A Novel Five-gene Signature Predicts Overall Survival in Hepatocellular Carcinoma

Zhigang Wang
The First People`s Hospital of Jingmen
Leyu Pan
The First People`s Hospital of Jingmen
Deliang Guo
Wuhan University Zhongnan Hospital
Xiaofeng Luo
The First People`s Hospital of Jingmen
Jie Tang
The First People`s Hospital of Jingmen
Weihua Yang
The First People`s Hospital of Jingmen
Yang Gu (791521131@qq.com)
The First People`s Hospital of Jingmen https://orcid.org/0000-0002-6584-7216
Yuxuan Pan
The First People`s Hospital of Jingmen

Primary research

Keywords: Hepatocellular carcinoma, Cox regression analysis, Prognosis, Nomogram, quantitative real-time PCR

Posted Date: June 1st, 2020

DOI: https://doi.org/10.21203/rs.3.rs-32144/v1

A Novel Five-gene Signature Predicts Overall Survival in Hepatocellular Carcinoma

Zhigang Wang[1†], Leyu Pan[1†], Deliang Guo[3], Xiaofeng Luo[1], Jie Tang[1], Weihua Yang[1], Yang Gu[1]*, Yuxuan Pan[2]*

[1]Department of Hepatobiliary and Pancreas, The First People`s Hospital of Jingmen, Jingmen, Hubei, China.

[2]Department of Blood Transfusion, The First People`s Hospital of Jingmen, Jingmen, Hubei, China.

[3]Department of Hepatobiliary and Pancreas, Zhongnan Hospital of Wuhan university, Wuhan, Hubei, China.

*To whom Correspondence should be addressed to Yang Gu. E-mail: 791521131@qq.com

Correspondence may also be addressed to Yuxuan Pan. E-mail: 974554423@qq.com

Abstract

Background: Hepatocellular carcinoma (HCC) is one of the most common challenges for public health worldwide. Due to its complex molecular and great heterogeneity, the effectiveness of existing HCC risk prediction models is unsatisfactory. Hence, more accurate prognostic models are pressingly needed.

Materials and methods: Differentially expressed mRNAs (DEMs) between HCC and normal tissues were identified after downloading GSE1450 from gene omnibus (GEO) database. We randomly divided all patients into training and testing sets. Univariate

Cox regression, lasso Cox regression and multivariable Cox regression analysis were used to constructed the prognostic gene signature in training set. Our study utilized Kaplan-Meier plot, time-dependent receiver operating characteristic (ROC), multivariable Cox regression analysis with clinical information, nomogram and decision curve analysis (DCA) to evaluate the predictive ability for overall survival of the novel gene signature in training, testing and whole sets. We also validated the prognostic capacity of the five-gene signature in an external validation set. The information of mutation of each gene was explored on cBioPortal online website. We performed gene set enrichment analysis (GSEA) to explore underlying mechanisms in the high and low risk group. Finally, quantitative real-time PCR was conducted to validate the expression tendency between 12 paired HCC and adjacent normal tissues.

Results: Our study constructed a novel five-gene signature (*CNIH4*, *SOX4*, *SPP1*, *SORBS2* and *CCL19*) for predicting overall survival of HCC. Time-dependent ROC curve indicated admirable ability in survival prediction in two datasets. Multivariable Cox regression analysis indicated that both this five-gene signature and TNM stage were two independent prognostic factors for overall survival of HCC patients. Combined with TNM stage clinical pathological parameters, the predictive capacity of nomogram had a decent improvement. The mutation of the five genes had no obvious variation. Plenty pathways were enriched by GSEA, including cell cycle and various metabolism. Furthermore, the mRNA levels of these five genes had significantly different expressions between HCC tissues and adjacent normal tissues by quantitative real-time PCR.

Conclusions: A five-gene prognostic model and nomogram were constructed and validated for predicting prognostic of HCC patients. And the five-gene risk score with TNM stage models might help various HCC patients to customize individual therapies.

Keywords: Hepatocellular carcinoma, Cox regression analysis, Prognosis, Nomogram, quantitative real-time PCR

Background

Hepatocellular carcinoma (HCC) is the third leading cause of cancer-related death, which parallels with gastric cancer and only behind colorectal cancer and lung cancer worldwide[1-3]. The major risk factors for HCC are hepatitis B virus (HBV) infection, hepatitis C virus (HCV) infection, cirrhosis, aflatoxin contamination and so on[4]. Since HCC patients are usually asymptomatic at an early stage and most HCC patients likely lose the best treatment period with an advanced stage. Although therapeutic methods for HCC have massively improved in recent decades, the ten year overall survival of HCC patients remains unsatisfactory[5, 6]. Conventional clinical parameter, such as TNM stage, histologic grade and portal vein tumor thrombus (PVTT), could help predict the prognosis of HCC patients[7]. Because of the enormous heterogeneity of HCC, the predictive effect of the traditional model is not satisfactory yet. It is vital and urgent to establish a sensitive and reliable prognostic signature for optimizing the clinical treatment decision for individuals.

With the development of genome-sequencing technology, plenty of studies showed that gene signature had huge potential in predicting HCC prognosis, including mRNA, lncRNA and microRNAs[8-11]. By using public free genomic data, people could identify efficient gene signature to predict prognosis of patients. However, it is limited that individual gene signature discards clinical parameters for predicting overall survival. So, it is necessary for us to identify novel gene signatures combined with clinical information.

In this study, plenty bioinformatics methods, such as differential expression mRNAs (DEMs), univariate Cox regression, least absolute shrinkage and selection operator (LASSO) Cox regression, multivariable Cox regression analysis, Kaplan-Meier plot, time-dependent ROC, decision curve analysis (DCA) and gene set enrichment analysis (GSEA), were used to build and validate a five-gene signature model which could combine with TNM stage to foresee the prognosis of HCC patients. Meanwhile, GSEA was conducted to explore underlying mechanisms in the high and low risk group. Quantitative real-time PCR was also operated to validate the expression tendency between 12 paired HCC and adjacent normal tissues.

Materials and methods

Data source

The GSE14520 dataset containing genes expression and clinical information was downloaded from GEO database in NCBI (https://www.ncbi.nlm.nih.gov/pmc/)[12]. We used the GPL3921 to re-annotate these probes. We visited UCSC Xena online website(http://xena.ucsc.edu/public/) for acquiring The Cancer Genome Atlas—Liver Hepatocellular Carcinoma (TCGA-LIHC) validation set which contained gene expression profile, clinical and survival information.

Differential expression mRNAs analysis

To obtain differential expression mRNAs (DEMs) between HCC and non-tumor samples, the "limma" software package in R was used to normalize and analyze the GSE14520 dataset[13]. The $|\log(FC)| > 1$ and p value < 0.05 were considered for next study.

