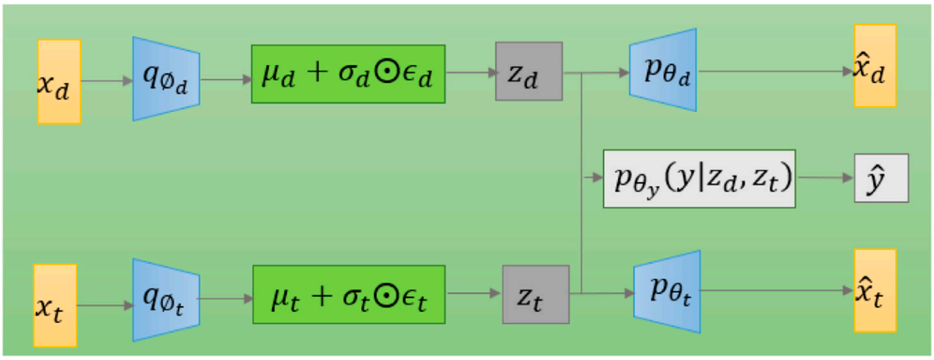


FIGURE 2
The graphical structure of the VAE model for DPIs prediction.

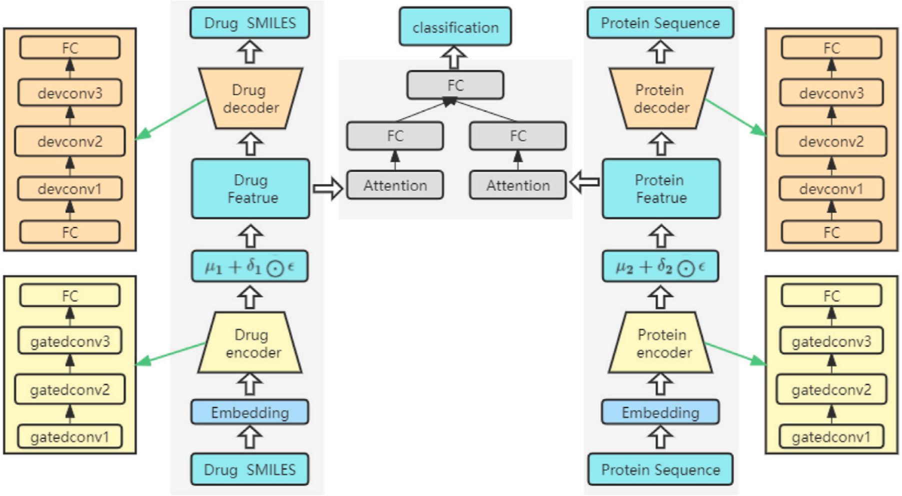


FIGURE 3
The framework structure of the model.

With the theoretical support of the VAE, we apply it to the prediction of DPIs. Here, x_d and x_t represent the drug and protein vectors, respectively, and y represents the interaction relationship or affinity value between drug and protein. We assume that the hidden variables of x_d and x_t are z_d and z_t, respectively. Our aim is to learn a model that predicts drug and protein interactions.

The model diagram for applying variational autoencoding to DPIs prediction is shown in Figure 2. The model has two encoders, which are mainly used to generate latent variables z_d and z_t for drug x_d and protein x_t, respectively. Three important decoders are used to generate x_d and x_t from the latent variables z_d and z_t and generate the drug-protein relationship y. Finally, for each drug-protein pair, the objective function of the VAE model is to maximize $\mathcal{L}'$.

$$\mathcal{L}' = \mathcal{L}_{DrugVAE} + \mathcal{L}_{PorteinVAE}$$
$$= \left\{-D_{KL}\left(q_{\varnothing_d}, p_d\right) + \mathbb{E}_{q_{\varnothing_d}}\left[logp_{\theta_d}\left(X_d|z_d\right)\right]\right\}$$
$$+\left\{-D_{KL}\left(q_{\varnothing_t}, p_t\right) + \mathbb{E}_{q_{\varnothing_t}}\left[logp_{\theta_t}\left(X_t|z_t\right)\right]\right\}$$

Attention mechanism for feature extraction

Attention mechanisms, as effective means of feature screening and enhancement, have been widely used in many fields of deep learning. A structural model based on an attention mechanism can not only record the positional relationships between pieces of information but also measure the importance levels of different information features according to the weight of the information. Dynamic weight parameters are established by making relevant and irrelevant choices for the information features to strengthen the key information and weaken the useless information, thereby improving the efficiency of deep learning algorithms and improving some of the defects of traditional deep learning techniques.

Utilizing an attention mechanism for the prediction of DPIs can enable effective atomic feature extraction because the structures of the molecular sequences of drugs and proteins are very similar to the structures of natural language sentences, and the context information of atoms is very important for understanding molecular features (Jastrzębski et al., 2018). In detail, we should pay attention to the interaction information of each atom and its adjacent atoms; each atom is also connected to the simplified molecular-input line-entry system (SMILES (Weininger, 1988), which is a symbol for molecular structure encoding). Information about the interactions of atoms that are farther away in the sequence can also have an impact on the predicted results. The molecular sequences of proteins are very long, and the best way to extract features is to use an attention mechanism.

Attention mechanisms are widely used in the natural language processing (NLP) field, and they have also been shown to be powerful for processing textual data. The core of such a mechanism is an attention function (Vaswani et al., 2017). The attention function can be described as mapping a query (Q) and a set of key-value (K-V) pairs to an output. Among them, dot product attention with Q, K and V is a widely used attention approach. Let the dimensions of Q and K be d_k and the dimensionality of V be d_v. Then, attention can be expressed by the following formula.

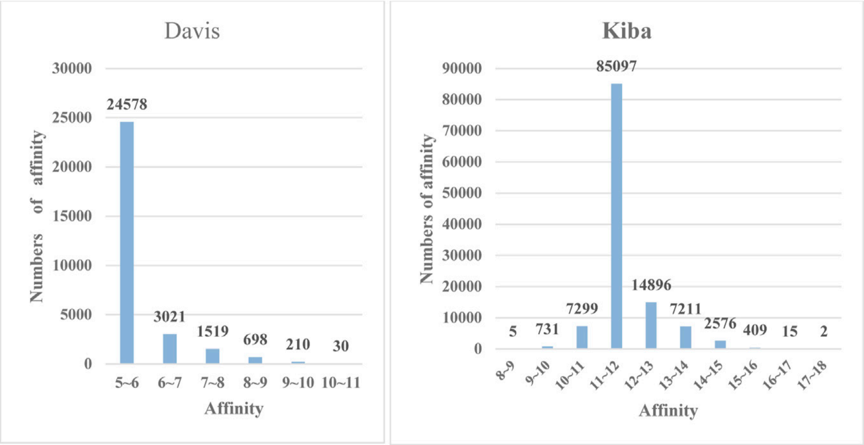


FIGURE 4
The frequency histograms of the affinities in the Davis and KIBA datasets. The horizontal axis denotes the affinity values of drugs and proteins, and the vertical axis represents the numbers of affinity values in certain intervals.

TABLE 1 Model parameters.

Parameter	Value
Number of filters in the encoder and decoder	32
Filter length (drug molecules)	5
Filter length (protein sequences)	7
Number of epochs	300
Batch size	256
Learning rate	0.001, 0.0001

$$Attenton(Q,K,V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V$$

where $Q \in \mathbb{R}^{n\times d_k}, K \in \mathbb{R}^{m\times d_k}$ and $V \in \mathbb{R}^{m\times d_v}$. The attention measures the similarity between the inner product of the matrixQand V, which is nonlinearly transformed by a softmax function with the matrix V. Here, d_k prevents the forward propagation of the network due to excessive data content.

The network structure of the model

The network structure of the model is shown in Figure 3. It consists of three key parts: an encoder, a decoder and a prediction module. Both the encoder and the decoder serve to predict the interactions between drugs and proteins. Among them, the overall structure extracts the features of drugs and proteins, sends the extracted features into the attention block to focus on the important parts, and finally sends them to the FC layer to predict the DPIs. The feature extraction process for drugs is the same as that for proteins. Before being fed into the encoder, drugs and proteins are sequences of text strings, which need to be converted into digital vectors. According to the existing character dictionary, each character is converted into an integer type, and then each sample is converted into an embedding matrix through embedding. In our model, three GatedCNNs are included in the coding layer, and a rectified linear unit (RELU) (Nair and Hinton, 2010) activation function is present after each layer of CNNs. The filters of the last two CNNs are the first CNNs, which are filtered two times and three times. A max pooling layer is appended after the third GatedCNN to compress the extracted features.

In the decoders for drugs and proteins, the input source data are reconstructed through deconvolutional networks (Zeiler et al., 2010). Each decoder has an FC layer and three deconvolution layers. The last deconvolution appends an FC layer to convert the output into drug and target sequences with the same size as that of the input.

In the DPIs prediction module, two FC layers are used to represent the features of drugs and proteins. To further extract the high-level features of drugs and proteins, a self-attention mechanism is introduced after the FC layers, focusing on important features in drug sequences or protein sequences and ignoring unnecessary features. Then, the final extracted high-level features are spliced and sent to a network containing three FC layers. A ReLU activation function and a dropout function are placed after the first two FC layers, and the dropout function is mainly used to prevent the network from overfitting. The final output can be used to predict DPIs.

The model parameters used in this experiment are shown in Table 1. Among them, we select several values [16,32,64] for the number of CNN filters in the encoder and decoder and find that

TABLE 2 AUC, precision, recall, and F1 values obtained under different methods.

Models	AUC	Precision	Recall	F1
BindingDB dataset				
K-NN	0.776	0.762	0.791	0.776
RF	0.742	0.834	0.600	0.698
L2	0.737	0.784	0.646	0.709
SVM	0.805	0.770	0.858	0.811
BridgeDPI	**0.960**	0.883	**0.903**	**0.893**
Ours	0.913	**0.888**	0.822	0.854
C.elegans dataset				
K-NN	0.858	0.801	0.827	0.814
RF	0.902	0.821	0.844	0.832
L2	0.892	0.890	0.877	0.883
SVM	0.894	0.785	0.818	0.801
BridgeDPI	**0.996**	**0.980**	**0.970**	**0.975**
Ours	0.925	0.927	0.897	0.912
Human dataset				
K-NN	0.860	0.798	0.927	0.858
RF	0.940	0.861	0.897	0.879
L2	0.911	0.891	0.913	0.902
SVM	0.910	**0.966**	0.950	0.958
BridgeDPI	**0.990**	0.962	**0.965**	**0.963**
Ours	0.914	0.934	0.862	0.897

The best result was highlighted in bold, while the suboptimal is denoted by underline.

the effect of 32 filters was best and that the filter lengths of drugs and proteins are both in [5,7,9,11]. We choose the best results, and the final filter lengths of drugs and proteins are 5 and 7, respectively.

Experiment

Datasets

To verify the effectiveness of the proposed model and compare it with the base method, we conducted experiments on the following datasets: C.elegans and Human datasets, BindingDB dataset, Davis dataset and KIBA dataset.

C.elegans and human datasets

In the work of (Liu et al., 2015), the authors used a systematic scanning framework, and their dataset contained a large number of negative samples. They constructed two datasets, C.elegans and human. Following the requirements of (Tsubaki et al., 2019), we used a balanced dataset with an approximately 1:1 ratio of positive and negative samples. The C.elegans dataset includes 1876 protein targets and 1767 drug molecules, and it contains 7786 affinity sample pairs, 3893 positive samples, and 3893 negative samples. The human dataset contains 6728 affinity pairs, the number of protein targets is 2001, and the number of drug molecules is 2726.

BindingDB dataset

BindingDB is a public, web-accessible database of measured binding affinities that focuses chiefly on the interactions of proteins considered to be drug targets with small, drug-like molecules. In this experiment, the method described in the paper of (Gao et al., 2018) was used; the dataset contains 39,747 positive samples and 31,218 negative samples.

Davis

The Davis dataset contains affinity pairs measured by their K_d value (kinase dissociation constant); it includes 68 drug molecules and 442 proteins (Davis et al., 2011). The K_d value can reflect the affinity between a drug and a protein. It is a bridge for predicting affinity, and its affinity value range is [0.016, 10000]. The frequency distribution plot of affinity values for the Davis dataset is shown in Figure 4.

KIBA

Measuring the relationships between the drugs and proteins in the KIBA dataset is mainly achieved through the KIBA score (Tang et al., 2014) combined the biological activities of different sources of kinase inhibitors, such as K_d,K_i and $IC50$, to calculate KIBA affinity scores. We conducted experiments according to this environment. Then, we found that the dataset contains 52498 drug molecules, 476 protein sequences, and 246088 KIBA scores and that the numerical range of the KIBA score is [0, 17.2]. Values in this dataset with KIBA scores that were less than 10 were removed, leaving 2111 drug molecules and 229 proteins in the end. The frequency distribution plot of the affinity values for the KIBA dataset is shown in Figure 4.

Training details

The model was implemented based on Python 3.6 and PyTorch 1.10.2. The program ran on a GTX1060 GPU with 8 GB of memory. The network parameter initialization process was implemented by the *xavier_normal_()* function in the library. During training, the network used the Adam optimizer (Kingma and Ba, 2014) with a learning rate of 0.0001 for the Davis and KIBA datasets and a learning rate of 0.001 for the other datasets to adjust the network parameters. To prevent overfitting, L2 regularization was added to the loss function. Each batch contained 256 samples, and the samples were randomly scrambled. Three hundred epochs were executed. Finally, the model was trained by minimizing the cross-entropy loss function.

$$L(y, \hat{y}) = -(ylog\hat{y} + (1 - y)\log(1 - \hat{y})) + \lambda \sum \|\Theta\|^2$$

where y is the true value, $\hat{y}$is the predicted value, $\lambda$is the regularization coefficient, and Θ is the network parameter. On the basis of the parameters in Table 1, we search the optimal value of λ by grid searching scheme within range of [-5,+5]. Empirical experiments show that setting the value ofλ being -3 yield the best performances.

Results

First, we conducted experiments on the BindingDB dataset extracted by Gao et al. According to the environment they set for the dataset, we utilized the same division to ensure that the data in the validation set did not appear in the training set, so that the experiment was closer to the real-world situation. During the training process, to prevent overfitting, we set the termination criterion according to the ACC evaluation index of the validation set. When the ACC of the validation set iterated for a certain number of steps and did not increase, the program terminated. To demonstrate the superiority of the model, we made a comparison with k-nearest neighbors (K-NN), a random forest (RF), L2, a support vector machine (SVM) and the BridgeDPI model (Wu et al., 2022a). The machine learning results for these methods were derived from the source paper on C.elegans and Human dataset (Tsubaki et al., 2019). We conducted experiments on BindingDB dataset. The table shows that on the BindingDB dataset, the AUC, Precision, Recall, and F1 of the proposed model reached 0.913, 0.888, 0.822, and 0.854, respectively. Our model outperforms traditional machine learning methods in AUC, Precision, and F1. The unsupervised K-NN method yielded lower results than the other models and methods, with AUCs and F1 scores of 0.858/0.814 and 0.860/0.858 on the C.elegans and human datasets, respectively. The effects of the RF, L2, and the SVM based on supervised learning were better. The AUC on the C.elegans dataset reached approximately 0.9, and the AUC on the human data exceeded 0.9. Compared with traditional machine learning methods, our model achieved the highest evaluation indicators on the C.elegans dataset, and its AUC, precision, recall, and F1 were 2.3%, 3.7%, 2.0%, and 2.9% higher than those of the second-best approaches, respectively. At the same time, our method performed slightly better on the human dataset. Since models such as K-NN, the RF, L2, and the SVM cannot obtain high-quality feature information, it is not easy for them to learn complex nonlinear DPIs. However, deep learning has strong feature extraction capabilities. Our model benefits from that. The Table 2 shows that our model doesn't perform as well as BridgeDPI that extracts features from the biological perspective. Our model is similar to nature language processing in extracting featrues, and it's indeed not as effective as BridgeDPI. However, our model has some advantages: when dealing with drug features and protein features, an attention mechanism is introduced to realize the key sites of drug-protein binding, thereby ignoring irrelevant site information and saving the screening time of drug-protein interactions. This has contributed to experts to identify drug-protein interactions.

TABLE 3 Multi-classification results obtained on the Davis and KIBA datasets.

	ACC		AUC	
Models	Davis	KIBA	Davis	KIBA
K-NN	0.854	0.777	0.581	0.574
RF	**0.868**	0.811	0.610	0.591
L2	0.752	0.699	0.582	0.556
SVM	0.862	0.801	0.541	0.543
Ours	0.850	**0.841**	**0.705**	**0.813**

The best result was highlighted in bold, while the suboptimal is denoted by underline.

To further demonstrate the feature extraction advantages of deep models, we performed a multiclass prediction experiment on the Davis and KIBA datasets, and the results are shown in Table 3. The Davis and KIBA datasets possess continuous values, and the pK_d values of Davis and the scores of KIBA are distributed as shown in Figure 4. We found their meansμ and variances σ according to the relationship between the mean and variance. We divided the data into five categories: $[\mu - \sigma, \mu + \sigma]$, $[\mu - 2\sigma, \mu + 2\sigma]$, $[\mu - 3\sigma, \mu + 3\sigma]$, $[\mu - 4\sigma, \mu + 4\sigma]$, $and\, other$. According to our model, the classification effect of the test was relatively objective, and the ACC and AUC reached 0.850/0.705 and 0.841/0.813 on the Davis and KIBA datasets, respectively. AUC is the best on both Davis and KIBA datasets. ACC is lower on Davis dataset due to uneven data distribution and less data, which may affect our results. The results show that the proposed model is robust.

Conclusion

In this work, we propose a model based on VAEs and attention mechanisms to predict DPIs. The high-level features of drugs and proteins are further extracted by a CNN and an attention mechanism. Experiments show that our method outperforms some base methods on the testing datasets and illustrates the powerful ability of deep learning to extract features. To further verify the robustness of the model, we perform a multiclass prediction experiment on the Davis and KIBA datasets. The final results of the experiment yield good metric values.

Data availability statement

Publicly available datasets were analyzed in this study. This data can be found here: https://www.bindingdb.org/bind/index.jsp, http://staff.cs.utu.fi/~aatapa/data/DrugTarget/, https://wormbase.org//species/c_elegans#104–10.

Author contributions

YZ: Ideas; Formulation of overarching research goals and aims; Development or design of methodology; Creation of models; Data analysis; Review and revision of the first draft. YH: Implementation of the computer code; Original Draft; Preparation of experimental data. HL: Formal analysis; Preparation of experimental data; Review and revision of the first draft. XL: Experimental Supervision and Leadership; Review and revision of the first draft.

Funding

This work was partly supported by National Natural Science Foundation of China (62172112, 62172113).

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

Avorn, J. (2015). The $2.6 billion pill — methodologic and policy considerations. *N. Engl. J. Med.* 372, 1877–1879. doi:10.1056/NEJMp1500848

Ballester, P. J., and Mitchell, J. B. O. (2010). A machine learning approach to predicting protein–ligand binding affinity with applications to molecular docking. *Bioinformatics* 26, 1169–1175. doi:10.1093/bioinformatics/btq112

Bleakley, K., and Yamanishi, Y. (2009). Supervised prediction of drug–target interactions using bipartite local models. *Bioinformatics* 25, 2397–2403. doi:10.1093/bioinformatics/btp433

Chen, J., Peng, H., Han, G., Cai, H., and Cai, J. (2019). Hogmmnc: A higher order graph matching with multiple network constraints model for gene–drug regulatory modules identification. *Bioinformatics* 35, 602–610. doi:10.1093/bioinformatics/bty662

Davis, M. I., Hunt, J. P., Herrgard, S., Ciceri, P., Wodicka, L. M., Pallares, G., et al. (2011). Comprehensive analysis of kinase inhibitor selectivity. *Nat. Biotechnol.* 29, 1046–1051. doi:10.1038/nbt.1990

Faulon, J.-L., Misra, M., Martin, S., Sale, K., and Sapra, R. (2008). Genome scale enzyme–metabolite and drug–target interaction predictions using the signature molecular descriptor. *Bioinformatics* 24, 225–233. doi:10.1093/bioinformatics/btm580

Gao, K. Y., Fokoue, A., Luo, H., Iyengar, A., Dey, S., and Zhang, P. (2018). "Interpretable drug target prediction using deep neural representation," in *Proceedings of the twenty-seventh international joint conference on artificial intelligence, twenty-seventh international joint conference on artificial intelligence {IJCAI-18}* (Stockholm, Sweden: International Joint Conferences on Artificial Intelligence Organization), 3371–3377.

