frontiers | Frontiers in Molecular Biosciences

TYPE Original Research
PUBLISHED 29 September 2022
DOI 10.3389/fmolb.2022.1000847

OPEN ACCESS

EDITED BY
Peng Zhang,
University of Maryland, United States

REVIEWED BY
Yafeng Li,
The Fifth Hospital of Shanxi Medical University, China
Qianqian Song,
Wake Forest School of Medicine, United States
Ya-Li Chen,
Second Hospital of Hebei Medical University, China

*CORRESPONDENCE
Jun Cai,
caijun@fuwaihospital.org

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

RECEIVED 22 July 2022
ACCEPTED 06 September 2022
PUBLISHED 29 September 2022

CITATION
Gao Q, Fan L, Chen Y and Cai J (2022), Identification of the hub and prognostic genes in liver hepatocellular carcinoma via bioinformatics analysis.
Front. Mol. Biosci. 9:1000847.
doi: 10.3389/fmolb.2022.1000847

Identification of the hub and prognostic genes in liver hepatocellular carcinoma *via* bioinformatics analysis

Qiannan Gao[1], Luyun Fan[1], Yutong Chen[2] and Jun Cai[1,3]*

[1]State Key Laboratory of Cardiovascular Disease, FuWai Hospital, National Center for Cardiovascular Diseases, Peking Union Medical College, Chinese Academy of Medical Sciences, Beijing, China, [2]Health Science Center, Peking University International Cancer Institute, Peking University, Beijing, China, [3]Hypertension Center, FuWai Hospital, State Key Laboratory of Cardiovascular Disease, National Center for Cardiovascular Diseases, Peking Union Medical College, Chinese Academy of Medical Sciences, Beijing, China

Hepatocellular carcinoma (HCC) is a common malignancy. However, the molecular mechanisms of the progression and prognosis of HCC remain unclear. In the current study, we merged three Gene Expression Omnibus (GEO) datasets and combined them with The Cancer Genome Atlas (TCGA) dataset to screen differentially expressed genes. Furthermore, protein–protein interaction (PPI) and weighted gene coexpression network analysis (WGCNA) were used to identify key gene modules in the progression of HCC. Gene Ontology (GO) and Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway enrichment analyses indicated that the terms were associated with the cell cycle and DNA replication. Then, four hub genes were identified (*AURKA, CCNB1, DLGAP5,* and *NCAPG*) and validated via the expression of proteins and transcripts using online databases. In addition, we established a prognostic model using univariate Cox proportional hazards regression and least absolute shrinkage and selection operator (LASSO) regression. Eight genes were identified as prognostic genes, and four genes (*FLVCR1, HMMR, NEB,* and *UBE2S*) were detrimental gens. The areas under the curves (AUCs) at 1, 3 and 5 years were 0.622, 0.69, and 0.684 in the test dataset, respectively. The effective of prognostic model was also validated using International Cancer Genome Consortium (ICGC) dataset. Moreover, we performed multivariate independent prognostic analysis using multivariate Cox proportional hazards regression. The results showed that the risk score was an independent risk factor. Finally, we found that all prognostic genes had a strong positive correlation with immune infiltration. In conclusion, this study identified the key hub genes in the development and progression of HCC and prognostic genes in the prognosis of HCC, which was significant for the future diagnosis and prognosis of HCC.

KEYWORDS

HCC, GEO, TCGA, hub genes, prognostic model, ICGC

Introduction

Primary liver cancer is the sixth-most frequently occurring cancer in the world and the third-most common cause of cancer mortality (Sung et al., 2021). Hepatocellular carcinoma (HCC) is the most common form of liver cancer and accounts for ~90% of cases (Llovet et al., 2021). Although the strategy of treatment of HCC, including resection, liver transplantation, image-guided tumor ablation, image-guided transcatheter tumor therapy and systemic treatment, was effective for HCC patients, treatment indication still should be evaluated individually (Forner et al., 2018). Moreover, the methods of diagnosis of HCC remain poor except for histology for lesions and radiologic tests (Yang and Heimbach, 2020). However, some biomarkers could be utilized as diagnostic genes. For example, a set of immunostaining markers, such as glypican 3, heat shock protein 70, and glutamine synthetase, could increase diagnostic accuracy (Villanueva, 2019). Therefore, it is urgent to identify novel genes for the diagnosis of HCC and the precision medicine of HCC.

Recently, it was reported that some molecular drivers were involved in the development of HCC (Llovet et al., 2018). Studies have shown that *TERT* and *CTNNB1* mutations are associated with malignant transformation in <10% of cases (Nault et al., 2017). Other frequent mutations or genetic alterations were found in *TP53*, *RB1*, *CCNA2*, *CCNE1*, *PTEN*, *ARID1A*, *ARID2*, *RPS6KA3* or *NFE2L2*, all of which altered cell cycle control (Llovet et al., 2021). In addition, two major molecular subtypes of HCC were proposed (Zucman-Rossi et al., 2015). One was the proliferation gene class involved in cell proliferation and survival. It was demonstrated that *TP53* inactivation and *FGF19* and/or *CCND1* amplifications were involved (Wang et al., 2013). The other was the nonproliferation gene class, which activated the canonical *WNT* signaling pathway owing to the mutation of *CTNNB1* (Lachenmayer et al., 2012). Moreover, genome-wide gene expression studies demonstrated that some pathways, including *TGFβ*, the cell cycle, interferon, *MYC*, *PI3K/AKT*, and *MET*, were aberrantly activated in HCC (Rebouissou and Nault, 2020). Thus, it is extremely important to identify novel genes involved in the occurrence of tumors and determine the molecular mechanism of the progression of HCC.

The development of next-generation sequencing (NGS) technologies and bioinformatic tools has been widely used to search for novel targets and biomarkers for the diagnosis and precision medicine of cancer. In the current study, three Gene Expression Omnibus (GEO) datasets and The Cancer Genome Atlas (TCGA) dataset were merged and combined with bioinformatic analysis to screen differentially expressed genes (DEGs) in HCC. Then, a protein–protein interaction (PPI) network was constructed to select candidate hub genes. Moreover, DEGs of TCGA were used to screen the key gene modules using weighted gene coexpression network analysis (WGCNA). Gene Ontology (GO) and Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway enrichment analyses were used to perform functional annotation and identify potential pathways in HCC, and the protein and transcript expression of hub genes was validated using the human protein atlas database and gene expression profiling interactive analysis (GEPIA) database. Additionally, univariate Cox regression analysis, LASSO regression analysis and multivariate Cox regression analysis were used to identify prognostic genes. And the effective of prognostic model also was validated using International Cancer Genome Consortium (ICGC) dataset. Overall, our work identified novel genes involved in the progression and prognosis of HCC, which was of significance for the diagnosis and treatment of HCC.

Materials and methods

Data collection and processing

The workflow of the current study is shown in Figure 1. The gene expression profiles of the GSE84402, GSE101685, and GSE113996 datasets, including 42 normal samples and 58 tumor samples in total, were downloaded from the GEO database. The datasets were merged and the batch effect was eliminated using the R package "sva" (Leek et al., 2012). Differentially expressed genes (DEGs) were analyzed using the R package "limma" (Ritchie et al., 2015). *p*-value were adjusted using the false discovery rate (FDR) correction method. The cutoff for DEGs was set as $|\log_2 FC| > 1$ and adjusted *p*-value < 0.05.

RNA sequencing (counts) and clinical data for TCGA liver hepatocellular carcinoma (LIHC) patients, including 51 normal samples and 371 tumor samples, were downloaded and analyzed using the R package "TCGAbiolinks" (Colaprico et al., 2016). The cutoff for DEGs was set as $|\log_2 FC| > 1$ and adjusted *p*-value < 0.05. Visualization of overlapping genes in the Venn diagram was performed using the R package "VennDiagram."

RNA sequencing (counts) and clinical data for ICGC Liver Cancer—RIKEN, JP (LIRI-JP) patients, including 232 tumor samples, were downloaded. Two samples were excluded because of infinite values in RNA sequencing data, and 230 samples were included for further analysis. LIRI-JP was used to validate the performance of the prognostic model constructed from TCGA dataset.

Protein–protein interaction network construction and analysis of modules

DEGs were used to build a PPI network using the Search Tool for the Retrieval of Interaction Genes (STRING), and visualized using Cytoscape software. Five analysis methods in CytoHubba were used to select the key genes in PPI: edge percolated component (EPC), maximal clique centrality (MCC), maximal

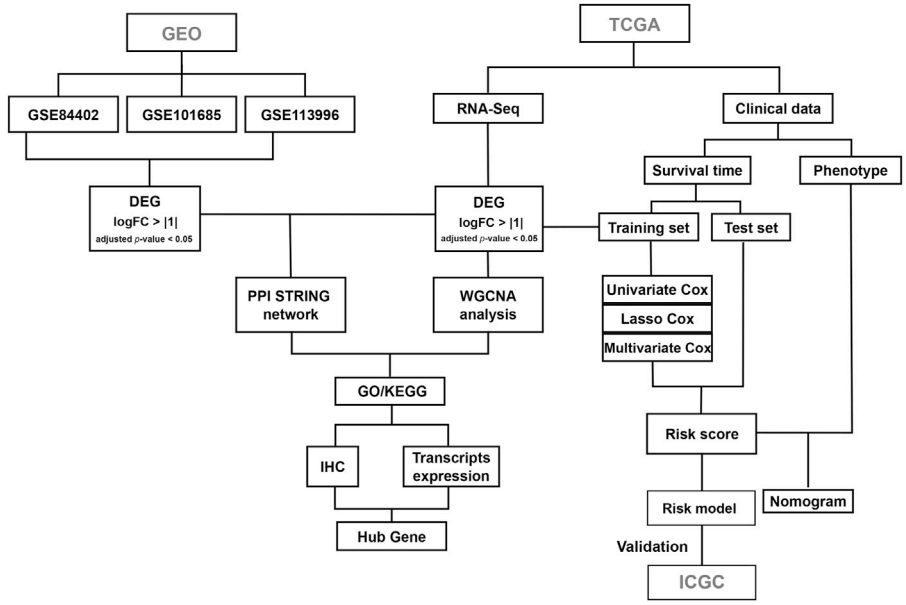

FIGURE 1
The workflow of the study design.

neighborhood component (MNC), node connect degree (degree), and node connect closeness (closeness) (Gao et al., 2020). The top 30 genes for each method were selected, and overlapping genes were identified as candidate hub genes. Molecular Complex Detection (MCODE) is a Cytoscape plugin for module analysis. Modules of interest were selected using a cutoff MCODE score >2, number of nodes >3, and confidence score >0.4.