Construction of prognostic gene signature

We eliminated the samples whose overall survival time was less than one month for subsequent operation. With p value < 0.05, we conducted univariate Cox regression analysis to explore DEMs which had significantly correlated with overall survival. By using "caret" package in R, the patients were randomly divided into two groups, training and testing sets. Next, LASSO regression analysis was used for vital prognostic DEMs. We used "glmnet" package to conduct this step[14]. Then, the vital prognostic mRNAs with expression and survival information were devoted into multivariable Cox regression analysis to obtain a prognostic risk formula. The risk score = ($Coefficient_{mRNA1}$ * expression value of mRNA1) + ($Coefficient_{mRNA2}$ * expression value of mRNA2) + ($Coefficient_{mRNA3}$ * expression value of mRNA3) + ⋯ + ($Coefficient_{mRNAn}$ * expression value of mRNAn). We utilized the "surv_cutpoint" function of "survminer" package to explore the optimum cut-off risk score to separate

patients into high and low risk groups. The Kaplan-Meier survival curve and log-rank test were used to examine the survival condition of two different groups. To check the predictive ability of the prognostic gene signature, we utilized “timeROC” package to calculate the area under curve (AUC) values of 1-year, 3-year and 5-year[15]. Finally, the coefficients of all prognostic DEMs were used to testing and whole sets.

Independent validation of the prognostic gene signature and expression

In order to examine the prognostic gene signature, we used TCGA-LIHC dataset to validate. The risk score of each patient was acquired by the same formula like before. Next, we also chose the optional cut-off value to separate patients into high and low risk groups using the “surv_cutpoint” function. The Kaplan-Meier curve and time-dependent ROC curve were utilized to examine the ability for predicting overall survival of the risk gene signature in validation dataset. The expression profile of mRNAs in the risk gene signature between tumor and non-tumor tissues in 50 paired tissues was calculated by paired t-test. The p value < 0.05 was considered statistically significant.

Independent prognostic role of the gene signature by multivariate analysis

To identify the capacity of the prognostic gene signature, we put all clinical information [gender, age, alanine transaminase (ALT), tumor size, multinodular, cirrhosis, serum alpha fetoprotein (AFP) and tumor node metastasis (TNM) stage] in GSE14520 into univariate and multivariate analysis. Meanwhile, we also had the same operation in TCGA-LIHC clinical information [gender, age, body mass index (BMI), tumor grade, cancer status and TNM stage]. The p value <0.05 was considered significant.

A predictive nomogram building and validating

Various cancer prognosis has been extensively predicted by Nomogram[16, 17]. The nomogram was built to research the probability of 1-year, 3-year and 5-year overall survival in GSE14520 using all independent prognostic factors calculated by multivariate Cox analysis. Meanwhile, the predictive capacity of nomogram was

evaluated by calibration and discrimination. We obtained the concordance index (C-index) to evaluate the discrimination of the nomogram by a bootstrap method with 1000 resamples. Next, the calibration curve was plotted to watch the influence of observation rates in different years on the prediction probability of nomogram. Next, we compared different models and combined models by C-index, time-dependent ROC curve and DCA in R[18]. Subsequently, we conducted the same operation in the independent validation set.

Gene set enrichment analysis

After downloaded The Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway in C2 genes sets from gene set enrichment analysis (GSEA) online website (https://www.gsea-msigdb.org/gsea/index.jsp)[19], we performed GSEA to probe the promising function of the established prognostic gene signature in GSE14520 or TCGA-LIHC set. The p value < 0.05 and false discovery rate (FDR) q value < 0.25 were remained.

Tissues collection and quantitative real-time PCR

We collected 12 pairs of HCC and corresponding adjacent tissues from 12 different HCC patients after surgery. The pathological type of these tissues was confirmed by the pathology department in The First People`s Hospital of Jingmen. The tissues of HCC patients used to quantitative real-time PCR were based on the following criteria: (1) patients treated in The First People`s Hospital of Jingmen; (2) none of the patients received any neoadjuvant therapy before surgery. Informed consent was obtained from each patient. The research was performed in accordance with relevant guidelines/regulations.

Total RNA was extracted using TRIzol reagent (Invitrogen, Carlsbad, CA, United States). The ratio of absorbance at A260/A280 was nearly 2.0 for reverse transcription using the NanoDrop spectrophotometer (Thermo Scientific Inc.). We used the PrimeScript RT reagent Kit with gDNAEraser (Takara, Tokyo, Japan) to proceed the reverse transcription. After setting up the proper protocol, quantitative real-time PCR

was conducted by the CFX Connect Real-Time PCR Detection System (Bio-Rad, United States) with the SYBR Green PCR kit (Toyobo, Osaka, Japan). We used *GAPDH* as an internal control. The $2^{-\Delta\Delta Ct}$ method was used to analyze the outcome. All the primers of genes were designed by the website PrimerBank (https://pga.mgh.harvard.edu/primerbank/). The sequences and Tm values of all primers were listed in Supplementary Table 1.

Statistical analysis

All of the statistical analysis was performed by R (version 3.6.3). Categorical variables were analyzed using the Pearson $\chi 2$ test or Fisher`s exact test; paired tissues were used paired t-test. Wilcoxon test was used to unpaired samples when non-normal distribution. $P < 0.05$ was considered statistically significant.

Results

DEMs analysis

The overview of research design is showed in Figure 1. After normalized the expression matrix, the Principal Components Analysis (PCA) analysis showed that the tumor group and normal group could distinguish well (Supplementary Figure1A). A total number of 443 DEMs, including 110 up-regulated mRNAs and 333 down-regulated mRNAs was identified between 225 HCC tissues and 220 normal tissues by “limma” package. The heatmap was presented in Supplementary Figure 1B and the volcano plot was showed in Supplementary Figure 1C.

Construction of the five-gene prognostic gene signature

All the DEMs with overall survival information of patients was constructed a new matrix. A total number of 221 patients with overall survival longer than were remained. All these patients were randomly classified into two groups, training set (n = 112) and testing set (n = 109). The matrix of two groups was listed in Supplementary file 1. By setting p value < 0.05, univariate Cox regression selected 84 mRNAs that have

correlation with overall survival. Next, LASSO analysis was used to identify the real significant mRNAs in training set and seven mRNAs were identified by LASSO analysis (Supplementary file 2). Then, the multivariable Cox regression analysis was used to the seven mRNAs. Five key mRNAs including Cornichon Family AMPA Receptor Auxiliary Protein 4 (*CNIH4*), SRY-Box Transcription Factor 4 (*SOX4*), Secreted Phosphoprotein 1 (*SPP1*), Sorbin And SH3 Domain Containing 2 (*SORBS2*) and C-C Motif Chemokine Ligand 19 (*CCL19*) were identified in the end. The prognostic risk score formula was as follow: risk score = 0.3972 * (the expression value of *CNIH4*) + 0.1962 * (the expression value of *SOX4*) + 0.1122 * (the expression value of *SPP1*) + (-0.2912) * (the expression value of *SORBS2*) + (-0.3176) * (the expression value of *CCL19*). The risk scores were calculated for each patient and all patients were separated into high risk group (n=45) and low risk group (n=67) by "surv_cutpoint" function of "survminer" package in R. We utilized time-dependent ROC and Kaplan-Meier curve to evaluate the ability for predicting overall survival of the five-gene signature in training set. The distribution of risk scores and survival status of the patients were also shown. We performed similar operation in testing set and the whole set. All the p values of Kaplan-Meier plot between high and low risk groups in three sets were less than 0.001, which meant high risk group had poorer prognosis compared with low risk group. The area under curves (AUCs) for 1-year, 3-year and 5-year overall survival in three sets were higher than 0.70 (Figures2A-B, Figure3A). These results indicated that the five-gene signature had good manifestation for overall survival prediction.