Gschwend, D. A., Good, A. C., and Kuntz, I. D. (1996). Molecular docking towards drug discovery. *J. Mol. Recognit.* 9, 175–186. doi:10.1002/(sici)1099-1352(199603)9:2<175::aid-jmr260>3.0.co;2-d

Huang, J., Chen, J., Zhang, B., Zhu, L., and Cai, H. (2021). Evaluation of gene–drug common module identification methods using pharmacogenomics data. *Brief. Bioinform.* 22, bbaa087. doi:10.1093/bib/bbaa087

Jastrzębski, S., Leśniak, D., and Czarnecki, W. M. (2018). *Learning to SMILE(S).* Tokyo, Japan: The Institute of Electronics, Information and Communication Engineers.

Keum, J., and Nam, H. (2017). SELF-BLM: Prediction of drug-target interactions via self-training SVM', *PLOS ONE. PLoS One* 12, e0171839. doi:10.1371/journal.pone.0171839

Kingma, D., and Ba, J. (2014). *Adam: A method for stochastic optimization* in 3rd International Conference on Learning Representations, ICLR 2015-Conference Track Proceedings. International Conference on Learning Representations, ICLR (San Diego, CA: OpenReview.net).

Kingma, D. P., and Welling, M. (2014). *Auto-encoding variational bayes*. in 2nd International Conference on Learning Representations, ICLR 2014 - Conference Track Proceedings. International Conference on Learning Representations, ICLR (Banff, Canada: OpenReview.net).

Liu, H., Sun, J., Guan, J., Zheng, J., and Zhou, S. (2015). Improving compound–protein interaction prediction by building up highly credible negative samples. *Bioinformatics* 31, i221–i229. doi:10.1093/bioinformatics/btv256

Liu, M.-Y., Breuel, T., and Kautz, J. (2018). *Unsupervised image-to-image translation networks* in Advances in neural information processing systems (Red Hook, NY, United States: Curran Associates Inc), 700–708.

Liu, Y., Wu, M., Miao, C., Zhao, P., and Li, X.-L. (2016). Neighborhood regularized logistic matrix factorization for drug-target interaction prediction. *PLOS Comput. Biol.* 12, e1004760. doi:10.1093/bioinformatics/btaa577

Mayr, A., Klambauer, G., Unterthiner, T., Steijaert, M., Wegner, K. J., Ceulemans, H., et al. (2018). Large-scale comparison of machine learning methods for drug target prediction on ChEMBL. *Chem. Sci.* 9, 5441–5451. Royal Society of Chemistry. doi:10.1039/c8sc00148k

Min, S., Lee, B., and Yoon, S. (2016). Deep learning in bioinformatics. *Brief. Bioinform.* 18, 851–869. doi:10.1093/bib/bbw068

Nascimento, A. C. A., Prudêncio, R. B. C., and Costa, I. G. (2016). A multiple kernel learning algorithm for drug-target interaction prediction. *BMC Bioinforma.* 17, 46. doi:10.1186/s12859-016-0890-3

Nguyen, T., Le, H., Quinn, T. P., Nguyen, T., Le, T. D., and Venkatesh, S. (2019). *GraphDTA: Predicting drug–target binding affinity with graph neural networks*, preprint. *Bioinformatics* 37, 1140–1147. doi:10.1093/bioinformatics/btaa921

Oprea, T. I., and Mestres, J. (2012). Drug repurposing: Far beyond new targets for old drugs. *AAPS J.* 14, 759–763. doi:10.1208/s12248-012-9390-1

Öztürk, H., Özgür, A., and Ozkirimli, E. (2018). DeepDTA: Deep drug–target binding affinity prediction. *Bioinformatics* 34, i821–i829. doi:10.1093/bioinformatics/bty593

Öztürk, H., Ozkirimli, E., and Özgür, A. (2019). *WideDTA: Prediction of drug-target binding affinity*. (Tokyo, Japan: The Institute of Electronics, Information and Communication Engineers).

Qiu, Y., Ching, W.-K., and Zou, Q. (2021a). Matrix factorization-based data fusion for the prediction of RNA-binding proteins and alternative splicing event associations during epithelial–mesenchymal transition. *Brief. Bioinform.* 22, bbab332. bbab332. doi:10.1093/bib/bbab332

Qiu, Y., Ching, W.-K., and Zou, Q. (2021b). Prediction of RNA-binding protein and alternative splicing event associations during epithelial–mesenchymal transition based on inductive matrix completion. *Brief. Bioinform.* 22, bbaa440. doi:10.1093/bib/bbaa440

Tang, J., Szwajda, A., Shakyawar, S., Xu, T., Hintsanen, P., Wennerberg, K., et al. (2014). Making sense of large-scale kinase inhibitor bioactivity data sets: A comparative and integrative analysis. *J. Chem. Inf. Model.* 54, 735–743. doi:10.1021/ci400709d

Tian, K., Shao, M., Wang, Y., Guan, J., and Zhou, S. (2016). Boosting compound-protein interaction prediction by deep learning. *Methods* 110, 64–72. doi:10.1016/j.ymeth.2016.06.024

Tsubaki, M., Tomii, K., and Sese, J. (2019). Compound–protein interaction prediction with end-to-end learning of neural networks for graphs and sequences', *Bioinformatics*. *Bioinformatics* 35, 309–318. doi:10.1093/bioinformatics/bty535

Unterthiner, T., and Mayr, A. (2014). *Deep learning as an opportunity in virtual screening* in Proceedings of the deep learning workshop at NIPS (Cambridge, MA, United States: MIT Press), 27, 1–9.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., et al. (2017). *Attention is all you need* in Advances in Neural Information Processing Systems30 (NIPS) (Red Hook, NY, United States: Curran Associates Inc), 5998–6008.

Walker, J., Marino, K., Gupta, A., and Hebert, M. (2017). "The pose knows: Video forecasting by generating pose futures," in *2017 IEEE international conference on computer vision (ICCV)* (Venice: IEEE), 3352–3361.

Wang, L., You, Z.-H., Chen, X., Xia, S.-X., Liu, F., Yan, X., et al. (2018). A computational-based method for predicting drug–target interactions by using stacked autoencoder deep neural network. *J. Comput. Biol.* 25, 361–373. doi:10.1089/cmb.2017.0135

Weininger, D. (1988). SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. *J. Chem. Inf. Model.* 28, 31–36. doi:10.1021/ci00057a005

Wu, Y., Gao, M., Zeng, M., Zhang, J., and Li, M. (2022a). BridgeDPI: A novel graph neural network for predicting drug–protein interactions', *bioinformatics*. *Bioinformatics* 38, 2571–2578. doi:10.1093/bioinformatics/btac155

Wu, Y., Zeng, M., Fei, Z., Yu, Y., Wu, F.-X., and Li, M. (2022b). Kaicd: A knowledge attention-based deep learning framework for automatic icd coding. *Neurocomputing* 469, 376–383. doi:10.1016/j.neucom.2020.05.115

Yamanishi, Y., Kotera, M., Kanehisa, M., and Goto, S. (2010). Drug-target interaction prediction from chemical, genomic and pharmacological data in an integrated framework. *Bioinformatics* 26, i246–i254. doi:10.1093/bioinformatics/btq176

Zeiler, M. D., Krishnan, D., Taylor, G. W., and Fergus, R. (2010). "Deconvolutional networks," in *2010 IEEE computer society conference on computer vision and pattern recognition, 2010 IEEE conference on computer vision and pattern recognition (CVPR)* (San Francisco, CA, USA: IEEE), 2528–2535.

Zeng, M., Zhang, F., Wu, F.-X., Li, Y., Wang, J., and Li, M. (2019). Protein–protein interaction site prediction through combining local and global features with deep neural networks. *Bioinformatics* 36, 1114–1120. doi:10.1093/bioinformatics/btz699

Zhang, F., Song, H., Zeng, M., Li, Y., Kurgan, L., and Li, M. (2019). DeepFunc: A deep learning framework for accurate prediction of protein functions from protein sequences and interactions. *PROTEOMICS* 19, 1900019. doi:10.1002/pmic.201900019

Zhang, F., Song, H., Zeng, M., Wu, F.-X., Li, Y., Pan, Y., et al. (2021). A deep learning framework for gene ontology annotations with sequence- and network-based information. *IEEE/ACM Trans. Comput. Biol. Bioinform.* 18, 2208–2217. doi:10.1109/TCBB.2020.2968882

frontiers | Frontiers in Genetics

TYPE Original Research
PUBLISHED 03 November 2022
DOI 10.3389/fgene.2022.1034832

Check for updates

OPEN ACCESS

EDITED BY
Jianzhen Xu,
Shantou University, China

REVIEWED BY
Meng Zhou,
Wenzhou Medical University, China
Dangdang Li,
Northeast Forestry University, China
Qi-En Yang,
Northwest Institute of Plateau Biology (CAS), China

*CORRESPONDENCE
Pengling Ge,
penglingge@126.com,
penglingge@yeah.net
Junzhi Wang,
drvolodia@163.com

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

RECEIVED 02 September 2022
ACCEPTED 21 October 2022
PUBLISHED 03 November 2022

CITATION
Zhang Y, Han D, Yu X, Shao X, Zong C, Zhang M, Wang J, Liang J and Ge P (2022), MiRNA-190a-5p promotes primordial follicle hyperactivation by targeting PHLPP1 in premature ovarian failure.
Front. Genet. 13:1034832.
doi: 10.3389/fgene.2022.1034832

MiRNA-190a-5p promotes primordial follicle hyperactivation by targeting PHLPP1 in premature ovarian failure

Yuchi Zhang[1,2], Dongwei Han[1], Xiaoyan Yu[1], Xinyu Shao[1,3], Chuju Zong[1,4], Manyu Zhang[1], Junzhi Wang[5]*, Jingwen Liang[1] and Pengling Ge[1]*

[1]Department of Pharmacology, School of Basic Medical Sciences, Heilongjiang University of Chinese Medicine, Harbin, China, [2]Department of Neurology, Faculty of Medicine, Shimane University, Izumo, Japan, [3]The First Affiliated Hospital of Qiqihar Medical University, Qiqihar, China, [4]Heilongjiang Institute for Drug Control, Harbin, China, [5]Department of Dermatology, First Affiliated Hospital, Heilongjiang University of Chinese Medicine, Harbin, China

We previously screened 6 differentially expressed miRNAs in ovarian tissues of 4-vinylcyclohexene diepoxide (VCD)-treated premature ovarian failure (POF) model in SD rats, including miRNA-190a-5p, miRNA-98-5p, miRNA-29a-3p, miRNA-144-5p, miRNA-27b-3p, miRNA-151-5p. In this study, to investigate the mechanisms causing the onset of POF, we first identified miRNAs with earlier differential expression at consecutive time points in the VCD-treated rat POF model and explored the mechanisms by which the target miRNAs promote POF. The SD rats were injected with VCD for 15 days to induce POF. Additionally, we collected rat blood and ovaries at the same time every day for 15 consecutive days, and luteinizing hormone (LH), follicle-stimulating hormone (FSH), Anti-Mullerian hormone (AMH), and estradiol (E_2) serum levels were detected by ELISA. Six miRNAs expression were measured in rat ovaries by qRT-PCR. Dual-luciferase reporter gene assays were employed to predict and verify the target gene (PHLPP1) of target miRNAs (miRNA-190a-5p). Western blot was examined to detect the expression levels of PHLPP1, AKT, p-AKT, FOXO3a, p-FOXO3a, and LHR proteins on the target gene PHLPP1 and its participation in the primordial follicular hyperactivation-related pathways (AKT-FOXO3a and AKT-LH/LHR). During the VCD modeling POF rat ovaries, miRNA-190a-5p was the first to show significant differential expression, i.e., 6th of VCD treating, and PHLPP1 was verified to be a direct downstream target of it. Starting from the 6th of VCD treatment, the more significant the up-regulation trend of miRNA-190a-5p expression, the more obvious the down-regulation trend of PHLPP1 and LHR mRNA and protein expression, accompanied by the more severe phosphorylation of AKT and FOXO3a proteins, thus continuously over-activating the rat primordial follicle to promote the development of POF. In conclusion, miRNA-190a-5p may become a potential biomarker for early screening of POF, and it can continuously activate primordial follicles in rats by targeting the expression of PHLPP1 and key proteins in the AKT-FOXO3a and AKT-LH/LHR pathways.

KEYWORDS

premature ovarian failure (POF), 4-vinylcyclohexene diepoxide (VCD), miRNA-190a-5p, PHLPP1, AKT, LHR, FOXO3a

1 Introduction

With the change in people's lifestyles and fertility concepts, women's childbearing age is moving backward, and the incidence of infertility is increasing year by year (Harumi, 2009). According to the World Health Organization (WTO) reported, infertility will become the third largest disease after cancer and cardiovascular disease (Vayena et al., 2002; Dong et al., 2021). Meanwhile, the latest survey showed that the incidence rate of infertility in the world is 15% to 20%, and the number of patients ranges from 48 million to 180 million (Thoma et al., 2021). Among them, premature ovarian failure (POF) is one of the important pathogenic factors that cause infertility in women of childbearing age, and it is a gynecological disease with unclear etiology (Shelling, 2010). Most patients are found to be in a state of POF due to menstrual disorders, amenorrhea, and infertility (Shelling, 2010). However, at present, the onset time of POF cannot be determined, and it has the characteristics of being irreversible or difficult to reverse (Santoro, 2003). Thus, early detection in the initial process of POF is an urgent medical research problem.

With the deepening of medical research, more and more evidence has confirmed that the over-activation of primordial follicles is an important pathological mechanism for POF, and it is also caused by the premature depletion of ovarian reserves (Adhikari et al., 2013; Zhou et al., 2017; Chakravarthi et al., 2020; Maidarti et al., 2020). Therefore, it is particularly important to find out the regulatory factors of the over-activation of primordial follicles to explore the mechanism of POF. Studies have shown that a common ovarian toxic chemical, 4-vinylcyclohexene diepoxide (VCD), accelerates ovarian failure by specifically activating primordial follicles (Hu et al., 2006; Fernandez et al., 2008; Kappeler and Hoyer, 2012; Lee et al., 2017). In our previous research, we replicated the POF rat model by VCD (Li et al., 2014), and detected 6 differentially expressed miRNAs between POF rats and normal rats by rat genome-wide miRNA expression profile technology, including miRNA-190a-5p, miRNA-98-5p, miRNA-29a-3p, miRNA-144-5p, miRNA-27b-3p, miRNA-151-5p (Kuang et al., 2014). Then, we speculated that the 6 differentially expressed miRNAs might be involved in the occurrence of POF. However, it is unknown whether these differentially expressed miRNAs are the cause of excessive activation of primordial follicles in early premature ovarian failure or the result of hyperactivation of primordial follicles. Thus, this study will detect the expression of 6 miRNAs at a continuous time point from the first day of the application of VCD to replicate the POF rat model, identify the target miRNAs with early differential expression, and verify whether they can induce POF.

2 Materials and methods

2.1 Animals

Twelve weeks old female Sprague-Dawley (SD) rats, weighing 200 ± 20 g, were supplied by the Liaoning Changsheng Biotechnology Co., Ltd. (Benxi, China) and were housed in the Heilongjiang University of Chinese Medicine with SPF conditions. The rats were housed in a temperature (20 ± 1°C)-and humidity (50 ± 5%)-controlled animal facility and maintained on a 12-h light/dark cycle and were acclimatized for 1 week before the experiment and allowed free access to a rodent diet and tap water. All experiments were approved by the Animal Experimental Ethical Committee of Heilongjiang University of Chinese Medicine (Heilongjiang, China) and performed by the Guide for the Care and Use of Laboratory Animals.

2.2 Experimental design

Two hundred female SD rats with an estrous cycle of 5 days were selected as experimental animals, we randomly divided them into two groups: The control groups (n = 100)

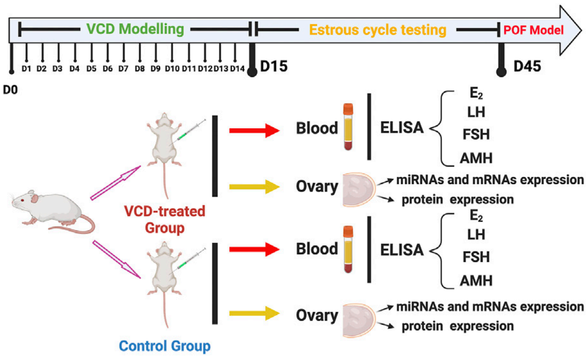


FIGURE 1
The experimental design of this study. The SD female rats divided into two groups: The VCD-treated group and Control group. The two group rats were sacrificed at the same time every day (D1-D15) for 15 consecutive days. After estrous cycle testing, the POF model rats also were collected. Each rat was detected with serum hormone levels (E_2, LH, FSH, and AMH) and screened for early differential expression of target miRNAs from those significantly differentially expressed on POF model rats (miRNA-190a-5p, miRNA-98-5p, miRNA-29a-3p, miRNA-144-5p, miRNA-27b-3p, and miRNA-151-5p). Subsequently, ovarian tissues from rats on the 6th (D6), 10th (D10), 15th (D15), and 45th (D45) of VCD injection were selected to detect the expression levels of target genes of target miRNAs and mRNAs and proteins on the primordial follicle hyperactivation signaling pathway.

TABLE 1 qRT-PCR primers of miRNAs.

Gene	Forward primer sequence [5′→3′]	Stem-loop primer sequence [5′→3′]	Accession ID	Primer melting temperature (T_m, °C)
rno-miRNA-190a-5p	CGCGCGTGATATGTTTGATATA	GTCGTATCCAGTGCAGGGTCCGAGGTATTCGCACTGGATACGACACCTAA	NR_031906	77.5
rno-miRNA-144-5p	GCGCGGGATATCATCATATACT	GTCGTATCCAGTGCAGGGTCCGAGGTATTCGCACTGGATACGACACTTAC	NR_031890	78
rno-miRNA-27b-3p	CGGGTTCACAGTGGCTAA	CCTGTTGTCTCCAGCCACAAAAGAGCACAATATTTCAGGAGACAACAGGGCAGAAC	NR_031832	78
rno-miRNA-29a-3p	CGGGTAGCACCATCTGAAA	CCTGTTGTCTCCAGCCACAAAAGAGCACAATATTTCAGGAGACAACAGGTAACCGA	NR_031836	77.2
rno-miRNA-98-5p	CGGGGTGAGGTAGTAAGTTG	CCTGTTGTCTCCAGCCACAAAAGAGCACAATATTTCAGGAGACAACAGGAACAATA	NR_031855	76
rno-miRNA-151-5p	CGGTCGAGGAGCTCACA	CCTGTTGTCTCCAGCCACAAAAGAGCACAATATTTCAGGAGACAACAGGACTAGAC	NR_031798	78
U6	CCTGCTTCGGCAGCACA			80.3

Note: Universal reverse primer in rno-miRNA-27b-3p, rno-miRNA-29a-3p, rno-miRNA-98-5p and rno-miRNA-151-5p was: CAGCCACAAAAGAGCACAAT., The other universal reverse primer in rno-miRNA-190a-5p and rno-miRNA-144-5p was: AGTGCAGGGTCCGAGGTATT., The universal reverse primer of U6 was AACGCTTCACGAATTTGCGT.

TABLE 2 qRT-PCR primers of mRNAs.

Gene	Forward primer sequence [5′→3′]	Reverse primer sequence [5′→3′]	Accession ID	Primer melting temperature (T_m, °C)	Product length
Phlpp1	AGACGCCAGGTCATTCTGTG	TTGACGCAGCCATCGTAAGT	NM_021657.1	60	216 bp
Akt1	CAAGATGTGTATGAGAAGAAGCTGA	GTTCACTGTCCACACACTCCA	NM_033230.2	60	154 bp
Foxo3a	GTCACGACAAGTTCCCCAGT	AGTTTGAGGGTCTGCTTTGCC	NM_001106395.1	60	261 bp
Lhr	TCGCCCTGTCTTCCTACTCA	TGGCGGAATAAAGCGTCTCG	NM_012978.1	60	213 bp
Gapdh	AGTGCCAGCCTCGTCTCATA	GATGGTGATGGGTTTCCCGT	NM_017008.3	60	248 bp

and the VCD-treated groups (n = 100). Then the rats were intraperitoneally injected with sesame oil or sesame oil plus VCD (80 mg/kg/day) for 15 consecutive days according to our previous study (Kuang et al., 2014). At the same time, we sacrificed the rats in the control group and VCD-treated group every day (D1-D15, n = 6) for 15 consecutive days, and collected the blood and ovaries. Then we used the vaginal cytology method to observe the changes in the estrous cycle of rats. After 30 consecutive days of observation, the rats with estrous cycle disorder were taken as the POF model group rats. After 30 days of estrous cycle testing, we also collected the blood and ovaries of the D45 control group (n = 6) and the VCD-treated group (n = 6) rats. At the end of the experimental period, all rats were sacrificed by CO_2 inhalation. All blood samples were centrifuged at 4,000 r/min and 4°C for 20 min to obtain blood serum samples for Enzyme-Linked Immunosorbent (ELISA) assays. After blood serum collection, all the ovaries were collected. From each rat, one ovary was rapidly frozen by liquid nitrogen and stored at 80°C for Quantitative real-time PCR (qRT-PCR), and the other ovary was stored at −80°C for Western blot analysis (Figure 1).

