European Review for Medical and Pharmacological Sciences 2022; 26: 8945-8958

Exploration and validation of the prognostic value of RNA-binding proteins in hepatocellular carcinoma

J. WANG[1], K. HAN[2], Y. LI[1], C. ZHANG[1], W.-H. CUI[1], L.-H. ZHU[1], T. LUO[1], C.-J. BIAN[1]

[1]Department of General Surgery, [2]Department of Thoracic Surgery, Xuanwu Hospital, Capital Medical University, Beijing, China

Abstract. – OBJECTIVE: Hepatocellular carcinoma (HCC) is one of the most common malignant tumors worldwide. Increasing evidence suggests that the dysregulation of RNA-binding proteins (RBPs) is involved in the development of various cancers. However, there is a paucity of studies investigating the roles of RBPs in HCC.

MATERIALS AND METHODS: Data on HCC samples were downloaded from The Cancer Genome Atlas (TCGA) and Gene Expression Omnibus (GEO) databases (available at: www.ncbi.nlm.nih.gov/geo), and data regarding human RBPs were integrated from SONAR, XRNAX, and CARIC results. We identified modules associated with prognosis using weighted gene co-expression network analysis (WGCNA) and performed functional enrichment analysis. Univariate and least absolute shrinkage and selection operator (LASSO) regression analyses were used to identify prognostic RBPs and establish a prediction model. According to the median risk score, we separated patients into high- and low-risk groups and investigated the differences in immune cell infiltration, somatic mutations, and gene set enrichment. Univariate and multivariate regression analyses were used to identify prognostic factors for HCC. A nomogram was constructed, and its performance was evaluated with calibration curves.

RESULTS: Sixteen RBPs (*MEX3A, TTK, MRPL53, IQGAP3, PFN2, IMPDH1, TCOF1, DYNC1LI1, EIF2B4, NOL10, GNL2, EIF1B, PSMD1, AHSA1, SEC61A1*, and *YBX1*) were identified as prognostic genes, and a prognostic model was established. Survival analysis indicated that the model had good predictive performance and that a high-risk score was significantly related to a poor prognosis. Additionally, there were significant differences in immune cell infiltration, somatic mutations, and gene set enrichment between the high- and low-risk groups. Univariate and multivariate regression analyses indicated that the RBP-based signature was an independent prognostic factor for HCC. The nomogram based on 16 RBPs performed well in predicting the overall survival of HCC patients.

CONCLUSIONS: The RBP-based signature is an independent prognostic factor for HCC, and this study could provide an innovative method for analyzing prognostic biomarkers and therapeutic targets for HCC.

Key Words:
Hepatocellular carcinoma, RNA-binding protein, WGCNA, Prognosis, Survival.

Introduction

Hepatocellular carcinoma (HCC), which accounts for 75-85% of primary liver cancers, is the sixth most common cancer and the third most common cause of cancer-related death worldwide[1]. Many factors contribute to the pathogenesis of HCC, including hepatitis B virus (HBV) or hepatitis C virus (HCV) infection, alcohol consumption, aflatoxin exposure, and obesity[2,3]. Despite advances in the diagnosis and treatment of HCC over the past few decades, the average 5-year relative survival rate of HCC is less than 20%[4]. Tumor-node-metastasis (TNM) staging is a classic prognostic model that helps predict HCC prognosis and is currently widely used in clinical practice[5]. However, due to the great heterogeneity and complexity of HCC, the predictive efficacy of the classic model is still unsatisfactory. Therefore, a systematic understanding of the molecular mechanisms underlying HCC is required to develop a more reliable prediction model for addressing risk stratification and achieving early detection.

RBP-binding proteins (RBPs) are inherently pleiotropic proteins that regulate different aspects of RNA metabolism and function at the posttran-

Corresponding Author: Chunjing Bian, MD; e-mail: aacj2020@163.net

scriptional level through processes such as RNA stability, localization, export, processing, splicing, degradation, and translation[6,7]. To conduct these functions, RBPs need to bind to their target RNAs through specific RNA-binding domains (RBDs) to form ribonucleoproteins[8]. Accumulating evidence has shown that RBPs are involved in the occurrence and development of various cancers via their influence on the physiological processes of cells. Considering that dysregulated RBPs are closely related to cancer initiation and progression, it is reasonable to believe that an in-depth study of the expression profiles of RBPs in cancers has immense potential in clinical practice.

An increasing number of studies[9,10] have demonstrated that RBP-based biomarkers have promising value in survival prediction for patients with HCC. Han et al[9] found not only that RBP sorbin and SH3 domain-containing 2 (SORBS2) inhibited HCC tumorigenesis and metastasis via posttranscriptional regulation of RAR related orphan receptor A (RORA) expression but also that downregulated expression of SORBS2 was strongly correlated with a poor clinical prognosis in HCC patients. Zhao et al[10] reported that the RBP RNA-binding motif protein 10 (RBM10) was expressed at low levels in HCC tissues and cell lines and that low RBM10 expression was an indicator of unfavorable patient survival. With the development of next-generation sequencing technology and the establishment of public databases, it is possible to acquire gene expression profiles, and develop a more reliable prediction model for the assessment of the prognosis of HCC using bioinformatics methods. To the best of our knowledge, there is a dearth of literature reporting RBP-based models for predicting the survival of patients with HCC. Thus, it is necessary to construct an RBP-based prediction model that can reliably predict HCC prognosis.

In this study, we first identified key modules highly associated with the survival of patients with HCC using weighted gene co-expression network analysis (WGCNA) and conducted functional enrichment analysis. Then, we performed univariate and LASSO regression analyses of the module genes, which confirmed the prognostic value of 16 RBPs and constructed a prediction model based on these RBPs. Finally, we estimated the predictive ability of the model using univariate and multivariate Cox regression analyses and found that it could serve as an independent prognostic factor in HCC.

We present the following article in accordance with the STROBE reporting checklist.

Materials and Methods

Data Processing

First, we integrated a total of 3,437 human RBPs from three sources: SONAR, XRNAX, and CARIC. Then, we downloaded data for 364 HCC samples, which included RNA-seq data, the corresponding clinicopathological features, and somatic mutation data from the UCSC Xena browser (http://xena.ucsc.edu/). Additionally, we downloaded the GSE54236 dataset and extracted 81 HCC samples with RNA-seq data and the corresponding clinicopathological features from the Gene Expression Omnibus (GEO) database (http://www.ncbi.nlm.nih.gov/geo) as a validation cohort. The 3,437 human RBPs were integrated with those from the TCGA database, and the overlapping RBPs were reserved for subsequent analysis.

Weighted Gene Co-expression Network Construction

The expression profiles of the 3,301 overlapping RBPs were used to construct a weighted co-expression network using the WGCNA package in R (version 4.2.0, The R Foundation for Statistical Computing, Vienna, Austria). WGCNA methodology analysis was used to reveal the correlation between the gene co-expression modules and the clinical features of interest. First, we converted the expression levels of RBPs into a similarity matrix based on Pearson's correlation value between paired genes. Second, we transformed the similarity matrix into an adjacency matrix using the optimal soft-threshold power (β). Selecting the appropriate β value can enhance strong correlations and penalize weak correlations at an exponential level. Third, the adjacency matrix was transformed into a topological matrix (TOM). To classify genes with similar expression patterns into the same modules, we performed average linkage hierarchical clustering according to the TOM-based dissimilarity measure with a minimum size of 30 for the gene dendrogram. In this study, we set a minimum module size of 30 for the gene dendrogram and selected a cutoff of 0.25 for the module dendrogram. Then, we merged some modules.

Identification of Prognosis-Related Modules

We first calculated the module eigengene (ME), which represents the first principal component of the gene expression matrix of each module. Then, we calculated the correlation between the clinical traits and ME in each module. Finally, we calculated the gene significance (GS) of each gene in the module, which could quantify the correlation between the genes and traits of interest. According to these two parameters, we selected key modules that significantly affected survival for further analysis.

Gene Ontology (GO) and Kyoto Encyclopedia of Genes and Genomes (KEGG) Functional Enrichment Analyses

To detect the biological functions of the key module, we performed GO and KEGG pathway analyses using the clusterProfiler package in R (version 4.2.0, The R Foundation for Statistical Computing, Vienna, Austria). The GO analysis results comprised three categories: cellular component (CC), molecular function (MF), and biological process (BP). A false discovery rate (FDR) < 0.05 was considered the threshold for significance.