External validation of the prognostic gene signature

To certify the capacity of the five-gene signature for predicting overall survival, we used the same formula to calculate risk score in TCGA-LIHC dataset. After chose appropriate cut-off value 2.88, all patients were split into high risk group (n=97) and low risk group (n=266). The Kaplan-Meier plot showed the same result with three sets. The AUCs for 1-year, 3-year and 5-year overall survival were respectively 0.728, 0.694 and 0.643 (Figure3B). In summary, the five-gene signature had certain ability to predict

overall survival in HCC.

Independent prognostic role of the prognostic gene signature

After we constructed the five-gene signature with high and low risk group, we formed all clinical information of GSE14520, including gender, age, ALT, tumor size, multinodular, cirrhosis, serum AFP and TNM stage into a table. The missing information used "Not Available" instead. Results from clinical studies indicated that higher risk score had significantly relation to advanced age, larger tumor size, cirrhosis, higher AFP and advanced TNM stage. Meanwhile, we conducted the same analysis to TCGA-LIHC clinical information. It showed that higher risk score was extremely correlated with advanced TNM stage and histologic grade (Table1). To confirm the significance of the five-gene signature, we used univariate and multivariate Cox regression analysis to analyze the GSE14520 and TCGA-LIHC datasets. Results showed that both TNM stage and risk prognostic model were the independent prognostic factors for overall survival (Figure 4). Two datasets showed that patients in high risk group had poorer overall survival compared with those in low risk group both in low TNM stage (I+II) and high TNM stage (III+IV) (Figures5A-D).

A predictive nomogram building and validating in GSE14520 and TCGA-LIHC sets

In GSE14520 set, we built the nomogram including prognostic model and TNM stage (Figure 6A). The 5-year nomogram was the most optimal in predicting overall survival according to the calibration analysis of 1-year, 3-year and 5-year (Figure 6C). The values of C-index were 0.716, 0.624 and 0.737 for prognostic model, TNM stage model and combined model, respectively. The ROCs of prognostic model, TNM stage model and combined model were also analyzed by "TimeROC" package (Figure 6D). The decision curve analysis (DCA) showed combined model had more net benefit for predicting overall survival of 5-year compared with prognostic model and TNM model (Figure 6B).

In TCGA-LIHC dataset, the nomogram was also built by the two independent

prognostic factors (Figure 7A). The calibration plots of 1-year, 3-year and 5-year indicated that presentation of the nomogram was best in predicting 5-year overall survival (Figure 7C). The values of C-index were 0.671, 0.590 and 0.667 for prognostic model, TNM stage model and combined model, respectively. The AUCs for 1-year, 3-year and 5-year were greater than 0.65 (Figure 7D). More net benefit for predicting overall survival of 5-year was demonstrated after combining the prognostic model with TNM module (Figure7B).

In summary, Combination of both prognostic model and TNM stage model could effectively enhance the sensitivity and specificity of prediction and acquire more net benefit and this combined model might enhance predictive function of overall survival for HCC patients in clinical.

Mutation information and gene set enrichment analysis

To find deeper value of the five-gene signature, we searched the cBioPortal online website to explore the mutation information of each gene. The result showed *CNIH4* had 6% amplification and *SORBS2* had 4% deep deletion and the other genes hardly changed (Figure 8A). By setting up cut-off p and q value, 37 significant KEGG pathways were enriched in GSE14520 by GSEA. Spliceosome, ribosome, cell cycle and basal transcription factors were enriched in high risk group. Plenty pathways of metabolism were enriched in low risk group, such as histidine metabolism, tyrosine metabolism, butanoate metabolism, fatty acid metabolism and so on (Figure 8B, Supplementary file3).

External validation in expression

We selected 50 paired samples to validate the expression tendency in TCGA-LIHC dataset. The results showed all of the five genes had significant differential expression between tumor and non-tumor tissues expect *SORBS2* (Supplementary Figure 2A*)*. We found risk score had significant differential expression between low and high TNM stage both in GSE14520 and TCGA-LIHC sets. Furthermore, risk score in low and high histologic grade also had significant differential expression (Supplementary Figure 2B).

The mRNA levels of these five genes had significantly different expressions between HCC tissues and adjacent normal tissues by quantitative real-time PCR in 12 paired HCC tissues (Supplementary Figure 2C). In summary, aberrant expression of the five genes was validated in HCC.

Discussion

Hepatocellular carcinoma is still one of the malignant tumors of high mortality worldwide[20]. Due to complicated pathogenesis, it is extremely hard to have a satisfying prediction model for overall survival in HCC. Traditional clinical indicators such as TNM stage, histologic grade and portal vein tumor thrombus (PVTT) could partly predict prognosis of HCC patients. Nevertheless, because of the enormous heterogeneity of HCC, it is necessary for people to find novel prognostic biomarkers and build more precise prognostic models. Compared with single biomarker, a system of multiple prognostic models could enhance predictive efficacy.

In this study, through mining the GSE14520 and validating in TCGA-LIHC dataset, we constructed a novel five-gene signature (*CNIH4*, *SOX4*, *SPP1*, *SORBS2* and *CCL19*) for prognosis prediction of HCC patients. The predictive capacity for overall survival of the five-gene signature performed well both in GSE14520 training, testing, whole set and TCGA-LIHC dataset. By univariate and multivariate analysis with clinical information in two datasets, the TNM stage and five-gene signature were two independent prognostic factors for overall survival in HCC. Meanwhile, patients in high risk group had significant poorer overall survival both in TNM stage I+II and III+IV in two datasets compared with those in low risk group. By constructing nomogram including five-gene signature model and TNM stage model, the prognosis prediction had a better improvement, which may subdivide HCC patients more accurately for individual treatment. In summary, all these results indicated that the five-gene signature model plays an important role for predicting overall survival of HCC patients. GSEA showed several significant enriched KEGG pathways for the five-gene signature, such as cell cycle and various kinds of metabolism, which might comprehend the underlying

molecular mechanisms. The 50 paired HCC tissues from TCGA-LIHC dataset showed these five genes had significant differential expression between tumor and non-tumor tissues expect *SORBS2*. And we used 12 paired HCC tissues of our hospital to validate expression tendency between tumor and non-tumor tissues.