2.3 Enzyme-Linked Immunosorbent assays (ELISA)

The levels of rat serum estradiol (E_2), anti-Mullerian hormone (AMH), follicle-stimulating hormone (FSH), and luteinizing hormone (LH) were determined by ELISA kits (Cloud-Clone Corp Inc. United States) following the manufacturer's instructions. Briefly, LH, AMH, E_2 or FSH standards at a final concentration of 8000, 2666.67, 888.89, 296.30, 98.77, and 0 pg/ml, 10000, 3333.3, 1111.1, 370.4, 123.5, and 0 pg/ml, 1000, 333.33, 111.11, 37.04, 12.35, and 0 pg/ml or 600, 200, 66.67, 22.22, 7.41, 2.4, and 0 ng/ml or diluted rat serum were added to anti-LH, AMH, E_2 or FSH antibody-coated wells and incubated for 30 min. After washing five times, the horseradish peroxide (HRP)-conjugated detection antibodies were added, followed by the addition of the substrate solution. The optical density (OD) value was determined at a wavelength of 450 nm.

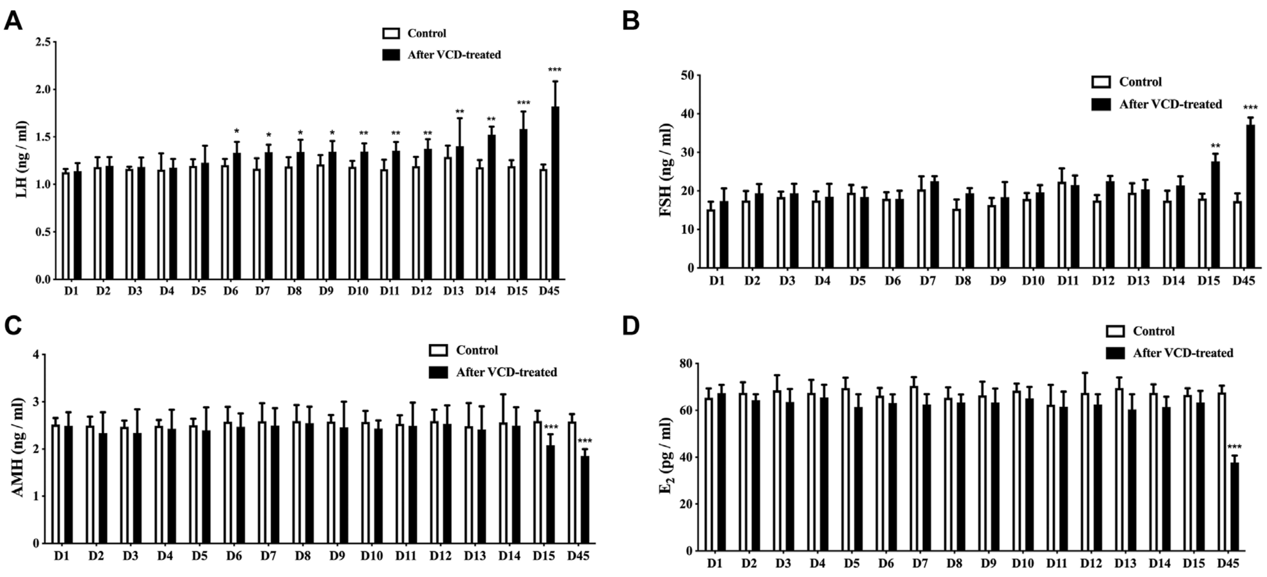


FIGURE 2
LH, FSH, AMH and E_2 serum levels. **(A)**LH, **(B)**FSH, **(C)**AMH, and **(D)**E_2 serum levels were detected by ELISA during the VCD-treated process (D1-D15, D45). Data are summarized and presented as mean ± SEM from four independent experiments. *$p < 0.05$, **$p < 0.01$ and ***$p < 0.001$ vs. control.

2.4 RNA extraction and quantitative real-time PCR (qRT-PCR)

According to the manufacturer's instructions, total RNA was extracted from whole ovaries by TRIzol reagent (Invitrogen, United States) and miRNA or mRNA easy Minikit (Qiagen, Valencia, CA, United States). In addition, RNA concentrations were determined by NanoDrop ND-1000 spectrophotometer (Wilmington, DE, United States).

For qRT-PCR experiments, the RNA was reverse-transcribed into cDNA with a RevertAid First Strand cDNA Synthesis Kit (Thermo Fisher, United States) following the manufacturer's instructions. Quantitative PCR reaction was run on a CFX Connect Real-Time PCR Detection Instrument (BioRad, United States). Expression of the selected miRNAs was quantified by qRT-PCR, after reverse transcription with miRNA-specific stem-loop primers. The primer sequences are shown in Tables 1, 2 to amplify fragments. The data were normalized to expression levels of the housekeeping genes U6 and GAPDH, respectively, and the $2^{-\Delta\Delta CT}$ method was used to calculate relative expression levels.

2.5 Dual-luciferase reporter gene assay

Dual-Luciferase Reporter Gene Assay Target genes of rno-miRNA-190a-5p were predicted using target gene prediction software miRBase (http://www.mirbase.org) and TargetScan (http://www.targetscan.org/vert_71/). Plasmids containing wild-type PHLPP1 Luciferase reporter gene vector (PHLPP1) and mutant PHLPP1 Luciferase reporter gene vector (PHLPP1-mut) were constructed. They were co-transfected with rno-miR-190a-5p mimics or negative control (NC), respectively into 293T cells. Luciferase activity was detected by the Dual-Luciferase Reporter Assay System (Promega, Madison, WI) after 24 h.

2.6 Western blot analysis

Three frozen rat ovaries tissues in each group were homogenized in RIPA buffer with 1% Phenylmethylsulfonyl fluoride (PMSF). Then the homogenates were centrifuged at 12,000× g for 10 min, and the supernatants were collected for western blotting. 12 μg of total protein samples were separated on SDS-PAGE gels, transferred to PVDF membranes (0.45 μm, Millipore, IPVH00010), and blocked with 5% non-fat milk in TBS containing 0.1% Tween 20 (TBST). The membranes were incubated with the following primary antibodies: anti-PHLPP1, anti-AKT, anti-p-AKT, anti-FOXO3a, anti-p-FOXO3a, anti-LHR and anti-GAPDH (1:1000, Abcam, United Kingdom), diluted in TBST-5% milk at 4°C overnight. The appropriate horseradish peroxidase-conjugated secondary antibodies were diluted 1:1000 in TBST-5% milk and incubated for 1 h at room temperature, and the protein bands were visualized with enhanced chemiluminescence (ECL).

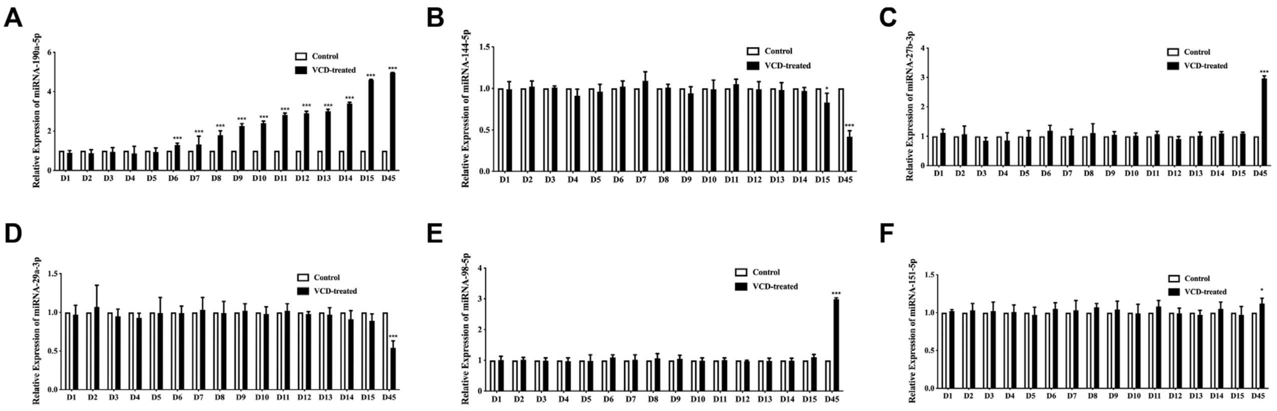


FIGURE 3
Screening of 6 miRNAs for early differential miRNAs in the VCD modeling process. **(A)** miRNA-190a-5p, **(B)** miRNA-144-5p, **(C)** miRNA-27b-3p, **(D)** miRNA-29a-3p, **(E)** miRNA-98-5p and **(F)** miRNA-151-5p expression levels during VCD-treated process (D1-D15, D45) in rat ovaries tissues were determined by qRT-PCR analysis. Data are summarized and presented as mean ± SEM from six independent experiments. *$p < 0.05$ and ***$p <$ 0.001 vs. control.

2.7 Statistical analysis

All results are represented as a mean ± standard error of the mean. Statistical analysis of these results was carried out by *t*-test or one-way ANOVA (with LSD or Dunnett's T3 correction for comparison of multiple means). *p* values <0.05 were considered statistically significant. All statistical analyses were done using IBM SPSS 19.0 Software.

3 Results

3.1 Identification of miRNAs that are differentially expressed earlier in when VCD treatment process

We first observed the changes of E_2, FSH, LH, and AMH hormone levels in rat serum by ELISA to determine the development of POF in rats and whether it was successfully created or not. We found that with the increase of VCD-treated days (D1-D15), the LH levels increased significantly increasing the number of VCD-treatment days starting from the D6 VCD-treated group ($p < 0.05$, $p < 0.01$, $p < 0.001$, Figure 2A), the FSH level increased significantly in D15 VCD-treated group ($p < 0.01$, Figure 2B), the AMH level decreased significantly in D15 VCD-treated group ($p < 0.001$, Figure 2C) and there were no significant changes in E_2 level (Figure 2D). On the 45th day (D45), which is the day to judge whether the POF animal model is successful or not, we found that the levels of LH and FSH level were significantly increased ($p < 0.001$, Figures 2A,B) and the levels of E_2 and AMH level in the VCD-treated groups were significantly decreased ($p < 0.001$, Figures 2C,D).

Then we used qRT-PCR analysis in rat ovarian tissue to observe the 6 significantly differentially expressed miRNAs (miRNA-190a-5p, miRNA-27b-3p, miRNA-29a-3p, miRNA-98-5p, miRNA-151-5p, and miRNA-144-5p), which were being screened in the POF model rat ovary by our previous research, expression level changes during VCD-treated process (D1-D15). In this study, we found that miRNA-190a-5p was significantly differentially up-regulated at the earliest in the D6 VCD-treated group, with the increase of VCD-treated days, the up-regulation trend of differential expression became more and more significantly ($p < 0.001$, Figure 3A) compared with other miRNAs, the expression of miRNA-144-5p was up-regulated on the D15 VCD-treated group ($p < 0.05$, Figure 3B), and there was no significant difference in other miRNAs (Figures 3C–F). Meanwhile, this study also verified our previous findings. In the D45 VCD-treated group, miRNA-190a-5p, miRNA-27b-3p, miRNA-98-5p and miRNA-151-5p were significantly up-regulated ($p < 0.001$, Figures 3A,C,E,F), while miRNA-144-5p and miRNA-29a-3p were significantly down-regulated ($p <$ 0.001, Figures 3B,D).

3.2 miRNA-190-5p directly targets PHLPP1 in the VCD-treating process

PHLPP1 was selected as a candidate miRNA-190a-5p target based on bioinformatics analysis (miRBase and TargetScan). As shown in Figure 4A, PHLPP1 was found to be a possible target gene of rno-miRNA-190a-5p. Then, the dual-luciferase reporter gene assay verified that miRNA-190a-5p impaired the luciferase activity of the wild-type PHLPP1 (PHLPP1) but not the mutant PHLPP1 3′-UTR (PHLPP1-mut) in cells (Figure 4B). The result

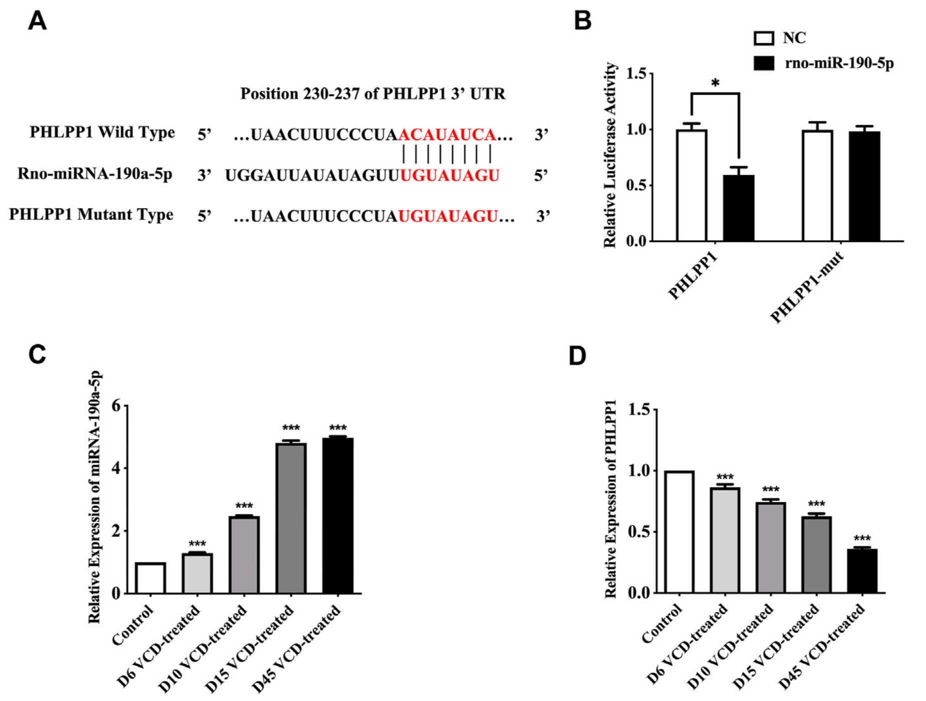


FIGURE 4
MiRNA-190-5p directly targets PHLPP1 during the VCD-treating process. **(A)** Diagram of putative miRNA-190a-5p binding sites of PHLPP1. Binding sequences for miRNA-190a-5p in the 3′-UTR of PHLPP1 and the mutations in the 3′-UTR of PHLPP1 are presented. **(B)** Luciferase activity of the wild-type PHLPP1 3′-UTR (PHLPP1) and mutant PHLPP1 3′-UTR (PHLPP1-mut) co-transfected with miRNA-190a-5p mimics (rno-miR-190a-5p) or negative control (NC) was measured. **(C)** and **(D)** qRT-PCR analysis of miRNA-190a-5p, PHLPP1 mRNA in rat ovaries tissues during the VCD treating process (D6, D10, D15, and D45). Data are summarized and presented as mean ± SEM from three independent experiments. **p* < 0.05 vs. NC and ****p* < 0.001 vs. control.

suggested that PHLPP1 was a target of miRNA-190a-5p. We then selected rat ovarian tissues from the VCD-treated groups on the 6th (D6), 10th (D10), 15th (D15), and 45th days (D45) to detect the expression changes of miRNA-190a-5p and PHLPP1 mRNA by qRT-PCR analysis. This data showed that with the increase of VCD treating days, the expression of miRNA-190a-5p was significantly up-regulated ($p < 0.001$, Figure 4C), while the expression trend of PHLPP1 was significantly down-regulated ($p < 0.001$,

4D). Taken together, all results verified that PHLPP1 was a direct downstream target of miRNA-190-5p in POF.

3.3 MiRNA-190a-5p targeting PHLPP1 to promote premature ovarian failure

To study the mechanism of miRNA-190a-5p to promote POF after VCD treatment. We detected the expression level of PHLPP1 mRNA and protein and its signaling pathway mRNAs and proteins, including AKT, FOXO3a, and LHR in the VCD-treated rat ovaries tissues. As shown in Figures 5, 6, with the increase of the VCD-treated process, from the VCD-treated groups on the 6th (D6), 10th (D10), 15th (D15), and 45th days (D45), the expression of PHLPP1 mRNA and protein in each model group decreased, showing a significant downward trend ($p < 0.05$, $p < 0.001$, Figure 4D and Figure 6B). And the LHR mRNA and protein expression levels decreased, showing a significant downward trend ($p < 0.001$, Figure 5A, Figure 6C). We also found the AKT (AKT1) mRNA expression was significantly decreased ($p < 0.001$, Figure 5B) and the ratio between phosphorylated and non-phosphorylated AKT protein (p-AKT/AKT) was increased, showing a significant upward trend ($p < 0.01$, $p < 0.001$), but the difference in the D6 VCD-treated group was not significant (Figure 6D). Meanwhile, FOXO3a mRNA expression was significantly decreased ($p < 0.001$, Figure 5C), and the ratio between phosphorylated and non-phosphorylated FOXO3a protein (p-FOXO3a/FOXO3a) was increased, showing a significant upward trend ($p < 0.001$, Figure 6E).

4 Discussion

Premature ovarian failure (POF) is characterized by amenorrhea, infertility, low estrogen levels, excess

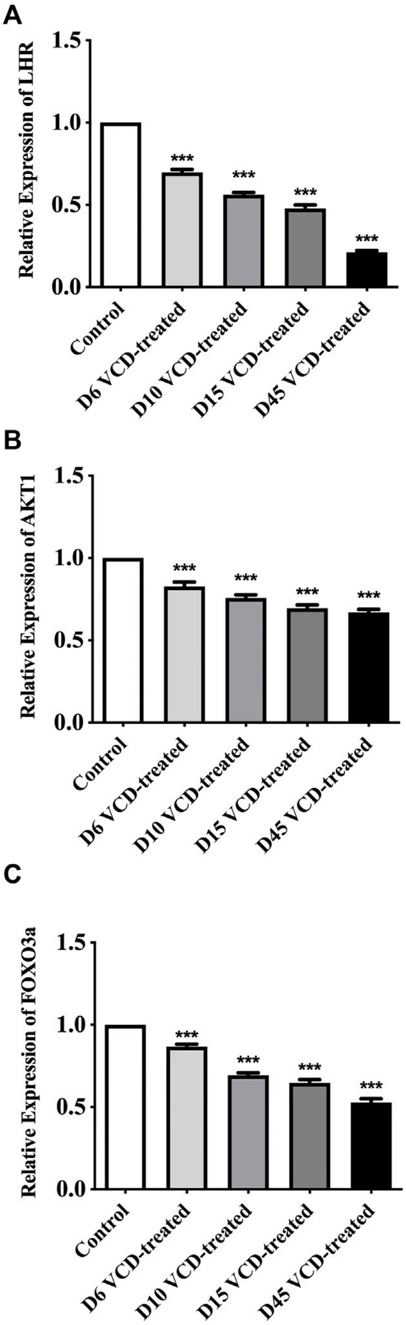


FIGURE 5
MiRNA-190a-5p targets PHLPP1 to regulate the mRNA expression levels of key genes on the primordial follicle hyperactivation signaling pathway. **(A)** LHR, **(B)** AKT1, and **(C)** FOXO3a mRNA expression during the VCD treating process (D6, D10, D15 and D45) in rat ovaries tissues were determined by qRT-PCR analysis. Data are summarized and presented as mean ± SEM from three independent experiments. ***$p < 0.001$ vs. control.

gonadotropins, and a lack of mature follicles before age 40 (Okeke et al., 2013). The occurrence of POF is mainly related to the continuous reduction of follicles in the ovary and the abnormal development of follicles. The size of the primordial follicles and the maintenance of the resting state determine the speed of ovarian aging. Once the primordial follicles are exhausted, the ovaries will be rapidly depleted after losing their function, and the body will enter a state of menopause. Studies have shown that 4-vinylcyclohexene diepoxie (VCD), a common ovarian toxic chemical, accelerates ovarian failure by specifically activating primordial follicles, and persists after drug withdrawal characteristics of toxicity. In animal experiments, the different stages of ovarian function from weakened to failure can be simulated by controlling the dosage and modeling time, so it can be widely used in the study of decreased ovarian reserve, POF, and menopause-related diseases (Hoyer et al., 2001; Frye et al., 2012; Kappeler and Hoyer, 2012). In addition, compared with other modeling methods, VCD only affects gonadal tissues such as ovaries, and hardly affects other tissues or organs (Tamura et al., 2009). And some experiments have confirmed that VCD can induce POF in experimental animal models (Hoyer et al., 2001; Wright et al., 2011). Therefore, in the early stage of our study (Kuang et al., 2014; Li et al., 2014), VCD was used to successfully induce the POF animal model.

