Original Article

Development and validation of an eight-gene signature based predictive model to evaluate the prognosis of hepatocellular carcinoma patients: a bioinformatic study

Jiehao Zhang[1#], Xin Fu[2#], Nannan Zhang[1#], Weizhen Wang[1,3], Hui Liu[3], Yibin Jia[4], Yongzhan Nie^

[1]State Key Laboratory of Cancer Biology and National Clinical Research Center for Digestive Diseases, Xijing Hospital of Digestive Diseases, Fourth Military Medical University, Xi'an, China; [2]National Center for International Research of Bio-targeting Theranostics, Guangxi Key Laboratory of Bio-targeting Theranostics, Collaborative Innovation Center for Targeting Tumor Diagnosis and Therapy, Guangxi Talent Highland of Bio-targeting Theranostics, Guangxi Medical University, Nanning, China; [3]College of Life Sciences, Northwest University, Xi'an, China; [4]Department of Neurosurgery, Xijing Hospital, Fourth Military Medical University, Xi'an, China

Contributions: (I) Conception and design: Y Nie, J Zhang, N Zhang; (II) Administrative support: None; (III) Provision of study materials or patients: None; (IV) Collection and assembly of data: Y Nie, X Fu, W Wang, H Liu, Y Jia; (V) Data analysis and interpretation: Y Nie, J Zhang, X Fu, N Zhang; (VI) Manuscript writing: All authors; (VII) Final approval of manuscript: All authors.

[#]These authors contributed equally to this work.

Correspondence to: Yongzhan Nie. State Key Laboratory of Cancer Biology and National Clinical Research Center for Digestive Diseases, Xijing Hospital of Digestive Diseases, Fourth Military Medical University, Xi'an 710032, China. Email: yongznie@fmmu.edu.cn.

Background: Hepatocellular carcinoma (HCC) is a malignant tumor with a poor prognosis, however, biomarkers for the prognostic assessment of HCC remain suboptimal. Consequently, we aimed to develop a reliable tool for prognostic estimation of HCC.

Methods: Differentially expressed genes (DEGs) between HCC and adjacent normal tissues in 3 Gene Expression Omnibus (GEO) datasets were identified, followed by hub gene selection and least absolute shrinkage and selection operator (LASSO) Cox regression to develop a prognostic gene signature. Kaplan-Meier survival analysis, univariate and multivariate Cox regression, time-dependent area under the curve (AUC), and integrated value of time-dependent AUC (iAUC) were used to assess the relationship between predictors and clinical outcomes in the training and validation datasets. Then we built nomograms including gene signature and clinicopathological factors to forecast the probability of death. Moreover, we performed quantitative real-time PCR (qPCR) to compare the expression of prognostic genes between HCC and adjacent normal tissues. Finally, the relationship between prognostic genes and tumor microenvironment (TME) was investigated using immune cell infiltration algorithms and single cell transcriptomic database.

Results: Eight prognostic genes (*CDC20*, *PTTG1*, *TOP2A*, *CXCL2*, *CXCL14*, *CYP2C9*, *MT1F*, and *GHR*) were finally identified to construct the gene signature. Each patient's risk score was calculated according to the gene signature. Patients with high-risk scores showed worse outcomes in the training set [hazard ratio (HR) =3.404, P<0.001]. Risk score, age, body mass index (BMI), and TNM stage were identified as independent prognostic factors for overall survival (OS) in the training set. The nomogram including risk score and other independent prognostic factors showed better performance as opposed to the clinicopathological model. In the validation dataset, we obtained the similar results as well. Moreover, we found a close relationship between risk score and immune cell infiltration. Patients with high-risk scores had elevated expression of immune checkpoint genes, indicating that these patients may be more suitable for immunotherapy.

Conclusions: We have established and validated an eight-gene based prognostic model, which could be an effective tool for the prognostic evaluation of HCC patients.

^ ORCID: 0000-0001-5201-7224.

Keywords: Hepatocellular carcinoma (HCC); prognosis; nomogram; tumor microenvironment (TME); bioinformatics

Submitted Mar 18, 2022. Accepted for publication May 09, 2022.
doi: 10.21037/atm-22-1934
View this article at: https://dx.doi.org/10.21037/atm-22-1934

Introduction

Hepatocellular carcinoma (HCC) is a commonly occurring malignant tumor with high incidence and mortality rates, ranking as the third cause of cancer-associated deaths, with approximately 906,000 new cases and 830,000 deaths in 2020 (1). While a small number of early-stage HCC can be treated with liver resection or liver transplantation, most HCC patients are at an advanced stage of disease at the time of diagnosis and are often untreatable even with careful monitoring. Moreover, the prognosis is still poor on account of metastasis probability and high recurrence. However, biomarkers for the prognostic assessment of HCC patients remain suboptimal (2). Therefore, it is critical to develop reliable prognostic tools to predict clinical outcomes of HCC and assist in the decision-making process.

A large number of studies have described clinicopathological factors including tumor size, tumor number, vascular invasion, alanine transaminase (ALT), aspartate transaminase (AST), and α-fetoprotein (AFP) to reveal the clinical outcomes of HCC patients (3,4). But these prognostic biomarkers for HCC remain suboptimal. For example, AFP has been used for many years as a serum marker for HCC diagnosis and screening. However, it has been recognized that AFP is less sensitive in detecting HCC, and AFP levels are often elevated in other chronic liver diseases such as chronic hepatitis and cirrhosis (5). Among these clinical variables, TNM stage is a universally acknowledged factor for predicting clinical outcomes. However, the performance of TNM stage is still far from satisfactory, which could be attributed to the reason that the prognosis of HCC is not only related to these clinicopathological indicators but also closely associated with the change in underlying molecular pathways (6). Consequently, considering the molecular biomarkers and alterations in molecular pathways may be a promising strategy. For example, Nault *et al.* developed a 5-gene score that was associated with disease-specific survival to predict outcomes of HCC patients treated by resection (7). Kim *et al.* identified a 65-gene based classifier to predict overall survival (OS) in HCC (8). Zhou *et al.* even constructed a plasma miRNA panel used for the diagnosis of hepatitis B virus-related HCC with a high accuracy (9).

With the development of transcriptomics, it is convenient for researchers to explore the mechanism of cancer progression using high throughput sequencing nowadays (10). In this study, we extracted HCC gene expression data from 4 datasets in the Gene Expression Omnibus (GEO) database and The Cancer Genome Atlas (TCGA) program. We assessed differentially expressed genes (DEGs) that were common in 3 GEO datasets and constructed a gene signature for prognostic estimation of HCC. Kaplan-Meier survival analysis, univariate and multivariate Cox regression, time-dependent area under the curve (AUC), and integrated value of time-dependent AUC (iAUC) were used to assess the association between gene signature and clinical outcomes in the training and validation datasets. Furthermore, we built nomograms including gene signature to forecast the probability of death. We present the following article in accordance with the TRIPOD reporting checklist (available at https://atm.amegroups.com/article/view/10.21037/atm-22-1934/rc).

Methods

Data collection and study design

The transcriptome profiles including GSE87630, GSE89377, and GSE121248 were obtained from the GEO database. GSE87630 contained 64 HCC samples as well as 30 adjacent non-tumor samples, GSE89377 contained 40 HCC samples and 13 non-tumor samples, and GSE121248 contained 70 HCC samples and 37 non-tumor samples. The gene expression matrix and clinicopathological data of 377 HCC tissues were extracted from the TCGA-liver hepatocellular carcinoma (TCGA-LIHC) database, which was used as a training set (11). The validation dataset GSE14520 was also downloaded from the GEO database. Detailed information is presented in Table S1. This study was a bioinformatics analysis. We identified common DEGs

between HCC and adjacent normal tissues in 3 GEO datasets and constructed a gene signature by LASSO. Then we explored the association between gene signature and clinical outcomes in the training and validation datasets. Nomograms including gene signature were also developed to forecast the probability of death. We also investigated the relationship between gene signature and tumor microenvironment (TME).

Screening out DEGs

We used GEO2R to find DEGs (adjusted P value <0.05 and $|\log_2$ fold-change$|$ >1.5) between HCC tissues and the adjacent normal tissues (12). Next, a Venn diagram was utilized to detect overlapping DEGs among the 3 datasets.

Identification of hub genes

PPI information for the abovementioned 69 DEGs was accessed via the Search Tool for the Retrieval of Interacting Genes (STRING) database (http://string-db.org) (13). Cytoscape software (version 3.6.0) was utilized to create a PPI network (12). The top 20 genes that had the highest degree scores as classified by CytoHubba were deemed as hub genes.

Construction of the gene signature

Based on the identified hub genes, we carried out LASSO Cox regression analysis in the TCGA-LIHC cohort (14). The optimal penalty parameter lambda and coefficients of prognostic genes were determined by the minimum set criteria via 10-fold cross-validation. Each patient's risk score was computed as follows: risk score = (coefficient1 × gene 1 expression) + (coefficient2 × gene 2 expressions) + … + (coefficient n × n gene expression). Thus, patients were categorized into 2 groups according to the median value of the risk score.

Evaluation of prognostic value

Kaplan-Meier survival analysis was used to compare the survival status between high-risk group and low-risk group. Sex, age, risk score, grade, BMI, pT, pN, pM, and TNM stage were subjected to univariate Cox regression analysis in TCGA-LIHC dataset. Characteristics which were significantly associated with OS in univariate Cox regression analysis were subjected to multivariate Cox regression analysis to determine independent prognostic factors. Time-dependent AUC compared the prognostic value between these prognostic factors. The model's predictive accuracy was also assessed via iAUC. Unbiased estimation was performed using 1,000× bootstrap resampling validation (15).

Nomogram

A nomogram integrating independent prognostic factors and risk score was determined to forecast the survival probabilities of patients with HCC in the training set. We named this model as the risk score model. The nomogram's performance was evaluated using calibration curves.

Comparison of predictive robustness between the risk score model and clinicopathological model

We also constructed a nomogram based on identified independent prognostic variables without the risk score, which was called the clinicopathological model. Calibration curves and ROC curves were employed for the comparison of the predictive performance of the 2 models. Furthermore, decision curve analysis (DCA) was used to quantify the clinical utility (16).

Validation of the gene signature in the GSE14520 cohort

The GSE14520 dataset includes microarray data obtained from 247 HCC patients. After 22 patients without detailed clinical information were removed, we used 225 HCC patients as a validation dataset to verify the gene signature's robustness for prognostic prediction. The same methods were utilized to evaluate the gene signature's predictive performance for OS and disease-free survival (DFS) in the GSE14520 validation dataset.

Gene set enrichment analysis (GSEA)

GSEA was carried out in the TCGA-LIHC training set to illustrate the molecular pathways (17). Gene sets which satisfied the following requirements were considered significant: normalized (NOM) P value <0.05 and false discovery rate (FDR) <0.25.

Immune infiltration analysis and immune checkpoint correlation analysis

We downloaded infiltration estimation results of all

TCGA tumors from TIMER 2.0 (18), which used the "immunedeconv" R package to evaluate immune infiltration status (19). Then, we selected the results of TCGA-LIHC samples. Subsequently, we performed correlation analysis between TIMER immune infiltration scores and the risk scores of each patient in TCGA-LIHC using Spearman correlation analysis. Furthermore, we compared the differences of CIBERSORT immune infiltration scores and immune checkpoint gene expression levels between 2 groups using the Wilcoxon rank sum test.

Tumor Immune Single Cell Hub (TISCH) Database

TISCH (http://tisch.comp-genomics.org/home/) is an online database of 76 tumor single-cell RNA sequencing datasets based on 27 types of tumors (20). In this study, we used TISCH to illustrate the relationship between prognostic genes and the TME of liver cancer.

Human liver tissue sample collection

A total of 15 paired HCC primary tissues and adjacent normal tissues were collected from Xijing Hospital, which was approved by the drug clinical trial Ethics Committee, Fourth Military Medical University (No. KY20193057). All the patients who donated tissue samples provided written informed consent. The study was conducted in accordance with the Declaration of Helsinki (as revised in 2013).