Cornichon Family AMPA Receptor Auxiliary Protein 4 (*CNIH4*) plays an important role in regulating G protein-coupled receptors (GPCRs) transporting from the endoplasmic reticulum (ER) to the functional site (cell surface). And overexpression and down-regulation of *CNIH4* resulted in the retention of GPCRs[21]. In colon cancer, *CNIH4* which encodes a member of the CORNICHON family of evolutionarily conserved TGFα exporters, is required for metastasis and is regulated by *TMED9* activity[22]. The role of *CNIH4* in HCC has not been reported yet. SRY-Box Transcription Factor 4 (*SOX4*), a member of a highly conserved transcription factor SOX (SRY-Box) family known to have a typical DNA-binding HMG domain[23]. It has been reported overexpressed *SOX4* could promote metastasis in HCC. Through technology of immunoprecipitation and gene ablation, two *SOX4* target genes which had influence on HCC metastasis were identified and validation[24]. A study demonstrated that the HMG box domain of *SOX4* interacted with p53, leading to the inhibition of p53-mediated transcription by the Bax promoter. More importantly, overexpressed *SOX4* led to a remarkably inhibition of p53-induced Bax expression and subsequent repression of p53-mediated apoptosis induced by gamma-irradiation in HCC[25]. Secreted Phosphoprotein 1 (*SPP1*) is a secreted arginine-glycine-aspartate (RGD)containing phosphoprotein. It might be an important molecule of tumor metastasis mediated by macrophage invasion and direct stimulation of macrophage migration[26]. It has been reported that genetic polymorphisms of the SPP1 gene are linked with HBV clearance and onset age of HCC in a large Korean HBV study, which might provide a way to elucidate the molecular mechanisms underlying HBV clearance and HCC progression[27]. Sorbin And SH3 Domain Containing 2 (*SORBS2*) is essential for regulating cell adhesion and actin/cytoskeletal organization. A recent reported *SORBS2* could suppress metastatic colonization in ovarian cancer by some mechanisms[28]. A study showed that expression of *SORBS2* in HCC was significantly

decreased, which was related to HCC metastasis, TNM stage and prognosis. Mechanistically, *SORBS2* contributed to the suppression of HCC tumourigenesis and metastasis via post-transcriptional regulation of *RORA* expression as an RNA-binding protein[29]. Another study indicated that *SORBS2* was down-regulated by *MEF2D* and inhibited HCC metastasis through the c-Abl /ERK signaling pathway, which might become a new prognostic marker or therapeutic target for HCC[30]. C-C Motif Chemokine Ligand 19 (*CCL19*) might play a role not only in inflammatory and immunological responses but also in normal lymphocyte recirculation and homing. A study showed knock-down levels of CC chemokine receptor like 1 (*CCRL1*) could inhibit expression of *CCL19* and *CCL21*. By isolating *CCL19* and *CCL21*, *CCRL1* reduced their binding to CCR7 and consequently reducing the harmful effects of CCR7, including Akt-GSK3 pathway activation in tumor cells[31]. The specific mechanism of *CCL19* in HCC remains to explore.

So far, our team identified a novel five-gene signature prognostic model and nomogram for predicting overall survival of HCC. Combined with TNM stage clinical pathological parameters, the capacity of prediction had a decent improvement, especially in 5-year overall survival. And by quantitative real-time PCR validation, these five genes had significant differential expression between tumor and non-tumor tissues. Although the five-gene signature was constructed and seemed to be a potential prognostic biomarker in clinical, there are some limitations. Firstly, the number of external validation dataset is no abundant. The second limitation is our study didn`t explore the expression and prognostic effects of the five genes at the protein level. Thirdly, the risk score model needs further validation from clinical trials. To verify the results of this study, further clinical studies are necessary. Finally, our study has not explored the underlying mechanical of the five genes at cellular and molecular levels.

Conclusion

In brief, our study constructed and validated a five-gene prognostic model and nomogram to predict overall survival of HCC. And the five-gene prognostic model with

TNM stage model might help various HCC patients to customize individual therapies.

Abbreviations

HCC, Hepatocellular Carcinoma; GEO, Gene Expression Omnibus; TCGA-LIHC, The Cancer Genome Atlas- Liver Hepatocellular Carcinoma; NCBI, National center for biotechnology information; DEMs, Differentially Expressed mRNAs; FC, Fold Change; LASSO, Least Absolute Shrinkage and Selection Operator; ROC, Receiver operating characteristic; AUC, Area Under Curve; ALT, alanine transaminase; AFP, alpha fetoprotein; TNM, tumor node metastasis; BMI, body mass index; C-index, concordance index; DCA, decision curve analysis; KEGG, Kyoto Encyclopedia of Genes and Genomes; GSEA, Gene Set Enrichment Analysis; FDR, false discovery rate; PCR, Polymerase Chain Reaction; PVTT, portal vein tumor thrombus.

Acknowledgements

No applicable.

Author contributions

Zhigang Wang and LeYu Pan are considered as co-first authors. Yang Gu, Yuxuan Pan, Zhigang Wang and Leyu Pan designed and analyzed the data. Deliang Guo conducted quantitative real-time PCR. Xiaofeng Luo, Jie Tang and Weihua Yang wrote the manuscript. All authors reviewed the manuscript.

Funding

No.

Availability of data and materials

The GSE14520 dataset was downloaded from GEO database in NCBI. The validation dataset TCGA-LIHC (mRNAs expression and corresponding clinical information) was downloaded from UCSC Xena online website.

Ethics approval and consent to participate

The studies involving human participants were reviewed and approved by the Protection of Human Subjects Committee of The First People`s Hospital of Jingmen. All patients provided their written informed consent to participate in this study.

Consent for publication

Not applicable.

Competing interests

The authors declare there are no conflict of interests.

Reference

1. Mortality GBD, Causes of Death C. Global, regional, and national life expectancy, all-cause mortality, and cause-specific mortality for 249 causes of death, 1980-2015: a systematic analysis for the Global Burden of Disease Study 2015. Lancet (London, England) 2016; 388: 1459-1544.
2. Forner A, Reig M, Bruix J. Hepatocellular carcinoma. Lancet 2018; 391: 1301-1314.
3. Bray F, Ferlay J, Soerjomataram I et al. Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. Ca-a Cancer Journal for Clinicians 2018; 68: 394-424.
4. El-Serag HB, Rudolph L. Hepatocellular carcinoma: Epidemiology and molecular carcinogenesis. Gastroenterology 2007; 132: 2557-2576.
5. Chapman WC, Klintmalm G, Hemming A et al. Surgical Treatment of Hepatocellular Carcinoma in North America: Can Hepatic Resection Still Be Justified? Journal of the American College of Surgeons 2015; 220: 628-637.
6. Gluer AM, Cocco N, Laurence JM et al. Systematic review of actual 10-year survival following resection for hepatocellular carcinoma. Hpb 2012; 14: 285-290.
7. Bruix J, Reig M, Sherman M. Evidence-Based Diagnosis, Staging, and Treatment of Patients With Hepatocellular Carcinoma. Gastroenterology 2016; 150: 835-853.
8. Liu GM, Zeng HD, Zhang CY, Xu JW. Identification of a six-gene signature predicting overall survival for hepatocellular carcinoma. Cancer Cell International 2019; 19: 13.
9. Zheng YJ, Liu YL, Zhao SF et al. Large-scale analysis reveals a novel risk score to predict overall survival in hepatocellular carcinoma. Cancer Management and Research 2018; 10: 6079-6096.
10. Wang S, Zhang JH, Wang H et al. A novel multidimensional signature predicts prognosis in hepatocellular carcinoma patients. Journal of Cellular Physiology 2019; 234: 11610-11619.
11. Zhang ZQ, Ouyang YL, Huang YY et al. Comprehensive bioinformatics analysis reveals potential lncRNA biomarkers for overall survival in patients with hepatocellular carcinoma: an on-line individual risk calculator based on TCGA cohort. Cancer Cell International 2019; 19.