Based on our previous study, we still used the same approach to establish the POF rat model in the present study. Even for the purpose of collecting tissues during VCD modeling (i.e., during primordial follicular hyperactivation). We tested blood samples and collected rat ovarian tissue after sacrificing rats at consecutive time points of VCD injection. Hormone level is an important and sensitive indicator of ovary function. The E_2, FSH, and LH levels are classical criteria for POF (Steiner, A. Z., 2013; Nie, X et al., 2018). Clinical studies showed that patients in the POF group were found to have significantly higher FSH and LH serum concentrations and lower E_2 serum levels when compared to healthy controls (Podfigurna, A et al., 2018; Qi et al., 2022). Our hormone level results are consistent with our previous findings (Kuang et al., 2014; Li et al., 2014). However, it is worth noting that in our result, we found that LH was the hormone with early differential changes in the VCD-treated process, that is, from the 6th day of treatment. In addition, some studies have shown that the change in LH level is an important link to trigger follicle ovulation in the process of follicle development (Holesh et al., 2021). Therefore, we speculate that the abnormal change in LH level caused by VCD modeling will damage the development of follicles and ovulation function.

In this study, we additionally observed the changes in serum Anti-Mullerian hormone (AMH) levels in rats during VCD-treated days. And we found that the AMH level decreased with the increase of VCD-treated time. Previous studies have shown that chemicals can destroy growing follicles, which indirectly causes a reduction in the primordial follicular pool (Wang et al., 2019), as the destruction of growing follicles leads to a decrease in AMH, which activates the primordial follicles, and an increase in follicular granulosa cells to produce AMH to inhibit

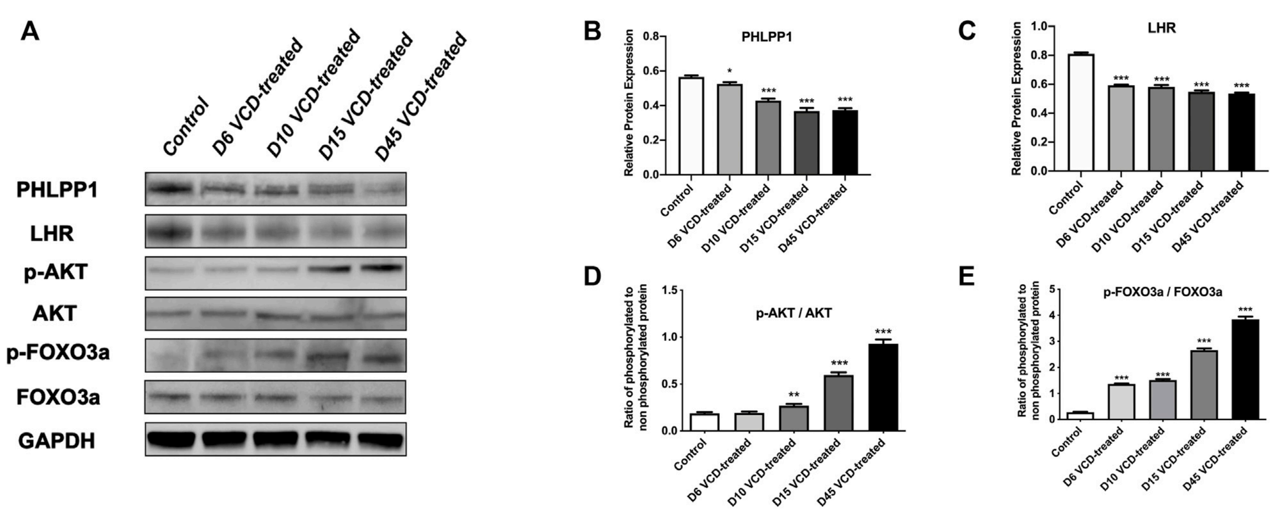


FIGURE 6
MiRNA-190a-5p targets PHLPP1 to regulate the protein expression levels of key genes on the primordial follicle hyperactivation signaling pathway. **(A)** Western blot analysis of PHLPP1, LHR, AKT, p-AKT, FOXO3a, and p-FOXO3a during the VCD treating process (D6, D10, D15, and D45) in rat ovaries tissues. GAPDH was used as a negative control. **(B)** and **(C)** Protein analysis comparing the concentrations of PHLPP1 and LHR with GAPDH used as a loading control. **(D)** and **(E)** Protein analysis comparing the concentrations of phosphorylated and total AKT, and FOXO3a with GAPDH used as a loading control, respectively. The ratio of phosphorylated to non-phosphorylated protein was calculated for each protein. Data are summarized and presented as mean ± SEM from four independent experiments. *$p < 0.05$, **$p < 0.01$ and ***$p < 0.001$ vs. control.

further primordial follicular activation (Durlinger et al., 1999). Thus, the activation and recruitment of primordial follicles may be accelerated in mice molded using the chemical induction method due to the sudden loss of growing follicles (Baker et al., 2011). Another study showed that when ovarian function is impaired and the number of growing follicles decreases leading to a decrease in AMH, its function of maintaining the dynamic balance between follicular atresia and follicular recruitment decreases, which induces the primordial follicles to be overactivated, and this dynamic balance, if not restored, can eventually lead to POF (Moolhuijsen and Visser, 2020). Therefore, our results speculate that the overactivation of primordial follicles becomes more pronounced as the days of VCD administration increase. Hence, it is therefore important to seek novel biomarkers of primordial follicle hyperactivation.

Based on successfully replicating the POF animal model by VCD-treated, our previous study detected the most significantly differentially expressed 6 miRNAs between POF and normal rats' ovaries tissues by using the rat genome-wide miRNA expression profile chip technology, including miRNA-190a-5p, miRNA-98-5p, miRNA-29a-3p, miRNA-144-5p, miRNA-27b-3p and miRNA-151-5p (Kuang et al., 2014). Other previous studies have shown that miRNAs are involved in the maturation of mouse oocytes and the development of follicles (Dehghan et al., 2021). Hence, our study continuously monitored the expression changes of these 6 miRNAs from the first day of VCD-treated and screened out potential biomarkers that can be used in the early stage of POF disease. In our study, we found that miRNA-190a-5p was abnormally upregulated at the early stage of VCD injection, i.e., the 6th day (D6), and showed a significant upward trend with the increase of VCD-treated days. Therefore, we speculated that miRNA-190a-5p might serve as a biomarker in the early stage of POF. Usually, we also know that miRNAs exert their functions by degrading mRNAs expression (Othman and Nagoor, 2019; Tian et al., 2019). Hence, we used the publicly available databases to seek the potential targets of miRNA-190a-5p. Previous research reported that miRNA-190a directly inhibits the PH domain leucine-rich repeat protein phosphatase (PHLPP) (Yu et al., 2014). Likewise, in this study, we firstly predicted by database and then verified that miRNA-190a-5p negatively regulates its downstream target gene PHLPP1 at the early stage of POF, i.e., from the 6th day of VCD-treated POF, by double-luciferase reporter gene assay and qRT-PCR assay in rat ovarian tissues. It is thus hypothesized that miRNA-190a-5p may play a role in promoting premature ovarian failure by targeting PHLPP1.

PHLPP1 is a tumor suppressor protein that inactivates the kinase AKT through Ser437 dephosphorylation (Gao et al., 2005). A recent study showed that trivalent arsenic (A^{3+}) induces the expression of miR-190 in human bronchial epithelial cells, which binds the 3′UTR of the PHLPP transcript, decreasing PHLPP protein levels. Subsequently, AKT activation and phosphorylation levels were increased (Beezhold et al., 2011). In addition, in chemical environment induced POF models, several studies have confirmed the presence of enhanced AKT kinase signaling pathways and lead to primordial follicular hyperactivation (Sobinoff et al., 2011; Sobinoff et al., 2012). It has even been shown that after direct or indirect activation of AKT, phosphorylated AKT can activate phosphorylation of a series of downstream proteins (FOXO3a, BCL-2, etc.), which can promote follicular cell growth and proliferation (Makker et al., 2014). Moreover, AKT activation by upstream molecules has been reported to hyperphosphorylated FOXO3a, transport it out of the nucleus, and stimulate primordial follicle initiation (John et al., 2008). Therefore, FOXO3a can also be used as a marker to determine if the primordial follicle is activated (Yan et al., 2018). Thus, in this study, we used qRT-PCR assay and Western Blot analysis to detect the effects of VCD modeling on the expression of PHLPP1, AKT, and FOXO3a mRNA and the phosphorylation levels of AKT and FOXO3a proteins in rat ovarian tissues, and observed that as the time of VCD modeling increased, the expression levels of PHLPP1 mRNA and protein decreased along with the expression levels of AKT and FOXO3a mRNA and proteins, while the phosphorylation levels of AKT and FOXO3a proteins showed a significant increase. Therefore, we hypothesized that the abnormal expression of miRNA-190a-5p was upregulated in VCD-treated rats followed by downregulation of PHLPP1, which further activated AKT and FOXO3a proteins on AKT-FOXO3a signaling pathway, thus hyperactivation primordial follicles leading to the development of POF.

As mentioned previously in this study the latest finding was that LH was the earlier hormone to show differential changes during the VCD-treated POF rat model and was significantly differentially expressed almost the same day as miRNA-190a-5p. It has been reported that LH itself induces the expression of luteinizing hormone receptor (LHR) and that the pulsatile release of LH maintains the levels of LHR and steroid hormone synthase (Plakkot et al., 2018). However, the experimental elevation of LH levels can desensitize hormonal signaling and lead to the downregulation of LHR expression (Dufau, 1998). It has been shown that mutations in the LHR gene lead to abnormal LHR function and failure to properly receive LH and stimulate hormonal signaling to the second messenger, resulting in impaired follicular maturation in the ovary, anovulation, delayed puberty, amenorrhea, infertility, with typical POF symptoms (Dixit et al., 2010). In addition, it has been shown that AKT knockout in mice is less responsive to LH processing, leading to a reduction in the number of primordial follicles in mouse ovaries (Maidarti et al., 2020). A recent study showed that cisplatin-induced primordial follicular hyperactivation in mice led to a significant reduction in the number of LHR expressions during POF (Chang et al., 2015). In our research, we found that the gene and protein expression of LHR in rat ovarian tissues showed a decreasing trend with the increase of VCD modeling time by qRT-PCR assay and Western Blot analysis. Therefore, we speculate that one of the alternative mechanisms by which VCD-induced miRNA-190a-5 promotes POF in rats may be that VCD induces miRNA-190a-5p to downregulate PHLPP1 after upregulation of abnormal expression at an early stage, which subsequently activates AKT, causing a decrease in AKT expression, leading directly or indirectly to an abnormal elevation of LH hormone and decrease in LHR receptor expression level, thus hyperactivation primordial follicles and ultimately leading to POF.

5 Conclusion

In this study, we found that miRNA-190a-5p may become a potential biomarker for early screening of POF, and it can continuously hyperactivation primordial follicles in rats by targeting the expression of PHLPP1 and key proteins in the AKT-FOXO3a and AKT-LH/LHR pathways.

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 animal study was reviewed and approved by the Animal Experimental Ethical Committee of Heilongjiang University of Chinese Medicine.

Author contributions

PG, YZ, and DH conceived and designed the study. YZ, XY, and XS wrote and edited the manuscript. DH helped with revising the manuscript. YZ, XS, CZ, MZ, and JL performed most of the experiments. YZ and XS collected and analyzed the data. PG provided financial support. JW contributed important reagents and provided technical assistance and supervised the experiment. All authors contributed to the article and approved the submitted version.

Funding

This work was supported in part by grants from the Natural Science Foundation of Heilongjiang province (No. JC2018025).

Acknowledgments

We would like to acknowledge the technical expertise of the Research and Experimental Center of the Second Affiliated Hospital of Harbin Medical University. We also thank Yuxin Liu, Shuying Yuan, Hui Wang, Jiaxin Li, Siqi Chen, Zhongmin

Lv, and Bo Wang for helping us with animal experiments and for assisting us in compiling and reviewing the relevant literature.

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

Adhikari, D., Risal, S., Liu, K., and Shen, Y. (2013). Pharmacological inhibition of mTORC1 prevents over-activation of the primordial follicle pool in response to elevated PI3K signaling. *PloS one* 8 (1), e53810. doi:10.1371/journal.pone.0053810

Baker, D. J., Wijshake, T., Tchkonia, T., LeBrasseur, N. K., Childs, B. G., Van De Sluis, B., et al. (2011). Clearance of p16Ink4a-positive senescent cells delays ageing-associated disorders. *Nature* 479 (7372), 232–236. doi:10.1038/nature10600

Beezhold, K., Liu, J., Kan, H., Meighan, T., Castranova, V., Shi, X., et al. (2011). miR-190-mediated downregulation of PHLPP contributes to arsenic-induced Akt activation and carcinogenesis. *Toxicol. Sci.* 123 (2), 411–420. doi:10.1093/toxsci/kfr188

Chakravarthi, V. P., Ghosh, S., Roby, K. F., Wolfe, M. W., and Rumi, M. A. (2020). A gatekeeping role of ESR2 to maintain the primordial follicle reserve. *Endocrinology* 161 (4), bqaa037. doi:10.1210/endocr/bqaa037

Chang, E. M., Lim, E., Yoon, S., Jeong, K., Bae, S., Lee, D. R., et al. (2015). Cisplatin induces overactivation of the dormant primordial follicle through PTEN/AKT/FOXO3a pathway which leads to loss of ovarian reserve in mice. *PLoS One* 10 (12), e0144245. doi:10.1371/journal.pone.0144245

Dehghan, Z., Mohammadi-Yeganeh, S., Rezaee, D., and Salehi, M. (2021). MicroRNA-21 is involved in oocyte maturation, blastocyst formation, and pre-implantation embryo development. *Dev. Biol.* 480, 69–77. doi:10.1016/j.ydbio.2021.08.008

Dixit, H., Rao, L., Padmalatha, V., Raseswari, T., Kapu, A. K., Panda, B., et al. (2010). Genes governing premature ovarian failure. *Reprod. Biomed. Online* 20 (6), 724–740. doi:10.1016/j.rbmo.2010.02.018

Dong, S. Q., Zhao, X. L., Sun, Y., and Zhang, J. W. (2021). Comparative efficacy and safety of complementary and alternative therapies for tubal obstructive infertility: A protocol for network meta-analysis. *Medicine* 100 (7), e24810. doi:10.1097/MD.0000000000024810

Dufau, M. L. (1998). The luteinizing hormone receptor. *Annu. Rev. Physiol.* 60 (1), 461–496. doi:10.1146/annurev.physiol.60.1.461

Durlinger, A. L., Kramer, P., Karels, B., de Jong, F. H., Uilenbroek, J. T. J., Grootegoed, J. A., et al. (1999). Control of primordial follicle recruitment by anti-Mullerian hormone in the mouse ovary. *Endocrinology* 140 (12), 5789–5796. doi:10.1210/endo.140.12.7204

Fernandez, S. M., Keating, A. F., Christian, P. J., Sen, N., Hoying, J. B., Brooks, H. L., et al. (2008). Involvement of the KIT/KITL signaling pathway in 4-vinylcyclohexene diepoxide-induced ovarian follicle loss in rats. *Biol. Reprod.* 79 (2), 318–327. doi:10.1095/biolreprod.108.067744

Frye, J. B., Lukefahr, A. L., Wright, L. E., Marion, S. L., Hoyer, P. B., and Funk, J. L. (2012). Modeling perimenopause in sprague–dawley rats by chemical manipulation of the transition to ovarian failure. *Comp. Med.* 62 (3), 193–202.

Gao, T., Furnari, F., and Newton, A. C. (2005). Phlpp: A phosphatase that directly dephosphorylates akt, promotes apoptosis, and suppresses tumor growth. *Mol. Cell.* 18 (1), 13–24. doi:10.1016/j.molcel.2005.03.008

Harumi, K. U. B. O. (2009). Epidemiology of infertility and recurrent pregnancy loss in society with fewer children. *JMAJ* 52 (1), 23–28.

Holesh, J. E., Bass, A. N., and Lord, M. (2021). *Physiology, ovulation*. Treasure Island: Statpearls.

Hoyer, P. B., Cannady, E. A., Kroeger, N. A., and Sipes, I. G. (2001). Mechanisms of ovotoxicity induced by environmental chemicals: 4-vinylcyclohexene diepoxide as a model chemical. *Adv. Exp. Med. Biol.* 500, 73–81. doi:10.1007/978-1-4615-0667-6_8

Hu, X., Roberts, J. R., Apopa, P. L., Kan, Y. W., and Ma, Q. (2006). Accelerated ovarian failure induced by 4-vinyl cyclohexene diepoxide in Nrf2 null mice. *Mol. Cell. Biol.* 26 (3), 940–954. doi:10.1128/MCB.26.3.940-954.2006

John, G. B., Gallardo, T. D., Shirley, L. J., and Castrillon, D. H. (2008). Foxo3 is a PI3K-dependent molecular switch controlling the initiation of oocyte growth. *Dev. Biol.* 321 (1), 197–204. doi:10.1016/j.ydbio.2008.06.017

Kappeler, C. J., and Hoyer, P. B. (2012). 4-vinylcyclohexene diepoxide: A model chemical for ovotoxicity. *Syst. Biol. Reprod. Med.* 58 (1), 57–62. doi:10.3109/19396368.2011.648820

Kuang, H., Han, D., Xie, J., Yan, Y., Li, J., and Ge, P. (2014). Profiling of differentially expressed microRNAs in premature ovarian failure in an animal model. *Gynecol. Endocrinol.* 30 (1), 57–61. doi:10.3109/09513590.2013.850659

Lee, J. H., Lee, M., Ahn, C., Kang, H. Y., Tran, D. N., and Jeung, E. B. (2017). Parabens accelerate ovarian dysfunction in a 4-vinylcyclohexene diepoxide-induced ovarian failure model. *Int. J. Environ. Res. Public Health* 14 (2), 161. doi:10.3390/ijerph14020161

Li, J., Fan, S., Han, D., Xie, J., Kuang, H., and Ge, P. (2014). Microarray gene expression profiling and bioinformatics analysis of premature ovarian failure in a rat model. *Exp. Mol. Pathol.* 97 (3), 535–541. doi:10.1016/j.yexmp.2014.10.015

Maidarti, M., Anderson, R. A., and Telfer, E. E. (2020). Crosstalk between PTEN/PI3K/akt signalling and DNA damage in the oocyte: Implications for primordial follicle activation, oocyte quality and ageing. *Cells* 9 (1), 200. doi:10.3390/cells9010200

Makker, A., Goel, M. M., and Mahdi, A. A. (2014). PI3K/PTEN/Akt and TSC/mTOR signaling pathways, ovarian dysfunction, and infertility: An update. *J. Mol. Endocrinol.* 53 (3), R103–R118. doi:10.1530/JME-14-0220

Moolhuijsen, L. M., and Visser, J. A. (2020). Anti-müllerian hormone and ovarian reserve: Update on assessing ovarian function. *J. Clin. Endocrinol. Metab.* 105 (11), dgaa513–3373. doi:10.1210/clinem/dgaa513

Nie, X., Dai, Y., Zheng, Y., Bao, D., Chen, Q., Yin, Y., et al. (2018). Establishment of a mouse model of premature ovarian failure using consecutive superovulation. *Cell. Physiol. biochem.* 51 (5), 2341–2358. doi:10.1159/000495895

Okeke, T. C., Anyaehie, U. B., and Ezenyeaku, C. C. (2013). Premature menopause. *Ann. Med. Health Sci. Res.* 3 (1), 90–95. doi:10.4103/2141-9248.109458

Othman, N., and Nagoor, N. H. (2019). Overexpression of miR-361-5p plays an oncogenic role in human lung adenocarcinoma through the regulation of SMAD2. *Int. J. Oncol.* 54 (1), 306–314. doi:10.3892/ijo.2018.4602

Plakkot, B., Saju, S. S., and Kanakkaparambil, R. (2018). A review article on gonadotropins and their significant contribution in ovarian follicle development. *J. Pharm. Innovation* 7 (11), 433–438.

Podfigurna, A., Stellmach, A., Szeliga, A., Czyzyk, A., and Meczekalski, B. (2018). Metabolic profile of patients with premature ovarian insufficiency. *J. Clin. Med.* 7, 374. doi:10.3390/jcm7100374

Qi, Y., Zhu, Y. M., and Li, B. (2022). Comparison of animal models for premature ovarian insufficiency induced by different doses of cyclophosphamide: A network meta-analysis. *Mapp. Intimacies.* doi:10.21203/rs.3.rs-1243777/v1

Santoro, N. (2003). Mechanisms of premature ovarian failure. *Ann. Endocrinol.* 64 (2), 87–92.