Weighted gene co-expression network analysis

WGCNA was used to find clusters (modules) of highly correlated genes (Langfelder and Horvath, 2008). DEGs in TCGA were subjected to WGCNA using the R package "WGCNA." Clinical data for 421 LIHC patients were processed, with age, sex, and tumor occurrence selected as clinical traits. The soft-threshold power was used to raise the absolute value of the correlation. Hierarchical clustering and dynamic tree cut methods were used to identify modules. Eigengene networks were used to study module relationships. The module–trait relationship was assessed by Pearson's correlation tests by attributing values of 0 and 1 to healthy individuals and tumor patients, respectively. Module membership (MM) was defined as the correlation of genes with modules of interest, where a high MMmodule score indicated that a gene was highly correlated with the module. Gene significance (GS) was defined as the correlation of genes with clinical traits. The cutoff values for hub genes were set as GS > 0.2 and MM > 0.8.

Functional enrichment analysis

Genes in the modules of interest from WGCNA were subjected to Gene Ontology (GO) and Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway enrichment analyses using The Database for Annotation, Visualization and Integrated Discovery (DAVID) database, with a comprehensive set of functional annotation tools (Dennis et al., 2003). GO terms included biological process (BP), cellular component (CC), and molecular function (MF). The cutoff was set as $p < 0.05$. Visualization of GO terms and KEGG pathways was performed using the R package "ggplot2."

Construction of prognostic model and survival analysis

Clinical data were downloaded using the R package "TCGAbiolinks," and samples with missing values in terms of overall survival data were removed. Finally, a total of 364 patient samples were randomly divided into a training dataset ($n = 273$) and a test dataset ($n = 91$). In the training

dataset, candidate prognostic genes were screened by univariate Cox proportional hazards regression analysis using the R package "survival." LASSO regression analysis was used to select the prognostic gene signature (Tibshirani, 1997) using the R package "glmnet." After performing 10-fold cross-validations 1,000 times, the minimum lambda value was confirmed. The risk score was identified as a prediction factor and calculated as follows:

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

where $Coef_i$ indicates the correlation coefficient of the prognostic gene signature, and X_i indicates the expression of the gene signature. Patients in the training and test datasets were then divided into high- and low-risk groups according to the median risk score. A heatmap of prognostic gene expression was drawn using the R package "ComplexHeatmap." Kaplan–Meier survival curves were plotted to evaluate the predictive effect of the model using the log-rank test. The performance of the model at different endpoints (1, 3, and 5 years) was then assessed via time-dependent receiver operating characteristic (ROC) curves using the R package "timeROC." Multivariate Cox proportional hazards regression analysis was then used to determine if the risk score and clinical information were risk factors using the R package "survminer," with a cutoff for risk factors of $p < 0.01$. The nomogram was analyzed and depicted using the R package "nomogram."

Validation of hub genes and prognostic genes

The protein expression of the hub genes was validated using the Human Protein Atlas database, transcript expression of hub genes in LIHC patients was validated using the GEPIA database, and the correlation between immune infiltration and prognostic genes was validated using the Tumor Immune Estimation Resource (TIMER) database.

Statistical analysis

Continuous variables were analyzed using Student's *t*-test, *U*-test, or nonparametric rank-sum test. Correlations between the quantitative data were expressed by Spearman's coefficient. Prognostic analyses were performed using univariate and multivariate Cox regression analyses. Overall survival was analyzed using Kaplan-Meier analysis, and survival differences between the high- and low-risk groups were compared by log-rank test. All statistical analyses were performed using RStudio, and $p < 0.05$ was considered significant.

Results

Identification of differentially expressed genes

We identified DEGs in HCC by analyzing transcriptome information from three GEO datasets (GSE101685, GSE113996, and GSE84402) and TCGA (TCGA-LIHC). The GEO datasets contained 100 samples, including 42 normal and 58 tumor samples (Supplementary Table S1). The batch effect was eliminated before the analysis of DEG. A heatmap of gene expression in the merged GEO datasets is shown in Figure 2A. The criteria for the identification of DEGs were $|\log_2 FC| > 1$ and adjusted p-value < 0.05. A total of 230 genes were identified, including 81 upregulated and 149 downregulated genes (Figure 2B). TCGA dataset contained 421 samples, including 50 normal and 371 tumor samples. A total of 6,311 genes were significantly expressed in TCGA dataset. Finally, a total of 189 overlapping genes were identified in the GEO and TCGA datasets (Figure 2C) and were subjected to further analyses.

protein–protein interaction network analysis of overlapping genes

Overlapping genes were used to perform PPI network analysis using the STRING database. The related genes were ranked and the top 30 genes were selected using five methods in cytoHubba, a plugin for rank nodes in Cytoscape. A total of nine genes overlapped and were identified as candidate hub genes (Figure 3A). Network analysis was then performed using the MCODE module in Cytoscape. The candidate hub genes were all in module 1, which was the most highly scored module (MCODE score 53.321), including 57 nodes and 2,986 edges (Figure 3B). Details of the MCODE scores are shown in Supplementary Table S2.

Weighted gene coexpression network analysis and key module identification

We identified the key gene modules in HCC by WGCNA using the TCGA-LIHC dataset. Samples were clustered, and TCGA.66.A9EV.01A and TCGA.DD.A3A6.01A were excluded according to their height (>160) in the hierarchical clustering tree (Supplementary Figure S1). The soft-threshold power was set as 13 based on the scale independence and mean connectivity (Figure 4A). A total of 21 modules were identified using the dynamic tree cut package (Figure 4C). The cluster of module eigengenes and the eigengene adjacency heatmap are shown in Figure 4B and Supplementary Figure S2A. We determined the correlations

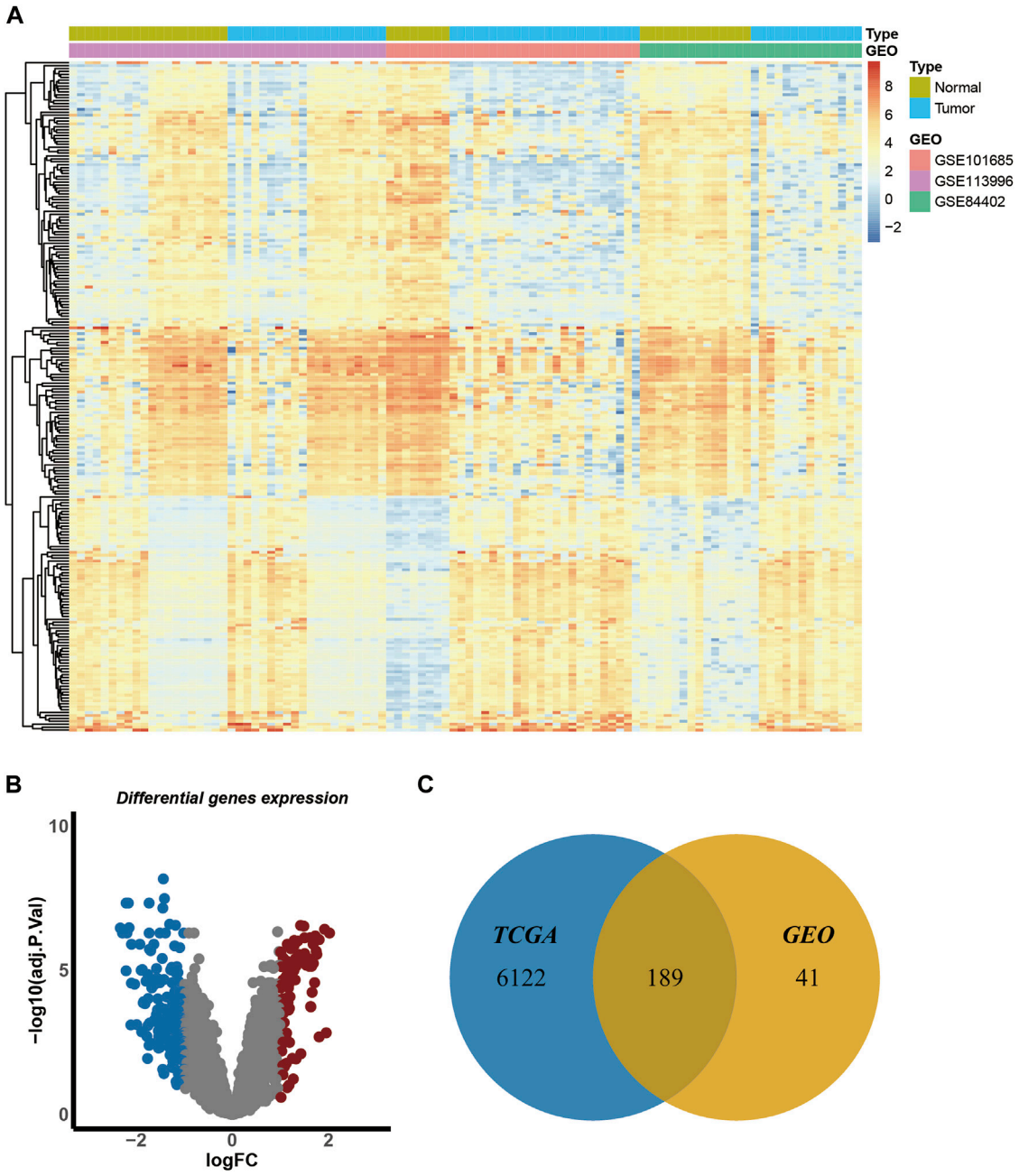

FIGURE 2
Identification of the differentially expressed genes of GEO and TCGA datasets. **(A)**. Heatmap of gene expression in merged GEO datasets. **(B)**. The volcano plot of the differentially expressed genes of merged GEO datasets. Blue dots indicate the distribution of downregulated genes, and red dots indicate the distribution of upregulated genes. **(C)**.Venn diagram of overlapping genes in GEO and TCGA datasets.

between the modules and the occurrence of tumors by establishing a module–trait relationship. The turquoise and purple modules (Figure 4D) were significantly correlated with tumor occurrence (coefficients 0.58 and 0.6, respectively), and the cyan module was significantly correlated with normal conditions (coefficient 0.67). In addition, high GS and high MM values were usually identified as features of hub genes. The gene distribution in the turquoise module showed that GS and MM were highly correlated, indicating that genes in this module were highly significantly associated with tumors (Figure 4E). The purple and light-cyan module are shown in Supplementary Figure S2B.