12. Edgar R, Domrachev M, Lash AE. Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. Nucleic Acids Research 2002; 30: 207-210.
13. Ritchie ME, Phipson B, Wu D et al. limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Research 2015; 43: 13.
14. Tibshirani R. The lasso method for variable selection in the cox model. Statistics in Medicine 1997; 16: 385-395.
15. Heagerty PJ, Lumley T, Pepe MS. Time-dependent ROC curves for censored survival data and a diagnostic marker. Biometrics 2000; 56: 337-344.
16. Zlotnik A, Abraira V. A general-purpose nomogram generator for predictive logistic regression models. Stata Journal 2015; 15: 537-546.
17. Iasonos A, Schrag D, Raj GV, Panageas KS. How to build and interpret a nomogram for cancer prognosis. Journal of Clinical Oncology 2008; 26: 1364-1370.
18. Vickers AJ, Elkin EB. Decision curve analysis: A novel method for evaluating prediction models. Medical Decision Making 2006; 26: 565-574.
19. Subramanian A, Tamayo P, Mootha VK et al. Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. Proceedings of the National Academy of Sciences of the United States of America 2005; 102: 15545-15550.
20. Siegel RL, Miller KD, Jemal A. Cancer Statistics, 2017. Ca-a Cancer Journal for Clinicians 2017; 67: 7-30.
21. Sauvageau E, Rochdi MD, Oueslati M et al. CNIH4 Interacts with Newly Synthesized GPCR and Controls Their Export from the Endoplasmic Reticulum. Traffic 2014; 15: 383-400.
22. Mishra S, Bernal C, Silvano M et al. The protein secretion modulator TMED9 drives CNIH4/TGF alpha/GLI signaling opposing TMED3-WNT-TCF to promote colon cancer metastases. Oncogene 2019; 38: 5817-5837.
23. Cheung M, Abu-Elmagd M, Clevers H, Scotting PJ. Roles of Sox4 in central nervous system development. Molecular Brain Research 2000; 79: 180-191.
24. Liao YL, Sun YM, Chau GY et al. Identification of SOX4 target genes using phylogenetic footprinting-based prediction from expression microarrays suggests that overexpression of SOX4 potentiates metastasis in hepatocellular carcinoma. Oncogene 2008; 27: 5578-5589.
25. Hur W, Rhim H, Jung CK et al. SOX4 overexpression regulates the p53-mediated apoptosis in hepatocellular carcinoma: clinical implication and functional analysis in vitro. Carcinogenesis 2010; 31: 1298-1307.
26. Oldberg A, Franzen A, Heinegard D. CLONING AND SEQUENCE-ANALYSIS OF RAT BONE SIALOPROTEIN (OSTEOPONTIN) CDNA REVEALS AN ARG-GLY-ASP CELL-BINDING SEQUENCE. Proceedings of the National Academy of Sciences of the United States of America 1986; 83: 8819-8823.
27. Shin HD, Park BL, Cheong HS et al. SPP1 polymorphisms associated with HBV clearance and HCC occurrence. International Journal of Epidemiology 2007; 36: 1001-1008.
28. Zhao LJ, Wang W, Huang S et al. The RNA binding protein SORBS2 suppresses metastatic colonization of ovarian cancer by stabilizing tumor-suppressive immunomodulatory transcripts. Genome Biology 2018; 19: 20.
29. Han LL, Huang C, Zhang SQ. The RNA-binding protein SORBS2 suppresses hepatocellular carcinoma tumourigenesis and metastasis by stabilizing RORA mRNA. Liver International 2019; 39: 2190-2203.

30. Yan B, Peng ZY, Xing CG. SORBS2, mediated by MEF2D, suppresses the metastasis of human hepatocellular carcinoma by inhibitiing the c-Abl-ERK signaling pathway. American Journal of Cancer Research 2019; 9: 2706-2718.
31. Shi JY, Yang LX, Wang ZC et al. CC chemokine receptor-like 1 functions as a tumour suppressor by impairing CCR7-related chemotaxis in hepatocellular carcinoma. Journal of Pathology 2015; 235: 546-558.

Figure legend:

Figure1. Flow chart of study design.

Figure2. Kaplan-Meier survival analysis, risk score analysis and time-dependent receiver operating characteristic (ROC) analysis in training and testing sets for the five-gene signature in hepatocellular carcinoma (HCC). The Kaplan-Meier plot, five mRNAs heatmap, cut-off value, survival states of patients and time-dependent ROC analysis in (A) training and (B) testing sets of GSE14520.

Figure3. Kaplan-Meier survival analysis, risk score analysis and time-dependent ROC analysis in whole and The Cancer Genome Atlas Liver Hepatocellular Carcinoma (TCGA-LIHC) sets for the five-gene signature in HCC. The Kaplan-Meier plot, five mRNAs heatmap, cut-off value, survival states of patients and time-dependent ROC analysis in (A) whole and (B) TCGA-LIHC sets.

Figure4. Forest plot of the univariate and multivariate Cox regression analysis in two sets. P value < 0.05 was considered significant. ALT: alanine transaminase; AFP: alpha fetoprotein; BMI: body mass index; TNM: tumor node metastasis.

Fig5. The Kaplan-Meier plots of the five-gene signature in different TNM stage of HCC patients. Patients in high risk group showed poorer overall survival compared with those in low risk group in (A, C) TNM stage I+II and (B, D) TNM stage III+IV in GSE14520 and TCGA-LIHC sets.

Fig6. A predictive nomogram building and validating in GSE14520 set. (A) The nomogram was built by two independent prognostic factors. (B) The decision curve analysis (DCA) of prognostic, TNM stage and combined models for 5-year overall survival. (C) The calibration plots for internal validation of the nomogram for 1-year, 3-year and 5-year survival, respectively. (D) The time-dependent ROC curves of the

nomograms compared for 1-year ,3-year and 5-year overall survival, respectively.

Fig7. A predictive nomogram building and validating in TCGA-LIHC set. (A) The nomogram was built based on two independent prognostic factors. (B) The decision curve analysis (DCA) of prognostic, TNM stage and combined models for 5-year overall survival. (C) The calibration plots for internal validation of the nomogram for 1-year, 3-year and 5-year survival, respectively. (D) The time-dependent ROC curves of the nomograms compared for 1-year ,3-year and 5-year overall survival, respectively.