Shelling, A. N. (2010). Premature ovarian failure. *Reproduction* 140 (5), 633–641. doi:10.1530/REP-09-0567

Sobinoff, A. P., Mahony, M., Nixon, B., Roman, S. D., and McLaughlin, E. A. (2011). Understanding the villain: DMBA-induced preantral ovotoxicity involves selective follicular destruction and primordial follicle activation through PI3K/akt and mTOR signaling. *Toxicol. Sci.* 123 (2), 563–575. doi:10.1093/toxsci/kfr195

Sobinoff, A. P., Nixon, B., Roman, S. D., and McLaughlin, E. A. (2012). Staying alive: PI3K pathway promotes primordial follicle activation and survival in response to 3MC-induced ovotoxicity. *Toxicol. Sci.* 128 (1), 258–271. doi:10.1093/toxsci/kfs137

Steiner, A. Z. (2013). Biomarkers of ovarian reserve as predictors of reproductive potential. *Semin. Reprod. Med.* 31 (06), 437–442. doi:10.1055/s-0033-1356479

Tamura, T., Yokoi, R., Okuhara, Y., Harada, C., Terashima, Y., Hayashi, M., et al. (2009). Collaborative work on evaluation of ovarian toxicity 2) Two-or four-week repeated dose studies and fertility study of mifepristone in female rats. *J. Toxicol. Sci.* 34, SP31–SP42. doi:10.2131/jts.34.s31

Thoma, M., Fledderjohann, J., Cox, C., and Adageba, R. K. (2021). "Biological and social aspects of human infertility: A global perspective," in *Oxford research encyclopedia of global public Health (Faculty of Arts & Social Sciences)* (New York: Oxford University Press).

Tian, Y., Chen, Y. Y., and Han, A. L. (2019). MiR-1271 inhibits cell proliferation and metastasis by targeting LDHA in endometrial cancer. *Eur. Rev. Med. Pharmacol. Sci.* 23 (13), 5648–5656. doi:10.26355/eurrev_201907_18300

Vayena, E., Rowe, P. J., and Griffin, P. D. (2002). *Current practices and controversies in assisted reproduction: Report of a meeting on medical, ethical and social aspects of assisted reproduction, held at WHO headquarters*. Geneva, Switzerland: World Health Organization.

Wang, Y., Liu, M., Johnson, S. B., Yuan, G., Arriba, A. K., Zubizarreta, M. E., et al. (2019). Doxorubicin obliterates mouse ovarian reserve through both primordial follicle atresia and overactivation. *Toxicol. Appl. Pharmacol.* 381, 114714. doi:10.1016/j.taap.2019.114714

Wright, L. E., Frye, J. B., Lukefahr, A. L., Marion, S. L., Hoyer, P. B., Besselsen, D. G., et al. (2011). 4-Vinylcyclohexene diepoxide (VCD) inhibits mammary epithelial differentiation and induces fibroadenoma formation in female Sprague Dawley rats. *Reprod. Toxicol.* 32 (1), 26–32. doi:10.1016/j.reprotox.2011.05.005

Yan, H., Zhang, J., Wen, J., Wang, Y., Niu, W., Teng, Z., et al. (2018). CDC42 controls the activation of primordial follicles by regulating PI3K signaling in mouse oocytes. *BMC Biol.* 16 (1), 73–16. doi:10.1186/s12915-018-0541-4

Yu, Y., Zhang, D., Huang, H., Li, J., Zhang, M., Wan, Y., et al. (2014). NF-κB1 p50 promotes p53 protein translation through miR-190 downregulation of PHLPP1. *Oncogene* 33 (8), 996–1005. doi:10.1038/onc.2013.8

Zhou, L., Xie, Y., Li, S., Liang, Y., Qiu, Q., Lin, H., et al. (2017). Rapamycin prevents cyclophosphamide-induced over-activation of primordial follicle pool through PI3K/Akt/mTOR signaling pathway *in vivo*. *J. Ovarian Res.* 10 (1), 56–11. doi:10.1186/s13048-017-0350-3

frontiers | Frontiers in Genetics

TYPE Review
PUBLISHED 13 March 2023
DOI 10.3389/fgene.2023.997383

OPEN ACCESS

EDITED BY
Panwen Wang,
Mayo Clinic Arizona, United States

REVIEWED BY
Teresa Colombo,
National Research Council (CNR), Italy
Tiziana Castrignanò,
University of Tuscia, Italy

*CORRESPONDENCE
Serghei Mangul,
✉ serghei.mangul@gmail.com

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

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

RECEIVED 18 July 2022
ACCEPTED 24 February 2023
PUBLISHED 13 March 2023

CITATION
Deshpande D, Chhugani K, Chang Y, Karlsberg A, Loeffler C, Zhang J, Muszyńska A, Munteanu V, Yang H, Rotman J, Tao L, Balliu B, Tseng E, Eskin E, Zhao F, Mohammadi P, P. Łabaj P and Mangul S (2023), RNA-seq data science: From raw data to effective interpretation. *Front. Genet.* 14:997383.
doi: 10.3389/fgene.2023.997383

RNA-seq data science: From raw data to effective interpretation

Dhrithi Deshpande[1†], Karishma Chhugani[1†], Yutong Chang[1], Aaron Karlsberg[2], Caitlin Loeffler[3], Jinyang Zhang[4], Agata Muszyńska[5,6], Viorel Munteanu[7], Harry Yang[8], Jeremy Rotman[2], Laura Tao[9], Brunilda Balliu[9], Elizabeth Tseng[10], Eleazar Eskin[3,9,11], Fangqing Zhao[4,12], Pejman Mohammadi[13], Paweł P. Łabaj[5,14] and Serghei Mangul[2,15]*

[1]Department of Pharmacology and Pharmaceutical Sciences, USC Alfred E. Mann School of Pharmacy and Pharmaceutical Sciences, Los Angeles, CA, United States, [2]Department of Clinical Pharmacy, USC Alfred E. Mann School of Pharmacy and Pharmaceutical Sciences, Los Angeles, CA, United States, [3]Department of Computer Science, University of California, Los Angeles, CA, United States, [4]Beijing Institutes of Life Science, Chinese Academy of Sciences, Beijing, China, [5]Małopolska Centre of Biotechnology, Jagiellonian University, Krakow, Poland, [6]Institute of Automatic Control, Electronics and Computer Science, Silesian University of Technology, Gliwice, Poland, [7]Department of Computers, Informatics and Microelectronics, Technical University of Moldova, Chisinau, Moldova, [8]Department of Microbiology, Immunology and Molecular Genetics, University of California Los Angeles, Los Angeles, CA, United States, [9]Department of Computational Medicine, David Geffen School of Medicine at UCLA, CHS, Los Angeles, CA, United States, [10]Pacific Biosciences, Menlo Park, CA, United States, [11]Department of Human Genetics, David Geffen School of Medicine at UCLA, Los Angeles, CA, United States, [12]Key Laboratory of Systems Biology, Hangzhou Institute for Advanced Study, University of Chinese Academy of Sciences, Hangzhou, China, [13]Department of Integrative Structural and Computational Biology, The Scripps Research Institute, La Jolla, CA, United States, [14]Department of Biotechnology, Boku University Vienna, Vienna, Austria, [15]Department of Quantitative and Computational Biology, USC Dornsife College of Letters, Arts and Sciences, Los Angeles, CA, United States

RNA sequencing (RNA-seq) has become an exemplary technology in modern biology and clinical science. Its immense popularity is due in large part to the continuous efforts of the bioinformatics community to develop accurate and scalable computational tools to analyze the enormous amounts of transcriptomic data that it produces. RNA-seq analysis enables genes and their corresponding *transcripts* to be probed for a variety of purposes, such as detecting novel exons or whole transcripts, assessing expression of genes and alternative transcripts, and studying alternative splicing structure. It can be a challenge, however, to obtain meaningful biological signals from raw RNA-seq data because of the enormous scale of the data as well as the inherent limitations of different sequencing technologies, such as *amplification bias* or *biases of library preparation*. The need to overcome these technical challenges has pushed the rapid development of novel computational tools, which have evolved and diversified in accordance with technological advancements, leading to the current myriad of RNA-seq tools. These tools, combined with the diverse computational skill sets of biomedical researchers, help to unlock the full potential of RNA-seq. The purpose of this review is to explain basic concepts in the computational analysis of RNA-seq data and define discipline-specific jargon.

KEYWORDS

RNA sequencing, transcriptome quantification, differential gene expression, high throughput sequencing, read alignment, bioinformatics

1 Introduction

High-throughput DNA sequencing technologies, including *next-generation sequencing* and the newly emerging *third-generation sequencing*, enable the gene sequences of living organisms to be probed in a cost-effective manner (Shendure and Ji, 2008). These sequencing technologies have also been adapted for RNA sequencing (RNA-seq), which enables the expression of various RNA populations, including mRNA and total RNA, to be detected and quantified. RNA-seq has reshaped biomedical research by expanding researchers' ability to analyze a vast range of biological data (Kukurba and Montgomery, 2015). To derive biological insights from RNA-seq data, researchers need to understand the steps involved in RNA-seq analysis and select appropriate tools to answer their research question.

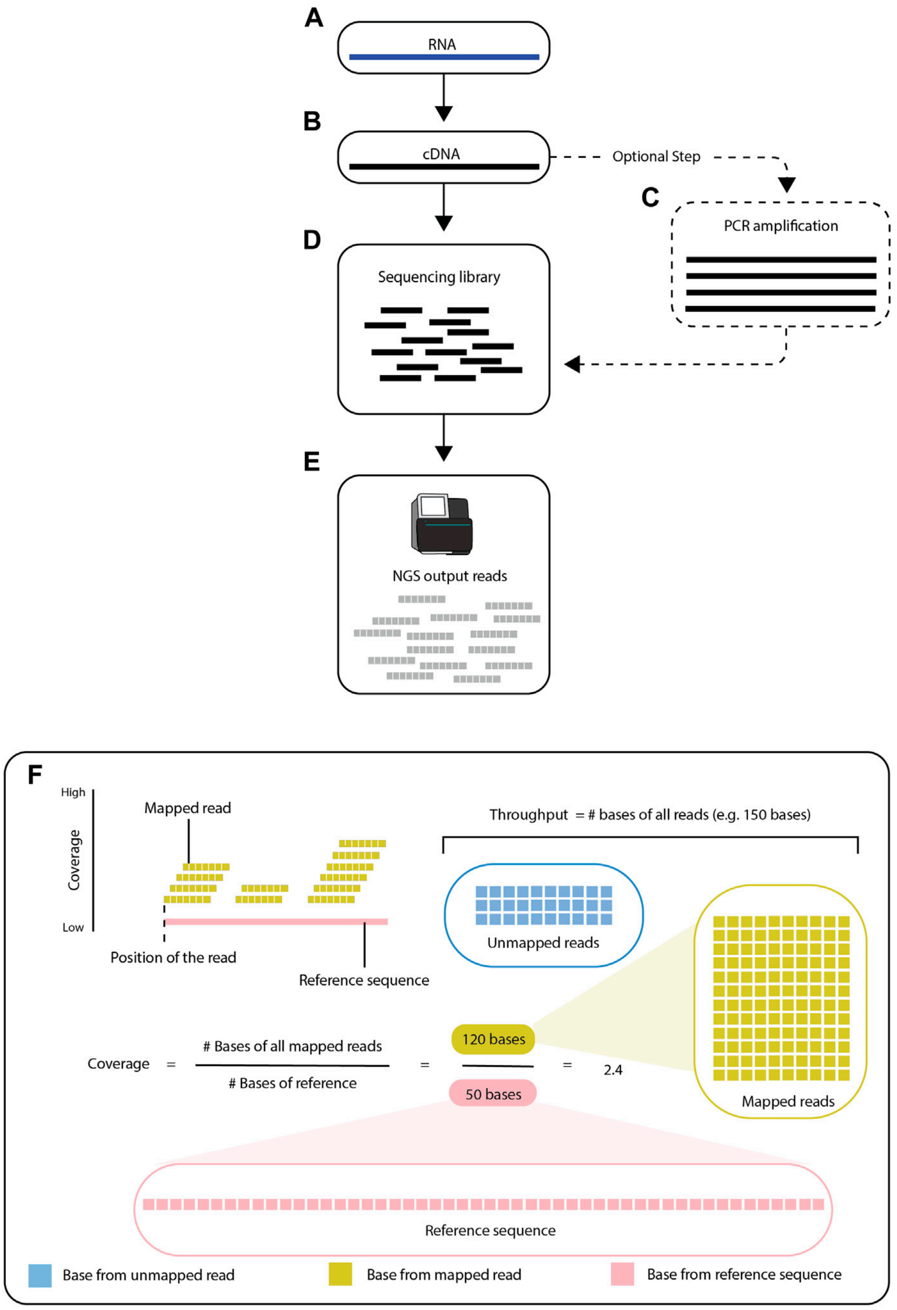


FIGURE 1
Overview of RNA-seq. RNA-seq is a process of creating short sequencing reads from RNA molecules. The steps consist of first converting the RNA **(A)** into cDNA **(B)**, then (optionally) amplifying the cDNA by PCR **(C)**, and finally fragmenting the cDNA into short pieces (known as fragments). After the sequencing library **(D)** is prepared, the fragments are used as input for next-generation sequencing **(E)**. The resulting sequence reads contained in FASTQ files are then aligned to a reference sequence **(F)**. Modern high-throughput sequencing machines can generate up to 150 million reads per run. The reference sequence, shown as a pink line, is known. The goal of the alignment is to find the *locus* in the reference sequence with the greatest match to each read. Reads are shown to align to the specific positions/locations and these mapped locations are recorded.

Biomedical researchers are often tasked with using computational methods for RNA-seq analysis, which are typically available wrapped as software tools and packages. In this review, we provide an overview of diverse methodologies for RNA-seq analyses that can be used to detect novel exons and transcripts, quantify gene expression and alternative splicing, and study alternative splicing structure. We discuss the steps from the generation of raw data using sequencing technologies to the effective interpretation and visualization of RNA-seq data using mapping and quantification techniques. By summarizing the biological and computational foundations of RNA-seq data generation, analysis, and software development, we hope this review will lead to a more deliberate use of existing computational tools.

2 RNA sequencing

RNA-seq uses high-throughput sequencing of nucleic acids to determine the nucleotide sequence of RNA molecules as well as the quantities of specific RNA species within populations of RNA molecules. RNA-seq analysis requires specialized computational tools that can account for the shortcomings of sequencing technologies, including the generation of *sequencing errors* (Le et al., 2013), *length biases* (Oshlack and Wakefield, 2009), and *fragmentation* (Tuerk et al., 2017). Computational analysis of RNA-seq data has led to many scientific advances, including novel therapeutic discoveries, detailed understanding of genetic regulatory regions, and identification of biomarkers and pathogenic mutations (Han et al., 2015).

Preparation of an RNA-seq library starts with extraction and isolation of RNA from a biological sample, such as a cell line or a frozen tissue sample. For RNA-seq performed with short-read sequencing (see Section 2.1), the isolated RNA is reverse-transcribed and converted into *cDNA*, which is then amplified by *polymerase chain reaction (PCR)* and fragmented into short sequences (either before or after PCR) (Prakash and Haeseler, 2017) (Figure 1). After the RNA molecules are processed, the RNA-seq library becomes the input for a sequencing platform (Kukurba and Montgomery, 2015), which generates reads (i.e., the sequenced fragments from the RNA-seq library).

2.1 High-throughput RNA-seq technologies

High-throughput sequencing techniques can derive millions of nucleotide sequences from an individual *transcriptome* (Stark et al., 2019). These nucleotide sequences provide multifold coverage of the whole transcriptome. High-resolution RNA-seq can identify which genes are actively transcribed in a sample and quantify the levels at which alternative transcripts of a gene are transcribed (Gerstein et al., 2007). The reads generated by different sequencing technologies have lengths ranging from hundreds of base pairs (usually referred to as short reads) to thousands of base pairs (referred to as long reads) (Shendure and Ji, 2008; Haas and Zody, 2010; Pollard et al., 2018). Illumina, Nanopore, and PacBio are among the most commonly used high-throughput sequencing platforms (Ye et al., 2015).

Illumina sequencing, considered a next-generation sequencing technology, is based on sequencing-by-synthesis chemistry and was first commercialized in 2006 (Shen and Shen, 2019). For Illumina RNA-seq, isolated RNAs are reverse-transcribed into single-stranded cDNA, which is then ligated to synthetic adapters, immobilized on a solid surface, and amplified by PCR. Then, a reaction mixture is added containing primers, DNA polymerase, and modified nucleotides. The modified *nucleotides have a fluorescent label* that serves as both a reversible terminator of DNA synthesis and an indicator of which nitrogenous base the nucleotide contains. As a new strand of DNA is synthesized using the immobilized cDNA as a template, each incorporated nucleotide is detected with a *charge-coupled device (CCD)* camera and identified by the color of the fluorescent label. The fluorescent label is then removed, and the next nucleotide is added in a new round of DNA synthesis. This cycle is repeated until each base in the cDNA is identified. The sequences of more than 10 million cDNA fragments can be simultaneously determined in parallel using the Illumina platform, giving rise to higher sequencing throughput compared with other sequencing platforms (Morganti et al., 2019; Workflows for RNA Sequencing, 2023).

Nanopore sequencing, which serves as the basis for the MinION, GridIOn, and PromethION platforms, was first introduced in 2014 by Oxford Nanopore Technologies. Nanopore sequencing can produce short or long reads from native DNA and RNA fragments of any length. Nanopores are very small holes in a membrane that can be created by pore-forming proteins or by non-biological means. The Nanopore sequencing method simultaneously sends an ionic current and a single strand of DNA or RNA through a nanopore. As the ionic current passes through each nucleotide that successively occupies the nanopore, it undergoes disruptions that are unique to the nitrogenous base. The patterns of disruption in the current can be interpreted to identify each base in the DNA or RNA strand that passes through the nanopore. Whereas short-read sequencing technologies such as Illumina require chemical modification or PCR amplification, Nanopore technology is capable of sequencing DNA or RNA without these additional steps, making it a third-generation sequencing technology (Bharagava et al., 2019).

PacBio sequencing, also known as SMRT (single-molecule, real-time) sequencing, was introduced in 2010 and generates full-length cDNA sequences (i.e., long reads) that characterize transcripts of targeted genes or across entire transcriptomes. Long reads generated by PacBio are accurate at the scale of a single molecule because they are generated by a process of circular consensus sequencing, in which the same cDNA is effectively read many times (Eid et al., 2009; Vierra et al., 2021). The comparatively high sensitivity of PacBio can be limited by external factors. For example, PacBio can produce full-length cDNA during the library preparation step; however, it can only generate high-quality reads if the target cDNA is short enough to be sequenced in multiple passes.