Prediction Model Construction and Evaluation

First, univariate Cox regression analysis was performed to screen the candidate RBPs ($p < 0.01$) in the key module. Then, to enhance the adaptability of the prediction model, LASSO regression analysis was used to further determine prognosis-related RBPs by filtering high-dimensional data. The risk score for each HCC patient was calculated using the following formula:

$$\text{Risk score} = \beta_1 * \text{Exp}_1 + \beta_2 * \text{Exp}_2 + \beta_i * \text{Exp}_i$$

where the regression coefficient (β) was derived from the LASSO regression analysis, and Exp represents the expression levels of RBPs.

To estimate the predictive ability of this prediction model, we divided the HCC patients from the TCGA dataset into high- and low-risk groups based on the median risk score. Then, we compared the difference in overall survival (OS) between the two groups via Kaplan-Meier analysis using the survival and survminer packages in R (version 4.2.0, The R Foundation for Statistical Computing, Vienna, Austria). Furthermore, a receiver operating characteristic (ROC) curve was generated to assess the accuracy of this model using the survivalROC R package (The R Foundation for Statistical Computing, Vienna, Austria). Additionally, 81 HCC samples extracted from the GSE54236 dataset were used as a validation cohort to confirm the predictive ability of this model.

Analysis of Differences in the Infiltration of 22 Immune Cell Types Between the High- and Low-Risk Groups

CIBERSORT (https://cibersort.stanford.edu/) was used to quantify the abundances of 22 immune cell types in a mixed cell population based on standardized gene expression data. In detail, the CIBERSORT online analytical platform was used to discriminate between 22 immune cell infiltrates based on the leukocyte gene signature matrix (547 genes). In this study, we utilized CIBERSORT to compare the differences in the infiltration levels of 22 immune cell types between the high- and low-risk groups.

Differential Analysis of Somatic Mutation Data Between High- and Low-Risk Groups

The somatic mutation profiles were downloaded from TCGA database in the form of a mutation annotation format (MAF) file. To investigate the differences in mutation data between the high- and low-risk groups, we analyzed and visualized mutation data using the maftools package in R (version 4.2.0, The R Foundation for Statistical Computing, Vienna, Austria).

Gene Set Enrichment Analysis (GSEA) Between High- and Low-Risk Groups

GSEA is a computational method that determines whether an *a priori* defined set of genes shows statistically significant, concordant differences between two biological states. In this study, we explored potential differences in biological functions between the high- and low-risk groups based on the Molecular Signatures Database (MSigDB) (available at: http://www.broadinstitute.org/gsea) collection using the clusterProfiler package in R (version 4.2.0, The R Foundation for Statistical Computing, Vienna, Austria). Gene set permutations were set at 1,000 for each analysis. Differences with an FDR value of < 0.05 were considered significant.

Identification of the Prognostic Factors for OS in HCC

To identify independent prognostic factors, four predominant clinicopathological characteristics, namely, age, sex, TNM stage, and the risk score of the RBP-based signature, were analyzed using univariate and multivariate Cox regression analyses. By using the "rms", "foreign", and "survival" packages in R (version 4.2.0, The R Foundation for Statistical Computing, Vienna, Austria), we constructed a nomogram consisting of relevant clinicopathological parameters and independent prognostic factors to assess the probability of 2-, 3-, and 5-year OS among HCC patients based on multivariate Cox regression analysis. Moreover, we plotted calibration curves to estimate the accuracy of the nomogram by comparing the actual observed survival rates with the predicted survival probability.

Statistical Analysis

All statistical analyses were conducted using R. Kaplan-Meier curves with the log-rank test were used to compare survival differences between the high- and low-risk groups. The predictive ability of the model was determined by ROC curve analysis. Then, univariate and multivariate Cox regression analyses were used to determine the independent prognostic factors. Finally, a nomogram was constructed to visualize the results of the multivariate Cox regression analysis. Calibration curves were plotted to assess the effectiveness of the nomogram. A p-value of < 0.05 was considered to indicate a significant difference.

Results

Identification of Prognosis-Related Modules by WGCNA

The workflow of this study is shown in Figure 1. A total of 3,301 overlapping RBPs obtained from the TCGA dataset were used to construct a weighted co-expression network using the WGCNA package in R (version 4.2.0). To build a scale-free network, we chose a soft threshold (β) of 8 (Figure 2A). We identified 10 co-expression modules based on the TOM and average linkage hierarchical clustering (Figure 2B). The number of genes in each module is shown in **Supplementary Table I**. The results of the correlation analysis between co-expression modules and clinical characteristics (including age, sex, OS rate, and OS time) are shown in Figure 2C and indicate that the blue module (including 545 genes) is most negatively correlated with the OS time of patients (correlation coefficient = -0.21, p = 5e-05). Therefore, we chose the blue module as the key module for further analysis.

Functional Enrichment Analysis for Key Modules

We performed GO and KEGG analyses to explore the biological functions of the blue module using the clusterProfiler package in R (version 4.2.0). The top six enriched GO terms (including BP, CC, and MF) and the top 12 KEGG pathways are shown in Figure 3. The results of GO analysis indicated that RBPs were significantly enriched in BPs associated with RNA splicing, ribonucleoprotein complex biogenesis, mRNA splicing, and RNA localization. For the CC category, RBPs were significantly enriched in the chromosomal region, nuclear speck, spliceosomal complex, spindle, nuclear periphery, and catalytic step 2 spliceosome. The MF analysis showed that RBPs were enriched in catalytic activity (acting on RNA and DNA), adenosine triphosphatase (ATPase) activity, helicase activity, single-stranded RNA binding, and mRNA 3'-untranslated region (UTR) binding. In addition, the results of the KEGG analysis indicated that the RBPs were enriched in the spliceosome, RNA transport, DNA replication, the cell cycle, ribosome biogenesis in eukaryotes, mismatch repair, base excision repair, amyotrophic lateral sclerosis, RNA degradation, the mRNA surveillance pathway, the Fanconi anemia pathway, and oocyte meiosis.

Identification of Prognosis-Related RBPs

Among all the blue module genes, 420 RBPs were selected as candidate RBPs with the criteria of $p < 0.01$ by univariate Cox regression analysis (**Supplementary Table II**). Subsequently, LASSO regression analysis was used to further assess the prognostic value of these 420 RBPs, and 16 RBPs were identified to have a good ability to predict the prognosis of patients with HCC (Figure 4).

Construction of a Predictive Risk Score Model

The 16 RBPs derived from LASSO regression were used to construct a prediction model. The risk score of each patient was calculated according to the following formula:

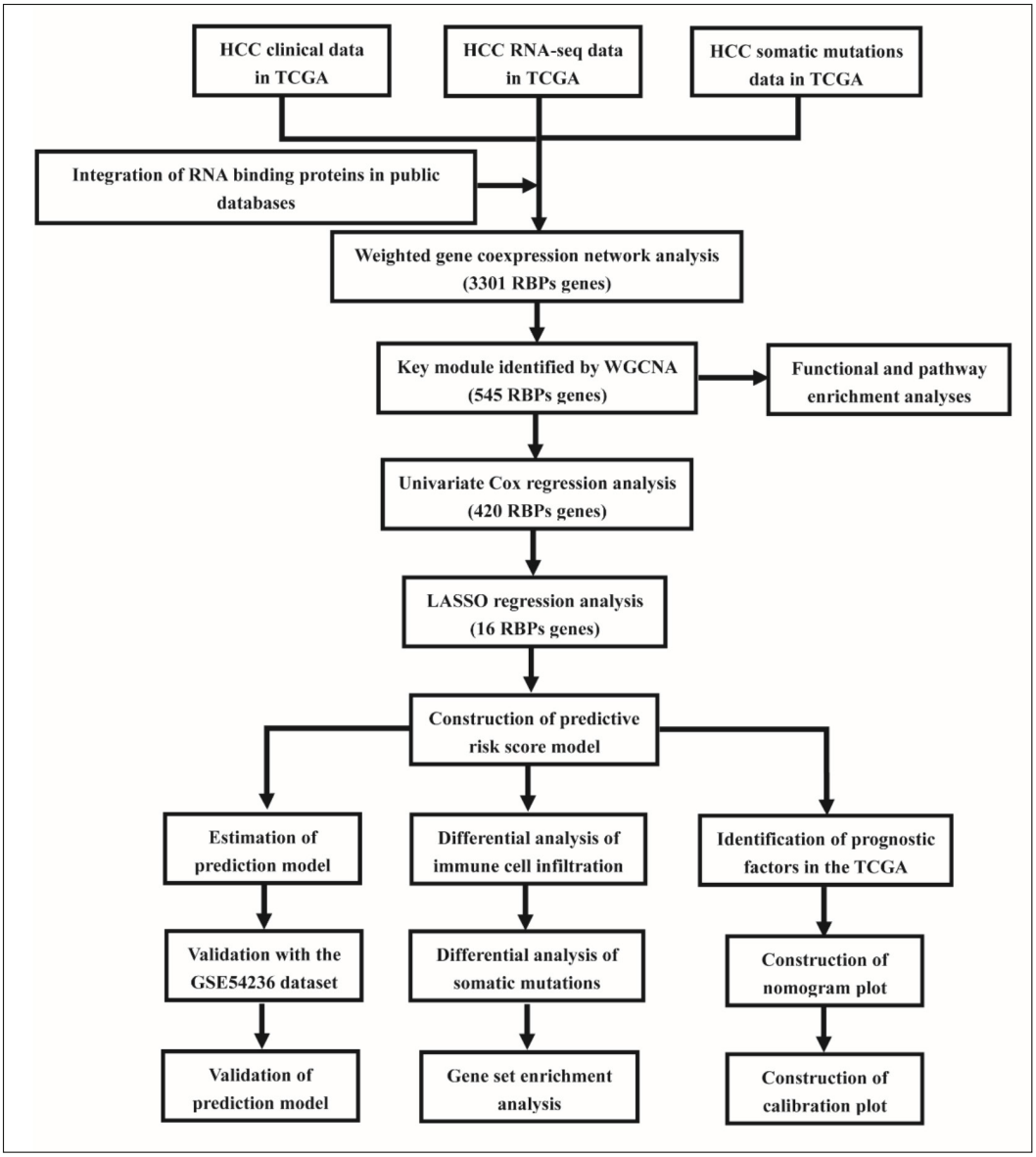

Figure 1. The workflow of the study.

Risk score = (0.07 * Exp*MEX3A*) + (0.01 * Exp*TTK*) + (0.02 * Exp*MRPL53*) + (0.12 * Exp*IQGAP3*) + (0.04 * Exp*PFN2*) + (0.01 * Exp*IMPDH1*) + (0.02 * Exp*TCOF1*) + (0.08 * Exp*DYNC1LI1*) + (0.14 * Exp*EIF2B4*) + (0.08 * Exp*NOL10*) + (0.13 * Exp*GNL2*) + (0.06 * Exp*EIF1B*) + (0.25 * Exp*PSMD1*) + (0.02 * Exp*AHSA1*) + (0.12 * Exp*SEC61A1*) + (0.19 * Exp*YBX1*).

Assessment and Validation of the Predictive Ability of the Risk Score Model

To assess the predictive ability of this model, we divided 364 HCC samples (obtained from TCGA database) into high- (n = 182) and low-risk (n = 182) groups according to the median risk score (median = 5.04) (Figure 5A). Similarly, we divided 81 HCC samples (obtained from the GSE54236 dataset) into high- (n = 40) and low-risk (n = 41) groups according to the median risk score (median = 4.36) (Figure 5B). The survival analysis of the two different datasets indicated that the high-risk groups had a shorter survival time than the low-risk groups (Figure 5C). To evaluate the predictive ability of this model, we performed ROC curve analysis based on the data obtained from the TCGA and GEO series (GSE)

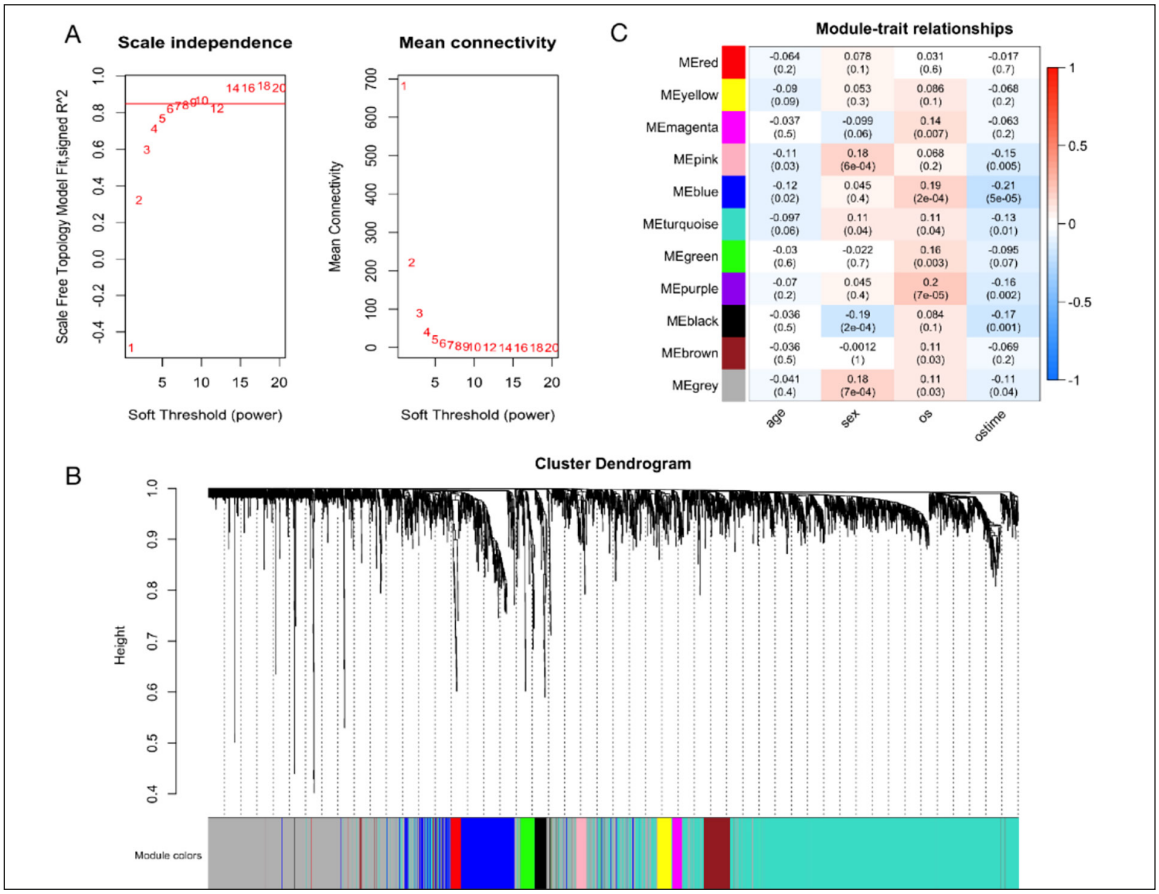

Figure 2. Key module identified by WGCNA. **A,** Determination of the soft-threshold power in WGCNA. The left panel represents the relationship between the soft threshold and scale-free R^2 values. The right panel represents the relationship between the soft threshold and mean connectivity. **B,** The hierarchical dendrogram shows the co-expression modules identified by WGCNA. Each branch in the figure represents one gene, and every color below represents one co-expression module. **C,** The heatmap shows the correlations between the modules and the clinical features. The values in the squares indicate the correlation coefficients and *p* values. A positive coefficient indicates a positive correlation. A negative coefficient indicates a negative correlation.