Fig8. Mutation information and gene set enrichment analysis (GSEA). (A)the mutation information of five prognostic genes in cBioPortal online website. (B) seven obvious Kyoto Encyclopedia of Genes and Genomes (KEGG) pathways enriched in high and low risk group in GSE14520 set.

Figures

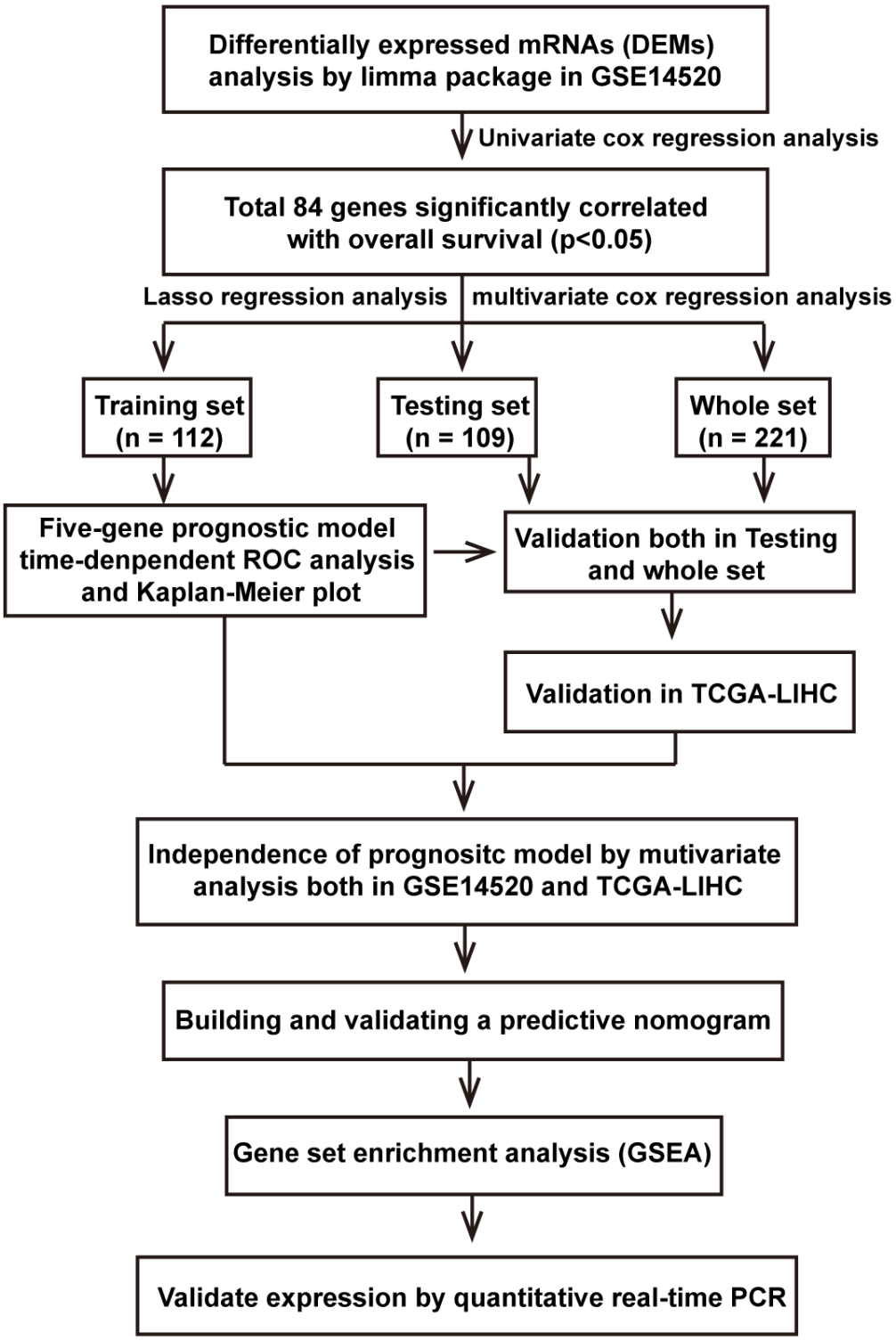

Figure 1

Flow chart of study design.