Extraction of RNA and quantitative real-time PCR (qPCR)

The mRNA isolation kit (cat# 74104, Qiagen, Germany) was employed to isolate total RNA. The final samples of RNA were kept at −80 ℃. The cDNA synthesis kit (Code No. 6215A, Takara, Japan) was utilized for cDNA synthesis based on 1,000 ng of total RNA, which was further analyzed by qPCR. Gene-specific primers were used to determine the relative mRNA amount. β-actin was the internal control. Table S2 displays the sequences of primers.

Statistical analysis

All statistical analyses were performed using SPSS software (version 24.0) and R software (version 3.6.1). P values <0.05 (two-sided) were considered significant.

Results

Identification of 69 genes shared by 3 GEO datasets

DEGs between tumor and non-tumor tissues were screened out in the GSE87630, GSE89377, and GSE121248 datasets. We found 403 DEGs in GSE87630, 151 DEGs in GSE89377, and 432 DEGs in GSE121248 (*Figure 1A-1C*). Among them, 82, 30, and 115 genes were upregulated in the GSE87630, GSE89377, and GSE121248 datasets, respectively. Additionally, 321, 121, and 317 genes were downregulated in the GSE87630, GSE89377, and GSE121248 datasets, respectively. Finally, 69 overlapping DEGs (10 upregulated, 59 downregulated) were used for the subsequent analyses (*Figure 1D*).

Identification of hub genes and construction of the gene signature

STRING and Cytoscape software were employed to create a PPI network for the 69 DEGs (*Figure 1E*). We used CytoHubba to select hub genes, and the top 20 genes with the highest degree scores are shown in Table S3. Subsequently, we carried out LASSO regression analysis and identified 8 genes that were significantly related to OS (*CDC20*, *PTTG1*, *TOP2A*, *CXCL2*, *CXCL14*, *MT1F*, *GHR*, *CYP2C9*) (*Figure 1F,1G*). Therefore, we constructed a gene signature based on prognosis-related genes and LASSO coefficients. Each patient's risk score was computed as follows: risk score = (0.0137 × expression of *CDC20*) + (0.0065 × expression of *PTTG1*) + (0.0071 × expression of *TOP2A*) + (−0.0005 × expression of *CXCL2*) + (−0.0047 × expression of *CXCL14*) + (−0.0013 × expression of *GHR*) + (−0.0014 × expression of *CYP2C9*) + (−0.0007 × expression of *MT1F*). The coefficients suggested that *CDC20*, *PTTG1*, and *TOP2A* were risk genes while *CXCL2*, *CXCL14*, *GHR*, *CYP2C9*, and *CYP2C9* played a protective role. Although only 8 genes were included in the model (*CDC20*, *PTTG1*, *TOP2A*, *CXCL2*, *CXCL14*, *GHR*, *MT1F*, *CYP2C9*), it still performed optimally.

Patients were categorized into high-risk and low-risk groups according to the median value of the risk score. Patients in the high-risk group showed worse outcomes. The heatmap illustrated that high-risk patients had higher expression of *CDC20*, *PTTG1*, and *TOP2A* and lower expression of *CXCL2*, *CXCL14*, *MT1F*, *GHR*, and *CYP2C9*

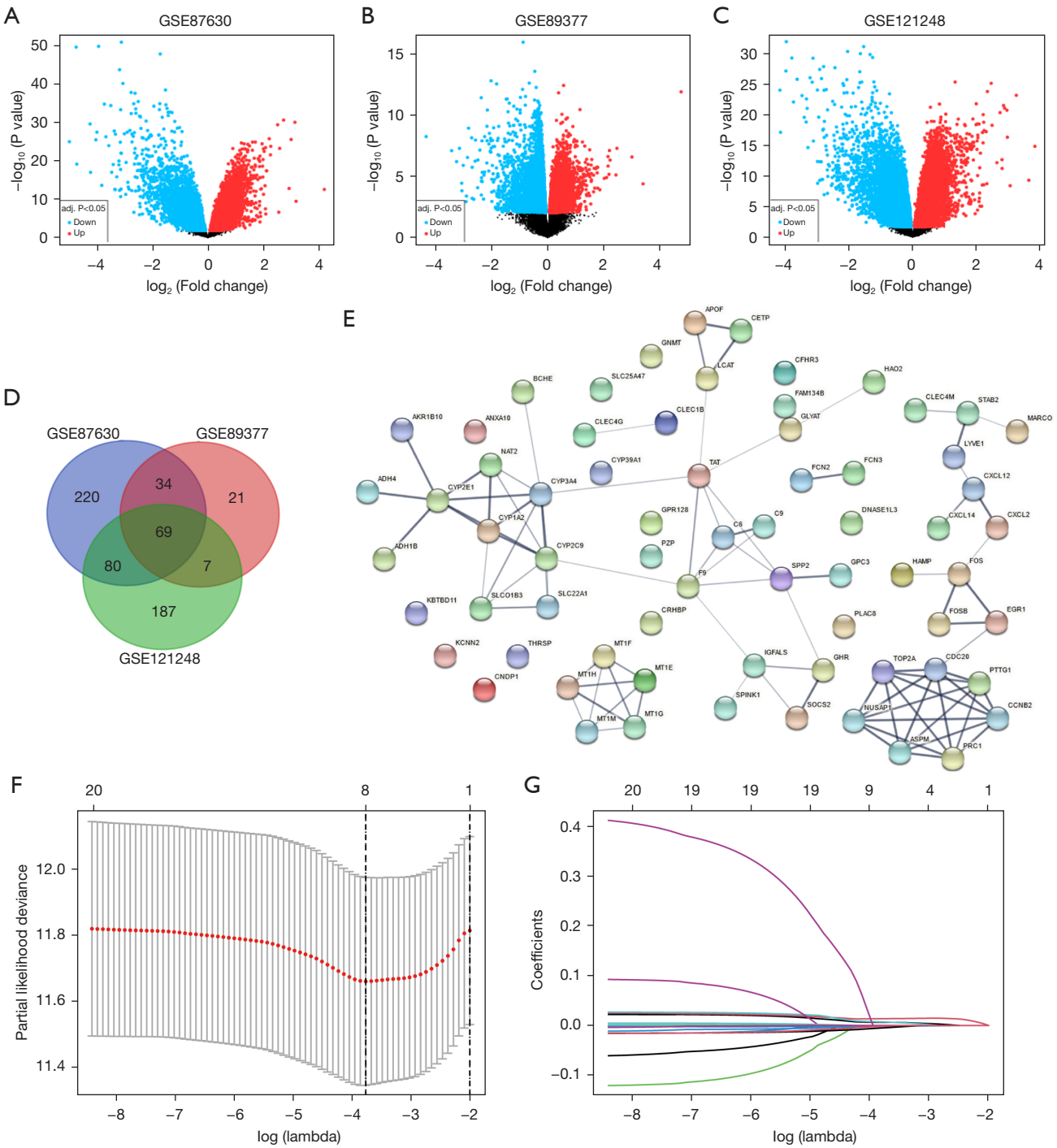

Figure 1 Identification of DEGs and development of gene signature. (A-C) Volcano plots of DEGs from GSE87630, GSE89377, and GSE121248. (D) Venn diagram of the overlapping DEGs between the 3 GEO datasets. (E) PPI network of the 69 identified DEGs. (F) Optimized lambda determined by LASSO. (G) The LASSO coefficients of 20 hub genes. DEG, differentially expressed gene; GEO, Gene Expression Omnibus; PPI, protein-protein interaction; LASSO, least absolute shrinkage and selection operator.

(*Figure 2A*). Kaplan-Meier survival analysis also illustrated that the survival rate of patients with high risk scores was distinctively lower (*Figure 2B*). Moreover, we validated the results in the validation dataset GSE14520. The performance of the gene signature in predicting OS (*Figure 2C*,*2D*) and DFS (Figure S1) in GSE14520 was excellent as well.

Appraisal of prognostic value in the TCGA-LIHC training cohort

Sex, age, risk score, grade, BMI, pT, pN, pM, and TNM stage were subjected to univariate Cox regression analysis in TCGA-LIHC dataset. According to *Figure 3A*, age [HR =1.502 (95% CI: 1.038–2.173), P=0.031], risk score [HR =3.404 (95% CI: 2.335–4.962), P<0.001], body mass index (BMI) [HR =1.554 (0.811–1.824), P=0.001], T stage [HR =1.516, (95% CI: 1.260–1.893), P<0.001], M stage [HR =3.849 (95% CI: 1.218–2.164), P=0.022], and TNM stage [HR =1.574 (95% CI: 1.291–1.918), P<0.001] were significantly correlated with OS in univariate Cox regression analysis. These factors were then included in a multivariate Cox proportional hazard model. As shown in *Figure 3B*, multivariate Cox regression showed that age [HR =1.582 (95% CI: 1.09–2.296), P=0.016], risk score [HR =1.878 (95% CI: 1.273–2.768), P=0.001], BMI [HR =1.683 (95% CI: 1.169–2.422), P=0.005], and TNM stage [HR =1.397 (95% CI: 1.138–1.715), P=0.001] should be considered as independent OS risk factors. Compared with TNM stage, age, and BMI, the novel gene signature showed significantly improved AUC at all time points (*Figure 3C*). Among the various clinical parameters, the risk score had the highest mean iAUC (*Figure 3D*). Overall, these results showed that the risk score of our novel gene signature was a noteworthy independent prognostic variable and had the best prognostic value compared with other clinical characteristics.

Validation in the GSE14520 cohort

To examine whether the gene signature could be effective in other datasets, we further examined the GSE14520 validation group. Sex, age, risk score, TNM stage, AFP, tumor size, and ALT were subjected to univariate Cox regression analysis. Results (*Figure 4A*) showed that, risk score [HR =1.974 (95% CI: 1.277–3.051), P=0.002], TNM stage [HR =2.329 (95% CI: 1.760–3.083), P<0.001], AFP [HR =1.556 (95% CI: 1.015–2.384), P=0.042], tumor size [HR =1.842 (95% CI: 1.264–2.983), P=0.042] were significantly correlated with OS. These factors were then subjected to multivariate Cox regression analysis. Results (*Figure 4B*) indicated that risk score [HR =1.5941 (95% CI: 1.09–2.493), P=0.03], TNM stage [HR =2.145 (95% CI: 1.552–2.964), P<0.001], and tumor size [HR =1.001 (95% CI: 0.599–1.675), P=0.038] were independent OS risk factors. Time-dependent AUC and iAUC illustrated that, compared with TNM stage and tumor size, the gene signature showed a significantly improved AUC at all time points (*Figure 4C*), and the risk score had the highest mean iAUC (*Figure 4D*). In comparison with OS, the DFS may be more specific in reflecting clinical benefits. Thus, we also used DFS as a clinical endpoint in the validation set GSE14520, and similar results could also be found for DFS (Figure S2A-S2D).

Construction of the predictive nomogram models

For better prediction of prognosis, we constructed a nomogram to forecast the death probability of HCC patients in the training set. The independent prognostic variables including risk score were incorporated in the prediction model, which was called the risk score model (Figure S3A). We also constructed a nomogram without risk score in the training set, which was called the clinicopathological model (Figure S3B). In the validation dataset GSE14520, we constructed the risk score model (Figure S4A) and clinicopathological model (Figure S4B) as well. Calibration curves could depict consistency between prediction and real observations. Thus, we used calibration curves to evaluate the predictive performance of the 2 models. Results showed that, at the time point of 3 years, the calibration curves of both models exhibited great concordance with the actual survival rate in the training and validation sets (Figure S5A,S5B). Then, we compared the predictive value between the clinicopathological model and risk score model. Time-dependent ROC curves showed that the AUC of the risk score model was higher than that of the clinicopathological model in the 2 datasets (Figure S5C,S5D), indicating that the accuracy of the risk score model was better than that of the clinicopathological model. DCA curves suggested that the risk score model had a better net benefit than the clinicopathological model, indicating that the risk score model can help clinicians make more accurate assessments of liver cancer prognosis (Figure S5E,S5F).