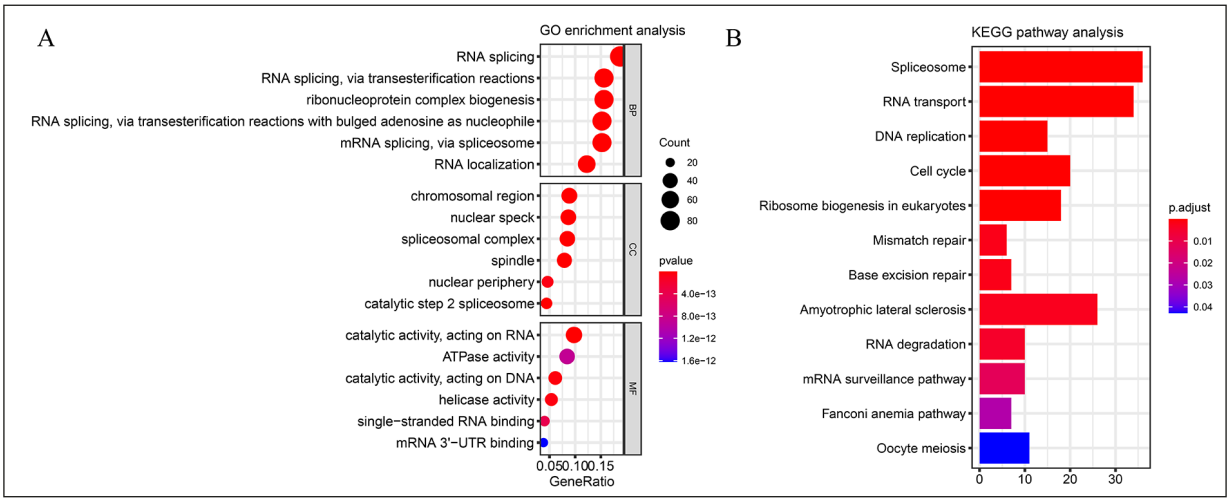

Figure 3. GO and KEGG pathway enrichment analyses of blue module genes. **A,** GO analysis results for the blue module genes showing the enriched BP, CC, and MF terms. The y-axis shows significantly enriched terms (FDR < 0.05). **B,** KEGG analysis of blue module genes. The y-axis shows significantly enriched pathways (FDR < 0.05).

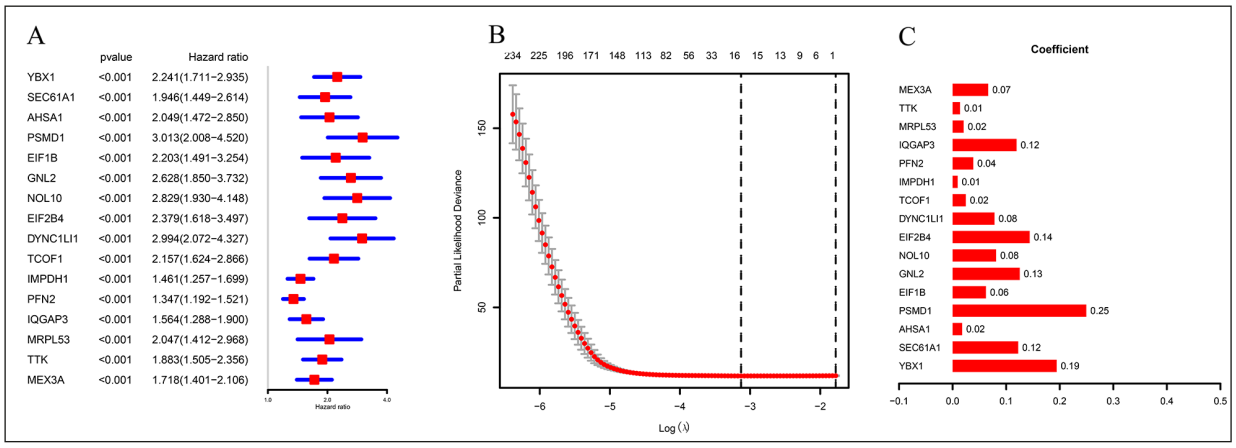

Figure 4. LASSO Cox regression model construction. **A,** Forest plot of hazard ratios demonstrating the prognostic values of RBPs by LASSO Cox regression analysis. **B,** Selection of the number of factors by LASSO Cox regression analysis. **C,** Determination of the LASSO model coefficient.

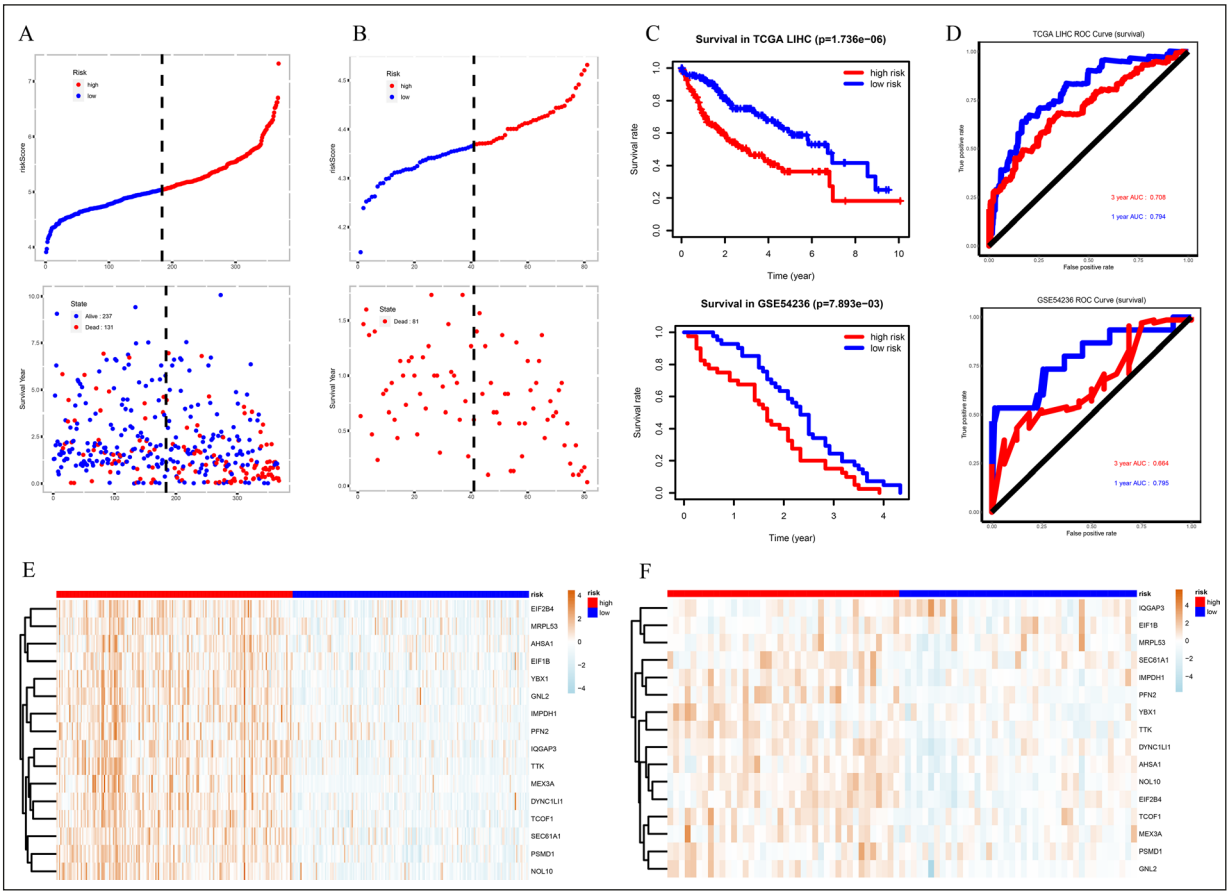

Figure 5. Survival analysis with the RBP-based prediction model in the TCGA and GSE databases. **A,** The risk score distribution (top) and survival status distribution (bottom) in the TCGA database. **B,** The risk score distribution (top) and survival status distribution (bottom) in the GSE dataset. **C,** Kaplan-Meier curves for the OS of the high- and low-risk groups in the TCGA (top) and GSE (bottom) databases. **D,** The ROC curves for predicting OS in patients according to the risk score in the TCGA (top) and GSE (bottom) datasets. **E,** Heatmap showing the expression profiles of the 16 RBPs for the high- and low-risk groups in the TCGA database. **F,** Heatmap showing the expression profiles of the 16 RBPs for the high- and low-risk groups in the GSE dataset.

datasets. The area under the ROC curve (AUC) values were 0.794 (TCGA, 1-year OS), 0.708 (TCGA, 3-year OS), 0.795 (GSE, 1-year OS), and 0.664 (GSE, 3-year OS), indicating a relatively good performance (Figure 5D). In addition, the expression levels of 16 RBPs were increased in the high-risk groups compared to the low-risk groups, indicating that high expression of these RBPs is associated with a poor prognosis in HCC patients (Figure 5E, 5F).