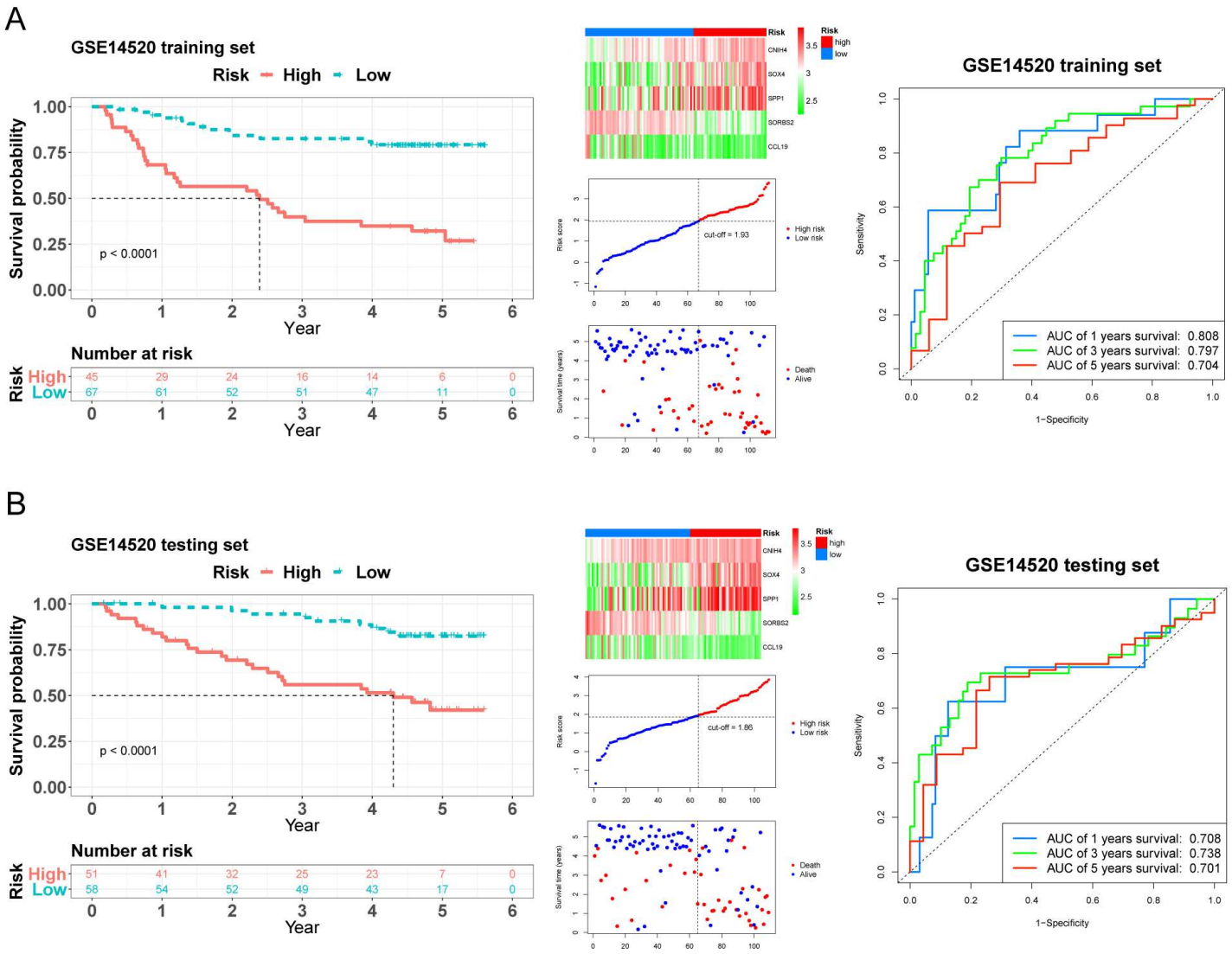

Figure 2

Kaplan-Meier survival analysis, risk score analysis and time-dependent receiver operating characteristic (ROC) analysis in training and testing sets for the five- gene signature in hepatocellular carcinoma (HCC). The Kaplan-Meier plot, five mRNAs heatmap, cut-off value, survival states of patients and time-dependent ROC analysis in (A) training and (B) testing sets of GSE14520.

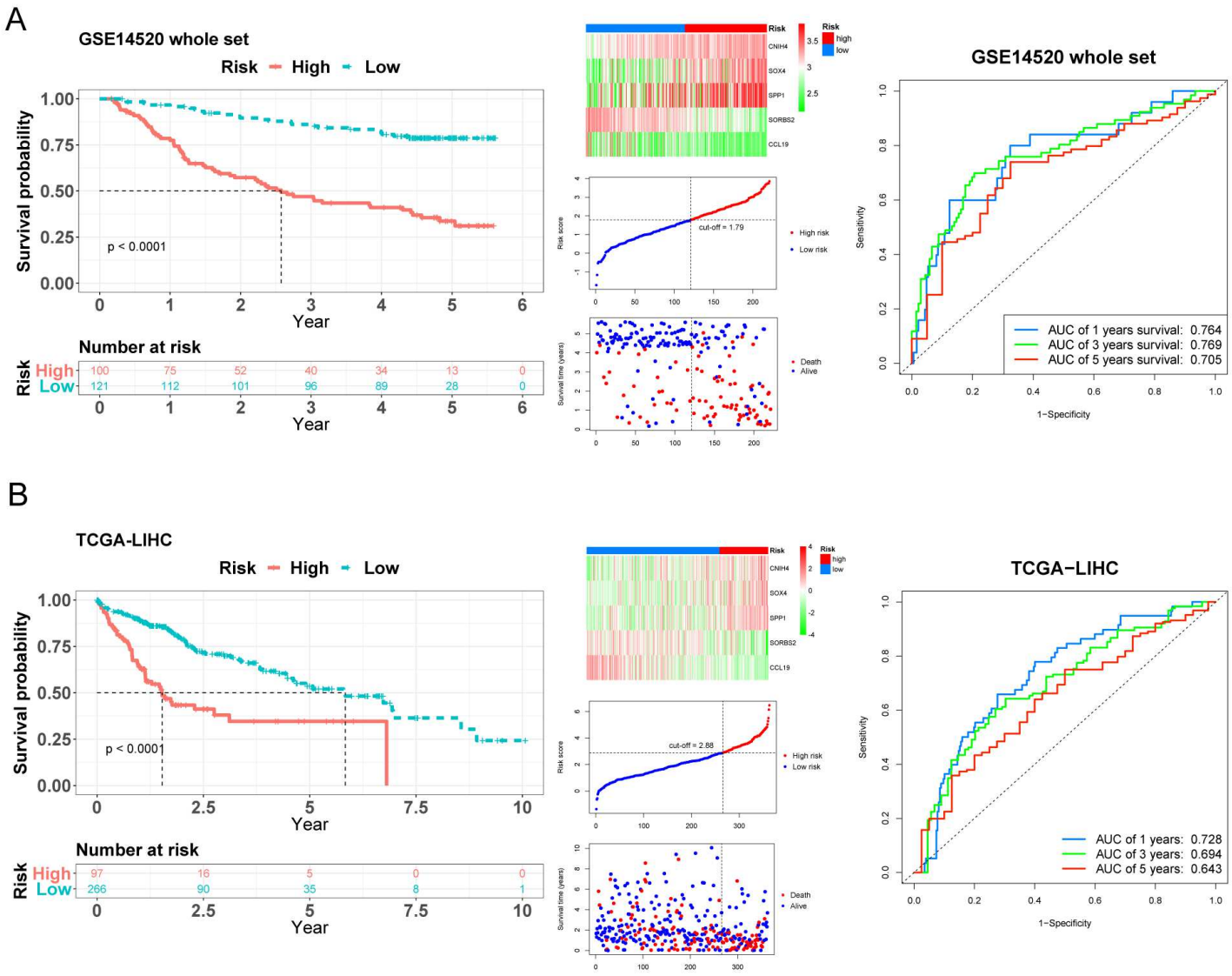

Figure 3

Kaplan-Meier survival analysis, risk score analysis and time-dependent ROC analysis in whole and The Cancer Genome Atlas Liver Hepatocellular Carcinoma (TCGA-LIHC) sets for the five-gene signature in HCC. The Kaplan-Meier plot, five mRNAs heatmap, cut-off value, survival states of patients and time-dependent ROC analysis in (A) whole and (B) TCGA-LIHC sets.