Validation of the expression levels of the 8 genes

The mRNA levels of *CDC20*, *PTTG1*, *TOP2A*, *CXCL2*,

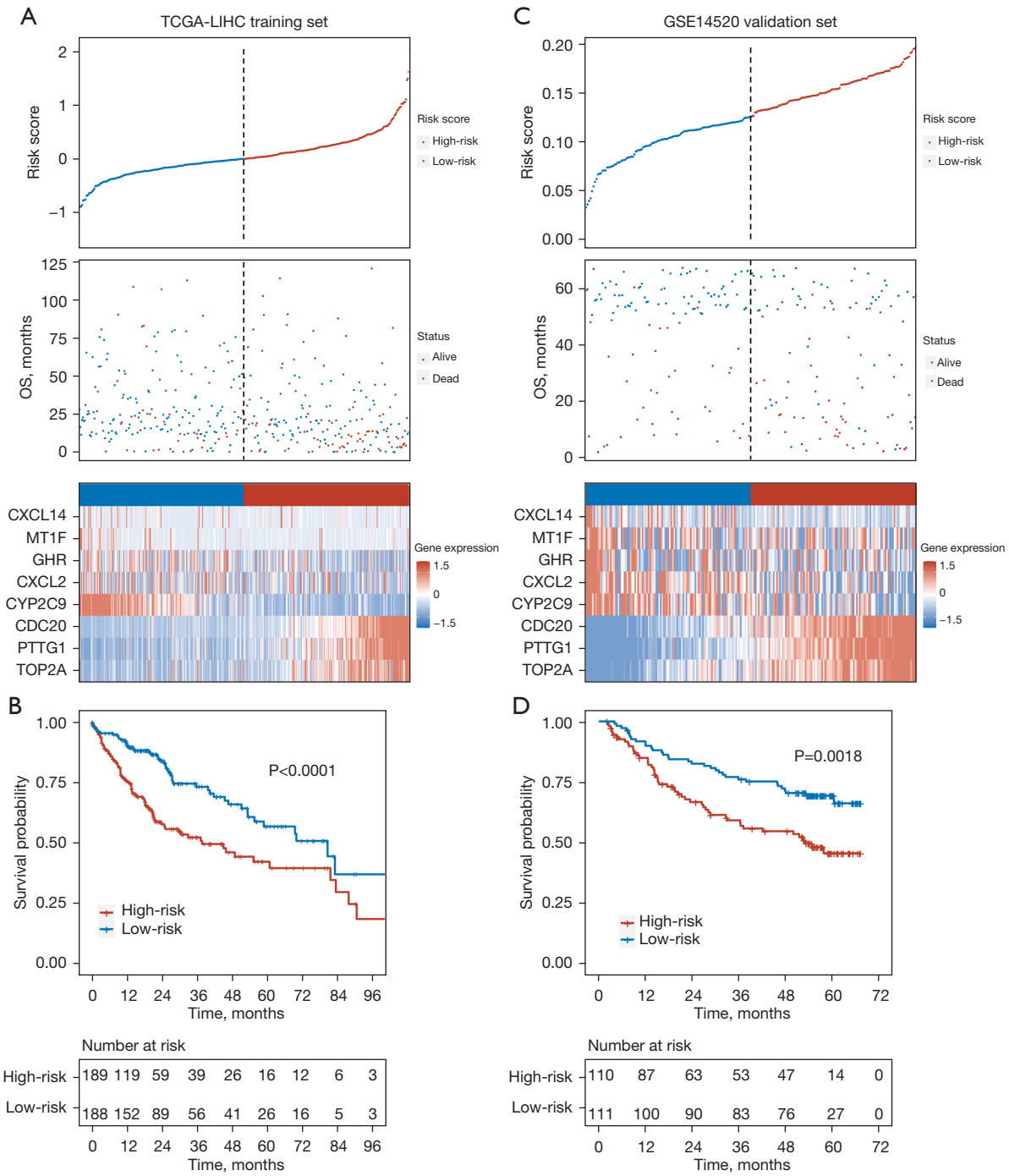

Figure 2 Evaluation of gene signature in TCGA-LIHC training set and GSE14520 validation set. (A) The distribution of risk score, survival status, and gene expression panels of HCC patients in the TCGA-LIHC training set. (B) Kaplan-Meier survival analysis between the 2 groups separated by the median value of the risk score in the TCGA-LIHC training set. The upper part illustrates the Kaplan-Meier curves for the 2 groups while the bottom illustrates the number of living patients. (C) The distribution of risk score, survival status, and gene expression panels of HCC patients in the GSE14520 validation set. (D) Kaplan-Meier survival analysis between high-risk group and low-risk group in the GSE14520 validation set. HCC, hepatocellular carcinoma; TCGA-LIHC, The Cancer Genome Atlas-liver hepatocellular carcinoma; OS, overall survival.

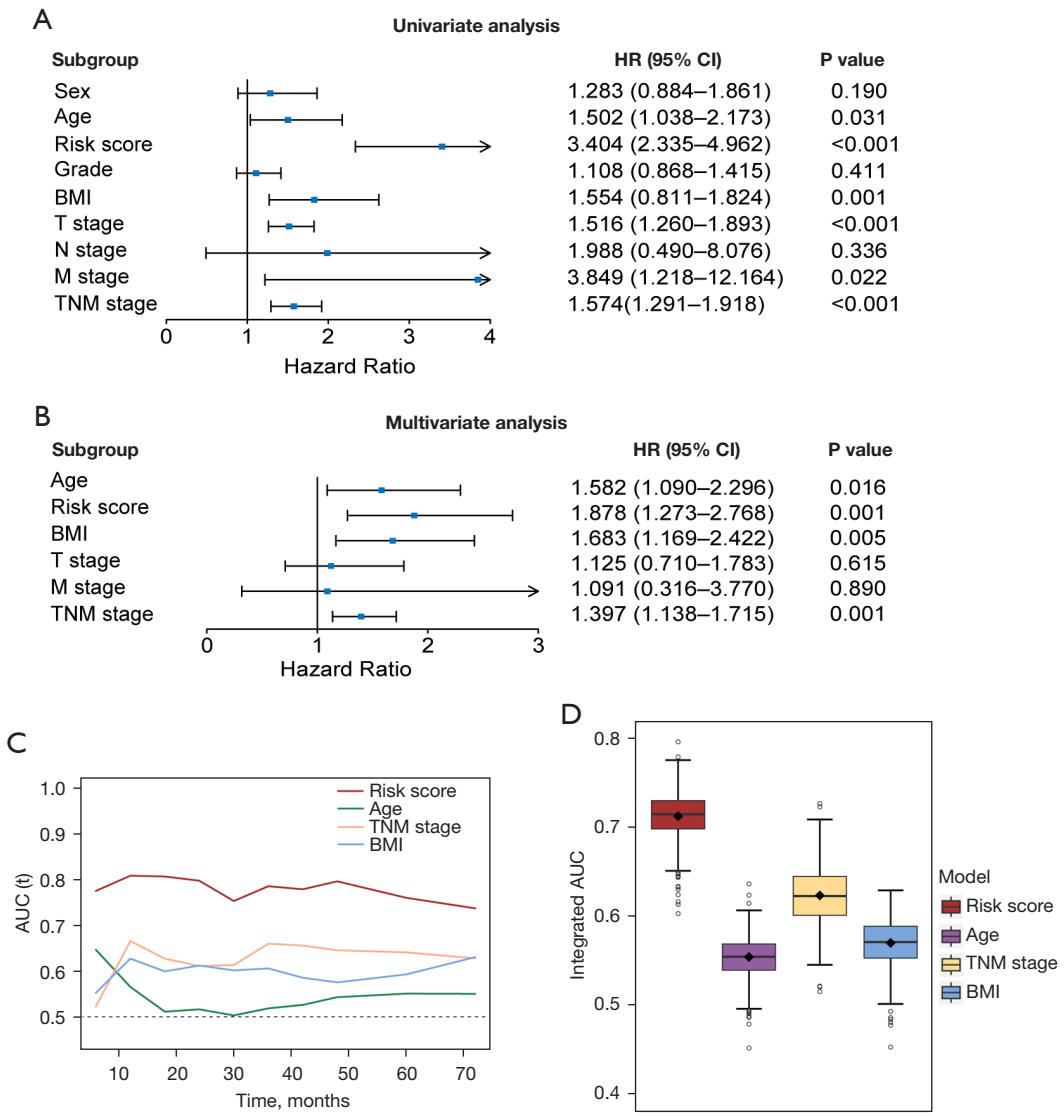

Figure 3 Identification of independent prognostic factors and comparison of predictive accuracy for OS in TCGA-LIHC cohort. (A,B) Univariate and multivariate Cox regression analysis of risk score and other clinicopathological characteristics in the TCGA-LIHC dataset. (C) Time-dependent AUC of risk score and other independent prognostic variables in the TCGA-LIHC dataset. (D) The predictive accuracy for OS based on the iAUC. The iAUC denotes the integrated area under the ROC curve. BMI, body mass index; TCGA-LIHC, The Cancer Genome Atlas-liver hepatocellular carcinoma; AUC, area under the curve; OS, overall survival; iAUC, integrated value of time-dependent AUC; ROC, receiver operating characteristic.

CXCL14, *GHR*, *CYP2C9*, and *MT1F* in 15 HCC tissues and adjacent normal tissues were evaluated using qPCR. As expected, the mRNA levels of *CDC20*, *PTTG1*, and *TOP2A* were significantly elevated in HCC tissues as opposed to the adjacent normal tissues (*Figure 5A-5C*). *CXCL2*, *CXCL14*, *GHR*, *CYP2C9*, and *MT1F* displayed higher expression in adjacent non-tumor tissues in comparison with HCC tissues (*Figure 5D-5H*). The same results were obtained from the GEPIA database. We also used GEPIA to perform survival analysis of the 8 genes. The outcomes showed that HCC patients who had higher expression of *CDC20*, *PTTG1*, and *TOP2A* experienced a worse outcome, while higher expression of *CXCL2*, *CXCL14*, *GHR*, *CYP2C9*, and *MT1F* represented better OS (Figure S6). Similar to the