A

Rank	Methods in cytoHubba				
	MCC	MNC	EPC	Degree	Closeness
1	**RACGAP1**	CDK1	ECT2	CDK1	NDC80
2	**CDC6**	**AURKA**	**RACGAP1**	**AURKA**	CDK1
3	**AURKA**	**CCNB1**	GINS1	**CCNB1**	HMMR
4	**RAD51AP1**	CDC20	SPC25	CDC20	**CCNB1**
5	**DLGAP5**	**CDC6**	FEN1	**CDC6**	**AURKA**
6	**NCAPG**	BUB1B	CENPW	BUB1B	TOP2A
7	**CCNB1**	CCNB2	E2F8	CCNB2	CDKN3
8	**KIF11**	TOP2A	CDKN2A	TOP2A	CDC20
9	KIAA0101	CDKN3	ANLN	**NCAPG**	EZH2
10	ZWINT	RRM2	TRIP13	CDKN3	CCNB2
11	BUB1B	NUSAP1	**CDC6**	RRM2	**CDC6**
12	RRM2	TPX2	**AURKA**	NUSAP1	BUB1B
13	TTK	**RACGAP1**	OIP5	HMMR	**NCAPG**
14	CDC20	**RAD51AP1**	**RAD51AP1**	TPX2	**KIF23**
15	NUSAP1	**DLGAP5**	**DLGAP5**	**RACGAP1**	RRM2
16	CCNB2	**NCAPG**	**NCAPG**	**RAD51AP1**	NUSAP1
17	CDCA8	**KIF23**	**CCNB1**	**DLGAP5**	TPX2
18	MELK	**KIF11**	**KIF23**	**KIF23**	MCM10
19	PBK	NDC80	**KIF11**	**KIF11**	**RACGAP1**
20	TPX2	KIAA0101	NDC80	NDC80	**RAD51AP1**
21	TOP2A	ZWINT	MCM6	KIAA0101	**DLGAP5**
22	CDK1	TTK	DEPDC1B	ZWINT	**KIF11**
23	ASPM	UBE2C	MCM2	TTK	KIAA0101
24	BUB1	CDCA8	CENPE	UBE2C	ZWINT
25	KIF20A	MELK	CCNE2	CDCA8	TTK
26	**KIF23**	PBK	EZH2	MELK	UBE2C
27	CENPE	ASPM	RFC4	PBK	CDCA8
28	HMMR	BUB1	MCM10	ASPM	MELK
29	UBE2C	KIF20A	UBE2T	BUB1	PBK
30	KIF4A	CENPE	CKS2	KIF20A	ASPM

B

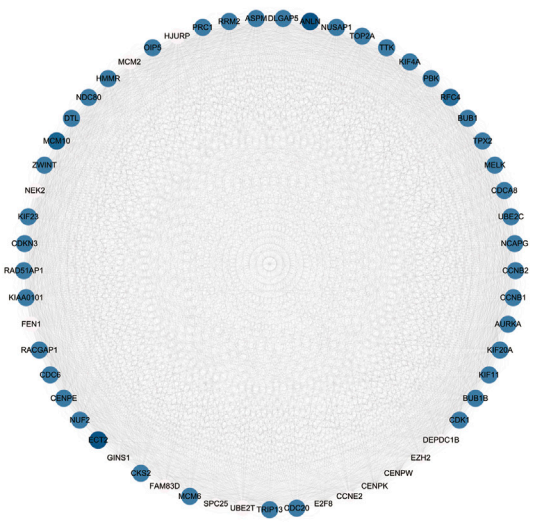

FIGURE 3
Identification of the candidate hub genes in the PPI network.**(A)**. Top 30 genes in cytoHubba.**(B)**.Gene interaction network of the most significant module in MCODE. The size of the dot is related to the degree of genes, and the gradation of the dot is related to the expression of genes.

10.3389/fmolb.2022.1000847

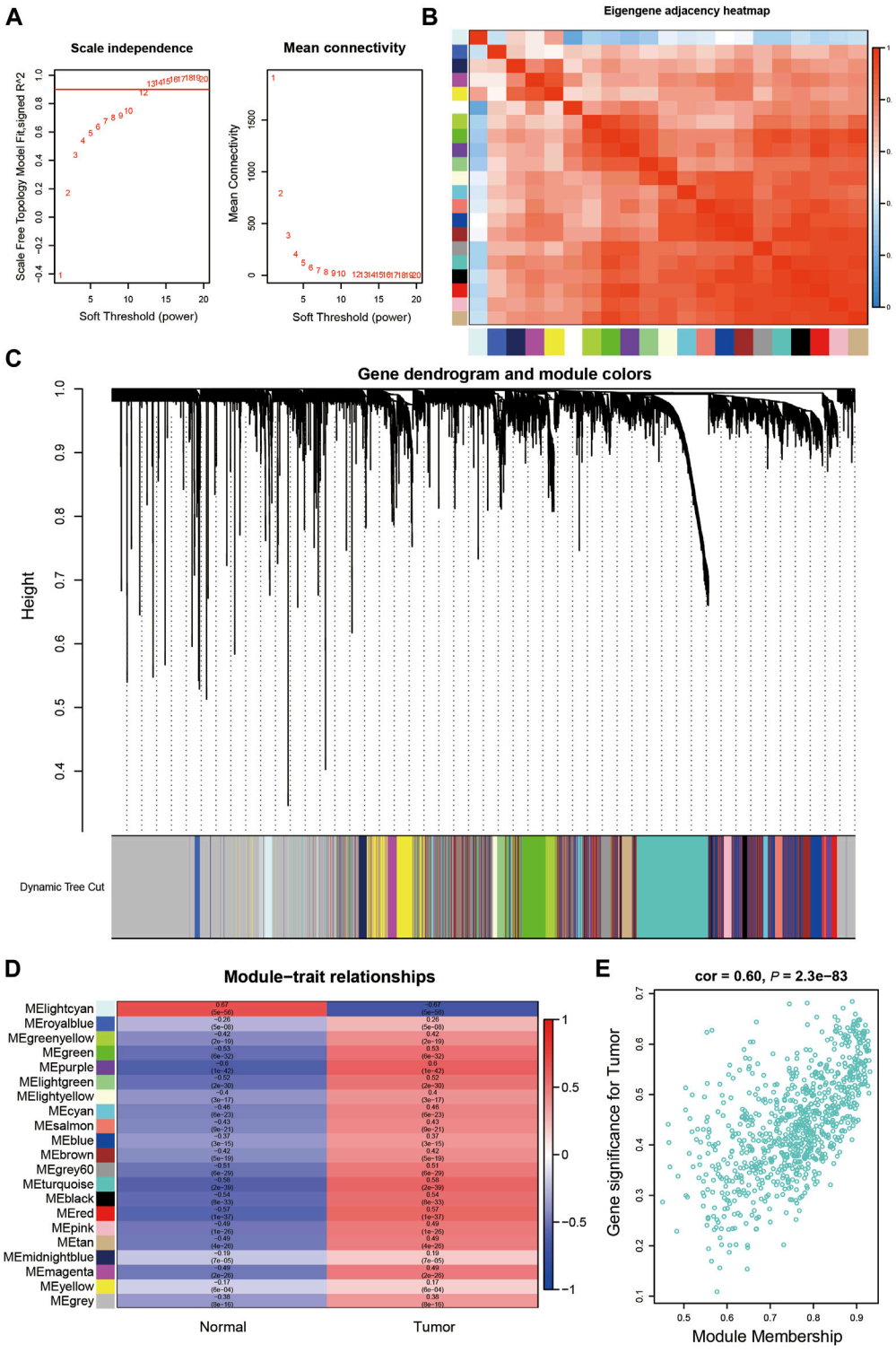

FIGURE 4
Identification of the key gene modules in WGCNA.**(A)**. Determination of the soft-thresholding power.**(B)**. The heatmap of Eigengene adjacency.**(C)**. Dendrogram of differentially expressed genes clustered based on a dissimilarity measure (1-TOM). **(D)**. The correlation of gene modules with clinical traits.**(E)** Gene correlation scatter plot of the turquoise module.