Each sequencing technology has inherent advantages and limitations, so no technology is best suited for all types of RNA-seq analysis (Box 1). Short-read technologies can generate data with a lower error rate and higher throughput than long-read technologies; however, the short-read length makes reconstruction and quantification of the transcriptome challenging (Korf, 2013; The RGASP Consortium et al., 2013a; The RGASP Consortium et al., 2013b; Amarasinghe et al., 2020). Long-read sequencing improves the accuracy of assembly (concatenation of individual reads to reassemble the transcriptome), or can even eliminate the need for assembly, as each read can cover an entire transcript. Long-read sequencing can also be used to produce complete, unambiguous information about

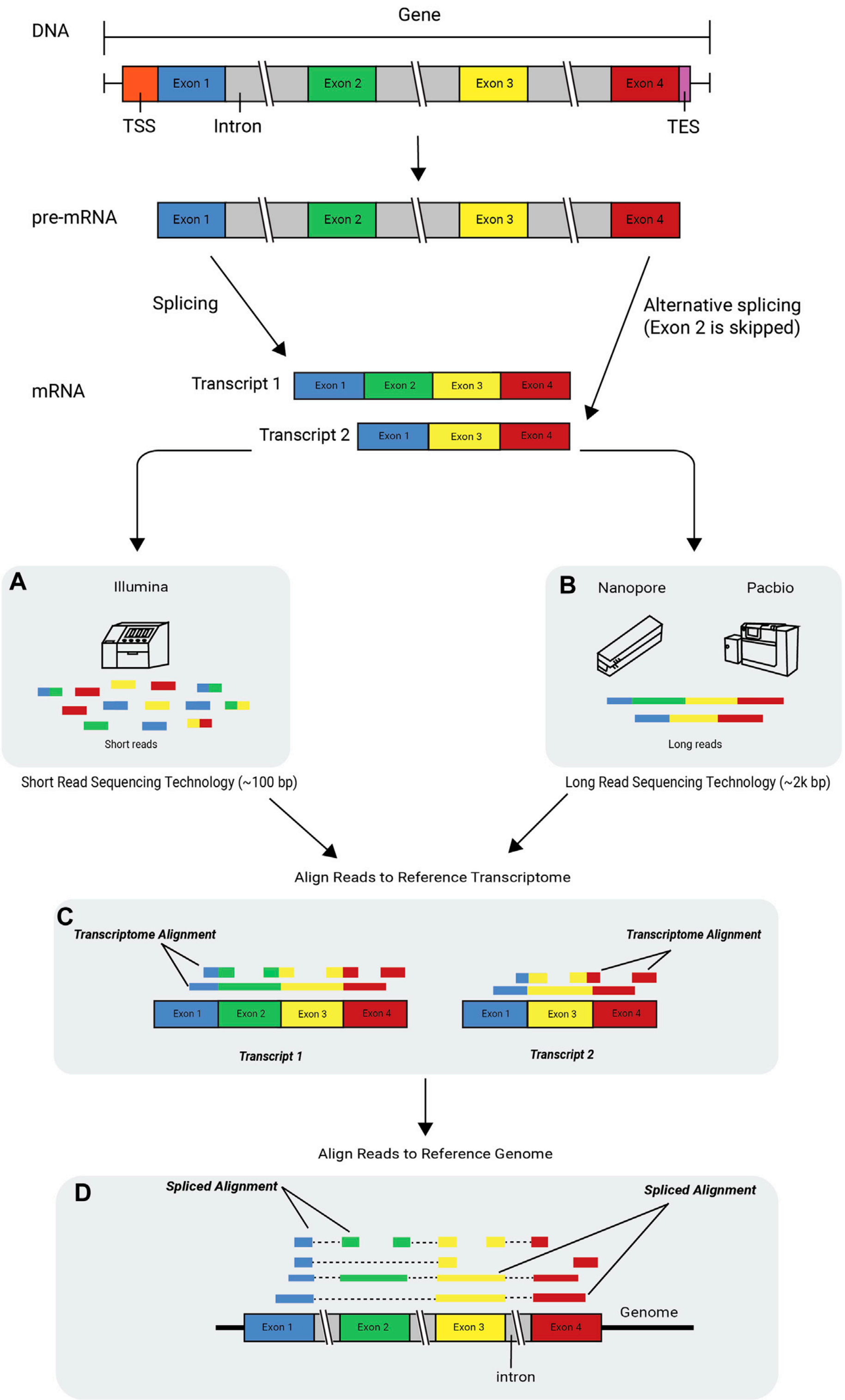


FIGURE 2
Alternative splicing and RNA-seq technologies. The flow of genetic information begins with DNA, which consists of introns and exons. DNA is transcribed into pre-mRNA and then further processed into mature mRNA by splicing out the introns and leaving the exons glued together. The mRNA is then translated into a protein. Transcripts with different arrangements of exons can be formed in a process called alternative splicing or exon skipping. An RNA-seq read is a short sequence sampled from a transcript. Reads are generated using sequencing technologies such as **(A)** the Illumina platform, which produces short reads, and the **(B)** Nanopore and PacBio platforms, which produce long reads. The figure depicts two scenarios in which uniquely mapped reads are aligned to a reference transcriptome **(C)** and a reference genome **(D)**, respectively. A few of the reads are multicolored, indicating that when aligned, they span across an exon-exon junction. Some of the shorter reads (single-colored) are aligned only to a single exon and do not span across the junction. *TSS*, transcription start site; *TES*, transcription end site.

Box 1 | Advantages and limitations of short and long reads

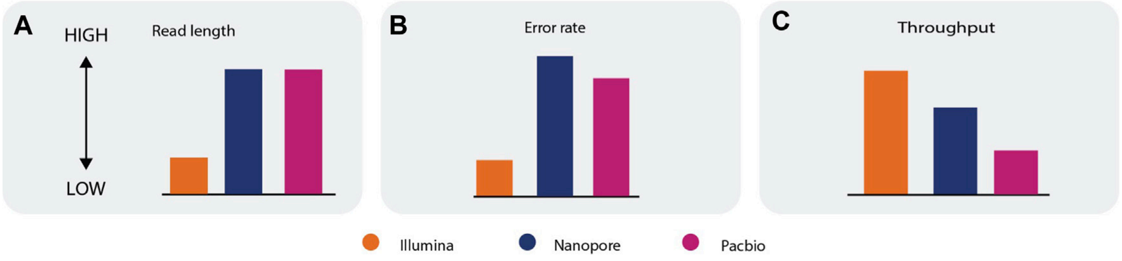
i. **Error rate**—Short read sequencing technologies have a lower error rate when compared to long read sequencing technologies **(a, b)**.

ii. **Throughput**—The throughput of long read sequencing technologies is typically lower than the throughput of short read sequencing technologies **(c)**.

iii. **Alignment**—Short reads suffer from multi-mapping issues, whereas longer reads, by nature of having more information, can be more accurately mapped to its origin. Due to a high error rate, pairwise alignment between the read, the reference transcriptome, and/or genome is more challenging for long reads compared to short reads.

iv. **Assemble novel transcripts**—Longer reads are preferred for *de novo* assembly, because they make the assembly step efficient. Most short reads do not span the shared region or shared exon junction, making the assembly step ambiguous. Full-length transcript sequencing eliminates the need for assembly.

v. **Estimate transcripts and gene expression**—Shorter reads are preferred for quantification of transcripts due to their higher throughput. However, assigning short reads to the transcripts requires more advanced probabilistic and statistical approaches. Longer reads have lower throughput, but they can usually cover the entire transcript and make determination of the transcript for each read a straightforward process.



alternative splicing, gene structure, regulatory elements, and coding regions. Long-read sequencing currently has a higher error rate and lower throughput compared with short-read sequencing, however (Figure 2) (Sedlazeck et al., 2018; De Maio et al., 2019; Mahmoud et al., 2019). Hybrid approaches that combine long reads and short reads can eliminate the limitations of each separate approach and can be used to accurately quantify and assemble known and novel transcripts (De Maio et al., 2019; Amarasinghe et al., 2020; Berbers et al., 2020), but they also have higher costs and more material requirements. Data gathered using Illumina, Nanopore, and PacBio sequencing technologies can be used to address a wide range of research areas, including transcriptome analysis, population-scale analysis, and clinical research (Wang et al., 2021).

3 RNA-seq data science: From raw data to effective interpretation

RNA-seq is multifaceted and can be used to uncover and expound new insights on, for example, a dysregulated gene or defective protein that has a downstream effect leading to a disease state (Costa et al., 2010). Computational analysis of RNA-seq data is central to decoding the biological complexities in the transcriptomes of living organisms, including humans (Costa et al., 2010). Here, we describe the major steps of computational analysis of RNA-seq data, beginning from the processing of raw data to the uncovering of biological insights.

3.1 Quality control of raw data

During the sequencing process, errors are introduced into reads that can bias the results of downstream analyses. Read trimming and data quality control to filter and assess the quality of raw reads (Yang et al., 2013) are therefore essential after the reads have been generated. Read trimming removes adapter sequences and portions of reads with low accuracy, as indicated by a low *PHRED quality score* (Martin, 2011; Dodt et al., 2012; Bolger et al., 2014). In addition, computational error correction can be applied to reduce the number of sequencing errors (Lima et al., 2020; Mitchell et al., 2020).

4 Read alignment

Read alignment is an essential step in RNA-seq downstream analysis. RNA-seq data typically lack information about the order and origin of the reads, including the specific part, homolog, or strand of the *genome* from which they originate. Computational alignment of the reads to an annotated reference transcriptome can establish where on the genome the reads originated (Figure 1) (Brown, 2002). Alignment of the reads to a reference sequence also reveals how many reads overlap each position on the reference sequence, which is known as the coverage. There are several bioinformatics tools (e.g., GenomeScope (Vurture et al., 2017), Smudgeplot (Ranallo-Benavidez et al., 2020), and Merqury (Rhie et al., 2020)) that can estimate the coverage without mapping the reads to a reference sequence (Ranallo-Benavidez et al., 2020; Rhie et al., 2020), as most of the overlap between reads is preserved with or without the reference sequence (Vurture et al., 2017) (Figure 1).

Alignment of RNA-seq reads to a complementary reference sequence can help determine which transcripts are expressed and the degree to which they are expressed, but the alignment approach is ill-equipped to discover transcripts that are missing from the reference sequence. Furthermore, even the human reference transcriptome remains incomplete (Nellore et al., 2016). Novel transcripts can be discovered by performing *de novo* assembly of RNA-seq reads to generate an entire transcriptome without alignment to a reference sequence; however, this can be challenging and requires large amounts of computational time and resources (Grabherr et al., 2011) As an alternative, RNA-seq reads can be aligned to curated databases of known transcripts such as RefSeq (Pruitt et al., 2007), UCSC genome browser, Ensembl, GENCODE (GENCODE, 2022), and AceView (Larsson et al., 2005), and reads that fail to align to known transcripts can then be aligned to a reference genome to identify novel transcripts.

One computational challenge in aligning RNA-seq reads to a reference genome is the handling of spliced junctions, where one part of the read maps to the end of one exon and the rest of the read maps to another exon, which may be located thousands of base pairs away from the first exon. Spliced junctions are the result of the removal of non-coding parts of a gene, called introns, and the splicing together of the coding parts of the gene, called exons. Genes can generate multiple mRNA transcripts through alternative splicing. As a result, exons are combined or skipped in different ways and have alternative start/end sites. These varying combinations create different transcripts, known as *isoforms*, from the same gene. As a biological process, alternative splicing is evolutionarily advantageous, because it enables the production of different protein variants from the same genetic information (Figure 2). When genome annotations are available, existing exon structures can be used to map reads across known splice junctions; however, this knowledge-guided approach may be biased towards mapping only known junctions while failing to discover novel ones.

In cases where reads align to multiple transcripts, it might not be possible to discern from which transcript the reads originate. Splice alignment software packages (Wang et al., 2010; Dobin et al., 2013; Kim et al., 2019) are designed to minimize multi-mapping by correctly aligning reads across the exon–intron junctions of the reference genome (Figure 2). This can be a crucial first step of reference-guided assembly, wherein transcripts that are present in the sample but not annotated in the reference are assembled using the spliced read alignments to the reference.

In some instances, reads do not perfectly align with the reference sequence but instead contain mismatches, which can be caused either by sequencing errors or by biological variation such as mutations (Mitchell et al., 2020). RNA-seq alignment tools are typically equipped with a customizable threshold for tolerating mismatches in the alignment; however, it is important to distinguish between sequencing errors and real variation between the transcripts and the reference sequence. Specialized computational tools (Abate et al., 2014; Fernandez-Cuesta et al., 2015) can identify and classify genes using strategies such as *de novo* assembly (assembly of reads without alignment to a reference sequence), identification of reads that span fusion junctions, and filtering of gene fusion candidates based on various criteria.

5 Quantitative analysis of gene expression

RNA-seq enables quantitative analysis of gene expression at the level of alternative transcripts. The sequence fragments derived from mRNA can reveal which genes are expressed and how strongly they are expressed. Additionally, differential expression (DE) analysis can show how expression levels change under different conditions or between different populations.

5.1 Estimation of transcript and gene expression

Computational methods can estimate expression levels of genes and transcripts by counting the number of reads that match individual reference transcripts. Tools like HT-Seq-count, Rcount, and featureCounts (Liao et al., 2014; Anders et al., 2015; Schmid and Grossniklaus, 2015) are highly robust and widely used for such analyses; however, counting-based tools are ill-equipped to estimate the expression levels of different isoforms of expressed genes using short reads, as the majority of isoforms share a large percentage of exons and cannot be uniquely assigned to individual transcripts (Figure 2). The shorter the reads, the greater the probability that they will match multiple transcripts. A conservative approach to tackle this challenge is to consider only the reads that uniquely map to a single transcript (e.g., reads that map to transcript-specific splicing junctions or exons) (Conesa et al., 2016). An alternative approach that utilizes a larger fraction of the RNA-seq reads is to probabilistically assign reads to the isoforms from which they likely originated (Li and Dewey, 2011; Nicolae et al., 2011; Trapnell et al., 2012; Pertea et al., 2015).

A number of approaches quantify gene expression using complete read alignment, which requires large amounts of computational power and time to compare each read to reference sequences base-by-base. Pseudoalignment methods have been developed as an alternative approach that has a much smaller computational burden. These methods forgo the base-by-base accuracy of alignment and determine an approximate alignment of the reads on the genome, which is still sufficiently accurate to quantify gene expression. Pseudoalignment algorithms leverage a pre-compiled library of unique k-mers (exact substrings of length k) contained in known transcripts and assign reads to transcripts by counting the k-mer occurrences in the reads, thus achieving up to 100 times faster quantification compared with alignment-based methods (Bray et al., 2016). Sailfish (the pioneer of pseudoalignment) (Patro et al., 2014), Salmon (Patro et al., 2017), and Kallisto (Bray et al., 2016) each utilize pseudo-alignment-based algorithms to quantify the isoforms of expressed transcripts (Alser et al., 2020), each providing comparable accuracy in expression quantification. A more detailed explanation of these tools can be found in Supplementary Material S2.

5.2 Differential gene expression analysis

After gene and transcript expression levels are estimated, statistical approaches are employed to detect differences in expression levels across experimental groups (e.g., different sexes or cohorts exposed to different environmental conditions) (Conesa et al., 2016). Expression levels measured for the same gene under different conditions cannot be directly compared, as each experiment represents a statistical sample, giving only the relative mRNA levels in comparison to the other mRNAs present in the sample. In addition, mRNA levels change over time, and reads can align to multiple places, making exact quantitation difficult. The purpose of statistical testing is to ensure that an observed change in mRNA levels is due to an actual difference in expression between experimental conditions.

To test whether the expression of a given gene is different between two groups, measurements are repeated in multiple replicates of the same experiments, and then a statistical test is applied. Through this process, the variation in expression between different conditions can be compared to the variation within replicates of the same condition. Each statistical test is based on a null hypothesis that the gene expression is the same between groups, which is usually true for the majority of genes. The value that indicates whether there is likely to be a true difference between groups is called the *p*-value, which gives the probability of observing a particular difference, or a more extreme difference, assuming that the null hypothesis is true. Small *p*-values give strong evidence against the null hypothesis. Genes with low *p*-values are considered to be differentially expressed, and the null hypothesis is rejected for those genes. The typical threshold for rejection of a null hypothesis is a *p*-value less than 0.05, but this cutoff is arbitrary and might need to be altered depending on how noisy the data are (Liu et al., 2006; Glaus et al., 2012; Shastry et al., 2020).

There are two types of error associated with statistical tests: Type I error and Type II error. A Type I error occurs if a test rejects a true null hypothesis. A Type II error occurs if a test accepts a false null hypothesis. The *p*-value indicates the probability of making a Type I error in a given test. For example, if the *p*-value threshold is set at 0.05 (i.e., 5%), and 20,000 genes are being tested, then 1,000 genes (5% · 20,000) will be wrongly considered to be differentially expressed because of Type I errors. There are two approaches to control Type I errors, also referred to as false positives. One approach is to control the family-wise error or the probability that there is at least one Type I error among all the rejected null hypotheses. The other approach is to control the false discovery rate, or the proportion of Type I errors among all the rejected null hypotheses. Both approaches involve calculation of an adjusted *p*-value (p-adj) for each gene, which can then be used for further analysis (Jafari and Ansari-Pour, 2018).

It is important to account for *noise* which includes sources of variation that are unrelated to the experimental variable of interest, when performing differential expression analysis. For example, *batch effects*, or confounding factors arising from samples being tested on different days, by different laboratory technicians, or in different laboratories (technical batch effects), can result in unwanted differences in measured values. In addition, variation due to intrinsic factors such as high GC content or gene body coverage evenness (biological batch effects) can affect the quantification of technical replicates of a sample. Existing statistical methods can effectively detect and adjust for hidden confounding factors (Li et al., 2014).

Other approaches to differential expression analysis that can produce more accurate results than conventional p-adj values use different metrics such as the minimum significant difference or the generalized linear model (GLM) framework (McCarthy et al., 2012), where a combination of *p*-values and log fold changes is applied to identify the genes or transcripts with the most significant differences in expression. Another alternative approach is the probability of positive log ratio (PPLR) (Liu et al., 2006), which was initially developed for microarray analysis and subsequently adjusted for RNA-seq data (Glaus et al., 2012). The PPLR uses a Bayesian hierarchical model to express the probability that the ratio of expression levels between two conditions is positive (i.e., the expression is upregulated in the second condition relative to the first). A PPLR value close to 1 means there is a very high probability that a given transcript is upregulated in the second condition (Liu et al., 2006). When the PPLR value is close to 0, there is a very low probability of upregulation, and consequently a high probability of downregulation, in the second condition relative to the first. There is no direct relation between PPLRs and *p*-values, as they look at the problem from different perspectives (i.e., in the probabilistic approach an uncertainty propagation between successive stages of analysis is possible and desired). Both approaches are capable of identifying large numbers of differentially expressed genomic features. If the number of differentially expressed features is too large, a more stringent cutoff for statistical significance can be applied to make the analysis more manageable.

Depending on the type of *normalization* performed on RNA-seq data, machine-learning approaches can be used to identify differentially expressed genes with classification models based on discrete or continuous distributions. Machine learning approaches have been used to manage, model, and categorize biological data, enabling high-impact discoveries in the field of biomedicine (Shastry et al., 2020). RNA-seq data are discrete in nature. The two most common ways to normalize RNA-seq data for machine learning-based differential expression analysis are to model the data as a Poisson or negative binomial distribution or transform the data to be similar to a distribution of microarray data. The Bioconductor MLSeq (Goksuluk et al., 2019) package is a comprehensive source of combinations of different normalization and machine-learning methods for RNA-seq analysis. After the data are normalized, genes or alternative transcripts (features) can be ranked, or standard sample classification can be performed, and the features that make the strongest contributions to the assignment of samples to particular groups can be extracted (Goksuluk et al., 2019). With a deep learning approach, it is also possible to predict differences in gene expression from histone modification signals (Sekhon et al., 2018).

Differential expression analysis can be complemented by *expression quantitative trait loci (eQTL)* analysis, which formally compares the expression levels of a given gene between groups with different copy numbers (0, 1, or 2) of the minor allele. Each read alignment technique produces different results, which may impact which genes are identified as differentially expressed (Castel et al., 2015). The power to detect differentially expressed genes and eQTLs depends on the sequencing depth of the sample, the minor allele frequency of the gene being tested, the expression level of the gene, and the length of the gene (McKenna et al., 2010). The magnitude of the eQTL can be quantified by the log allelic fold change (Hu et al., 2015), and its significance is tested using a binomial distribution or over-dispersed generalizations (Kumasaka et al., 2016; Knowles et al., 2017; Mohammadi et al., 2019; Zou et al., 2019; Wang et al., 2020). Some of the popular approaches to detect eQTLs use transformation and linear regression models (Shabalin, 2012; Ongen et al., 2016; Taylor-Weiner et al., 2019).

The results of differential expression analyses can be validated using independent techniques such as *quantitative PCR (qPCR)*, which is statistically assessable (Skelly et al., 2011). Measurements of gene expression obtained by qPCR are relatively similar to measurements obtained by RNA-seq analysis, where a value can be calculated for the concentration of a target region in a given sample (Harvey et al., 2015; Romanel et al., 2015; Xie et al., 2019). Additional information about quantification of RNA splicing and splicing QTL (sQTL) analyses can be found in Supplementary Material S3.

6 Measurement of allele-specific expression

RNA-seq can measure allele-specific expression (ASE or allelic expression) to uncover the cis-regulatory effects of genetic variants (McKenna et al., 2010; Castel et al., 2015; Raghupathy et al., 2018). ASE represents gene expression measured independently for the paternal and maternal alleles of a gene. In a typical RNA-seq experiment, ASE can be measured only in genes that contain a heterozygous *single-nucleotide polymorphism (SNP)* within the transcribed region. This SNP, referred to as the aseSNP, can be used as a tag to identify reads that originate from each copy of the gene (Figure 3).