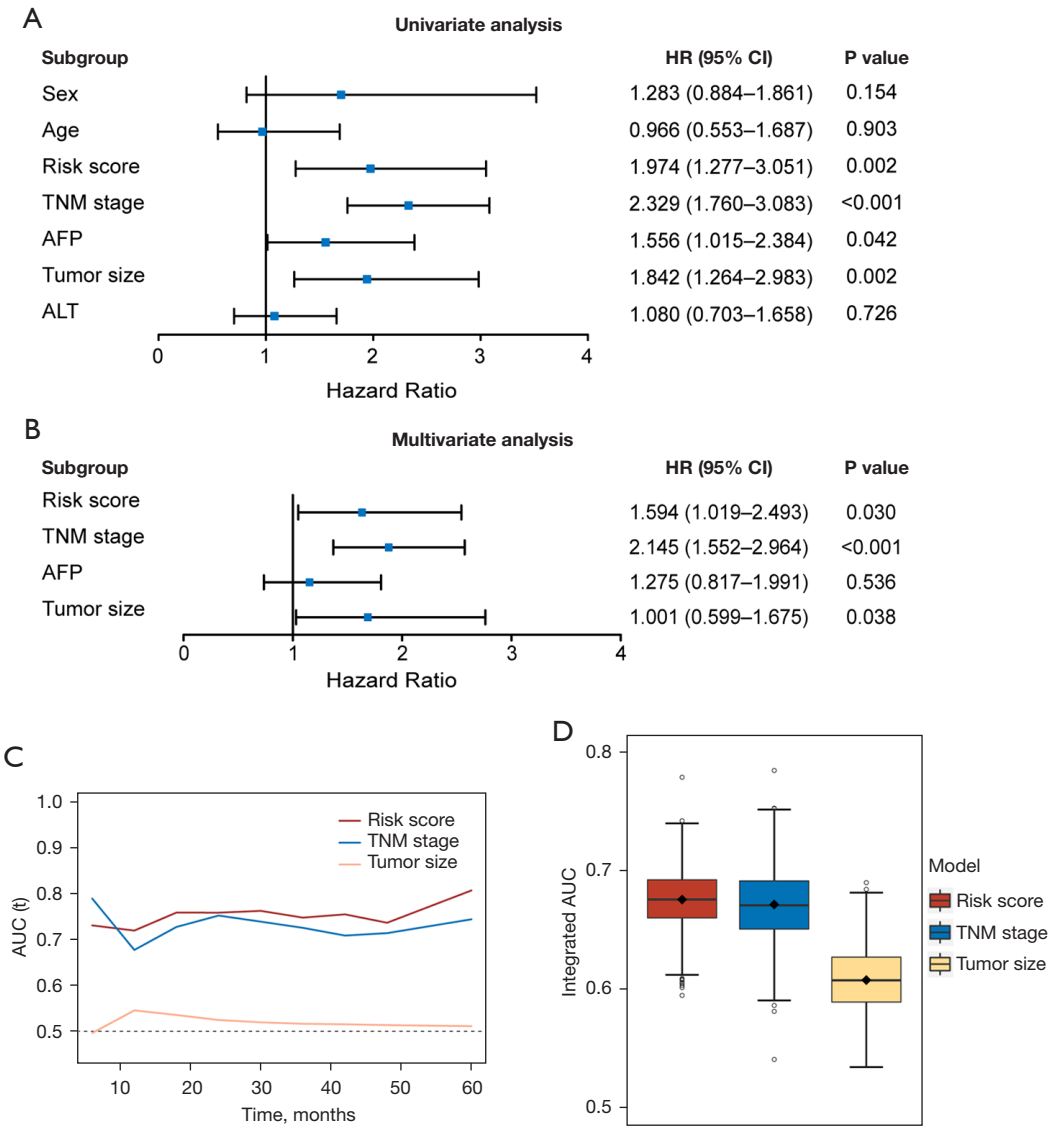

Figure 4 Identification of independent prognostic factors and comparison of predictive accuracy for OS in GSE14520 cohort. (A,B) Univariate and multivariate Cox regression analysis of risk score and other clinicopathological characteristics in the GSE14520 validation set. (C) Time-dependent AUC of risk score and independent prognostic variables in the GSE14520 validation set. (D) The predictive accuracy for OS according to the iAUC. AFP, α-fetoprotein; ALT, alanine transaminase; AUC, area under the curve; OS, overall survival; iAUC, integrated value of time-dependent AUC.

conclusions drawn from the LASSO coefficients, the results of GEPIA showed that *CDC20*, *PTTG1*, and *TOP2A* may be the risk genes for HCC, while *CXCL2*, *CXCL14*, *GHR*, *CYP2C9*, and *MT1F* may be protective.

Molecular pathways of prognostic genes

Results of GSEA illustrated that the high-risk group was linked to oncogenic pathways, such as DNA repair [normalized enrichment score (NES) =2.0, nominal P<0.001], MYC targets (NES =2.0, nominal P<0.001), G2M checkpoint (NES =2.1, nominal P<0.001), and E2F targets (NES =2.2, nominal P<0.001) (*Figure 5I*). In contrast, the low-risk group was associated with bile acid metabolism (NES =−2.1, nominal P=0.002), xenobiotic metabolism (NES =−2.1, nominal P<0.001), coagulation (NES =−1.8,

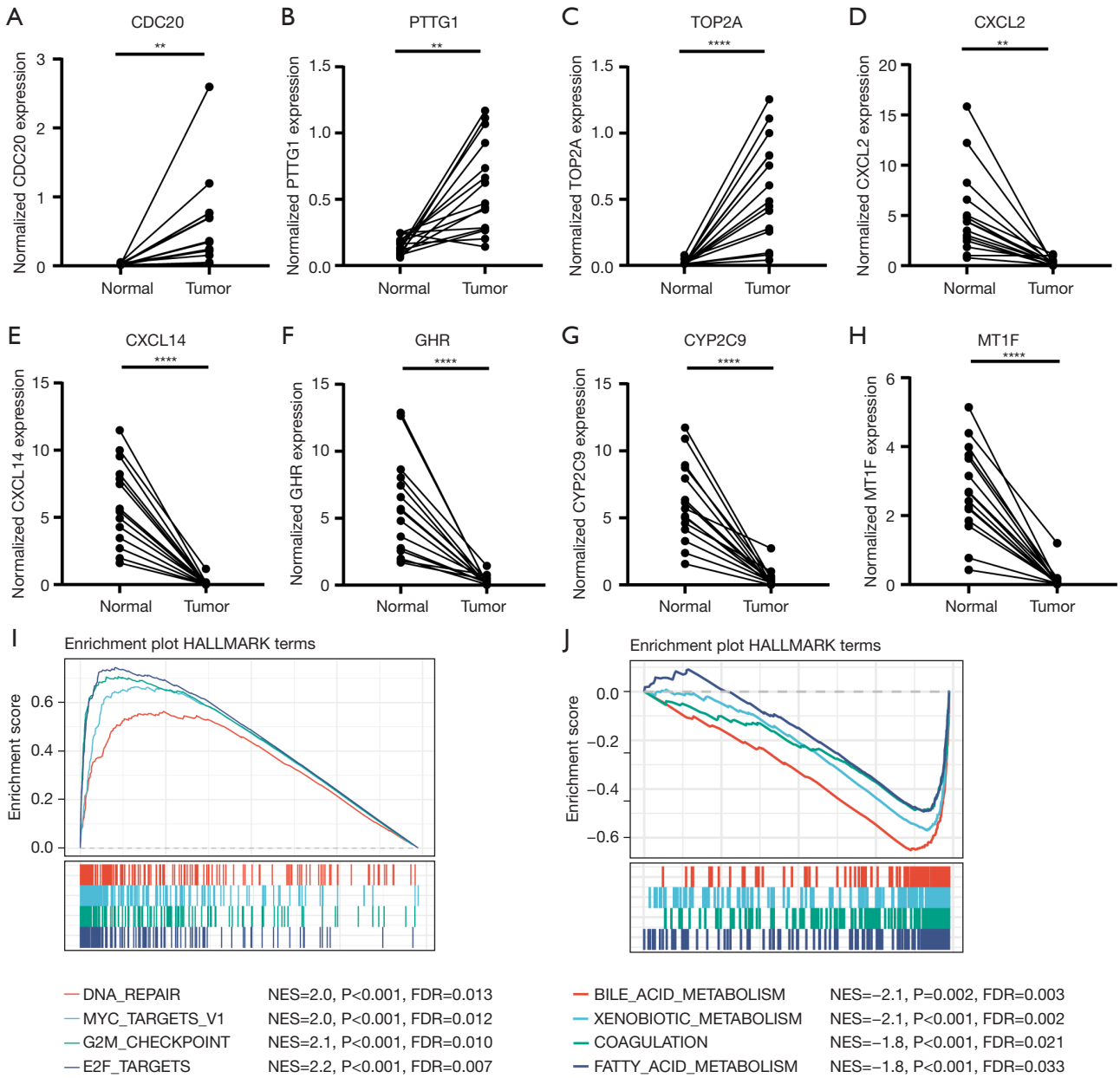

Figure 5 mRNA expression levels of 8 genes and significantly enriched pathways identified by GSEA. (A-C) qPCR shows increased expression of *CDC20*, *PTTG1*, and *TOP2A* in 15 HCC tissues comparative to adjacent non-tumor tissues. (D-H) qPCR shows decreased expression of *CXCL2*, *CXCL14*, *GHR*, *CYP2C9*, and *MT1F* in 15 HCC tissues comparative to adjacent non-tumor tissues. (I) Upregulated pathways in the high-risk group identified by GSEA. (J) Downregulated pathways in the high-risk group. **, P<0.01; ****, P<0.0001. NES, normalized enrichment score; FDR, false discovery rate; qPCR, quantitative real-time PCR; HCC, hepatocellular carcinoma; GSEA, gene set enrichment analysis.

nominal P<0.001), and fatty acid metabolism (NES =–1.8, nominal P<0.001) (*Figure 5J*).

Relationship between the gene signature and the TME

Correlation analysis showed that risk score was positively correlated with the infiltration of B cell, $CD4^+$ T cell, macrophage, neutrophil, and dendritic cell in the hepatocellular TME, while there was no significant correlation between risk score and $CD8^+$ T cell (*Figure 6A-6F*). CIBERSORT indicated that, as opposed to the low-risk group, high-risk patients had more infiltration of memory B cells, plasma B cells, activated $CD4^+$ memory T cells, T follicular helper cells, regulatory T cells, and M0 macrophages. The low-risk group attracted more naïve B cells, activated $CD4^+$ memory T cells, resting natural killer (NK) cells, monocytes, and activated mast cells (*Figure 6G*). Furthermore, we compared the expression of classical immune checkpoint genes in the 2 groups. The results demonstrated that *CTLA4*, *PD-1* (*PDCD1*), *TIM-3* (*HAVCR2*), *LAG3*, and *TIGIT* were highly enriched in the high-risk group (*Figure 6H*), which indicated that high-risk patients are inclined to benefit from immune checkpoint inhibitor (ICI) treatment.

Meanwhile, we used TISCH, a single-cell RNA sequencing database, to explore the single-cell transcriptome profiles of liver cancer. The relationship between the 8 prognostic genes and immune cells of the TME was investigated in the LIHC_GSE140228_10X dataset. The cells in the LIHC_GSE140228_10X dataset were categorized into 12 types. Figure S7A,S7B depict the distribution and number of various immune cells. Results (Figure S7C-S7J) showed that *CDC20*, *PTTG1*, and *TOP2A* were highly expressed in proliferating T cells, which suggested that *CDC20*, *PTTG1*, and *TOP2A* may have an imperative function in T cell proliferation. *CXCL2*, which is a classical chemokine, was highly expressed in macrophages, indicating that macrophages in the TME may be attracted by the *CXCL2-CXCR2* axis. *MT1F* was shown to be highly expressed in proliferating T cells, followed by exhausted $CD8^+$ T cells, implying that MT1F is an important regulator in T cell development. While these results need further experiments for verification, these findings suggest that these 8 genes are closely related to the TME in liver cancer.

Discussion

HCC is a highly heterogeneous cancer with a high mortality rate and recurrence rate. Nevertheless, there are as yet no accurate and reliable prognostic biomarkers for predicting HCC risk in clinical practice. As such, it is imperative to develop novel prognostic biomarkers for patients with HCC (21). In this research, an 8-gene signature for HCC was developed based on transcriptome analysis. The risk score computed by the sum of the weighted expression levels of the 8 genes, including *CDC20*, *PTTG1*, *TOP2A*, *CXCL2*, *CXCL14*, *CYP2C9*, *MT1F*, and *GHR*, could be employed in predicting the clinical outcome of HCC patients. Patients that have a relatively high risk score may have a shorter DFS and OS time; thus, more aggressive therapies are necessary for these patients. Furthermore, this 8-gene signature is an independent prognostic factor, and the prognostic model based on the 8-gene signature showed great predictive performance.

These 8 genes are important regulators of the cell cycle, inflammation, and cellular metabolism, which are 3 important hallmarks of cancer (22). For example, *CDC20*, *PTTG1*, and *TOP2A* are involved in cell cycle regulation. *CDC20* is a cell cycle progression hub gene that contributes to the progression of many cancers (23). In HCC, *CDC20* stimulates PHD3 ubiquitination and activates the HIF-1 pathway, thus accelerating cancer cell proliferation (24). *PTTG1* is highly expressed in several tumors and is correlated with tumor differentiation, invasion, and metastasis (25). Fujii *et al.* reported that *PTTG1* was a biomarker for DFS and OS in HCC. *PTTG1* may be involved in the progression of HCC through promoting angiogenesis (26). *TOP2A* is another cell cycle-related gene that encodes an enzyme called DNA topoisomerase 2-alpha to alter the DNA strand topology states in the replication process. *TOP2A* is overexpressed in many types of cancers (27). *TOP2A* overexpression in HCC was reported to correlate with onset at an early age, shorter OS, and resistance to chemotherapy (28). In this research, the expression of *CDC20*, *PTTG1*, and *TOP2A* was upregulated in the high-risk group and indicative of poor prognosis as well.