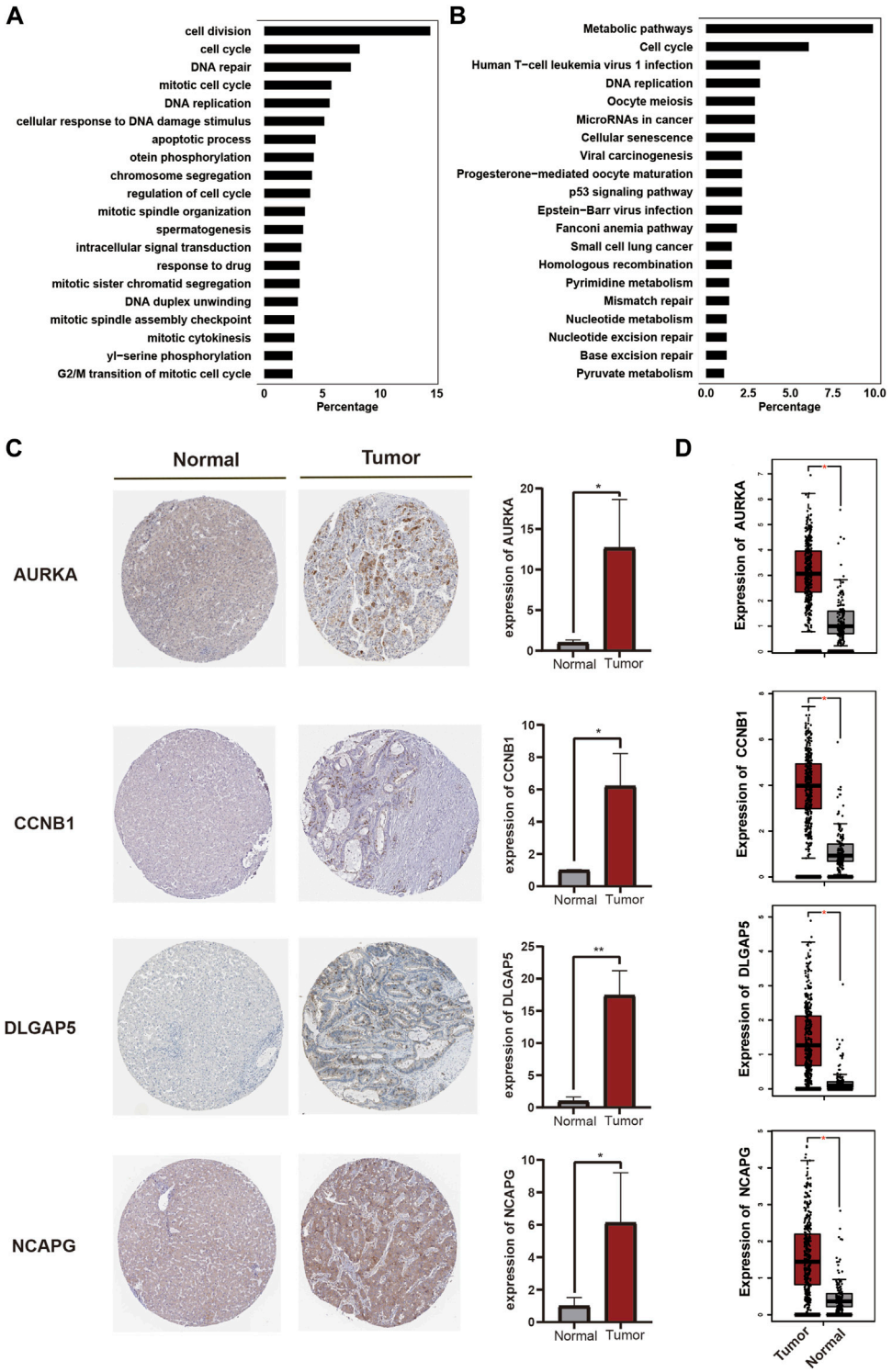

FIGURE 5
Validation of hub genes. **(A)**. GO-BP term enrichment of genes in the turquoise module. **(B)**. KEGG pathway enrichment of genes in the turquoise module. **(C)**. The protein expression of hub genes in tumor and normal samples using validation of immunohistochemistry **(D)**. Transcripts expression of hub genes in tumor and normal samples.

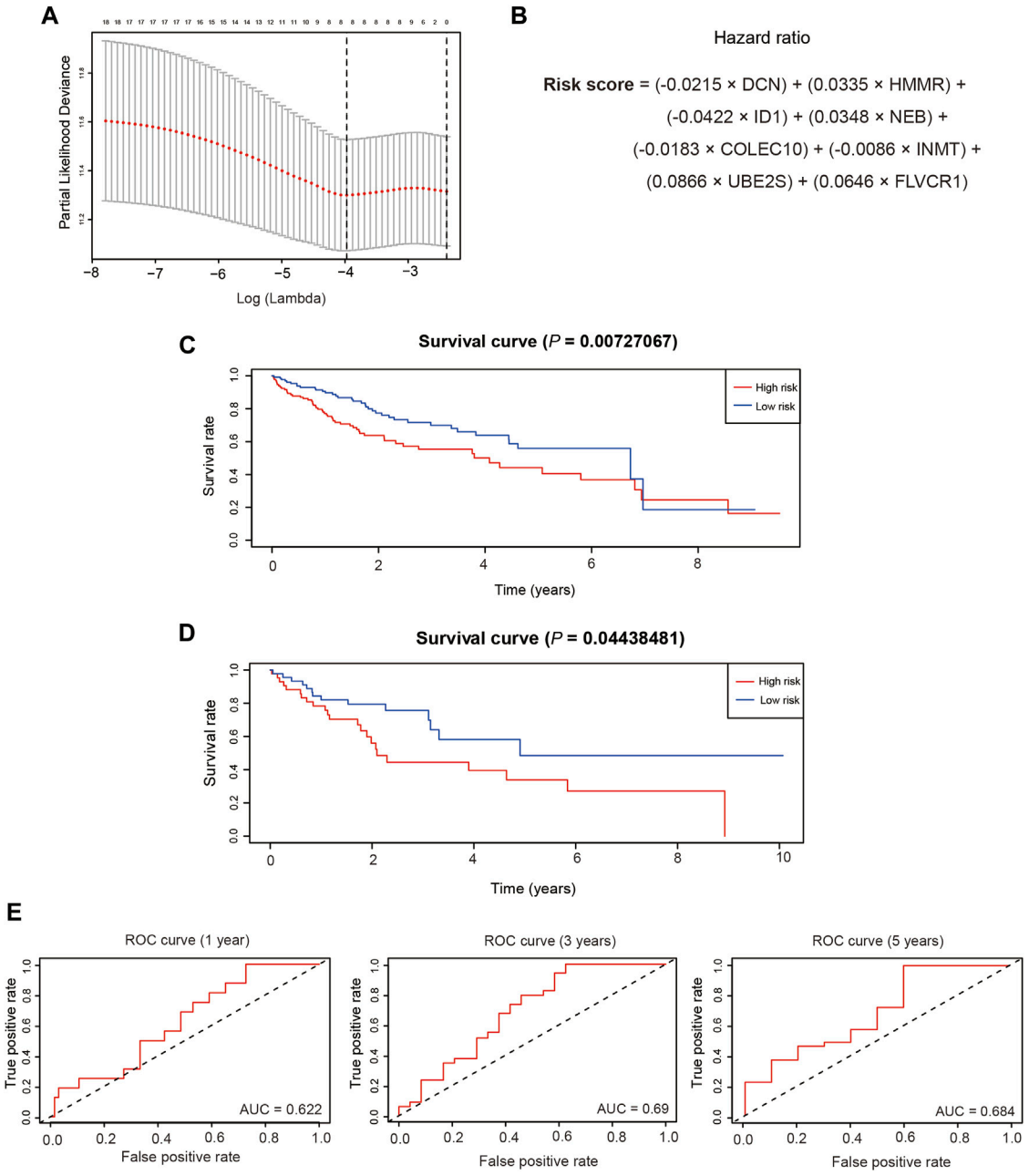

FIGURE 6
Establishment of the prognostic model. **(A)**. The selection of the minimum lambda of the lasso model via 10 folds of cross-validation. **(B)**. The calculation formula of the risk score.**(C)**.The overall survival of different risk groups in the training dataset. **(D)**. The overall survival of different risk groups in the test dataset.**(E)**. The time-dependent ROC curve of the performance of the prognostic model at 1, 3 and 5 years in the test dataset.

Genes in the turquoise module were subsequently subjected to GO and KEGG analyses. These genes were highly enriched in GO-BP terms containing cell division, cell cycle, mitotic cell cycle, and DNA replication (Figure 5A), and were enriched in GO-CC and GO-MF terms containing nucleus, cytosol, protein binding, and ATP binding (Supplementary Figure S3). KEGG pathway analysis showed that the genes were enriched in pathways including metabolic pathways, cell cycle and human T-cell leukemia virus 1 infection, and DNA replication (Figure 5B).

Moreover, the above candidate hub genes (Figure 3A) were all included in the turquoise module, suggesting that these genes played an important role in the progression of LIHC. We therefore validated the related transcript and protein