Allelic imbalance—the ratio between paternal and maternal allele expression—identifies genetic cis-regulatory differences between two haplotypes. The log allelic fold change can also be calculated to quantify the magnitude of allelic imbalance (Hu et al., 2015). An aseSNP is not itself a regulatory variant and should not induce an imbalanced ASE signal. However, there can be a bias in ASE data that falsely suggests that the haplotype carrying the reference allele for the aseSNP has slightly higher expression across all genes. This issue, known as allelic bias or reference bias, can be mitigated in two ways: by aligning the RNA-seq reads to a personalized reference genome that excludes likely biased sites (Dobin et al., 2013; van de Geijn et al., 2015; Gao and Zhao, 2018; Kristensen et al., 2019; Ferraro et al., 2020), or by aggregating the ASE signal from multiple aseSNPs in each gene (Chen et al., 2021). ASE data can also be used to improve statistical power for identifying eQTLs (Gao and Zhao, 2018; Kristensen et al., 2019; Zou et al., 2019; Ferraro et al., 2020) and to map the causal regulatory variants in eQTL data (Kim and Salzberg, 2011; Gao et al., 2018; Haas et al., 2019). Furthermore, ASE data are inherently robust to noise, so they are useful for identifying gene-by-environment interaction effects (Li, 2013) or the effects of rare genetic variants on gene expression to improve diagnostic accuracy for Mendelian diseases (Hoffmann et al., 2014; Ji et al., 2019).

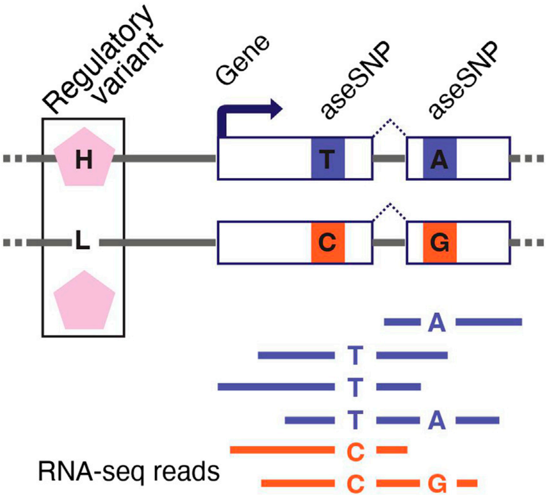


FIGURE 3
Measuring allele-specific expression with RNA-seq. RNA-seq can be used to generate allele-specific expression (ASE) data for genes with a heterozygous single-nucleotide polymorphism in the transcribed region (aseSNP). The aseSNP enables sequencing reads to be mapped to the haplotype from which they originate. Imbalance in ASE data is a functional indicator of a cis-regulatory difference between the two haplotypes that is driven by heterozygous regulatory variants. Data from multiple aseSNPs can be aggregated to improve ASE data quality. The non-coding regulatory variant depicted here has two alleles inducing higher (H) and lower (L) expression of the target gene.

7 Profiling circular RNA with RNA-seq

Circular RNA (circRNA) is a large class of RNA molecules with a covalently closed circular structure that plays important roles in various biological processes and metabolic mechanisms (Wu et al., 2020). In recent years, a variety of computational tools have been developed for circRNA study (Gao and Zhao, 2018; Chen et al., 2021). Identification of circRNAs is based on detection of reads spanning the circle junction, termed the back-splice junction (BSJ). Most tools (Cheng et al., 2016; Zhang et al., 2016; Gao et al., 2018) employ aligners (Humphreys et al., 2019; Wu et al., 2019; Zheng et al., 2019) to detect putative back-splicing events from fusion reads or split alignment results, whereas other splice-aware aligners (Wang et al., 2010; Zheng and Zhao, 2020) can align circular reads and detect BSJs directly.

Considering that most circRNAs are derived from exonic regions (Ji et al., 2019; Wu et al., 2020) where computational methods cannot accurately distinguish linear and circular reads, the BSJ read count is the most reliable measurement of circRNA expression levels. The BSJ read count is inferred from alignment results, and different filters and statistical strategies have been employed to improve its accuracy and sensitivity (Mangul et al., 2019; Zhang et al., 2020). Alternative approaches using pseudoalignment-based tools for circRNA quantification (Li et al., 2017) can substantially increase the computational efficiency compared with regular alignment-based methods. To compare the expression levels of circRNAs and their host genes, the junction ratio, defined as the ratio of BSJ reads and linear reads mapped to the BSJ site, is often used for comparative analysis. Several computational methods have been developed for accurate estimation of junction ratios (Reimers and Carey, 2006; The Comprehensive R Archive Network, 2022). In addition, circRNAs exhibit alternative splicing patterns, and a number of specific tools have been developed for circular transcript assembly (Gao et al., 2016; Zhang et al., 2016; Wu et al., 2019; Zheng et al., 2019), internal structure visualization (Li et al., 2016; Mose et al., 2016), and differential expression analysis (Zhang et al., 2020; The Comprehensive R Archive Network, 2022). Several comprehensive databases have been constructed for circRNA annotation and prioritization analysis (Dong et al., 2018; Xia et al., 2018; Wu et al., 2020).

8 Discussion

As technology advances, RNA-seq methods have become increasingly popular and have revolutionized modern biology and clinical applications, driven by continuous efforts of the bioinformatics community to develop accurate and scalable computational tools. In addition, advancements in sequencing technologies have provided an unprecedented ability to analyze a wide range of biological data, enabling new explorations of novel and existing biological problems. To increase access to RNA-seq methods among new users and young scientists, we provided an overview of the fundamentals of RNA-seq and its associated computational methods and discussed the advantages and limitations of various applications.

Computational analysis of RNA-seq data can be used to tackle important biological problems such as estimating gene expression profiles across various phenotypes and conditions or detecting novel alternative splicing on specific exons. Specialized analyses of RNA-seq data can also help to detect changes in the concentration, function, or localization of transcription factors that affect splicing and can cause the onset of neurodegenerative diseases and cancers (Ozsolak and Milos, 2011; Szabo and Salzman, 2016). Some recently developed computational tools (Xu et al., 2014; Bolotin et al., 2015; Li et al., 2016; Mose et al., 2016; Mandric et al., 2020) are even capable of repurposing RNA-seq data to characterize the individual adaptive immune repertoire and microbiome (Varadhan and Roland, 2008). Additionally, computational deconvolution can be applied to RNA-seq data to study cell-type compositions in tissue samples (Melsted et al., 2017; Kang et al., 2019).

The interdisciplinary nature of RNA-seq applications and related analytic methods and software development introduces a host of terms that can challenge researchers in the wider scientific and medical research communities. The literature on RNA-seq methods has traditionally assumed that readers are familiar with the fundamental concepts of RNA-seq and related bioinformatics analyses (Nariai et al., 2013; Srivastava et al., 2016; Zakeri et al., 2017; Green et al., 2018; Li et al., 2018; Vaquero-Garcia et al., 2018). These methods may require diverse computational skills to be used effectively. A lack of computational skills can therefore limit the ability of biomedical researchers to unlock the full potential of RNA-seq, highlighting the need for a review that explains basic RNA-seq concepts and defines discipline-specific jargon.

Author contributions

SM conceived of the idea presented and supervised the project. DD, KC and SM led the project. DD, KC, YC, AK, CL, JZ, AM, VM, HY, JR, LT, BB, ET, EE, FZ, PM, PL and SM contributed to the writing ofthe manuscript. DD, KC, YC, and PM produced figures in the main text. KC and DD created Supplementary Materials and the box. All authors discussed the text and commented on the manuscript.All authors read and approved the final manuscript.

Funding

Serghei Mangul's work has been supported by the National Science Foundation Grants (2041984 and 2135954). P. Mohammadi's work has been supported by the National Institute of General Medical Sciences (R01GM140287). Fangqing Zhao's work has been supported by the National Natural Science Foundation of China (31722031, 91640117). Agata M's work has been co-funded by the European Union through the European Social Fund grant (POWR.03.02.00-00-I029).

Acknowledgments

We thank Dr. Patro for his valuable comments and suggestions on the manuscript.

Conflict of interest

ET was employed by the company Pacific Biosciences (United States).

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

Publisher's note

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

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2023.997383/full#supplementary-material

References

Abate, F., Zairis, S., Ficarra, E., Acquaviva, A., Wiggins, C. H., Frattini, V., et al. (2014). Pegasus: A comprehensive annotation and prediction tool for detection of driver gene fusions in cancer. *BMC Syst. Biol.* 8, 97. doi:10.1186/s12918-014-0097-z

Alser, M., Rotman, J., Deshpande, D., Taraszka, K., Shi, H., Baykal, P. I., et al. (2020). Technology dictates algorithms: Recent developments in read alignment. *Genome Biol.* 22, 249. doi:10.1186/s13059-021-02443-7

Amarasinghe, S. L., Su, S., Dong, X., Zappia, L., Ritchie, M. E., and Gouil, Q. (2020). Opportunities and challenges in long-read sequencing data analysis. *Genome Biol.* 21, 30. doi:10.1186/s13059-020-1935-5

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

Berbers, B., Saltykova, A., Garcia-Graells, C., Philipp, P., Arella, F., Marchal, K., et al. (2020). Combining short and long read sequencing to characterize antimicrobial resistance genes on plasmids applied to an unauthorized genetically modified Bacillus. *Sci. Rep.* 10, 4310. doi:10.1038/s41598-020-61158-0

Bharagava, R. N., Purchase, D., Saxena, G., and Mulla, S. I. (2019). Applications of metagenomics in microbial bioremediation of pollutants. *Microb. Divers. Genomic Era* 2019, 459–477. doi:10.1016/B978-0-12-814849-5.00026-5

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

Bolotin, D. A., Poslavsky, S., Mitrophanov, I., Shugay, M., Mamedov, I. Z., Putintseva, E. V., et al. (2015). MiXCR: Software for comprehensive adaptive immunity profiling. *Nat. Methods* 12, 380–381. doi:10.1038/nmeth.3364

Bray, N. L., Pimentel, H., Melsted, P., and Pachter, L. (2016). Near-optimal probabilistic RNA-seq quantification. *Nat. Biotechnol.* 34, 525–527. doi:10.1038/nbt.3519

Brown, T. A. (2002). *Understanding a genome sequence. Genomes*. 2nd edition. Hoboken, NJ, USA: Wiley-Liss.

Castel, S. E., Levy-Moonshine, A., Mohammadi, P., Banks, E., and Lappalainen, T. (2015). Tools and best practices for data processing in allelic expression analysis. *Genome Biol.* 16, 195. doi:10.1186/s13059-015-0762-6

Chen, L., Wang, C., Sun, H., Wang, J., Liang, Y., Wang, Y., et al. (2021). The bioinformatics toolbox for circRNA discovery and analysis. *Brief. Bioinform.* 22, 1706–1728. doi:10.1093/bib/bbaa001

Cheng, J., Metge, F., and Dieterich, C. (2016). Specific identification and quantification of circular RNAs from sequencing data. *Bioinformatics* 32, 1094–1096. doi:10.1093/bioinformatics/btv656

Conesa, A., Madrigal, P., Tarazona, S., Gomez-Cabrero, D., Cervera, A., McPherson, A., et al. (2016). A survey of best practices for RNA-seq data analysis. *Genome Biol.* 17, 13. doi:10.1186/s13059-016-0881-8

Costa, V., Angelini, C., De Feis, I., and Ciccodicola, A. (2010). Uncovering the complexity of transcriptomes with RNA-seq. *Biomed. Res. Int.* 2010, e853916. doi:10.1155/2010/853916

De Maio, N., Shaw, L. P., Hubbard, A., George, S., Sanderson, N. D., Swann, J., et al. (2019). Comparison of long-read sequencing technologies in the hybrid assembly of complex bacterial genomes. *Microb. Genomics* 5, e000294. doi:10.1099/mgen.0.000294

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

Dodt, M., Roehr, J., Ahmed, R., and Dieterich, C. (2012). FLEXBAR—flexible barcode and adapter processing for next-generation sequencing platforms. *Biology* 1, 895–905. doi:10.3390/biology1030895

Dong, R., Ma, X. K., Li, G. W., and Yang, L. (2018). CIRCpedia v2: An updated database for comprehensive circular RNA annotation and expression comparison. *Genomics Proteomics Bioinforma.* 16, 226–233. doi:10.1016/j.gpb.2018.08.001

Eid, J., Fehr, A., Gray, J., Luong, K., Lyle, J., Otto, G., et al. (2009). Real-time DNA sequencing from single polymerase molecules. *Science* 323, 133–138. doi:10.1126/science.1162986

Fernandez-Cuesta, L., Sun, R., Menon, R., George, J., Lorenz, S., Meza-Zepeda, L. A., et al. (2015). Identification of novel fusion genes in lung cancer using breakpoint assembly of transcriptome sequencing data. *Genome Biol.* 16, 7. doi:10.1186/s13059-014-0558-0

Ferraro, N. M., Strober, B. J., Einson, J., Abell, N. S., Aguet, F., Barbeira, A. N., et al. (2020). Transcriptomic signatures across human tissues identify functional rare genetic variation. *Science* 369, eaaz5900. doi:10.1126/science.aaz5900

Gao, Y., Wang, J., Zheng, Y., Zhang, J., Chen, S., and Zhao, F. (2016). Comprehensive identification of internal structure and alternative splicing events in circular RNAs. *Nat. Commun.* 7, 12060. doi:10.1038/ncomms12060

Gao, Y., Zhang, J., and Zhao, F. (2018). Circular RNA identification based on multiple seed matching. *Brief. Bioinform.* 19, 803–810. doi:10.1093/bib/bbx014

Gao, Y., and Zhao, F. (2018). Computational strategies for exploring circular RNAs. *Trends Genet.* 34, 389–400. doi:10.1016/j.tig.2017.12.016

GENCODE (2022). *GENCODE - home page*. Available at: https://www.gencodegenes.org/.

Gerstein, M. B., Bruce, C., Rozowsky, J. S., Zheng, D., Du, J., Korbel, J. O., et al. (2007). What is a gene, post-ENCODE? History and updated definition. *Genome Res.* 17, 669–681. doi:10.1101/gr.6339607

Glaus, P., Honkela, A., and Rattray, M. (2012). Identifying differentially expressed transcripts from RNA-seq data with biological variation. *Bioinformatics* 28, 1721–1728. doi:10.1093/bioinformatics/bts260

Goksuluk, D., Zararsiz, G., Korkmaz, S., Eldem, V., Zararsiz, G. E., Ozcetin, E., et al. (2019). MLSeq: Machine learning interface for RNA-sequencing data. *Comput. Methods Programs Biomed.* 175, 223–231. doi:10.1016/j.cmpb.2019.04.007

Grabherr, M. G., Haas, B. J., Yassour, M., Levin, J. Z., Thompson, D. A., Amit, I., et al. (2011). Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nat. Biotechnol.* 29, 644–652. doi:10.1038/nbt.1883

Green, C. J., Gazzara, M. R., and Barash, Y. (2018). MAJIQ-SPEL: Web-tool to interrogate classical and complex splicing variations from RNA-seq data. *Bioinformatics* 34, 300–302. doi:10.1093/bioinformatics/btx565

Haas, B. J., Dobin, A., Li, B., Stransky, N., Pochet, N., and Regev, A. (2019). Accuracy assessment of fusion transcript detection via read-mapping and de novo fusion transcript assembly-based methods. *Genome Biol.* 20, 213. doi:10.1186/s13059-019-1842-9

Haas, B. J., and Zody, M. C. (2010). Advancing RNA-Seq analysis. *Nat. Biotechnol.* 28, 421–423. doi:10.1038/nbt0510-421

Han, Y., Gao, S., Muegge, K., Zhang, W., and Zhou, B. (2015). Advanced applications of RNA sequencing and challenges. *Bioinforma. Biol. Insights* 9, BBI.S28991. doi:10.4137/bbi.s28991

Harvey, C. T., Moyerbrailean, G. A., Davis, G. O., Wen, X., Luca, F., and Pique-Regi, R. (2015). QuASAR: Quantitative allele-specific analysis of reads. *Bioinformatics* 31, 1235–1242. doi:10.1093/bioinformatics/btu802

Hoffmann, S., Otto, C., Doose, G., Tanzer, A., Langenberger, D., Christ, S., et al. (2014). A multi-split mapping algorithm for circular RNA, splicing, trans-splicing and fusion detection. *Genome Biol.* 15, R34. doi:10.1186/gb-2014-15-2-r34

Hu, Y. J., Sun, W., Tzeng, J. Y., and Perou, C. M. (2015). Proper use of allele-specific expression improves statistical power for cis-eQTL mapping with RNA-seq data. *J. Am. Stat. Assoc.* 110, 962–974. doi:10.1080/01621459.2015.1038449

Humphreys, D. T., Fossat, N., Demuth, M., Tam, P. P. L., and HoUlarcirc, J. W. K. (2019). Ularcirc: Visualization and enhanced analysis of circular RNAs via back and canonical forward splicing. *Nucleic Acids Res.* 47, e123. doi:10.1093/nar/gkz718

Jafari, M., and Ansari-Pour, N. (2018). Why, when and how to adjust your P values? *Cell. J. Yakhteh* 20, 604–607. doi:10.22074/cellj.2019.5992

Ji, P., Wu, W., Chen, S., Zheng, Y., Zhou, L., Zhang, J., et al. (2019). Expanded expression landscape and prioritization of circular RNAs in mammals. *Cell. Rep.* 26, 3444–3460. doi:10.1016/j.celrep.2019.02.078

Kang, K., Meng, Q., Shats, I., Umbach, D. M., Li, M., Li, Y., et al. (2019). CDSeq: A novel complete deconvolution method for dissecting heterogeneous samples using gene expression data. *PLOS Comput. Biol.* 15, e1007510. doi:10.1371/journal.pcbi.1007510

Kim, D., Paggi, J. M., Park, C., Bennett, C., and Salzberg, S. L. (2019). Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. *Nat. Biotechnol.* 37, 907–915. doi:10.1038/s41587-019-0201-4

Kim, D., and Salzberg, S. L. (2011). TopHat-fusion: An algorithm for discovery of novel fusion transcripts. *Genome Biol.* 12, R72. doi:10.1186/gb-2011-12-8-r72

Knowles, D. A., Davis, J. R., Edgington, H., Raj, A., Fave, M. J., Zhu, X., et al. (2017). Allele-specific expression reveals interactions between genetic variation and environment. *Nat. Methods* 14, 699–702. doi:10.1038/nmeth.4298

Korf, I. (2013). Genomics: The state of the art in RNA-seq analysis. *Nat. Methods* 10, 1165–1166. doi:10.1038/nmeth.2735

Kristensen, L. S., Andersen, M. S., Stagsted, L. V. W., Ebbesen, K. K., Hansen, T. B., and Kjems, J. (2019). The biogenesis, biology and characterization of circular RNAs. *Nat. Rev. Genet.* 20, 675–691. doi:10.1038/s41576-019-0158-7

Kukurba, K. R., and Montgomery, S. B. (2015). RNA sequencing and analysis. *Cold Spring Harb. Protoc.* 2015, 951–969. doi:10.1101/pdb.top084970

Kumasaka, N., Knights, A. J., and Gaffney, D. J. (2016). Fine-mapping cellular QTLs with RASQUAL and ATAC-seq. *Nat. Genet.* 48, 206–213. doi:10.1038/ng.3467

Larsson, T. P., Murray, C. G., Hill, T., Fredriksson, R., and Schiöth, H. B. (2005). Comparison of the current RefSeq, Ensembl and EST databases for counting genes and gene discovery. *FEBS Lett.* 579, 690–698. doi:10.1016/j.febslet.2004.12.046

Le, H. S., Schulz, M. H., McCauley, B. M., Hinman, V. F., and Bar-Joseph, Z. (2013). Probabilistic error correction for RNA sequencing. *Nucleic Acids Res.* 41, e109. doi:10.1093/nar/gkt215

Li, B., and Dewey, C. N. (2011). Rsem: Accurate transcript quantification from RNA-seq data with or without a reference genome. *BMC Bioinforma.* 12, 323. doi:10.1186/1471-2105-12-323

Li, B., Li, T., Pignon, J. C., Wang, B., Wang, J., Shukla, S. A., et al. (2016). Landscape of tumor-infiltrating T cell repertoire of human cancers. *Nat. Genet.* 48, 725–732. doi:10.1038/ng.3581

Li, H. (2013). Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *Genomics* 1303. doi:10.48550/ARXIV.1303.3997

Li, M., Xie, X., Zhou, J., Sheng, M., Yin, X., Ko, E. A., et al. (2017). Quantifying circular RNA expression from RNA-seq data using model-based framework. *Bioinformatics* 33, 2131–2139. doi:10.1093/bioinformatics/btx129

Li, S., Labaj, P. P., Zumbo, P., Sykacek, P., Shi, W., Shi, L., et al. (2014). Detecting and correcting systematic variation in large-scale RNA sequencing data. *Nat. Biotechnol.* 32, 888–895. doi:10.1038/nbt.3000

Li, Y. I., Knowles, D. A., Humphrey, J., Barbeira, A. N., Dickinson, S. P., Im, H. K., et al. (2018). Annotation-free quantification of RNA splicing using LeafCutter. *Nat. Genet.* 50, 151–158. doi:10.1038/s41588-017-0004-9