C-X-C motif chemokine ligands (CXCLs) are important in the immune regulation of cancer, among which CXCL2 and CXCL14 are essential members (29). CXCL2 is a cytokine produced by macrophages and monocytes, and is chemotactic for hematopoietic stem cells as well as polymorphonuclear leukocytes (29). Ding *et al.* found that CXCL2 was downregulated in almost all HCC issues as opposed to adjacent normal tissues. Furthermore, CXCL2 overexpression impedes proliferation and stimulates

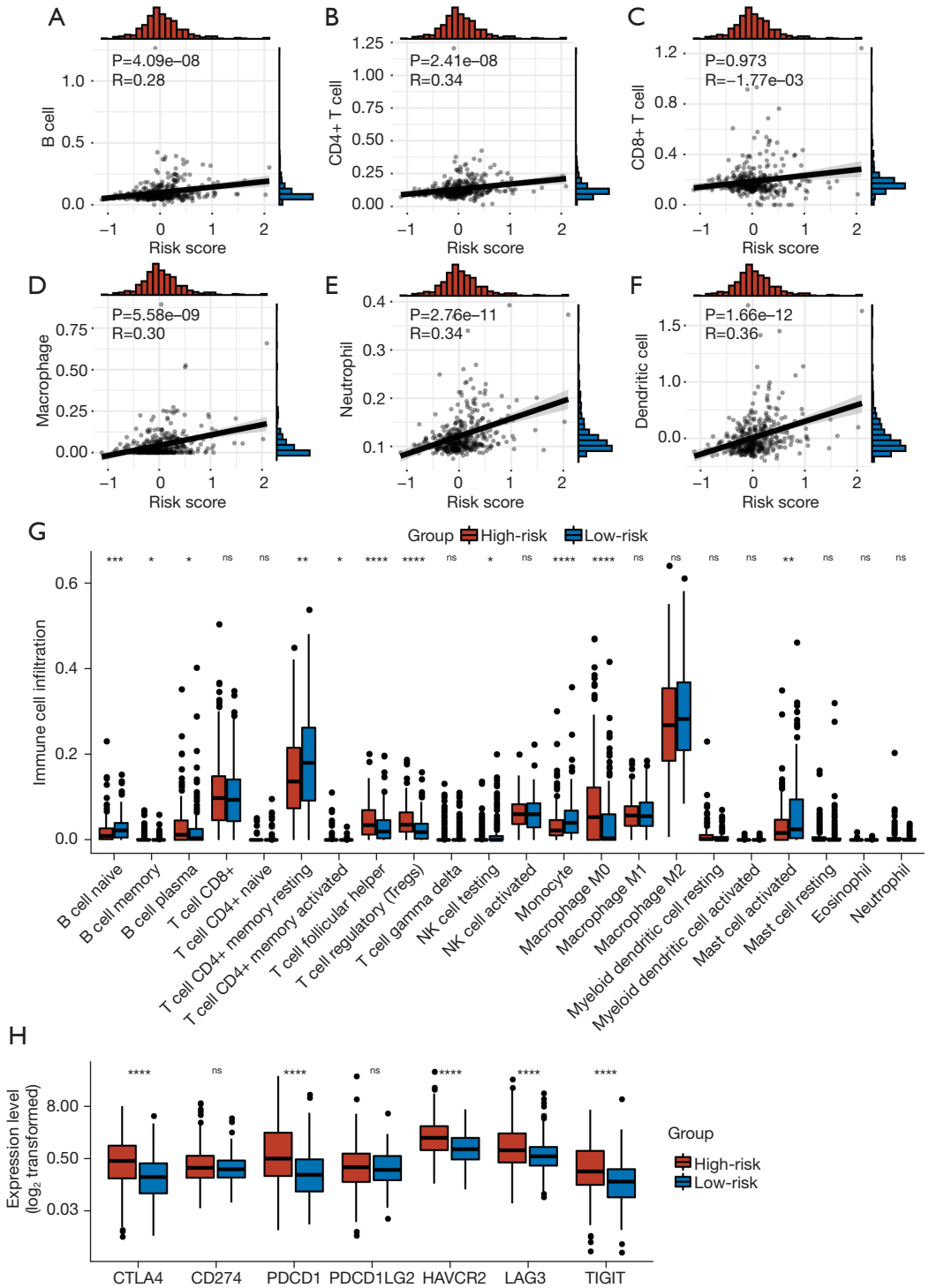

Figure 6 Immune cell infiltration and expression of immune checkpoint genes between high-risk and low-risk groups. (A-F) Correlation analysis between risk score and immune infiltration score calculated by TIMER. (G) Box plot of the infiltration of 22 immune cells in the high-risk group and low-risk group calculated by CIBERSORT. (H) Expression (log_2 transformed FPKM) of immune checkpoint molecules in the 2 groups. *, P<0.05; **, P<0.01; ***, P<0.001; ****, P<0.0001; ns, not statistically significant. NK cell, natural killer cell; FPKM, fragments per kilobase of exon model per million mapped fragments.

the apoptosis of HCC cell lines (30). CXCL14 is also a cytokine, which is constitutively expressed at high levels in many normal tissues (31). However, it is reduced or absent in cancer. This chemokine is chemotactic for and activates monocytes, dendritic cells, and NK cells. It was reported that overexpression of CXCL14 inhibits the angiogenesis of HCC. Lin *et al.* suggested that CXCL2/10/12/14 can be utilized as prognostic biomarkers for HCC (32). In this research, we found that downregulation of CXCL2 and CXCL14 in HCC to some extent changed immune cell infiltration to promote inflammation and tumor growth. However, this conclusion needs further confirmation.

CYP2C9 participates in the metabolism of xenobiotics and fatty acids in the liver. Downregulation of CYP2C9 may be a biomarker of HCC (33). Yu *et al.* showed that CYP2C9 suppression by hsa-miR-128-3p is involved in the pathogenesis of HCC (34), which is implicated in the etiology of HCC. The *GHR* gene encodes growth hormone receptor, which is embedded in the cell membrane and is most abundant in liver cells. Growth hormone and growth hormone receptors are critical for cell proliferation and metabolism in the liver. It was reported that GHR participated in the pathogenesis of HCC with chronic hepatitis C (35). GHR downregulation was found to be a new HCC biomarker. Furthermore, Gao *et al.* showed that GHR was involved in the sorafenib resistance of HCC cell lines (36). The last gene, *MT1F*, encodes an enzyme called metallothionein-1F involved in the maintenance of metal ion homeostasis. MT1F participates in cell proliferation as well as apoptosis (37), and serves as a tumor suppressor. In HCC tissue, MT1F shows downregulated expression, and overexpression of MT1F inhibits the growth of the HepG2 cell line (38). As reprogrammed energy metabolism is a significant hallmark of cancer, and the genes we identified (*CYP2C9*, *GHR*, and *MT1F*) are associated with metabolism, combing these genes in our model delivers good performance.

Prognostic biomarkers of tumors have been widely studied based on transcriptional profiles in recent years. For example, Yuan *et al.* constructed a metabolism-related gene signature (including *FABP6*, *ELOVL3*, *CSPG5*, *HMMR*, *AKR1B15*, and *G6PD*) for HCC prognosis prediction (39). Lin *et al.* constructed an inflammatory response gene signature for HCC, which was also proven to have an impact on immune cell infiltration (40). Zhang *et al.* developed a gene signature including cell cycle-related genes such as *CDC20*, *PTTG1*, *CCNB1*, *CDK1*, and *CCNA2* (41). However, these gene signatures mainly focus on 1 typical characteristic of tumors, and cannot fully reflect all the hallmarks of cancer. Consequently, we integrated genes representing cell proliferation, tumor-promoting inflammation, and aberrant metabolism into 1 model. Our model showed better predictive performance than the clinicopathological model, which proves the importance of these 8 genes. Moreover, we showed that high-risk patients have elevated expression of immune checkpoint genes, which may make them more suitable for ICI treatment. Single-cell RNA sequencing indicated that *CDC20*, *PTTG1*, and *TOP2A* are correlated with T cell proliferation, while *CXCL2* may be involved in macrophage recruitment in HCC. Our study demonstrated that these 8 genes could be probable therapeutic targets for HCC in the future.

Although the gene signature has great predictive performance, our model still has limitations. Firstly, the relationship between the 8 genes and OS should be explored at the protein level as well. Secondly, our study should be validated prospectively in independent HCC cohorts in the future. Thirdly, the gene signature must also be evaluated in cohorts with larger sample sizes. Additionally, although we carried out GSEA analysis to clarify the mechanism of this model for predicting HCC prognosis, the underlying mechanism is still unclear, so further experiments need to be conducted in the future. Lastly, significant efforts are required to translate this study into clinical application.

Conclusions

In conclusion, this research developed a novel 8-gene signature based predictive model for HCC prognostic prediction based on GEO and TCGA datasets. The nomogram based on the gene signature showed better performance as opposed to the clinicopathological model. This study might provide potential biomarkers for liver cancer. However, gene signature validation in clinical cohorts and functional experiments of these genes are warranted.

Acknowledgments

Funding: None.

Footnote

Reporting Checklist: The authors have completed the TRIPOD reporting checklist. Available at https://atm.amegroups.com/article/view/10.21037/atm-22-1934/rc

Data Sharing Statement: Available at https://atm.amegroups.com/article/view/10.21037/atm-22-1934/dss

Conflicts of Interest: All authors have completed the ICMJE uniform disclosure form (available at https://atm.amegroups.com/article/view/10.21037/atm-22-1934/coif). The authors have no conflicts of interest to declare.

Ethical Statement: The authors are accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved. This study was approved by the drug clinical trial Ethics Committee, Fourth Military Medical University (No. KY20193057). The study was conducted in accordance with the Declaration of Helsinki (as revised in 2013). All the patients who donated tissue samples provided written informed consent.