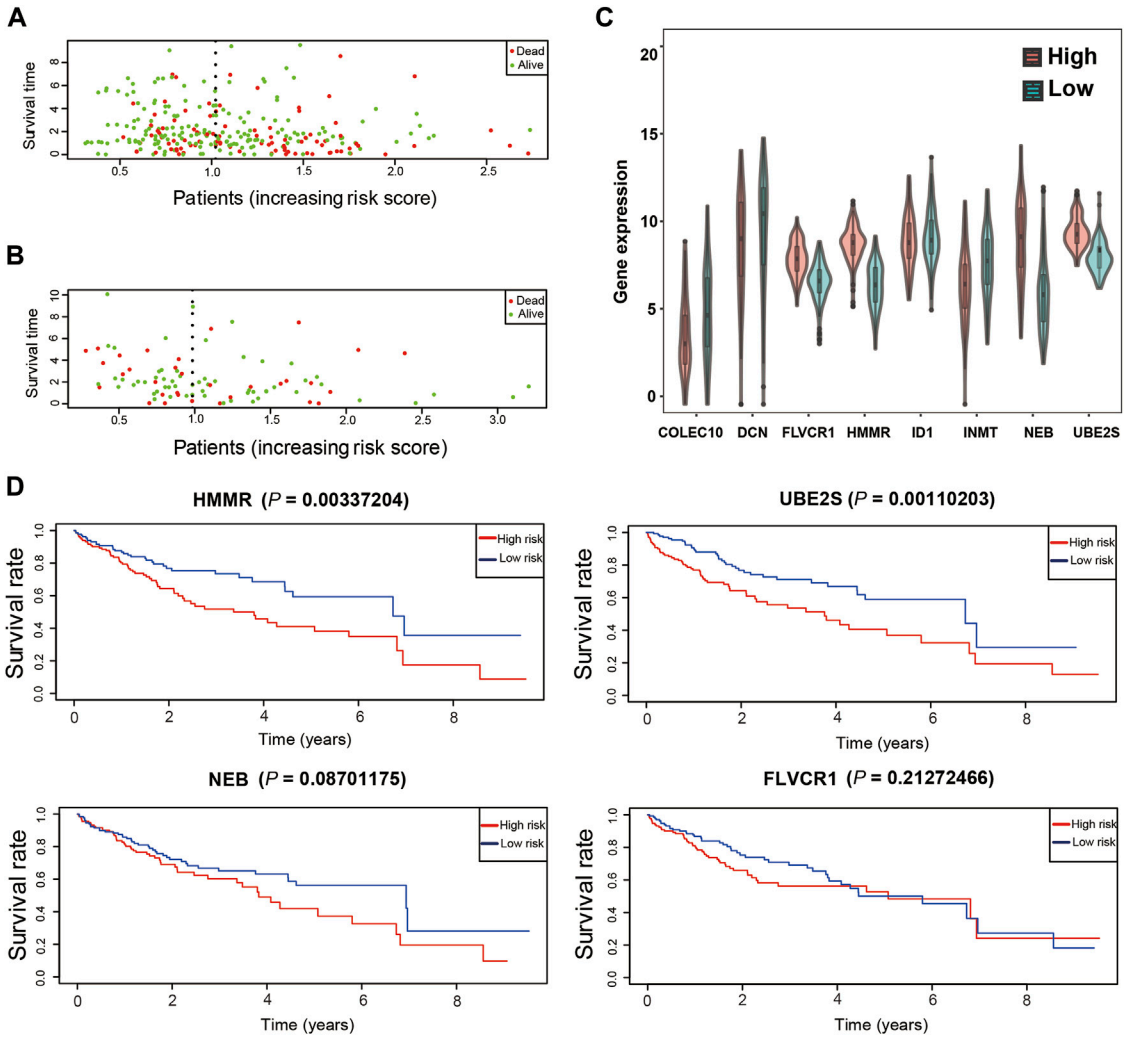

FIGURE 7
Validation of the prognostic model. **(A)**. The distribution of survival duration in the training dataset. **(B)**. The distribution of survival duration in the test dataset. **(C)**. The expression of detrimental prognostic genes in different risk groups in the training dataset. **(D)**. The overall survival of detrimental prognostic genes.

expression of the candidate hub genes using the human protein atlas and Gene Expression Profiling Interactive Analysis (GEPIA) databases. Four hub genes were finally identified. The results of immunohistochemistry showed that protein expression levels of *AURKA*, *CCNB1*, *DLGAP5,* and *NCAPG* were upregulated in tumor tissue compared with normal tissue (Figure 5C). In addition, transcript levels of the four genes were also significantly upregulated in LIHC patients compared with healthy subjects (Figure 5D). The GS and MM values and the combined scores of the four hub genes are shown in Supplementary Tables S3,S4. These results thus indicated that these four genes (*AURKA*, *CCNB1*, *DLGAP5,* and *NCAPG*) were key hub genes involved in the development and progression of HCC.

Construction of a prognostic model of Hepatocellular carcinoma

To establish a prognostic model of HCC, we randomly divided the subjects into a training dataset ($n = 273$) and a test dataset ($n = 91$). The training dataset was subjected to univariate Cox proportional hazards regression analysis followed by LASSO regression to screen for prognostic genes, and by multivariate Cox proportional hazards regression. A total of 19 genes were identified by univariate analysis, and LASSO regression identified eight genes with the minimum lambda value of 0.0188 (Figure 6A). The risk score was calculated as the sum of the gene coefficients multiplied by each gene expression level

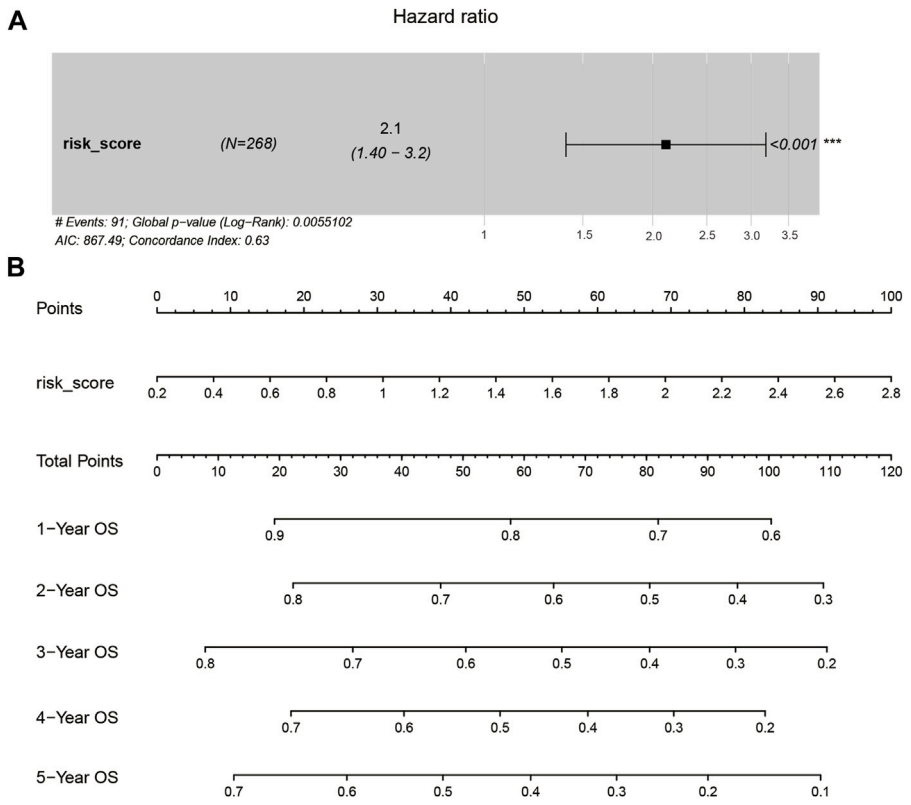

FIGURE 8
The identification of the risk score was an independent risk factor. **(A)**. The hazard ratio of risk score using multivariate Cox regression analysis. **(B)**. The nomogram of prognostic model judgment.

(Figure 6B), and the high- and low-risk patient groups were subsequently classified by the median risk score. Survival curves in the training and test datasets were examined using the Kaplan-Meier method, which showed that the low-risk group had a higher survival probability than the high-risk group in both the training and test datasets (Figures 6C,D). In addition, we predicted the overall survival in the training and test datasets at 1, 3, and 5 years. The respective areas under the time-dependent ROC curves (AUCs) were 0.622, 0.69, and 0.684 in the test dataset (Figure 6E) and 0.677, 0.645, and 0.63 in the training dataset (Supplementary Figure S4A).

Moreover, patients with high risk scores were more likely to die in the training and test datasets (Figures 7A,B). *FLVCR1*, *HMMR*, *NEB,* and *UBE2S* expression levels were significantly upregulated in the high-risk groups compared with the low-risk groups (Figure 7C), while *COLEC10*, *DCN*, *ID1,* and *INMT* were significantly downregulated. This was in accordance with the coefficients of the LASSO model and consistent with the heatmap of gene expression in the training and test datasets (Supplementary Figure S4B). In addition, *HMMR* and *UBE2S*, but not *FLVCR1* and *NEB*, were highly associated with poor survival probability in the training dataset (Figure 7D). However, we found that only *FLVCR1* were highly associated with poor survival probability in test dataset (Supplementary Figure S4C).

Multivariate Cox proportional hazards regression analysis showed that the risk score was an important factor strongly associated with the prediction of overall survival at 1, 2, 3, 4, and 5 years (Figures 8A,B). Increasing risk score was associated with a decreasing probability of overall survival in the subsequent 1–5 years. These results thus indicated that the established prognostic model could effectively predict the prognosis in patients with HCC.

Validation of prognostic genes in International Cancer Genome Consortium dataset

We validated the effectiveness of the prognostic model constructed from TCGA dataset in another dataset, LIRI-JP from the ICGC database. A total of 230 LIRI-JP RNA sequencing and clinical data were merged for further analysis. High- and low-risk patient groups were classified according to the median risk score, calculated as for the training dataset of