Analysis of the Differences in the nfiltration of 22 Immune Cell Types Between the High- and Low-Risk Groups

To estimate the differences in the infiltration of 22 immune cell types between the high- and low-risk groups, we analyzed the gene expression profiles from the TCGA database with the CIBERSORT platform (available at: https://cibersortx.stanford.edu). As shown in Figure 6, the infiltration levels of activated memory cluster of differentiation 4 $(CD4)^+$ T cells, M0 macrophages and resting dendritic cells were significantly higher in the high-risk group than in the low-risk group, and the infiltration levels of resting memory $CD4^+$ T cells, activated natural killer (NK) cells and resting mast cells were significantly higher in the low-risk group. Further study of the infiltration of 22 immune cell types in HCC patients may contribute to the development of immunotherapy.

Analysis of Differences in Somatic Mutations Between the High- and Low-Risk Groups

To investigate the differences between the high- and low-risk groups at the somatic mutation level, we analyzed somatic mutation data from the TCGA database. The top 30 genes of the high- and low-risk groups are illustrated in Figure 7A, 7B. The results indicated that the proportion of *TP53* mutations was significantly higher in the high-risk group (48%) than in the low-risk group (11%). The most common types of *TP53* mutations were missense mutations, frameshift deletions, nonsense mutations, and splice sites in both the high- and low-risk groups.

GSEA of the High- and Low-Risk Groups

To further identify the differentially activated signaling pathways between the high- and low-risk groups, we performed GSEA. Figure 8 shows the top 10 significant signaling pathways associ-

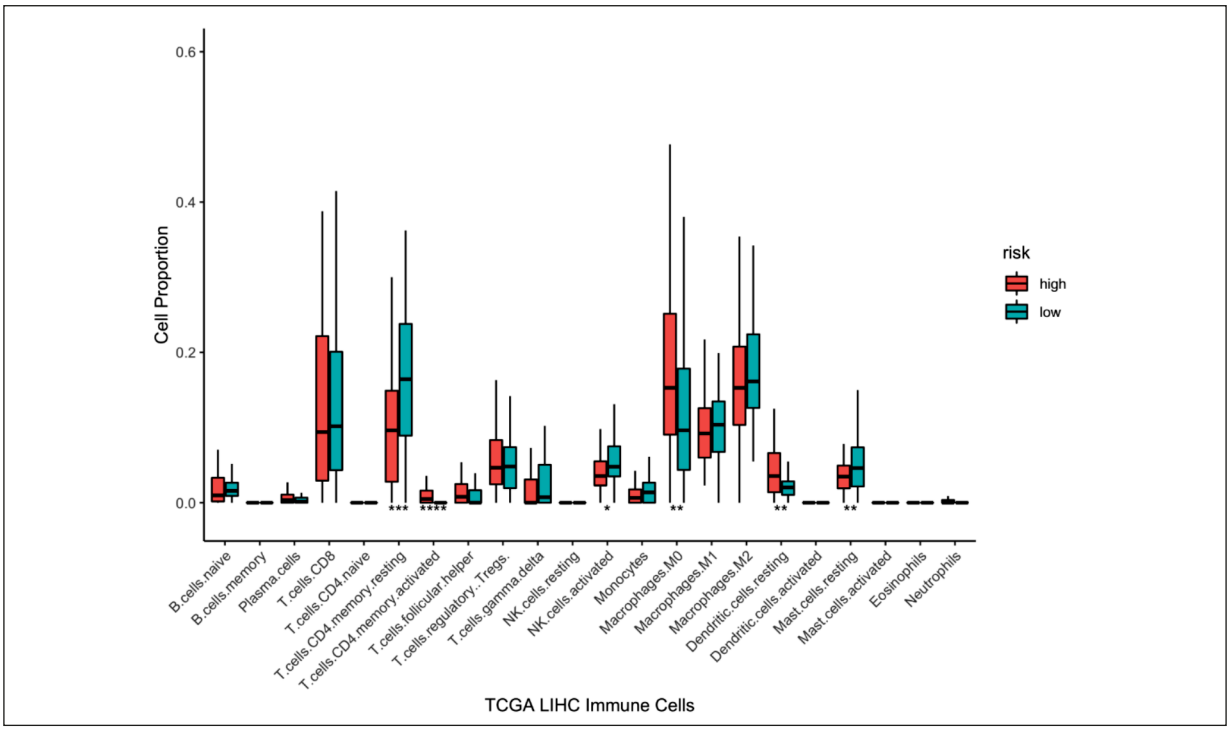

Figure 6. Relative proportions of 22 infiltrating immune cell types between the high- and low-risk groups in the TCGA database. Red represents the high-risk group, and blue represents the low-risk group.

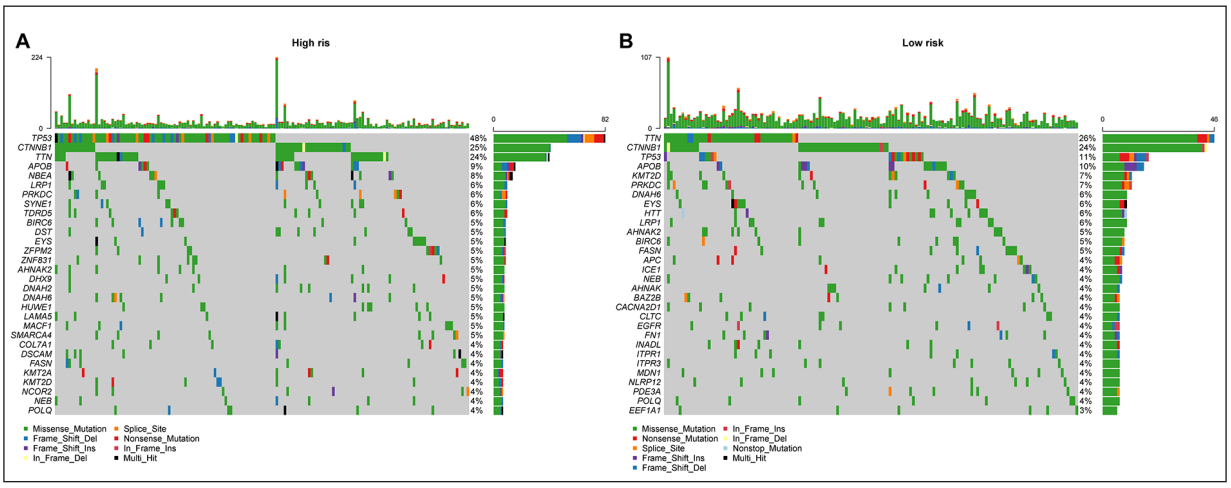

Figure 7. Association between the risk score and somatic mutations in the TCGA dataset. Differentially mutated genes between the **(A)** high- and **(B)** low-risk groups.