References

1. Sung H, Ferlay J, Siegel RL, et al. Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries. CA Cancer J Clin 2021;71:209-49.
2. Villanueva A. Hepatocellular Carcinoma. N Engl J Med 2019;380:1450-62.
3. Bouattour M, Mehta N, He AR, et al. Systemic Treatment for Advanced Hepatocellular Carcinoma. Liver Cancer 2019;8:341-58.
4. Duvoux C, Roudot-Thoraval F, Decaens T, et al. Liver transplantation for hepatocellular carcinoma: a model including α-fetoprotein improves the performance of Milan criteria. Gastroenterology 2012;143:986-94.e3; quiz e14-5.
5. Debes JD, Romagnoli PA, Prieto J, et al. Serum Biomarkers for the Prediction of Hepatocellular Carcinoma. Cancers (Basel) 2021;13:1681.
6. Baj J, Bryliński Ł, Woliński F, et al. Biomarkers and Genetic Markers of Hepatocellular Carcinoma and Cholangiocarcinoma-What Do We Already Know. Cancers (Basel) 2022;14:1493.
7. Nault JC, De Reyniès A, Villanueva A, et al. A hepatocellular carcinoma 5-gene score associated with survival of patients after liver resection. Gastroenterology 2013;145:176-87.
8. Kim SM, Leem SH, Chu IS, et al. Sixty-five gene-based risk score classifier predicts overall survival in hepatocellular carcinoma. Hepatology 2012;55:1443-52.
9. Zhou J, Yu L, Gao X, et al. Plasma microRNA panel to diagnose hepatitis B virus-related hepatocellular carcinoma. J Clin Oncol 2011;29:4781-8.
10. Kanehisa M, Bork P. Bioinformatics in the post-sequence era. Nat Genet 2003;33 Suppl:305-10.
11. Liu J, Lichtenberg T, Hoadley KA, et al. An Integrated TCGA Pan-Cancer Clinical Data Resource to Drive High-Quality Survival Outcome Analytics. Cell 2018;173:400-416.e11.
12. Shannon P, Markiel A, Ozier O, et al. Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome Res 2003;13:2498-504.
13. Szklarczyk D, Franceschini A, Wyder S, et al. STRING v10: protein-protein interaction networks, integrated over the tree of life. Nucleic Acids Res 2015;43:D447-52.
14. Gao J, Kwan PW, Shi D. Sparse kernel learning with LASSO and Bayesian inference algorithm. Neural Netw 2010;23:257-64.
15. Heagerty PJ, Zheng Y. Survival model predictive accuracy and ROC curves. Biometrics 2005;61:92-105.
16. Vickers AJ, Elkin EB. Decision curve analysis: a novel method for evaluating prediction models. Med Decis Making 2006;26:565-74.
17. Subramanian A, Tamayo P, Mootha VK, et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc Natl Acad Sci U S A 2005;102:15545-50.
18. Li T, Fu J, Zeng Z, et al. TIMER2.0 for analysis of tumor-infiltrating immune cells. Nucleic Acids Res 2020;48:W509-14.
19. Sturm G, Finotello F, Petitprez F, et al. Comprehensive evaluation of transcriptome-based cell-type quantification methods for immuno-oncology. Bioinformatics 2019;35:i436-45.
20. Sun D, Wang J, Han Y, et al. TISCH: a comprehensive web resource enabling interactive single-cell transcriptome

visualization of tumor microenvironment. Nucleic Acids Res 2021;49:D1420-30.
21. Williams R, Alexander G, Aspinall R, et al. Gathering momentum for the way ahead: fifth report of the Lancet Standing Commission on Liver Disease in the UK. Lancet 2018;392:2398-412.
22. Hanahan D, Weinberg RA. Hallmarks of cancer: the next generation. Cell 2011;144:646-74.
23. Wang L, Zhang J, Wan L, et al. Targeting Cdc20 as a novel cancer therapeutic strategy. Pharmacol Ther 2015;151:141-51.
24. Shi M, Dai WQ, Jia RR, et al. APCCDC20-mediated degradation of PHD3 stabilizes HIF-1a and promotes tumorigenesis in hepatocellular carcinoma. Cancer Lett 2021;496:144-55.
25. Vlotides G, Eigler T, Melmed S. Pituitary tumor-transforming gene: physiology and implications for tumorigenesis. Endocr Rev 2007;28:165-86.
26. Fujii T, Nomoto S, Koshikawa K, et al. Overexpression of pituitary tumor transforming gene 1 in HCC is associated with angiogenesis and poor prognosis. Hepatology 2006;43:1267-75.
27. Kisling SG, Natarajan G, Pothuraju R, et al. Implications of prognosis-associated genes in pancreatic tumor metastasis: lessons from global studies in bioinformatics. Cancer Metastasis Rev 2021;40:721-38.
28. Wong N, Yeo W, Wong WL, et al. TOP2A overexpression in hepatocellular carcinoma correlates with early age onset, shorter patients survival and chemoresistance. Int J Cancer 2009;124:644-52.
29. Van Sweringen HL, Sakai N, Tevar AD, et al. CXC chemokine signaling in the liver: impact on repair and regeneration. Hepatology 2011;54:1445-53.
30. Ding J, Xu K, Zhang J, et al. Overexpression of CXCL2 inhibits cell proliferation and promotes apoptosis in hepatocellular carcinoma. BMB Rep 2018;51:630-5.
31. Gowhari Shabgah A, Haleem Al-Qaim Z, Markov A, et al. Chemokine CXCL14; a double-edged sword in cancer development. Int Immunopharmacol 2021;97:107681.
32. Lin T, Zhang E, Mai PP, et al. CXCL2/10/12/14 are prognostic biomarkers and correlated with immune infiltration in hepatocellular carcinoma. Biosci Rep 2021;41:BSR20204312.
33. Tsunedomi R, Iizuka N, Hamamoto Y, et al. Patterns of expression of cytochrome P450 genes in progression of hepatitis C virus-associated hepatocellular carcinoma. Int J Oncol 2005;27:661-7.
34. Yu D, Green B, Marrone A, et al. Suppression of CYP2C9 by microRNA hsa-miR-128-3p in human liver cells and association with hepatocellular carcinoma. Sci Rep 2015;5:8534.
35. Lin CC, Liu TW, Yeh ML, et al. Significant down-regulation of growth hormone receptor expression revealed as a new unfavorable prognostic factor in hepatitis C virus-related hepatocellular carcinoma. Clin Mol Hepatol 2021;27:313-28.
36. Gao S, Ni Q, Wu X, et al. GHR knockdown enhances the sensitivity of HCC cells to sorafenib. Aging (Albany NY) 2020;12:18127-36.
37. Si M, Lang J. The roles of metallothioneins in carcinogenesis. J Hematol Oncol 2018;11:107.
38. Lu DD, Chen YC, Zhang XR, et al. The relationship between metallothionein-1F (MT1F) gene and hepatocellular carcinoma. Yale J Biol Med 2003;76:55-62.
39. Yuan C, Yuan M, Chen M, et al. Prognostic Implication of a Novel Metabolism-Related Gene Signature in Hepatocellular Carcinoma. Front Oncol 2021;11:666199.
40. Lin Z, Xu Q, Miao D, et al. An Inflammatory Response-Related Gene Signature Can Impact the Immune Status and Predict the Prognosis of Hepatocellular Carcinoma. Front Oncol 2021;11:644416.
41. Zhang H, Liu R, Sun L, et al. Comprehensive Analysis of Gene Expression Changes and Validation in Hepatocellular Carcinoma. Onco Targets Ther 2021;14:1021-31.

(English Language Editor: C. Betlazar-Maseh)

Cite this article as: Zhang J, Fu X, Zhang N, Wang W, Liu H, Jia Y, Nie Y. Development and validation of an eight-gene signature based predictive model to evaluate the prognosis of hepatocellular carcinoma patients: a bioinformatic study. Ann Transl Med 2022;10(9):524. doi: 10.21037/atm-22-1934

Supplementary

Table S1 The clinical data of the 5 independent cohorts

Characteristic	TCGA-LIHC	GSE14520	GSE87630	GSE89377	GSE121248
Platform	Illumina HiSeq2000	GPL571 GPL3921	GPL6947	GPL6947	GPL570
Samples	427	488	94	53	107
Normal	50	241	30	13	37
Tumor	377	247	64	40	70
Survival status					
Death	128	96	NA	NA	NA
Survival	249	146	NA	NA	NA
Age, years					107
≤65	235	216	NA	NA	65
>65	141	26	NA	NA	42
Gender					107
Female	122	31	NA	NA	15
Male	255	211	NA	NA	92
TNM stage					
I	180	96	NA	NA	NA
II	89	78	NA	NA	NA
III	76	51	NA	NA	NA
IV	8	0	NA	NA	NA
T classification					
T1	182	NA	NA	NA	NA
T2	91	NA	NA	NA	NA
T3	63	NA	NA	NA	NA
T4	17	NA	NA	NA	NA
N classification					
N0	349	NA	NA	NA	NA
N1	4	NA	NA	NA	NA
M classification					
M0	349	NA	NA	NA	NA
M1	4	NA	NA	NA	NA

https://dx.doi.org/10.21037/atm-22-1934

Table S2 The primer sequences of eight prognostic genes and β-actin

Primer	Forward sequence	Reverse sequence
CDC20	CGGAAGACCTGCCGTTACATTC	CAGAGCTTGCACTCCACAGGTA
PTTG1	GCTTTGGGAACTGTCAACAGAGC	CTGGATAGGCATCATCTGAGGC
TOP2A	GTGGCAAGGATTCTGCTAGTCC	ACCATTCAGGCTCAACACGCTG
CXCL2	GGCAGAAAGCTTGTCTCAACCC	CTCCTTCAGGAACAGCCACCAA
CXCL14	AGATCCGCTACAGCGACGTGAA	GCAGTGCTCCTGACCTCGGTA
CYP2C9	CAGAGACGACAAGCACAACCCT	ATGTGGCTCCTGTCTTGCATGC
GHR	GCAGCTATCCTTAGCAGAGCAC	AAGTCTCTCGCTCAGGTGAACG
MT1F	GACTGATGCCAGGACAACCT	AGGAATGTAGCAAATGGGTCA
β-actin	CACCAACTGGGACGACAT	ACAGCCTGGATAGCAACG

Table S3 Top 20 genes ranked by the degree method

Rank	Gene	Score	Rank	Gene	Score
1	*PTTG1*	6	10	*IGFALS*	2
1	*CDC20*	6	10	*STAB2*	2
1	*CCNB2*	6	10	*CXCL2*	2
1	*NUSAP1*	6	10	*CXCL12*	2
1	*TOP2A*	6	10	*CYP1A2*	2
1	*PRC1*	6	10	*CYP2C9*	2
1	*ASPM*	6	10	*MT1G*	2
8	*FOS*	3	10	*MT1H*	2
8	*CYP2E1*	3	19	*CXCL14*	1
10	*MT1F*	2	19	*GHR*	1

https://dx.doi.org/10.21037/atm-22-1934

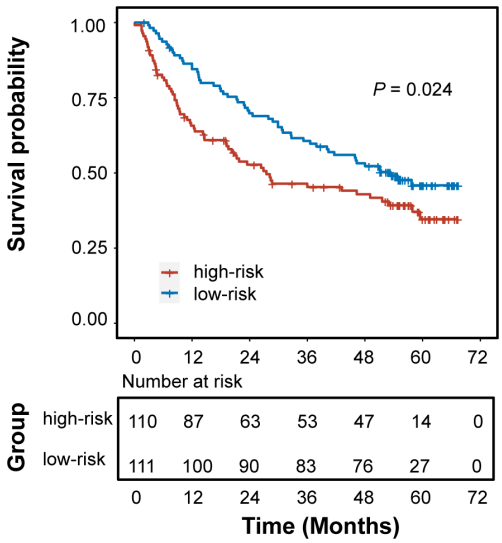

Figure S1 Kaplan-Meier analysis shows that HCC patients with high risk scores have poorer DFS in the GSE14520 validation set.