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

Lima, L., Marchet, C., Caboche, S., Da Silva, C., Istace, B., Aury, J. M., et al. (2020). Comparative assessment of long-read error correction software applied to Nanopore RNA-sequencing data. *Brief. Bioinform.* 21, 1164–1181. doi:10.1093/bib/bbz058

Liu, X., Milo, M., Lawrence, N. D., and Rattray, M. (2006). Probe-level measurement error improves accuracy in detecting differential gene expression. *Bioinformatics* 22, 2107–2113. doi:10.1093/bioinformatics/btl361

Mahmoud, M., Gobet, N., Cruz-Davalos, D. I., Mounier, N., Dessimoz, C., and Sedlazeck, F. J. (2019). Structural variant calling: The long and the short of it. *Genome Biol.* 20, 246. doi:10.1186/s13059-019-1828-7

Mandric, I., Rotman, J., Yang, H. T., Strauli, N., Montoya, D. J., Van Der Wey, W., et al. (2020). Profiling immunoglobulin repertoires across multiple human tissues using RNA sequencing. *Nat. Commun.* 11, 3126. doi:10.1038/s41467-020-16857-7

Mangul, S., Mosqueiro, T., Abdill, R. J., Duong, D., Mitchell, K., Sarwal, V., et al. (2019). Challenges and recommendations to improve the installability and archival stability of omics computational tools. *PLOS Biol.* 17, e3000333. doi:10.1371/journal.pbio.3000333

Martin, M. (2011). Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.J.* 17, 10–12. doi:10.14806/ej.17.1.200

McCarthy, D. J., Chen, Y., and Smyth, G. K. (2012). Differential expression analysis of multifactor RNA-Seq experiments with respect to biological variation. *Nucleic Acids Res.* 40, 4288–4297. doi:10.1093/nar/gks042

McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., et al. (2010). The genome analysis toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* 20, 1297–1303. doi:10.1101/gr.107524.110

Melsted, P., Hateley, S., Joseph, I. C., Pimentel, H., Bray, N., and Pachter, L. (2017). *Fusion detection and quantification by pseudoalignment*. 166322 Preprint. doi:10.1101/166322

Mitchell, K., Brito, J. J., Mandric, I., Wu, Q., Knyazev, S., Chang, S., et al. (2020). Benchmarking of computational error-correction methods for next-generation sequencing data. *Genome Biol.* 21, 71. doi:10.1186/s13059-020-01988-3

Mohammadi, P., Castel, S. E., Cummings, B. B., Einson, J., Sousa, C., Hoffman, P., et al. (2019). Genetic regulatory variation in populations informs transcriptome analysis in rare disease. *Science* 366, 351–356. doi:10.1126/science.aay0256

Monlong, J., Calvo, M., Ferreira, P. G., and Guigó, R. (2014). Identification of genetic variants associated with alternative splicing using sQTLseekeR. *Nat. Commun.* 5, 4698. doi:10.1038/ncomms5698

Morganti, S., Tarantino, P., Ferraro, E., D'Amico, P., Duso, B. A., and Curigliano, G. (2019). Next generation sequencing (ngs): A revolutionary technology in pharmacogenomics and personalized medicine in cancer. *Adv. Exp. Med. Biol.* 1168, 9–30. doi:10.1007/978-3-030-24100-1_2

Mose, L. E., Selitsky, S. R., Bixby, L. M., Marron, D. L., Iglesia, M. D., Serody, J. S., et al. (2016). Assembly-based inference of B-cell receptor repertoires from short read RNA sequencing data with V'DJer. *Bioinformatics* 32, 3729–3734. doi:10.1093/bioinformatics/btw526

Nariai, N., Hirose, O., Kojima, K., and Nagasaki, M. (2013). Tigar: Transcript isoform abundance estimation method with gapped alignment of RNA-seq data by variational bayesian inference. *Bioinformatics* 29, 2292–2299. doi:10.1093/bioinformatics/btt381

Nellore, A., Collado-Torres, L., Jaffe, A. E., Alquicira-Hernandez, J., Wilks, C., Pritt, J., et al. (2016). Rail-RNA: Scalable analysis of RNA-seq splicing and coverage. *Bioinformatics* 33, 4033–4040. doi:10.1093/bioinformatics/btw575

Nicolae, M., Mangul, S., Măndoiu, I. I., and Zelikovsky, A. (2011). Estimation of alternative splicing isoform frequencies from RNA-Seq data. *Algorithms Mol. Biol.* 6, 9. doi:10.1186/1748-7188-6-9

Ongen, H., Buil, A., Brown, A. A., Dermitzakis, E. T., and Delaneau, O. (2016). Fast and efficient QTL mapper for thousands of molecular phenotypes. *Bioinformatics* 32, 1479–1485. doi:10.1093/bioinformatics/btv722

Oshlack, A., and Wakefield, M. J. (2009). Transcript length bias in RNA-seq data confounds systems biology. *Biol. Direct* 4, 14. doi:10.1186/1745-6150-4-14

Ozsolak, F., and Milos, P. M. (2011). RNA sequencing: Advances, challenges and opportunities. *Nat. Rev. Genet.* 12, 87–98. doi:10.1038/nrg2934

Patro, R., Duggal, G., Love, M. I., Irizarry, R. A., and Kingsford, C. (2017). Salmon provides fast and bias-aware quantification of transcript expression. *Nat. Methods* 14, 417–419. doi:10.1038/nmeth.4197

Patro, R., Mount, S. M., and Kingsford, C. (2014). Sailfish enables alignment-free isoform quantification from RNA-seq reads using lightweight algorithms. *Nat. Biotechnol.* 32, 462–464. doi:10.1038/nbt.2862

Pertea, M., Pertea, G. M., Antonescu, C. M., Chang, T. C., Mendell, J. T., and Salzberg, S. L. (2015). StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. *Nat. Biotechnol.* 33, 290–295. doi:10.1038/nbt.3122

Pollard, M. O., Gurdasani, D., Mentzer, A. J., Porter, T., and Sandhu, M. S. (2018). Long reads: Their purpose and place. *Hum. Mol. Genet.* 27, R234–R241. doi:10.1093/hmg/ddy177

Prakash, C., and Haeseler, A. V. (2017). An enumerative combinatorics model for fragmentation patterns in RNA sequencing provides insights into nonuniformity of the expected fragment starting-point and coverage profile. *J. Comput. Biol.* 24, 200–212. doi:10.1089/cmb.2016.0096

Pruitt, K. D., Tatusova, T., and Maglott, D. R. (2007). NCBI reference sequences. (RefSeq): A curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Res.* 35, D61–D65. doi:10.1093/nar/gkl842

Raghupathy, N., Choi, K., Vincent, M. J., Beane, G. L., Sheppard, K. S., Munger, S. C., et al. (2018). Hierarchical analysis of RNA-seq reads improves the accuracy of allele-specific expression. *Bioinformatics* 34, 2177–2184. doi:10.1093/bioinformatics/bty078

Ranallo-Benavidez, T. R., Jaron, K. S., Schatz, M. C., and GenomeScope, 2. (2020). GenomeScope 2.0 and Smudgeplot for reference-free profiling of polyploid genomes. *Nat. Commun.* 11, 1432. doi:10.1038/s41467-020-14998-3

Reimers, M., and Carey, V. J. (2006). Bioconductor: An open source framework for bioinformatics and computational biology. *Methods Enzymol.* 411, 119–134. doi:10.1016/S0076-6879(06)11008-3

Rhie, A., Walenz, B. P., Koren, S., and Phillippy, A. M. (2020). Merqury: Reference-free quality, completeness, and phasing assessment for genome assemblies. *Genome Biol.* 21, 245. doi:10.1186/s13059-020-02134-9

Risso, D., Ngai, J., Speed, T. P., and Dudoit, S. (2014). Normalization of RNA-seq data using factor analysis of control genes or samples. *Nat. Biotechnol.* 32, 896–902. doi:10.1038/nbt.2931

Romanel, A., Lago, S., Prandi, D., Sboner, A., and Demichelis, F. (2015). Aseq: Fast allele-specific studies from next-generation sequencing data. *BMC Med. Genomics* 8, 9. doi:10.1186/s12920-015-0084-2

Schmid, M. W., and Grossniklaus, U. (2015). Rcount: Simple and flexible RNA-seq read counting. *Bioinformatics* 31, 436–437. doi:10.1093/bioinformatics/btu680

Sedlazeck, F. J., Lee, H., Darby, C. A., and Schatz, M. C. (2018). Piercing the dark matter: Bioinformatics of long-range sequencing and mapping. *Nat. Rev. Genet.* 19, 329–346. doi:10.1038/s41576-018-0003-4

Sekhon, A., Singh, R., and Qi, Y. (2018). DeepDiff: DEEP-learning for predicting DIFFerential gene expression from histone modifications. *Bioinformatics* 34, i891–i900. doi:10.1093/bioinformatics/bty612

Shabalin, A. A. (2012). Matrix eQTL: Ultra fast eQTL analysis via large matrix operations. *Bioinformatics* 28, 1353–1358. doi:10.1093/bioinformatics/bts163

Shastry, K. A., and Sanjay, H. A. (2020). "Machine learning for bioinformatics," in *Statistical modelling and machine learning principles for bioinformatics techniques, tools, and applications*. Editors K. G. Srinivasa, G. M. Siddesh, and S. R. Manisekhar (Berlin, Germany: Springer), 25–39. doi:10.1007/978-981-15-2445-5_3

Shen, C.-H. (2019). "Chapter 11 - techniques in sequencing," in *Diagnostic molecular biology*. Editor C. H. Shen (Cambridge, MA, USA: Academic Press), 277–302. doi:10.1016/B978-0-12-802823-0.00011-0

Shendure, J., and Ji, H. (2008). Next-generation DNA sequencing. *Nat. Biotechnol.* 26, 1135–1145. doi:10.1038/nbt1486

Simoneau, J., Gosselin, R., and Scott, M. S. (2020). Factorial study of the RNA-seq computational workflow identifies biases as technical gene signatures. *Nar. Genomics Bioinforma.* 2, lqaa043. doi:10.1093/nargab/lqaa043

Skelly, D. A., Johansson, M., Madeoy, J., Wakefield, J., and Akey, J. M. (2011). A powerful and flexible statistical framework for testing hypotheses of allele-specific gene expression from RNA-seq data. *Genome Res.* 21, 1728–1737. doi:10.1101/gr.119784.110

Srivastava, A., Sarkar, H., Gupta, N., and Patro, R. (2016). RapMap: A rapid, sensitive and accurate tool for mapping RNA-seq reads to transcriptomes. *Bioinforma. Oxf. Engl.* 32, i192–i200. doi:10.1093/bioinformatics/btw277

Stark, R., Grzelak, M., and Hadfield, J. (2019). RNA sequencing: The teenage years. *Nat. Rev. Genet.* 20, 631–656. doi:10.1038/s41576-019-0150-2

Szabo, L., and Salzman, J. (2016). Detecting circular RNAs: Bioinformatic and experimental challenges. *Nat. Rev. Genet.* 17, 679–692. doi:10.1038/nrg.2016.114

Taylor-Weiner, A., Aguet, F., Haradhvala, N. J., Gosai, S., Anand, S., Kim, J., et al. (2019). Scaling computational genomics to millions of individuals with GPUs. *Genome Biol.* 20, 228. doi:10.1186/s13059-019-1836-7

The Comprehensive R Archive Network (2022). *The comprehensive R archive Network*. Available at: https://cran.r-project.org/.

The RGASP ConsortiumAbril, J. F., Engstrom, P. G., Kokocinski, F., Hubbard, T. J., Guigó, R., et al. (2013a). Assessment of transcript reconstruction methods for RNA-seq. *Nat. Methods* 10, 1177–1184. doi:10.1038/nmeth.2714

The RGASP ConsortiumSteijger, T., Sipos, B., Grant, G. R., Kahles, A., Ratsch, G., et al. (2013b). Systematic evaluation of spliced alignment programs for RNA-seq data. *Nat. Methods* 10, 1185–1191. doi:10.1038/nmeth.2722

Trapnell, C., Roberts, A., Goff, L., Pertea, G., Kim, D., Kelley, D. R., et al. (2012). Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. *Nat. Protoc.* 7, 562–578. doi:10.1038/nprot.2012.016

Tuerk, A., Wiktorin, G., and Güler, S. (2017). Mixture models reveal multiple positional bias types in RNA-Seq data and lead to accurate transcript concentration estimates. *PLOS Comput. Biol.* 13, e1005515. doi:10.1371/journal.pcbi.1005515

van de Geijn, B., McVicker, G., Gilad, Y., and Pritchard, J. K. (2015). Wasp: Allele-specific software for robust molecular quantitative trait locus discovery. *Nat. Methods* 12, 1061–1063. doi:10.1038/nmeth.3582

van Ijzendoorn, D. G. P., Szuhai, K., Briaire-de Bruijn, I. H., Kostine, M., Kuijjer, M. L., and Bovee, J. V. M. G. (2019). Machine learning analysis of gene expression data reveals novel diagnostic and prognostic biomarkers and identifies therapeutic targets for soft tissue sarcomas. *PLOS Comput. Biol.* 15, e1006826. doi:10.1371/journal.pcbi.1006826

Vaquero-Garcia, J., Norton, S., and Barash, Y. (2018). *LeafCutter vs. MAJIQ and comparing software in the fast moving field of genomics*. 463927 Preprint. doi:10.1101/463927

Varadhan, R., and Roland, C. (2008). Simple and globally convergent methods for accelerating the convergence of any EM algorithm. *Scand. J. Stat.* 35, 335–353. doi:10.1111/j.1467-9469.2007.00585.x

Vierra, M., Kingan, S., Tseng, E., Hon, T., Rowell, W., Mountcastle, J., et al. (2021). *From RNA to full-length transcripts: The PacBio Iso-Seq method for transcriptome analysis and genome annotation - PacBio*. Available at: https://www.pacb.com/proceedings/from-rna-to-full-length-transcripts-the-pacbio-iso-seq-method-for-transcriptome-analysis-and-genome-annotation/.

Vurture, G. W., Sedlazeck, F. J., Nattestad, M., Underwood, C. J., Fang, H., Gurtowski, J., et al. (2017). GenomeScope: Fast reference-free genome profiling from short reads. *Bioinformatics* 33, 2202–2204. doi:10.1093/bioinformatics/btx153

Wang, A. T., Shetty, A., O'Connor, E., Bell, C., Pomerantz, M. M., Freedman, M. L., et al. (2020). Allele-specific QTL fine mapping with PLASMA. *Am. J. Hum. Genet.* 106, 170–187. doi:10.1016/j.ajhg.2019.12.011

Wang, K., Singh, D., Zeng, Z., Coleman, S. J., Huang, Y., Savich, G. L., et al. (2010). MapSplice: Accurate mapping of RNA-seq reads for splice junction discovery. *Nucleic Acids Res.* 38, e178. doi:10.1093/nar/gkq622

Wang, Y., Zhao, Y., Bollas, A., Wang, Y., and Au, K. F. (2021). Nanopore sequencing technology, bioinformatics and applications. *Nat. Biotechnol.* 39, 1348–1365. doi:10.1038/s41587-021-01108-x

Workflows for RNA Sequencing (2023). *A guide to Illumina solutions for next-generation RNA sequencing applications*.

Wu, J., Wang, C., Cui, Y., Xu, T., Wang, C., Wang, X., et al. (2019). CircAST: Full-length assembly and quantification of alternatively spliced isoforms in circular RNAs. *Genomics Proteomics Bioinforma.* 17, 522–534. doi:10.1016/j.gpb.2019.03.004

Wu, W., Ji, P., and Zhao, F. (2020). CircAtlas: An integrated resource of one million highly accurate circular RNAs from 1070 vertebrate transcriptomes. *Genome Biol.* 21, 101. doi:10.1186/s13059-020-02018-y

Xia, S., Feng, J., Chen, K., Ma, Y., Gong, J., Cai, F., et al. (2018). Cscd: A database for cancer-specific circular RNAs. *Nucleic Acids Res.* 46, D925–D929. doi:10.1093/nar/gkx863

Xie, J., Ji, T., Ferreira, M. A. R., Li, Y., Patel, B. N., and Rivera, R. M. (2019). Modeling allele-specific expression at the gene and SNP levels simultaneously by a Bayesian logistic mixed regression model. *BMC Bioinforma.* 20, 530. doi:10.1186/s12859-019-3141-6

Xu, G., Strong, M. J., Lacey, M. R., Baribault, C., Flemington, E. K., and Taylor, C. M. (2014). RNA CoMPASS: A dual approach for pathogen and host transcriptome analysis of RNA-seq datasets. *PLOS ONE* 9, e89445. doi:10.1371/journal.pone.0089445

Yang, Q., Hu, Y., Li, J., and Zhang, X. (2017). ulfasQTL: an ultra-fast method of composite splicing QTL analysis. *BMC Genomics* 18, 963. doi:10.1186/s12864-016-3258-1

Yang, X., Liu, D., Liu, F., Wu, J., Zou, J., Xiao, X., et al. (2013). HTQC: A fast quality control toolkit for Illumina sequencing data. *BMC Bioinforma.* 14, 33. doi:10.1186/1471-2105-14-33

Ye, H., Meehan, J., Tong, W., and Hong, H. (2015). Alignment of short reads: A crucial step for application of next-generation sequencing data in precision medicine. *Pharmaceutics* 7, 523–541. doi:10.3390/pharmaceutics7040523

Zakeri, M., Srivastava, A., Almodaresi, F., and Patro, R. (2017). Improved data-driven likelihood factorizations for transcript abundance estimation. *Bioinformatics* 33, i142–i151. doi:10.1093/bioinformatics/btx262

Zhang, J., Chen, S., Yang, J., and Zhao, F. (2020). Accurate quantification of circular RNAs identifies extensive circular isoform switching events. *Nat. Commun.* 11, 90. doi:10.1038/s41467-019-13840-9

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

Zheng, Y., Ji, P., Chen, S., Hou, L., and Zhao, F. (2019). Reconstruction of full-length circular RNAs enables isoform-level quantification. *Genome Med.* 11, 2. doi:10.1186/s13073-019-0614-1

Zheng, Y., and Zhao, F. (2020). Visualization of circular RNAs and their internal splicing events from transcriptomic data. *Bioinforma. Oxf. Engl.* 36, 2934–2935. doi:10.1093/bioinformatics/btaa033

Zou, J., Hormozdiari, F., Jew, B., Castel, S. E., Lappalainen, T., Ernst, J., et al. (2019). Leveraging allelic imbalance to refine fine-mapping for eQTL studies. *PLOS Genet.* 15, e1008481. doi:10.1371/journal.pgen.1008481

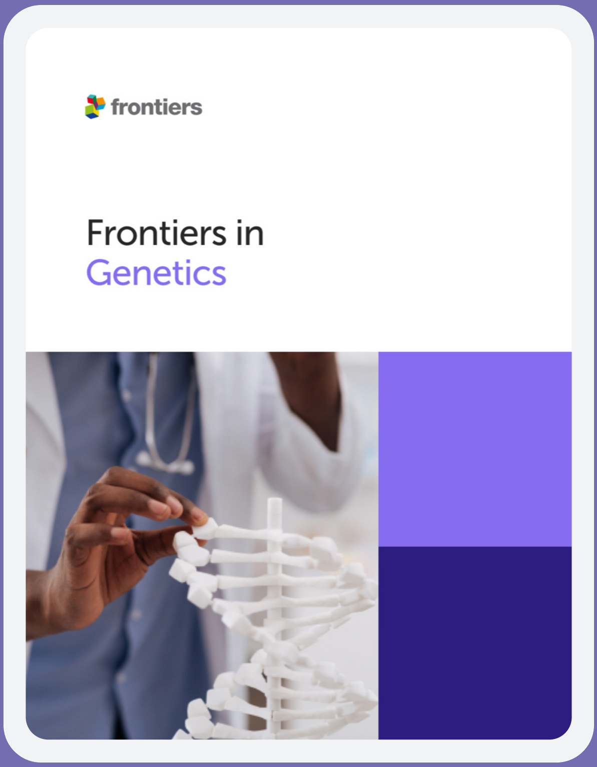
Frontiers in
Genetics

Highlights genetic and genomic inquiry relating to all domains of life

The most cited genetics and heredity journal, which advances our understanding of genes from humans to plants and other model organisms. It highlights developments in the function and variability of the genome, and the use of genomic tools.

Discover the latest Research Topics

See more →



Frontiers

Avenue du Tribunal-Fédéral 34
1005 Lausanne, Switzerland
frontiersin.org

Contact us

+41 (0)21 510 17 00
frontiersin.org/about/contact