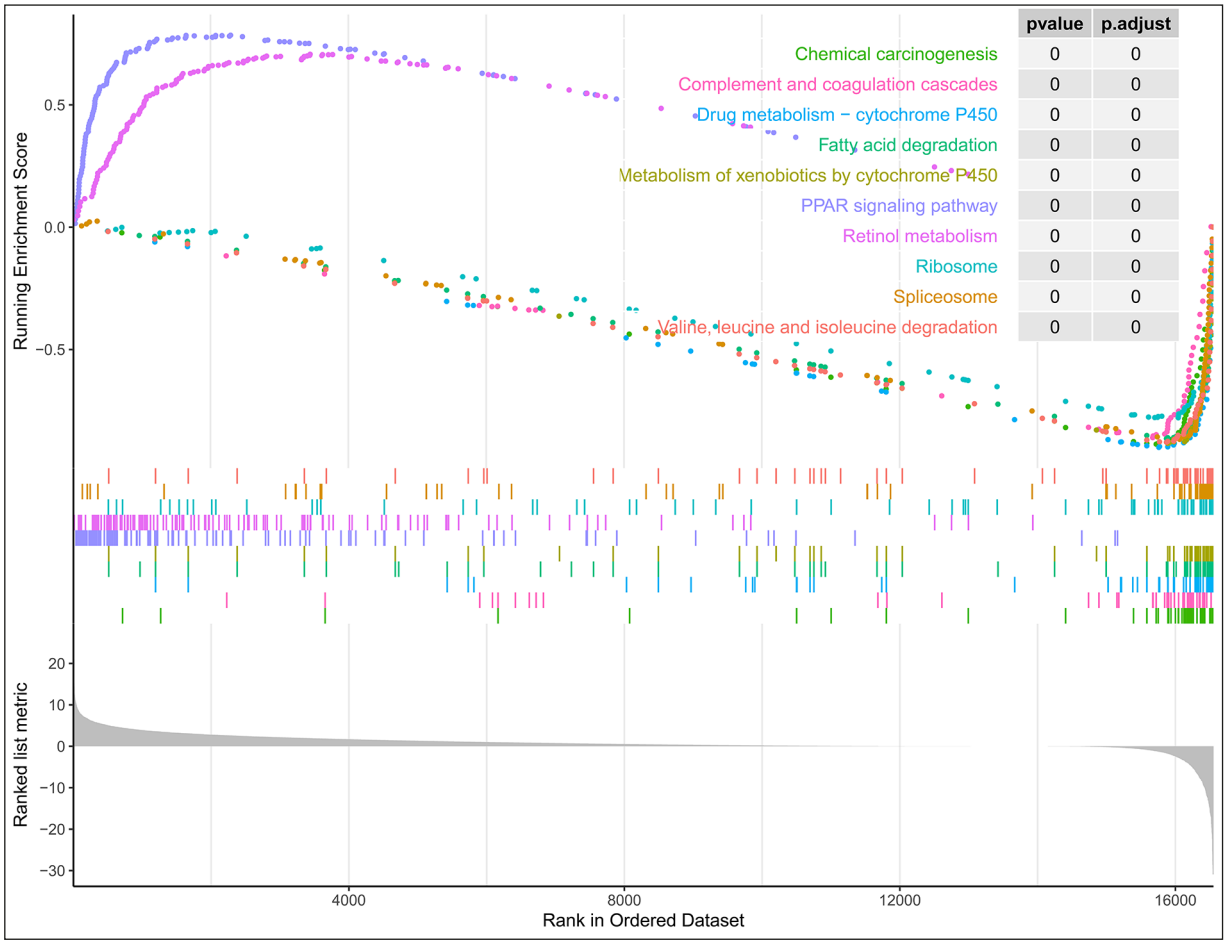

Figure 8. Results of GSEA of the high- and low-risk groups in the TCGA database. The broken line in the upper part of the figure shows the enrichment fraction of the 10 pathways broken line, and the lines in the lower part of the figure correspond to the genes of each pathway.

ated with the risk score. These included chemical carcinogenesis, complement and coagulation cascades, drug metabolism-cytochrome P450, fatty acid degradation, metabolism of xenobiotics by cytochrome P450, the peroxisome proliferator-activated receptor (PPAR) signaling pathway, retinol metabolism, ribosome, spliceosome, and valine, leucine and isoleucine degradation.

Identification of the Prognostic Factors for OS in HCC

To identify prognostic factors for OS in patients with HCC from the TCGA database, we performed univariate and multivariate Cox regression analyses. The results of both the univariate and multivariate Cox regression analyses showed that TNM stage and risk score had significant prognostic value for OS in HCC ($p < 0.05$) (Figure 9A, 9B). To develop a quantitative method for predicting prognosis, we built a nomogram plot based on the multivariate Cox regression analysis results (Figure 9C). This allowed us to calculate the estimated probabilities of 2-, 3-, and 5-year OS by drafting a vertical line between the total point axis and each prognosis axis. Moreover, we drew calibration curves to assess the predictive performance of the nomogram. As shown in Figure 9D, satisfactory agreement was observed between the predicted and observed outcomes for 2-, 3-, and 5-year OS, which indicated that the nomogram had good efficacy in predicting the survival of patients with HCC.

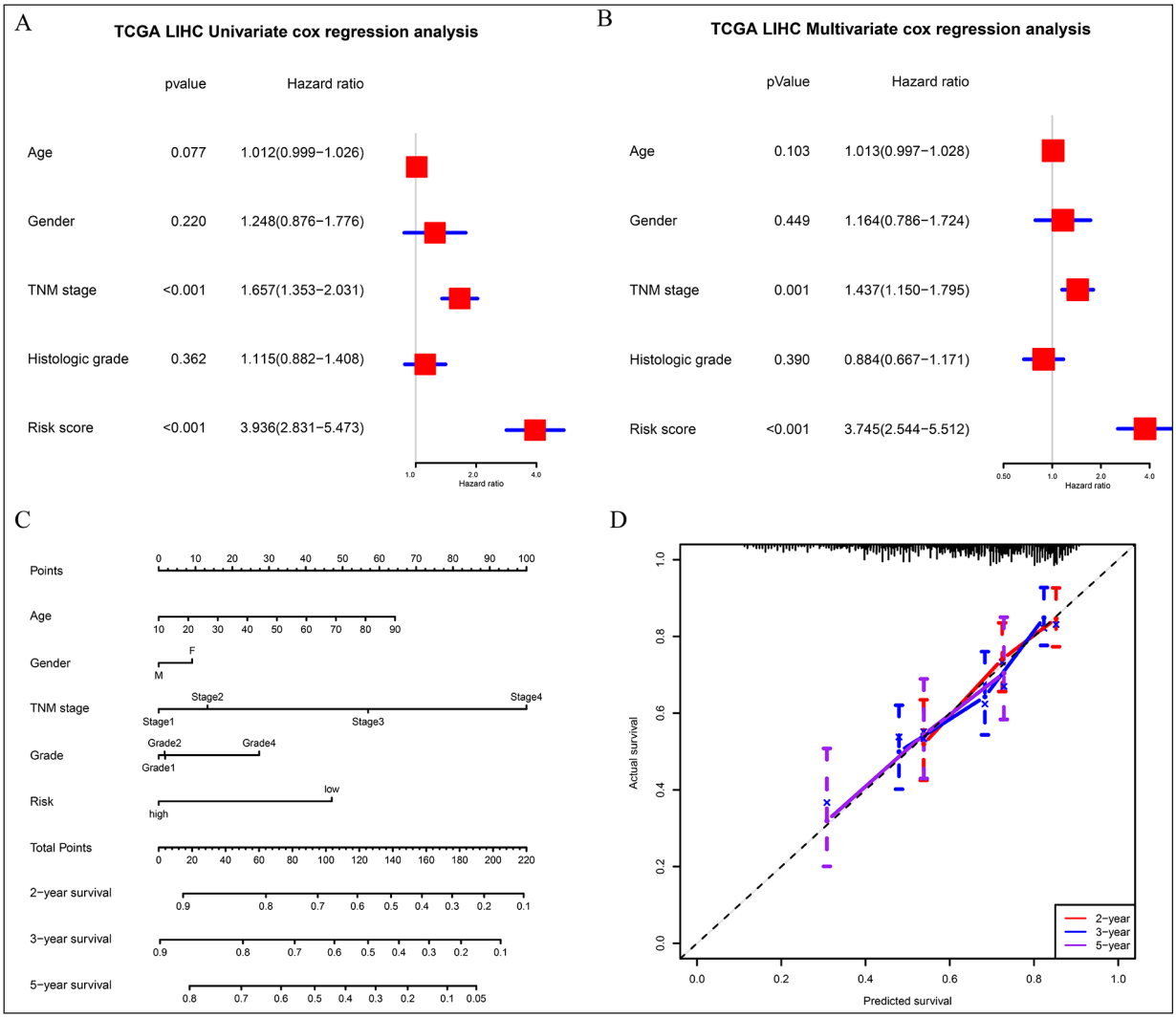

Figure 9. Identification of prognostic factors in HCC patients in the TCGA database. Univariate **(A)** and multivariate **(B)** Cox regression analyses of factors associated with OS in HCC patients from the TCGA database. **C,** The nomogram to predict the 2-, 3-, and 5-year OS of HCC patients in the TCGA database. **D,** The calibration curves of the nomogram to predict OS at 2, 3 and 5 years.

Discussion

HCC is the most common type of primary liver cancer and is characterized by uncontrolled cell proliferation, which mainly results from the activation of proto-oncogenes and the inactivation of tumor suppressor genes. Recently, emerging evidence has demonstrated that RBPs play an important role in the development and progression of various cancers by regulating RNA metabolism and function at the posttranscriptional level. However, little is currently known about the specific functional roles of RBPs in HCC.