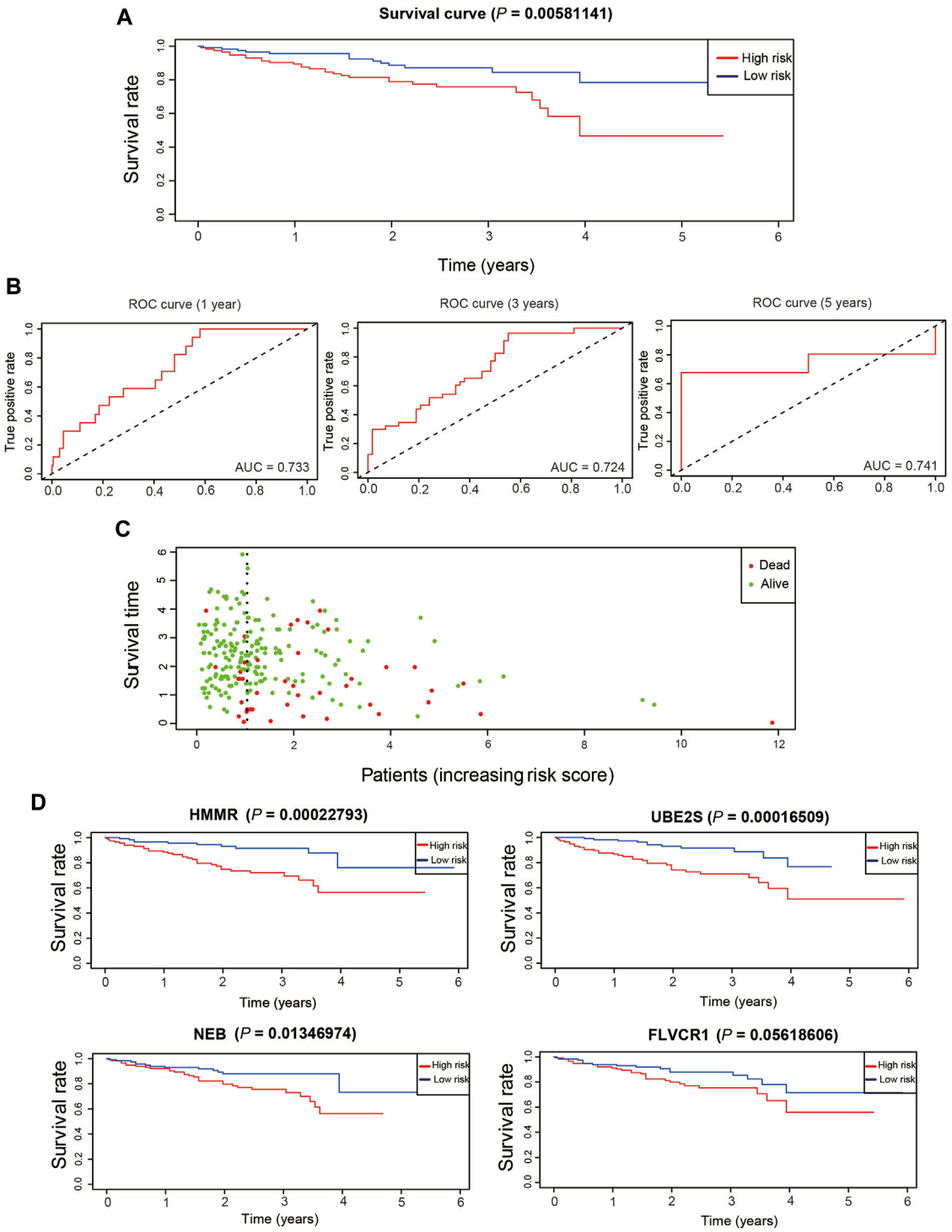

FIGURE 9
Validation of prognostic genes in ICGC dataset. **(A)**. The overall survival of different risk groups in ICGC dataset. **(B)**. The time-dependent ROC curve of the performance of the prognostic model at 1, 3 and 5 years in ICGC dataset. **(C)**. The distribution of survival duration in ICGC dataset. **(D)**. The overall survival of detrimental prognostic genes in ICGC dataset.

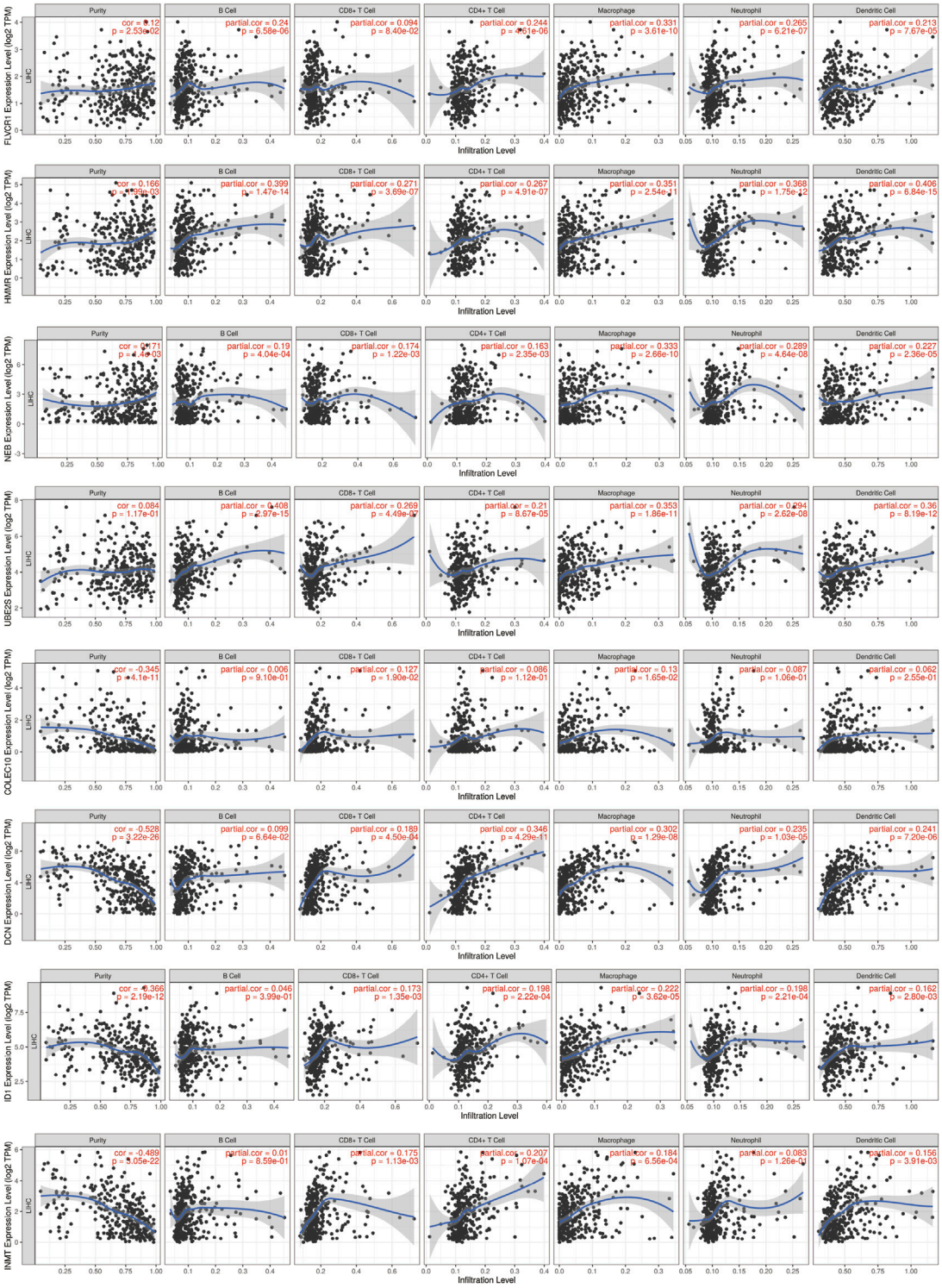

FIGURE 10
The correlation of prognostic genes with immune infiltration.

TCGA. The low-risk group had a higher survival probability than the high-risk group (Figure 9A), and the AUCs were 0.733, 0.724, and 0.741 for predicting overall survival at 1, 3, and 5 years, respectively (Figure 9B). In addition, patients with high risk scores in the ICGC dataset were more likely to die (Figure 9C), consistent with the results for the training and test datasets. *HMMR*, *NEB*, and *UBE2S* were highly associated with poor survival probability in the ICGC dataset (Figure 9D). Taken together, our results demonstrated that the prognostic model had effective and robust performance for HCC.

Validation of the prognostic model

To validate the prognostic model, we analyzed the association with genes and immune infiltration. The results showed that *FLVCR1* expression was strongly positively correlated with B cells (cor = 0.24, p = 6.58e-06), $CD4^+$ T cells (cor = 0.244, p = 4.61e-06), macrophages (cor = 0.331, p = 3.61e-10), neutrophils (cor = 0.265, p = 6.21e-07) and dendritic cells (cor = 0.213, p = 7.67e-05). *HMMR* expression was strongly positively correlated with B cells (cor = 0.399, p = 1.47e-14), $CD8^+$ T cells (cor = 0.271, p = 3.69e-07), $CD4^+$ T cells (cor = 0.267, p = 4.91e-07), macrophages (cor = 0.351, p = 2.54e-11), neutrophils (cor = 0.368, p = 1.75e-12) and dendritic cells (cor = 0.406, p = 6.84e-15). *NEB* expression was strongly positively correlated with B cells (cor = 0.19, p = 4.04e-04), $CD8^+$ T cells (cor = 0.174, p = 1.22e-03), $CD4^+$ T cells (cor = 0.163, p = 2.35e-03), macrophages (cor = 0.333, p = 2.66e-10), neutrophils (cor = 0.289, p = 4.64e-08) and dendritic cells (cor = 0.227, p = 2.63e-05). *UBE2S* expression was strongly positively correlated with B cells (cor = 0.408, p = 2.97e-15), $CD8^+$ T cells (cor = 0.269, p = 4.49e-07), $CD4^+$ T cells (cor = 0.21, p = 8.67e-05), macrophages (cor = 0.353, p = 1.86e-11), neutrophils (cor = 0.294, p = 2.62e-08) and dendritic cells (cor = 0.36, p = 8.19e-12). *COLEC10* expression was strongly positively correlated with $CD8^+$ T cells (cor = 0.127, p = 1.90e-02) and macrophages (cor = 0.13, p = 1.65e-02). *DCN* was strongly positively correlated with $CD8^+$ T cells (cor = 0.189, p = 4.50e-04), $CD4^+$ T cells (cor = 0.346, p = 4.29e-11), macrophages (cor = 0.302, p = 1.29e-08), neutrophils (cor = 0.235, p = 1.03e-05) and dendritic cells (cor = 0.241, p = 7.20e-06). *ID1* expression was strongly positively correlated with $CD8^+$ T cells (cor = 0.173, p = 1.35e-03), $CD4^+$ T cells (cor = 0.198, p = 2.22e-04), macrophages (cor = 0.222, p = 3.62e-05), neutrophils (cor = 0.198, p = 2.21e-04) and dendritic cells (cor = 0.162, p = 2.80e-03). *INMT* expression was strongly positively correlated with $CD8^+$ T cells (cor = 0.175, p = 1.13e-03), $CD4^+$ T cells (cor = 0.207, p = 1.07e-04), macrophages (cor = 0.184, p = 6.56e-04) and dendritic cells (cor = 0.156, p = 3.91e-03) (Figure 10). Thus, the results above indicated that the prognostic model we established had potential and effective prediction for the prognosis of HCC.