https://dx.doi.org/10.21037/atm-22-1934

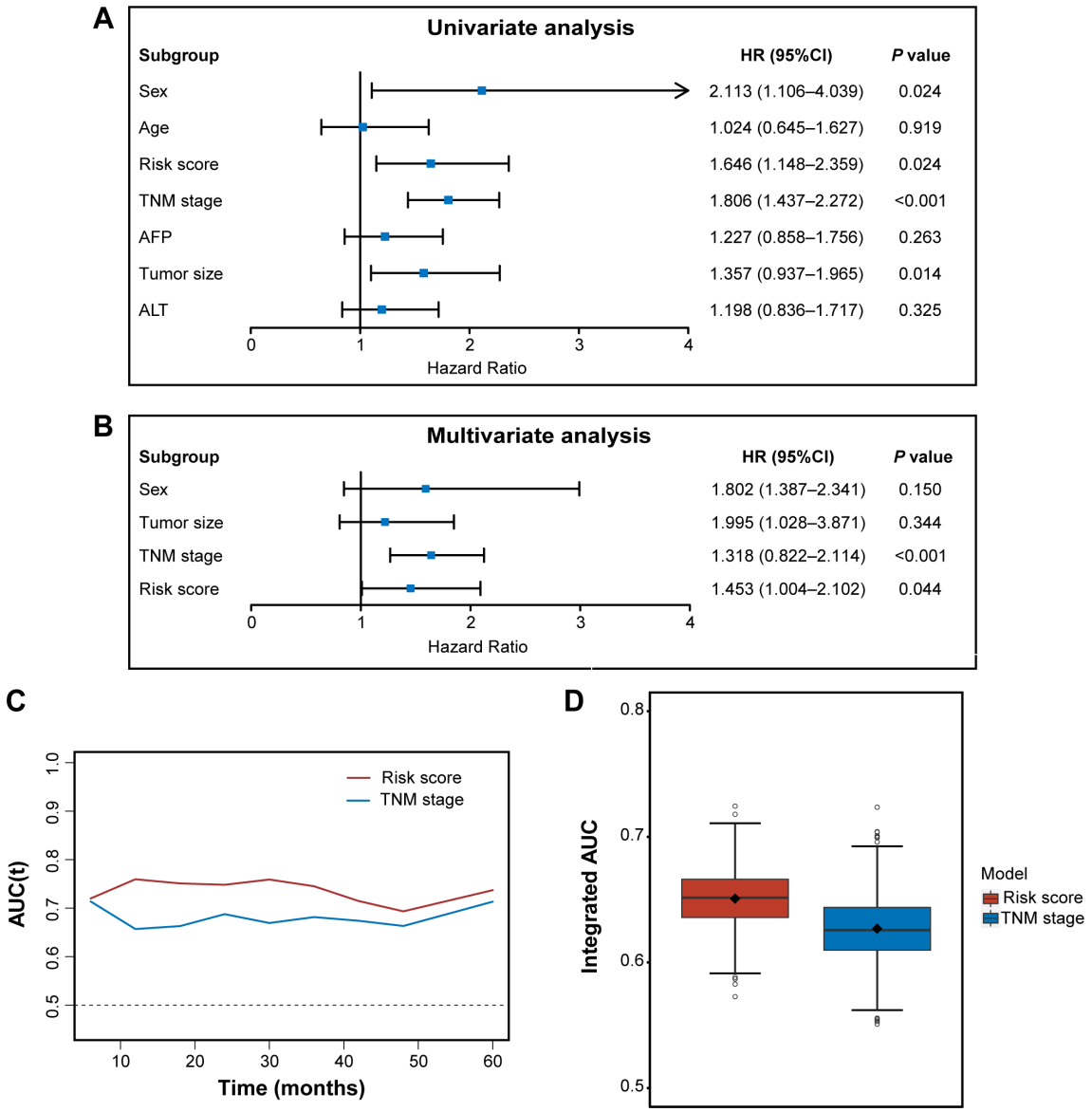

Figure S2 (A,B) Univariate and multivariate Cox regression analysis for DFS in the GSE14520 validation set. (C) Time-dependent AUC for DFS shows that risk score has better predictive accuracy than TNM stage in the GSE14520 validation set. (D) The iAUC indicates integrated area under the ROC curve, which shows that risk score has better predictive performance than TNM stage. DFS, disease-free survival; AUC, area under the curve; iAUC, integrated value of time-dependent AUC.

https://dx.doi.org/10.21037/atm-22-1934

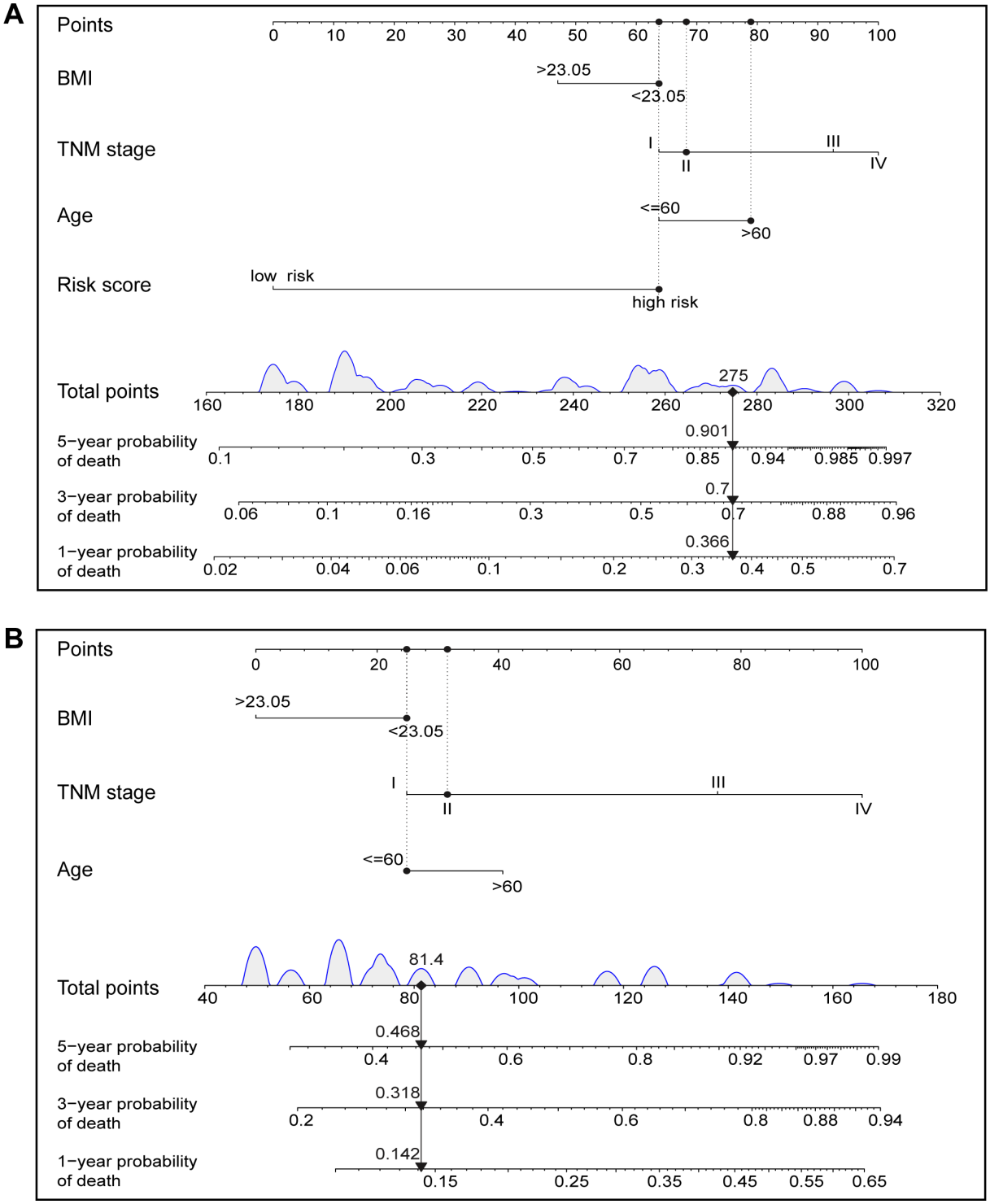

Figure S3 (A) The nomogram to forecast the 1-, 3-, and 5-year death likelihood of HCC patients in the TCGA-LIHC training set. The nomogram model is constructed based on BMI, TNM stage, age, and risk score of the 8 risk genes. (B) The nomogram without risk score in the TCGA-LIHC training set.

https://dx.doi.org/10.21037/atm-22-1934

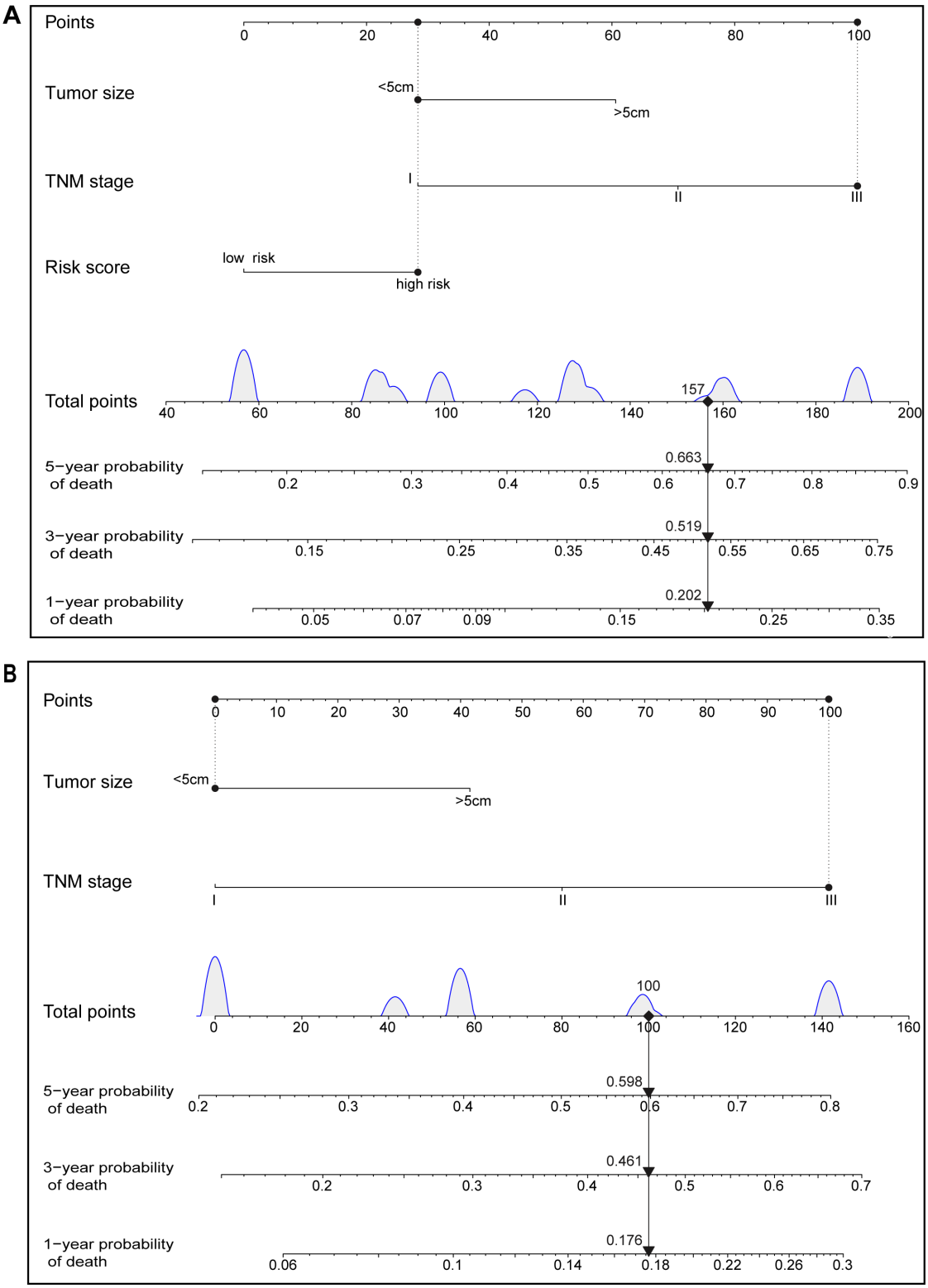

Figure S4 (A) The nomogram to forecast the death likelihood of HCC patients at 1, 3, and 5 years in the GSE14520 validation set. The nomogram model is constructed based on TNM stage, tumor size, and risk score of the 8 risk genes. (B) The nomogram without risk score constructed in the GSE14520 validation set.