In this study, we first identified the blue module as a module that is significantly associated with survival in HCC by WGCNA and performed functional enrichment analysis of this module. Then, we performed univariate and LASSO regression analyses of the blue module genes and identified 16 RBPs, with which we constructed a prediction model. We estimated the predictive ability of this model using Kaplan-Meier curves, ROC curves, risk score plots, and heatmaps based on the median risk value. Furthermore, we investigated the differences in immune cell infiltration, somatic mutations, and gene set enrichment between the high- and low-risk groups. Finally, we performed univariate and multivariate Cox regression analyses and identified that the TNM stage and RBP-based risk score were independent prognostic factors for OS in HCC.

The GO enrichment analysis demonstrated that the blue module genes were significantly related to the following terms: RNA splicing, ribonucleoprotein complex biogenesis, RNA localization, chromosomal region, nuclear speck, spliceosomal complex, spindle, nuclear periphery, catalytic step 2 spliceosome, catalytic activity (acting on RNA and DNA), ATPase activity, helicase activity, single-stranded RNA binding, and mRNA 3'-UTR binding. Previous studies[11-14] have demonstrated that aberrant RNA splicing contributes to tumorigenesis in humans. Moreover, RBPs need to bind to their target mRNAs through specific RBDs to form ribonucleoprotein complexes and regulate mRNA translation[15]. Dong et al[16] found that the RNA binding motif protein 3 (RBM3) can promote HCC cell proliferation by regulating the formation of circular RNA SCD-circRNA 2. Zhao et al[17] identified that the expression of the RNA-binding ribosomal protein S3 (RPS3) was increased in HCC patients compared to control individuals and that overexpression of RPS3 was significantly linked with HCC progression and aggressive clinicopathological features via upregulation of its target oncogene silent information regulator 1 (SIRT1). Additionally, a causal relationship between the dysregulation of ribonucleoprotein complex biogenesis and elevated cancer risk has been established[18]. Zhou et al[19] demonstrated that heterogeneous nuclear ribonucleoprotein AB (HNRNPAB) is overexpressed in HCC tissues and promotes the epithelial-mesenchymal transition (EMT) and metastasis of HCC cells, which predicts poor clinical outcomes. The KEGG pathway analysis showed that these RBPs were enriched in pathways related to the spliceosome, RNA transport, DNA replication, the cell cycle, ribosome biogenesis in eukaryotes, mismatch repair, base excision repair, RNA degradation, and mRNA surveillance.

Then, we performed univariate Cox regression and LASSO regression analyses, which identified 16 RBPs that were used to construct the predictive risk score model. The results of the ROC curve analysis revealed that the prediction model had a relatively good predictive ability. Among the 16 RBPs, *MEX3A* is highly expressed in HCC tissues, and its expression is positively correlated with a poor histological grade and a poor prognosis[20]. *TTK*, a mitotic spindle checkpoint gene, is frequently upregulated in HCC tissues compared with adjacent nontumor tissues and promotes HCC cell proliferation and resistance to sorafenib[21]. *IQGAP3* is markedly upregulated in HCC tissues and enhances HCC cell migration, invasion and EMT by activating the transforming growth factor-β (TGF-β) signaling pathway in HCC. Additionally, elevated expression of *IQGAP3* is significantly associated with aggressive clinicopathological features and a poor prognosis[22]. Hu et al[23] identified *NOL10* as a potential prognostic marker for patients with HCC by bioinformatics analysis and confirmed that the high expression of *NOL10* contributed to HCC cell proliferation and metastasis *in vitro* and *in vivo*. Tan et al[24] demonstrated that *PSMD1* promotes the proliferation of HCC cells by facilitating the accumulation of cellular lipid droplets, providing energy and membrane components for tumor cells. Additionally, it was positively associated with a poor prognosis in HCC patients. *YBX1*, which regulates protein expression at the transcriptional and translational levels, is highly expressed in HCC and induces drug resistance and immune escape[25]. Current reports on *DYNC1LI1*, *AHSA1* and *SEC61A1* in relation to HCC are primarily derived from bioinformatics mining

of public databases[26-28]. To our knowledge, three studies[29-31] reported that the expression levels of the three genes are significantly increased in HCC tissues compared to control tissues and that they are potential prognostic markers for patients with HCC. Although the remaining genes, including *MRPL53*, *PFN2*, *IMPDH1*, *TCOF1*, *EIF2B4*, *GNL2*, and *EIF1B*, have been scarcely studied in HCC, some of them have been reported to be involved in other cancers. *PFN2* is a negative regulator of colorectal cancer metastasis, and overexpression of *PFN2* may inhibit the EMT process[29]. Ruan et al[30] found that the expression level of *IMPDH1* is increased in clear cell renal cell carcinoma (ccRCC) tissues compared to normal control tissues and that *IMPDH1* positively regulates metastasis and the EMT signaling pathway. *GNL2*, which regulates nucleotide binding/metabolism, plays a role in resistance to 5-fluorouracil in colon cancer[31].

Furthermore, we investigated the differences in immune cell infiltration, somatic mutations, and gene set enrichment between the high- and low-risk groups of HCC patients from the TCGA database. The CIBERSORT results revealed that the infiltration levels of activated memory $CD4^+$ T cells, M0 macrophages and resting dendritic cells were high in the high-risk group, while those of resting $CD4^+$ T cells, activated NK cells and resting mast cells were low in the high-risk group. A recent study[30] demonstrated that tumor-infiltrating lymphocytes, including activated NK cells, resting memory CD4 T cells, eosinophils, and activated mast cells, were significantly correlated with HCC survival, suggesting that the differential infiltration of immune cells can predict the prognosis of HCC. Our results are also similar to the findings of a previous study[32], which revealed that activated NK cells exerted remarkably high cytotoxicity against HCC cells. The results of the somatic mutation analysis showed that the proportion of *TP53* mutations was significantly higher in the high-risk group than in the low-risk group. It has been reported that HCC patients with *TP53* mutations and upregulated *TP53* expression in tumor tissue have shorter OS and recurrence-free survival (RFS) times than patients with wild-type *TP53* and low/undetectable *TP53* expression levels[33]. Moreover, Long et al[34] analyzed the relationship between *TP53* mutations and the immune response in HCC and found that the immune response of HCC patients without *TP53* mutations was markedly stronger than that of HCC patients with *TP53* mutations. Additionally, GSEA showed differences in 10 signaling pathways, including the chemical carcinogenesis pathway, between the high- and low-risk groups.

Finally, using univariate and multivariate Cox regression analyses, we identified that TNM stage and the risk score were significant independent factors for predicting the OS of patients with HCC. Moreover, we constructed a nomogram plot to calculate the estimated OS probabilities of patients with HCC. The corresponding calibration curves showed that the nomogram had good efficacy in predicting the survival of patients with HCC. Although TNM stage was also a significant independent factor for the OS of patients with HCC in our study, the survival outcomes of patients within the same stage often differ in the clinic. This indicates that the current staging system is insufficient for effective prediction. Therefore, it is necessary to find more accurate biomarkers that can be used as prognostic and therapeutic indicators.

Overall, our study provides a new understanding of how RBPs affect the tumorigenesis and progression of HCC and indicates that RBPs may be used in clinical treatment decision making for HCC patients. Moreover, our prediction model based on 16 RBPs shows good performance for survival prediction in patients with HCC and may present new prognostic factors for HCC. Nonetheless, this study has several limitations. It is designed on the basis of a retrospective analysis, and a prospective study should be performed to verify these results. In addition, the datasets did not provide much clinical information, which may lead to an increased uncertainty of prediction.

Conclusions

In conclusion, we successfully constructed a prediction model based on 16 RBPs identified by bioinformatics methods that can serve as an independent prognostic factor for the OS of patients with HCC. Moreover, immune cell infiltration, somatic mutations and multiple signaling pathways may be involved in the differential OS outcomes of the high- and low-risk groups. This study provides an innovative analysis of prognostic biomarkers and therapeutic targets for HCC.

Conflict of Interest
The Authors declare that they have no conflict of interests.

Funding
Beijing Hospitals Authority Youth Program (QMS20200803 to C.Z). National Natural Science Foundation of China (No. 81800483 to C.Z.).

Availability of Data and Materials
The datasets and codes used or analyzed for the current study can be accessible from the corresponding author upon reasonable request if needed. Additionally, the datasets used during the present study are available in the following public repository: UCSC Xena browser: http://xena.ucsc.edu/Gene Expression Omnibus (GEO): www.ncbi.nlm.nih.gov/geo CIBERSORT: https://cibersort.stanford.edu/Molecular Signatures Database v7.5.1 (MSigDB): http://www.broadinstitute.org/gsea.