Discussion

In the current study, we merged three GEO datasets and TCGA datasets and combined them with bioinformatics analysis to screen and identify hub genes and prognostic genes in the development and progression of HCC. We identified four hub genes (*AURKA, CCNB1, DLGAP5,* and *NCAPG*) using WGCNA and PPI network analysis based on the clustering of the key modules and biological functional annotation. Moreover, we established a prognostic model and identified four detrimental prognostic genes (*FLVCR1, HMMR, NEB,* and *UBE2S*) using Lasso-Cox regression. Therefore, these genes could be potential biomarkers and prediction factors in the future diagnosis and treatment of HCC.

Correlation networks are increasingly being used in bioinformatics applications. WGCNA is a systems biology method for exploring the module structure in a network, measuring the relationships between genes and modules and the relationships among modules (Langfelder and Horvath, 2008). Using WGCNA, we found that the turquoise module was strongly positively correlated with the occurrence of tumors, and most genes in the turquoise module overlapped with those in the PPI networks, which indicated that these genes were highly associated with the development of HCC. Finally, we identified four hub genes (*AURKA, CCNB1, DLGAP5,* and *NCAPG*) and validated them in HCC samples.

Aurora A (*AURKA*) is a type of Aurora kinase that belongs to the family of serine/threonine kinases and plays essential roles in regulating cell division during mitosis. It was reported that the expression of *AURKA* was aberrantly high in HCC (Du et al., 2021). Increased *AURKA* expression and a positive correlation between *AURKA* and *MYC* expression were found in *TP53*-mutated human HCCs (Dauch et al., 2016), which indicated that *AURKA* was a potential druggable target. Additionally, overexpression of Aurora-A was associated with high-grade (grade II-IV) and high-stage (stage IIIB-IV) tumors, *p53* mutation, infrequent beta-catenin mutation, and poor outcome (Jeng et al., 2004).

Cyclin B1 (*CCNB1*) is a regulator involved in mitosis. *MYC* was reported to activate *WDR4* transcription, and *WDR4* promoted *CCNB1* mRNA stability and translation to enhance HCC progression (Xia et al., 2021). In addition, *CCNB1* could be a candidate biomarker and potential therapeutic target for HBV-related HCC recurrence after surgery (Weng et al., 2012). DLG-associated protein 5 (*DLGAP5*), also known as *HURP*, can enable microtubule binding activity. The molecular mechanism of *DLGAP5* in the development of HCC is limited, but known studies have reported that *DLGAP5* may be a potential biomarker in HCC (Hao et al., 2021). Non-SMC condensin I complex subunit G (*NCAPG*) was first isolated from HeLa cell nuclei and demonstrated to regulate the location of DNA on chromosomes (Sun et al., 2022). It was demonstrated that *NCAPG* was a novel mitotic gene involved in the proliferation

and migration of HCC cells (Zhang et al., 2018; Gong et al., 2019). Moreover, *AURKA* and *DLGAP5* were mitosis-associated genes. *DLGAP5* is also a substrate of *AURKA* (Wu et al., 2013), which is consistent with our results that *AURKA* is a hub gene. Although investigation of the association between these four hub genes was currently poor, they were involved in the regulation of cell proliferation and cell cycle and played important role in the development and progression of cancer.

Moreover, we established the prognostic model using Lasso Cox regression and identified eight prognostic genes, four of which were detrimental genes (*NEB, UBE2S, FLVCR1,* and *HMMR*). These four genes could predict the overall survival time in the high-risk groups and low-risk groups. In addition, the prognostic model showed excellent prediction performance. Therefore, four detrimental genes could be potential diagnostic and prognostic genes.

Nebulin (*NEB*) encodes a giant protein component of the cytoskeletal matrix that coexists with the thick and thin filaments within the sarcomeres of skeletal muscle. It was reported that the mutation of *NEB* was associated with many diseases, such as nemaline myopathy (Piga et al., 2016) and fetal akinesia deformation sequence/arthrogryposis multiplex congenita (Feingold-Zadok et al., 2017). Additionally, the mutation frequency of *NEB* at the amino acid 1,133 locus of thyroid cancer patients was much higher than that of the normal population (Wang et al., 2020). Although the relationship between *NEB* and HCC has not yet been reported, further research is needed to determine the molecular mechanism of *NEB* in HCC.

Ubiquitin conjugating enzyme E2S (*UBE2S*), also known as *E2EPF*, belongs to the E2 family of proteins and elongates the K11-linked polyubiquitin chain on APC/C substrates for 26 S proteasome-mediated degradation to promote cell division (Wu et al., 2010). *UBE2S* can promote the progression of many types of cancer, such as ovarian cancer (Hu et al., 2021), non-small cell lung cancer (Qin et al., 2020), colorectal cancer (Li et al., 2018) and prostate cancer (Peng et al., 2022). In addition, *UBE2S* promoted cell chemoresistance through *PTEN-AKT* signaling in HCC (Gui et al., 2021), which indicated that *UBE2S* may be a novel oncogene in the development of cancer. Feline leukemia virus subgroup C receptor 1 (*FLVCR1*) has been reported to have a crucial role in cell proliferation and cell death (Peng et al., 2018). A recent study showed that *FLVCR1* was significantly higher in the HCC cohort from ICGC than in normal samples (Tang et al., 2020), which was consistent with our results. Hyaluronan-mediated motility receptor (*HMMR*) is an oncogene involved in neoplastic progression of human leukemias and solid tumors (Tilghman et al., 2014). The overexpression of HMMR was strongly associated with the occurrence of HCC.

However, there are some limitations in this study. First, gene-based markers as biologic signatures were not enough to use as prognostic model for predicting patient outcomes. Network or subnetworks markers need to be developed to perform more meaningful and accurate prediction. Song et al. developed a method that identified survival prognostic subnetwork markers (SPNs), which had more accurate and effective performance for prediction of distant metastasis-free survival time and uncovered the biological mechanism in in breast cancer (Song et al., 2015). Additionally, Discrepancy of tumor immune microenvironment under differed treatments need to be resolved at single-cell level. It was reported that single-cell multi-omics gene co-regulatory algorithm (SMGR) was developed to discover cis-element elements and regulatory networks in mixed-phenotype acute leukemia cells by integrating single-cell RNA-sequencing and single-cell assay for transposase-accessible chromatin using sequencing (Song et al., 2022). Taken together, more comprehensive models and integrating methods need to be used for the validation and analysis of results.

In conclusion, our results indicated that *AURKA, CCNB1, DLGAP5,* and *NCAPG* were key hub genes and that *NEB, UBE2S, FLVCR1* and *HMMR* were crucial detrimental prognostic genes, which could be potential biomarkers and druggable targets in the diagnosis and treatment of HCC.

Data availability statement

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

Author contributions

JC and QG designed the study. Data collection and analysis were performed by QG and LF. The figures were drawn by QG and YC. The draft of the manuscript was written by QG and LF. All authors read and approved the final manuscript.

Funding

This work was supported by grants from the National Natural Science Foundation of China (Project ID. 81825002).

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/fmolb.2022.1000847/full#supplementary-material

References

Colaprico, A., Silva, T. C., Olsen, C., Garofano, L., Cava, C., Garolini, D., et al. (2016). TCGAbiolinks: An R/bioconductor package for integrative analysis of TCGA data. *Nucleic Acids Res.* 44 (8), e71. doi:10.1093/nar/gkv1507

Dauch, D., Rudalska, R., Cossa, G., Nault, J. C., Kang, T. W., Wuestefeld, T., et al. (2016). A MYC-aurora kinase A protein complex represents an actionable drug target in p53-altered liver cancer. *Nat. Med.* 22 (7), 744–753. doi:10.1038/nm.4107

Dennis, G., Jr., Sherman, B. T., Hosack, D. A., Yang, J., Gao, W., Lane, H. C., et al. (2003). David: Database for annotation, visualization, and integrated discovery. *Genome Biol.* 4 (5), R60. doi:10.1186/gb-2003-4-9-r60

Du, R., Huang, C., Liu, K., Li, X., and Dong, Z. (2021). Targeting AURKA in cancer: Molecular mechanisms and opportunities for cancer therapy. *Mol. Cancer* 20 (1), 15. doi:10.1186/s12943-020-01305-3

Feingold-Zadok, M., Chitayat, D., Chong, K., Injeyan, M., Shannon, P., Chapmann, D., et al. (2017). Mutations in the NEB gene cause fetal akinesia/arthrogryposis multiplex congenita. *Prenat. Diagn.* 37 (2), 144–150. doi:10.1002/pd.4977

Forner, A., Reig, M., and Bruix, J. (2018). Hepatocellular carcinoma. *Lancet (London, Engl.* 391 (10127), 1301–1314. doi:10.1016/S0140-6736(18)30010-2

Gao, M., Kong, W., Huang, Z., and Xie, Z. (2020). Identification of key genes related to lung squamous cell carcinoma using bioinformatics analysis. *Int. J. Mol. Sci.* 21 (8), E2994. doi:10.3390/ijms21082994

Gong, C., Ai, J., Fan, Y., Gao, J., Liu, W., Feng, Q., et al. (2019). NCAPG promotes the proliferation of hepatocellular carcinoma through PI3K/AKT signaling. *Onco. Targets. Ther.* 12, 8537–8552. doi:10.2147/OTT.S217916