https://dx.doi.org/10.21037/atm-22-1934

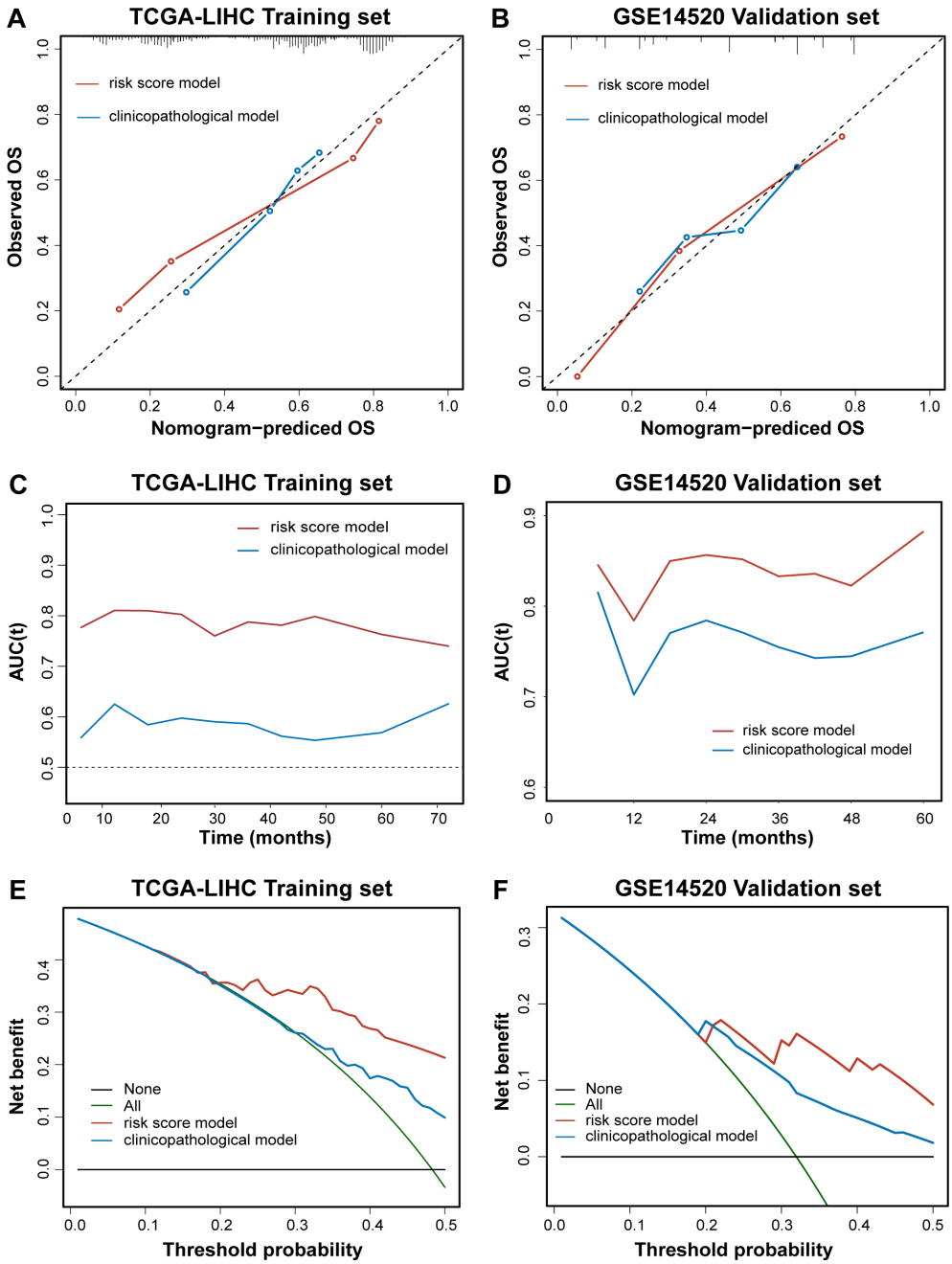

Figure S5 (A,B) The nomogram calibration curves in the TCGA-LIHC dataset and GSE14520 validation set. (C,D) Time-dependent AUC of the risk score model and clinicopathological model in the training set and validation set. (E,F) DCA curves of the risk score model and clinicopathological model in the training set and validation set.

https://dx.doi.org/10.21037/atm-22-1934

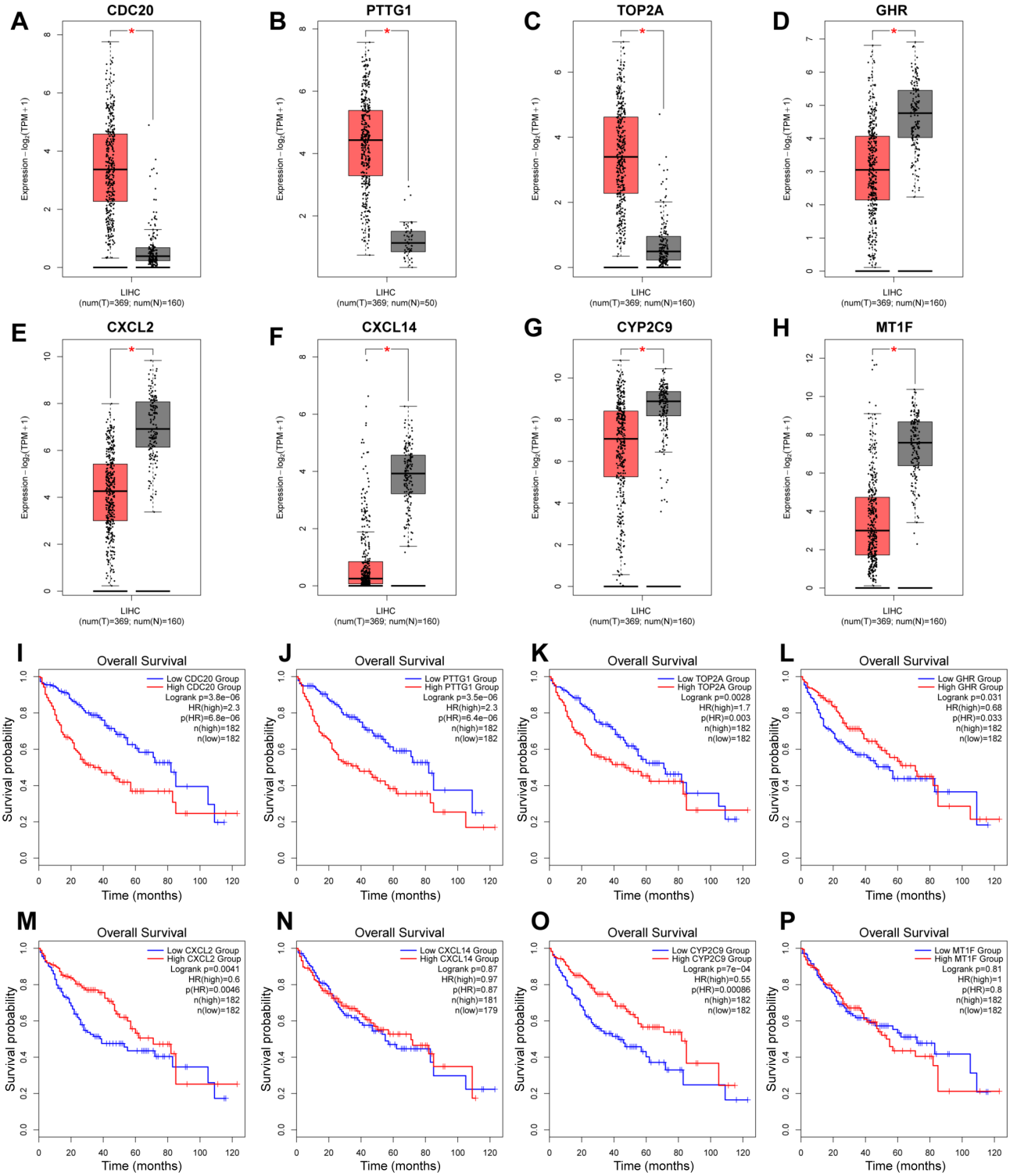

Figure S6 (A-H) Comparisons of mRNA expression of each gene in HCC tissues versus adjacent normal tissues in TCGA-LIHC via GEPIA. (I-P) Validation of the prognostic role of each gene by Kaplan-Meier survival analysis via GEPIA. *, P<0.05.

https://dx.doi.org/10.21037/atm-22-1934

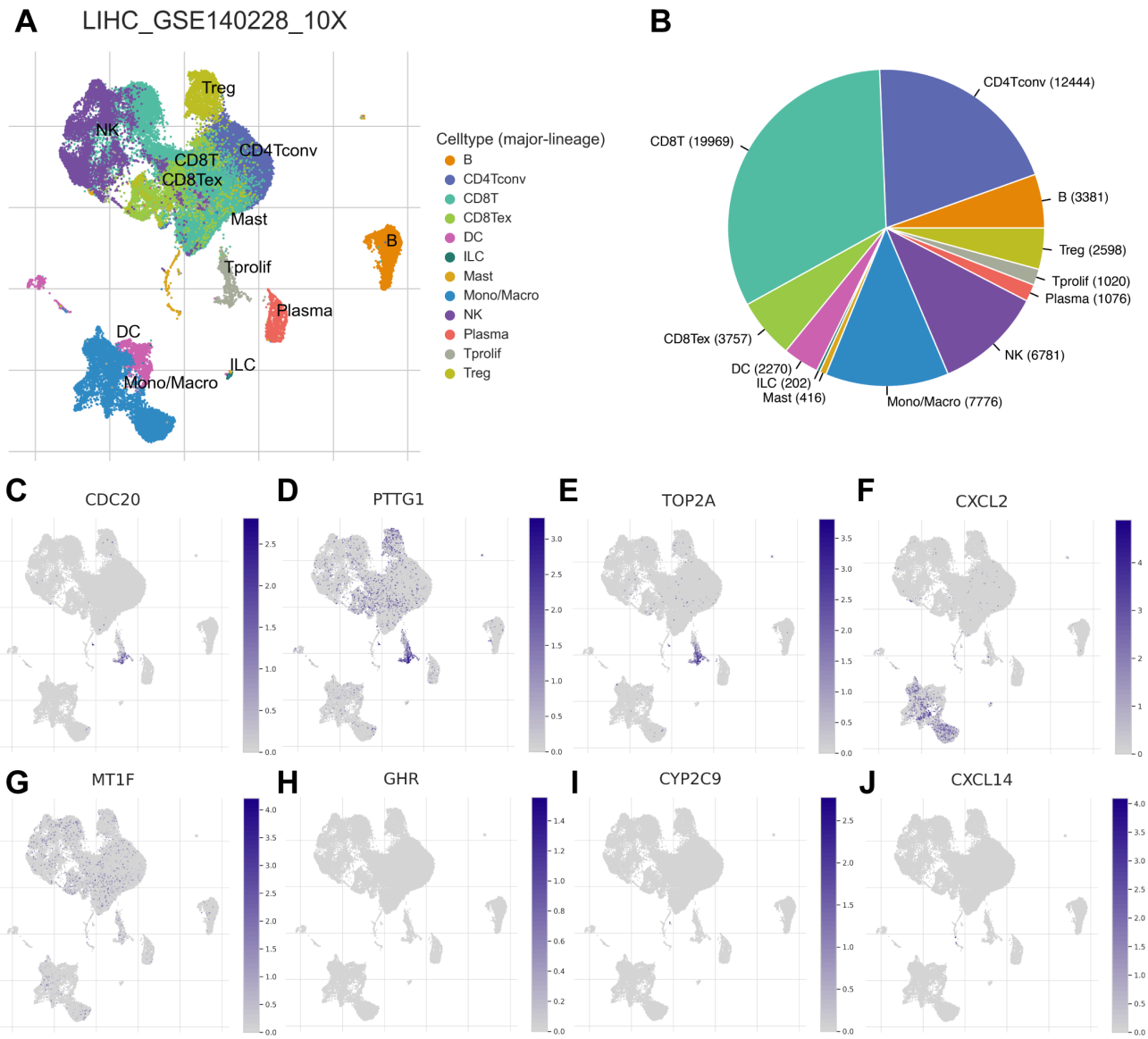

Figure S7 (A) The landscape of different cell types in the LIHC_GSE140228_10X single-cell RNA sequencing dataset. (B) The cell types as well as their distribution in the LIHC_GSE140228_10X dataset. (C-J) The expression of *CDC20*, *PTTG1*, *TOP2A*, *CXCL2*, *MT1F*, *GHR*, *CYP2C9*, and *CXCL14* in different cell types.

https://dx.doi.org/10.21037/atm-22-1934