Authors' Contribution
(I) conception and design: Jukun Wang, Chunjing Bian; (II) administrative support: Tao Luo, Linzhong Zhu; (III) provision of study materials or patients: Chao Zhang, Yu Li, Wenhao Cui; (IV) collection and assembly of data: Jukun Wang, Ke Han; (V) data analysis and interpretation: Jukun Wang, Ke Han; (VI) manuscript writing: all authors; and (VII) final approval of manuscript: all authors.

References

1) Bray F, Ferlay J, Soerjomataram I, Siegel RL, Torre LA, Jemal A. Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA Cancer J Clin 2018; 68: 394-424.
2) Yang JD, Roberts LR. Hepatocellular carcinoma: A global view. Nat Rev Gastroenterol Hepatol 2010; 7: 448-458.
3) Mittal S, El-Serag HB. Epidemiology of hepatocellular carcinoma: consider the population. J Clin Gastroenterol 2013; 47 Suppl: S2-6.
4) Sangiovanni A, Colombo M. Treatment of hepatocellular carcinoma: beyond international guidelines. Liver Int 2016; 36 Suppl 1: 124-129.
5) Bruix J, Reig M, Sherman M. Evidence-Based Diagnosis, Staging, and Treatment of Patients With Hepatocellular Carcinoma. Gastroenterology 2016; 150: 835-853.
6) van Kouwenhove M, Kedde M, Agami R. MicroRNA regulation by RNA-binding proteins and its implications for cancer. Nat Rev Cancer 2011; 11: 644-656.
7) Glisovic T, Bachorik JL, Yong J, Dreyfuss G. RNA-binding proteins and post-transcriptional gene regulation. FEBS Lett 2008; 582: 1977-1986.
8) Masuda K, Kuwano Y. Diverse roles of RNA-binding proteins in cancer traits and their implications in gastrointestinal cancers. Wiley Interdiscip Rev RNA 2019; 10: e1520.
9) Han L, Huang C, Zhang S. The RNA-binding protein SORBS2 suppresses hepatocellular carcinoma tumourigenesis and metastasis by stabilizing RORA mRNA. Liver Int 2019; 39: 2190-2203.
10) Zhao Z, Li J, Shen F. Protective effect of the RNA-binding protein RBM10 in hepatocellular carcinoma. Eur Rev Med Pharmacol Sci 2020; 24: 6005-6013.
11) Climente-Gonzalez H, Porta-Pardo E, Godzik A, Eyras E. The Functional Impact of Alternative Splicing in Cancer. Cell Rep 2017; 20: 2215-2226.
12) Lee SC, Abdel-Wahab O. Therapeutic targeting of splicing in cancer. Nat Med 2016; 22: 976-986.
13) Singh B, Eyras E. The role of alternative splicing in cancer. Transcription 2017; 8: 91-98.
14) Dvinge H, Kim E, Abdel-Wahab O, Bradley RK. RNA splicing factors as oncoproteins and tumour suppressors. Nat Rev Cancer 2016; 16: 413-430.
15) Gerstberger S, Hafner M, Tuschl T. A census of human RNA-binding proteins. Nat Rev Genet 2014; 15: 829-845.
16) Dong W, Dai ZH, Liu FC, Guo XG, Ge CM, Ding J, Liu H, Yang F. The RNA-binding protein RBM3 promotes cell proliferation in hepatocellular carcinoma by regulating circular RNA SCD-circRNA 2 production. EBioMedicine 2019; 45: 155-167.
17) Zhao L, Cao J, Hu K, Wang P, Li G, He X, Tong T, Han L. RNA-binding protein RPS3 contributes to hepatocarcinogenesis by post-transcriptionally up-regulating SIRT1. Nucleic Acids Res 2019; 47: 2011-2028.
18) Pelletier J, Thomas G, Volarevic S. Ribosome biogenesis in cancer: new players and therapeutic avenues. Nat Rev Cancer 2018; 18: 51-63.
19) Zhou ZJ, Dai Z, Zhou SL, Hu ZQ, Chen Q, Zhao YM, Shi YH, Gao Q, Wu WZ, Qiu SJ, Zhou J, Fan J. HNRNPAB induces epithelial-mesenchymal transition and promotes metastasis of hepatocellular carcinoma by transcriptionally activating SNAIL. Cancer Res 2014; 74: 2750-2762.
20) Yang D, Jiao Y, Li Y, Fang X. Clinical characteristics and prognostic value of MEX3A mRNA in liver cancer. PeerJ 2020; 8: e8252.
21) Liang XD, Dai YC, Li ZY, Gan MF, Zhang SR, Yin P, Lu HS, Cao XQ, Zheng BJ, Bao LF, Wang DD, Zhang LM, Ma SL. Expression and function analysis of mitotic checkpoint genes identifies TTK as a potential therapeutic target for human hepatocellular carcinoma. PLoS One 2014; 9: e97739.
22) Shi Y, Qin N, Zhou Q, Chen Y, Huang S, Chen B, Shen G, Jia H. Role of IQGAP3 in metastasis and epithelial-mesenchymal transition in human hepatocellular carcinoma. J Transl Med 2017; 15: 176.

23) Hu X, Bao M, Huang J, Zhou L, Zheng S. Identification and Validation of Novel Biomarkers for Diagnosis and Prognosis of Hepatocellular Carcinoma. Front Oncol 2020; 10: 541479.

24) Tan Y, Jin Y, Wu X, Ren Z. PSMD1 and PSMD2 regulate HepG2 cell proliferation and apoptosis via modulating cellular lipid droplet metabolism. BMC Mol Biol 2019; 20: 24.

25) Tao Z, Ruan H, Sun L, Kuang D, Song Y, Wang Q, Wang T, Hao Y, Chen K. Targeting the YB-1/PD-L1 Axis to Enhance Chemotherapy and Antitumor Immunity. Cancer Immunol Res 2019; 7: 1135-1147.

26) Wang X, Qiao J, Wang R. Exploration and validation of a novel prognostic signature based on comprehensive bioinformatics analysis in hepatocellular carcinoma. Biosci Rep 2020; 40: BSR20203263.

27) Li W, Lu J, Ma Z, Zhao J, Liu J. An Integrated Model Based on a Six-Gene Signature Predicts Overall Survival in Patients With Hepatocellular Carcinoma. Front Genet 2019; 10: 1323.

28) Li N, Zhao L, Guo C, Liu C, Liu Y. Identification of a novel DNA repair-related prognostic signature predicting survival of patients with hepatocellular carcinoma. Cancer Manag Res 2019; 11: 7473-7484.

29) Zhang H, Yang W, Yan J, Zhou K, Wan B, Shi P, Chen Y, He S, Li D. Loss of profilin 2 contributes to enhanced epithelial-mesenchymal transition and metastasis of colorectal cancer. Int J Oncol 2018; 53: 1118-1128.

30) Ruan H, Song Z, Cao Q, Ni D, Xu T, Wang K, Bao L, Tong J, Xiao H, Xiao W, Cheng G, Xiong Z, Liang H, Liu D, Wang L, Olivier T, Jane BH, Yang H, Zhang X, Chen K. IMPDH1/YB-1 Positive Feedback Loop Assembles Cytoophidia and Represents a Therapeutic Target in Metastatic Tumors. Mol Ther 2020; 28: 1299-1313.

31) De Angelis PM, Svendsrud DH, Kravik KL, Stokke T. Cellular response to 5-fluorouracil (5-FU) in 5-FU-resistant colon cancer cell lines during treatment and recovery. Mol Cancer 2006; 5: 20.

32) Kamiya T, Chang YH, Campana D. Expanded and Activated Natural Killer Cells for Immunotherapy of Hepatocellular Carcinoma. Cancer Immunol Res 2016; 4: 574-581.

33) Liu J, Ma Q, Zhang M, Wang X, Zhang D, Li W, Wang F, Wu E. Alterations of TP53 are associated with a poor outcome for patients with hepatocellular carcinoma: evidence from a systematic review and meta-analysis. Eur J Cancer 2012; 48: 2328-2338.

34) Long J, Wang A, Bai Y, Lin J, Yang X, Wang D, Yang X, Jiang Y, Zhao H. Development and validation of a TP53-associated immune prognostic model for hepatocellular carcinoma. EBioMedicine 2019; 42: 363-374.