Gui, L., Zhang, S., Xu, Y., Zhang, H., Zhu, Y., and Kong, L. (2021). UBE2S promotes cell chemoresistance through PTEN-AKT signaling in hepatocellular carcinoma. *Cell Death Discov.* 7 (1), 357. doi:10.1038/s41420-021-00750-3

Hao, L., Li, S., Peng, Q., Guo, Y., Ji, J., Zhang, Z., et al. (2021). Anti-malarial drug dihydroartemisinin downregulates the expression levels of CDK1 and CCNB1 in liver cancer. *Oncol. Lett.* 22 (3), 653. doi:10.3892/ol.2021.12914

Hu, W., Li, M., Chen, Y., and Gu, X. (2021). UBE2S promotes the progression and Olaparib resistance of ovarian cancer through Wnt/β-catenin signaling pathway. *J. Ovarian Res.* 14 (1), 121. doi:10.1186/s13048-021-00877-y

Jeng, Y. M., Peng, S. Y., Lin, C. Y., and Hsu, H. C. (2004). Overexpression and amplification of Aurora-A in hepatocellular carcinoma. *Clin. Cancer Res.* 10 (6), 2065–2071. doi:10.1158/1078-0432.ccr-1057-03

Lachenmayer, A., Alsinet, C., Savic, R., Cabellos, L., Toffanin, S., Hoshida, Y., et al. (2012). Wnt-pathway activation in two molecular classes of hepatocellular carcinoma and experimental modulation by sorafenib. *Clin. Cancer Res.* 18 (18), 4997–5007. doi:10.1158/1078-0432.CCR-11-2322

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

Leek, J. T., Johnson, W. E., Parker, H. S., Jaffe, A. E., and Storey, J. D. (2012). The sva package for removing batch effects and other unwanted variation in high-throughput experiments. *Bioinforma. Oxf. Engl.* 28 (6), 882–883. doi:10.1093/bioinformatics/bts034

Li, Z., Wang, Y., Li, Y., Yin, W., Mo, L., Qian, X., et al. (2018). Ube2s stabilizes β-Catenin through K11-linked polyubiquitination to promote mesendoderm specification and colorectal cancer development. *Cell Death Dis.* 9 (5), 456. doi:10.1038/s41419-018-0451-y

Llovet, J. M., Kelley, R. K., Villanueva, A., Singal, A. G., Pikarsky, E., Roayaie, S., et al. (2021). Hepatocellular carcinoma. *Nat. Rev. Dis. Prim.* 7 (1), 16018. doi:10.1038/nrdp.2016.18

Llovet, J. M., Montal, R., Sia, D., and Finn, R. S. (2018). Molecular therapies and precision medicine for hepatocellular carcinoma. *Nat. Rev. Clin. Oncol.* 15 (10), 599–616. doi:10.1038/s41571-018-0073-4

Nault, J. C., Couchy, G., Balabaud, C., Morcrette, G., Caruso, S., Blanc, J. F., et al. (2017). Molecular classification of hepatocellular adenoma associates with risk factors, bleeding, and malignant transformation. *Gastroenterology* 152 (4), 880–894. e6. doi:10.1053/j.gastro.2016.11.042

Peng, C., Song, Y., Chen, W., Wang, X., Liu, X., Wang, F., et al. (2018). FLVCR1 promotes the proliferation and tumorigenicity of synovial sarcoma through inhibiting apoptosis and autophagy. *Int. J. Oncol.* 52 (5), 1559–1568. doi:10.3892/ijo.2018.4312

Peng, S., Chen, X., Huang, C., Yang, C., Situ, M., Zhou, Q., et al. (2022). UBE2S as a novel ubiquitinated regulator of p16 and β-catenin to promote bone metastasis of prostate cancer. *Int. J. Biol. Sci.* 18 (8), 3528–3543. doi:10.7150/ijbs.72629

Piga, D., Magri, F., Ronchi, D., Corti, S., Cassandrini, D., Mercuri, E., et al. (2016). New mutations in NEB gene discovered by targeted next-generation sequencing in nemaline myopathy Italian patients. *J. Mol. Neurosci.* 59 (3), 351–359. doi:10.1007/s12031-016-0739-2

Qin, Y., Du, J., and Fan, C. (2020). Ube2S regulates Wnt/β-catenin signaling and promotes the progression of non-small cell lung cancer. *Int. J. Med. Sci.* 17 (2), 274–279. doi:10.7150/ijms.40243

Rebouissou, S., and Nault, J. C. (2020). Advances in molecular classification and precision oncology in hepatocellular carcinoma. *J. Hepatol.* 72 (2), 215–229. doi:10.1016/j.jhep.2019.08.017

Ritchie, M. E., Phipson, B., Wu, D., Hu, Y., Law, C. W., Shi, W., et al. (2015). Limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res.* 43 (7), e47. doi:10.1093/nar/gkv007

Song, Q., Wang, H., Bao, J., Pullikuth, A. K., Li, K. C., Miller, L. D., et al. (2015). Systems biology approach to studying proliferation-dependent prognostic subnetworks in breast cancer. *Sci. Rep.* 5, 12981. doi:10.1038/srep12981

Song, Q., Zhu, X., Jin, L., Chen, M., Zhang, W., and Su, J. (2022). Smgr: A joint statistical method for integrative analysis of single-cell multi-omics data. *Nar. Genom. Bioinform.* 4 (3), lqac056. doi:10.1093/nargab/lqac056

Sun, H., Zhang, H., Yan, Y., Li, Y., Che, G., Zhou, C., et al. (2022). NCAPG promotes the oncogenesis and progression of non-small cell lung cancer cells through upregulating LGALS1 expression. *Mol. Cancer* 21 (1), 55. doi:10.1186/s12943-022-01533-9

Sung, H., Ferlay, J., Siegel, R. L., Laversanne, M., Soerjomataram, I., Jemal, A., et al. (2021). Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *Ca. Cancer J. Clin.* 71 (3), 209–249. doi:10.3322/caac.21660

Tang, B., Zhu, J., Li, J., Fan, K., Gao, Y., Cheng, S., et al. (2020). The ferroptosis and iron-metabolism signature robustly predicts clinical diagnosis, prognosis and immune microenvironment for hepatocellular carcinoma. *Cell Commun. Signal.* 18 (1), 174. doi:10.1186/s12964-020-00663-1

Tibshirani, R. (1997). The lasso method for variable selection in the Cox model. *Stat. Med.* 16 (4), 385–395. doi:10.1002/(sici)1097-0258(19970228)16:4<385::aid-sim380>3.0.co;2-3

Tilghman, J., Wu, H., Sang, Y., Shi, X., Guerrero-Cazares, H., Quinones-Hinojosa, A., et al. (2014). HMMR maintains the stemness and tumorigenicity of glioblastoma stem-like cells. *Cancer Res.* 74 (11), 3168–3179. doi:10.1158/0008-5472.CAN-13-2103

Villanueva, A. (2019). Hepatocellular carcinoma. *N. Engl. J. Med.* 380 (15), 1450–1462. doi:10.1056/NEJMra1713263

Wang, H., Nie, X., Li, X., Fang, Y., Wang, D., Wang, W., et al. (2020). Bioinformatics analysis and high-throughput sequencing to identify differentially expressed genes in nebulin gene (NEB) mutations mice. *Med. Sci. Monit.* 26, e922953. doi:10.12659/MSM.922953

Wang, K., Lim, H. Y., Shi, S., Lee, J., Deng, S., Xie, T., et al. (2013). Genomic landscape of copy number aberrations enables the identification of oncogenic drivers in hepatocellular carcinoma. *Hepatol. Baltim. Md* 58 (2), 706–717. doi:10.1002/hep.26402

Weng, L., Du, J., Zhou, Q., Cheng, B., Li, J., Zhang, D., et al. (2012). Identification of cyclin B1 and Sec62 as biomarkers for recurrence in patients with HBV-related

hepatocellular carcinoma after surgical resection. *Mol. Cancer* 11, 39. doi:10.1186/1476-4598-11-39

Wu, J. M., Chen, C. T., Coumar, M. S., Lin, W. H., Chen, Z. J., Hsu, J. T., et al. (2013). Aurora kinase inhibitors reveal mechanisms of HURP in nucleation of centrosomal and kinetochore microtubules. *Proc. Natl. Acad. Sci. U. S. A.* 110 (19), E1779–E1787. doi:10.1073/pnas.1220523110

Wu, T., Merbl, Y., Huo, Y., Gallop, J. L., Tzur, A., and Kirschner, M. W. (2010). UBE2S drives elongation of K11-linked ubiquitin chains by the anaphase-promoting complex. *Proc. Natl. Acad. Sci. U. S. A.* 107 (4), 1355–1360. doi:10.1073/pnas.0912802107

Xia, P., Zhang, H., Xu, K., Jiang, X., Gao, M., Wang, G., et al. (2021). MYC-targeted WDR4 promotes proliferation, metastasis, and sorafenib resistance by inducing CCNB1 translation in hepatocellular carcinoma. *Cell Death Dis.* 12 (7), 691. doi:10.1038/s41419-021-03973-5

Yang, J. D., and Heimbach, J. K. (2020). New advances in the diagnosis and management of hepatocellular carcinoma. *BMJ Clin. Res. ed* 371, m3544. doi:10.1136/bmj.m3544

Zhang, Q., Su, R., Shan, C., Gao, C., and Wu, P. (2018). Non-SMC condensin I complex, subunit G (NCAPG) is a novel mitotic gene required for hepatocellular cancer cell proliferation and migration. *Oncol. Res.* 26 (2), 269–276. doi:10.3727/096504017X15075967560980

Zucman-Rossi, J., Villanueva, A., Nault, J. C., and Llovet, J. M. (2015). Genetic landscape and biomarkers of hepatocellular carcinoma. *Gastroenterology* 149 (5), 1226–1239. e4. doi:10.1053/j.gastro.2015.05.061