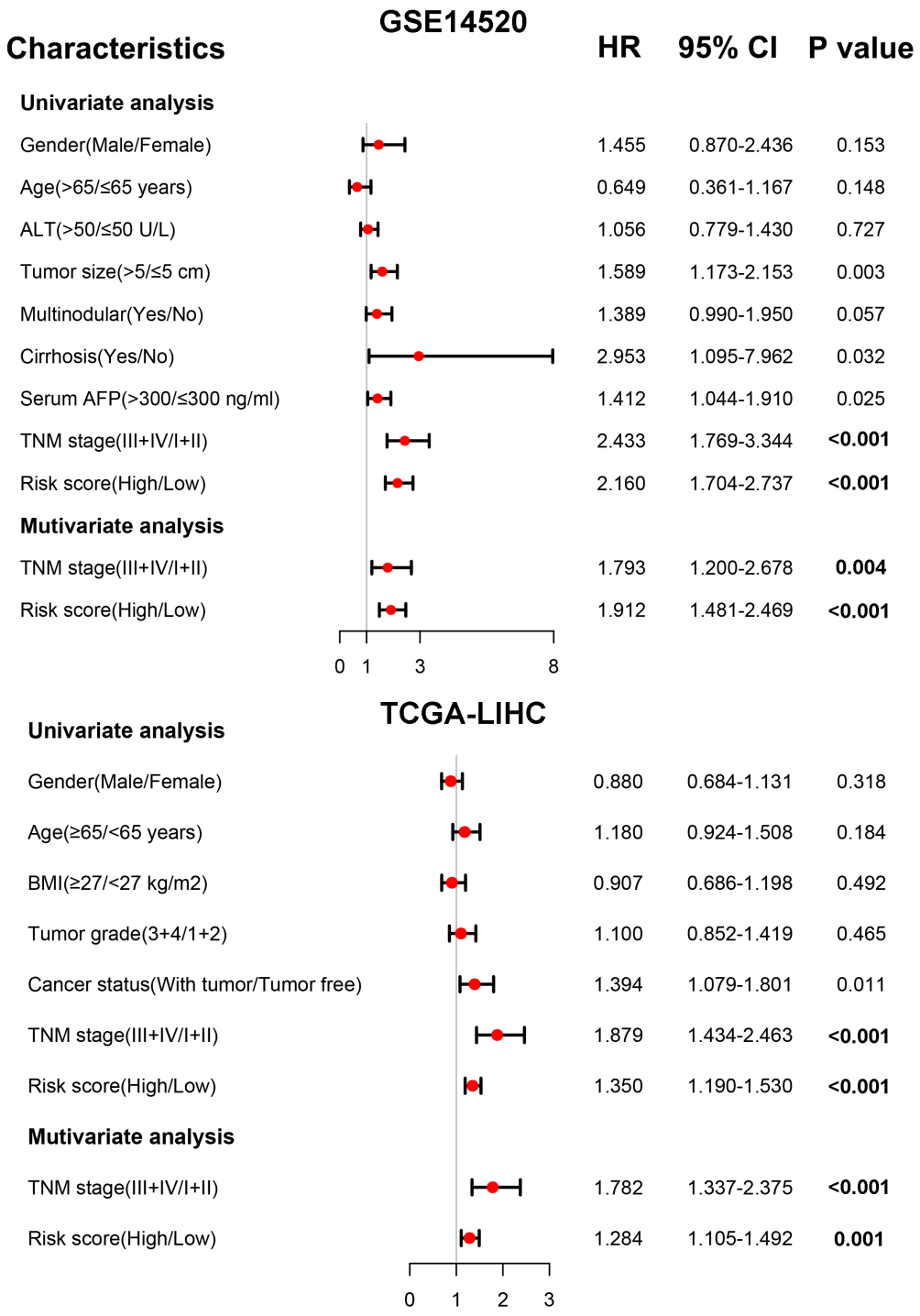

Figure 4

Forest plot of the univariate and multivariate Cox regression analysis in two sets. P value < 0.05 was considered significant. ALT: alanine transaminase; AFP: alpha fetoprotein; BMI: body mass index; TNM: tumor node metastasis.

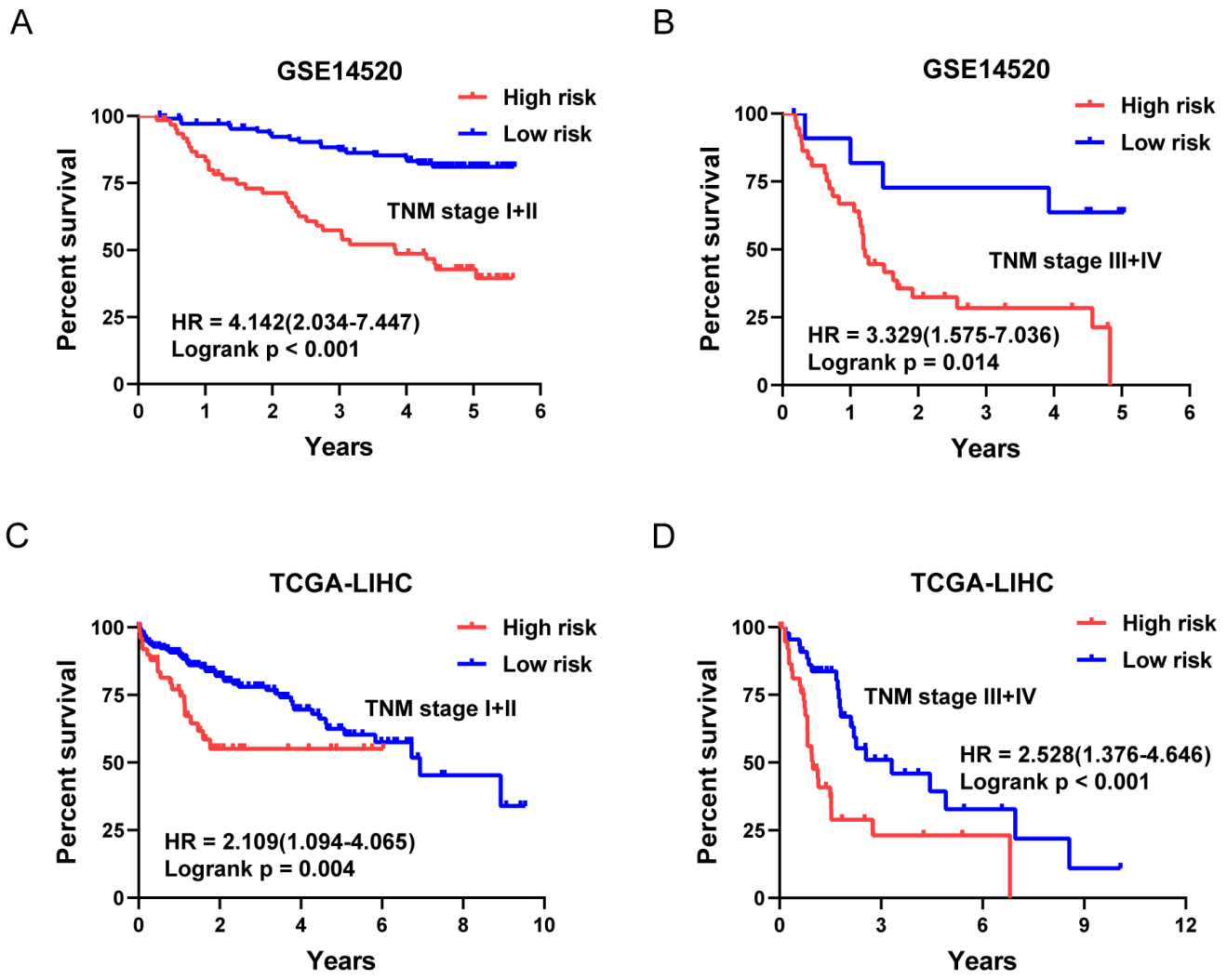

Figure 5

The Kaplan-Meier plots of the five-gene signature in different TNM stage of HCC patients. Patients in high risk group showed poorer overall survival compared with those in low risk group in (A, C) TNM stage I+II and (B, D) TNM stage III+IV in GSE14520 and TCGA-LIHC sets.

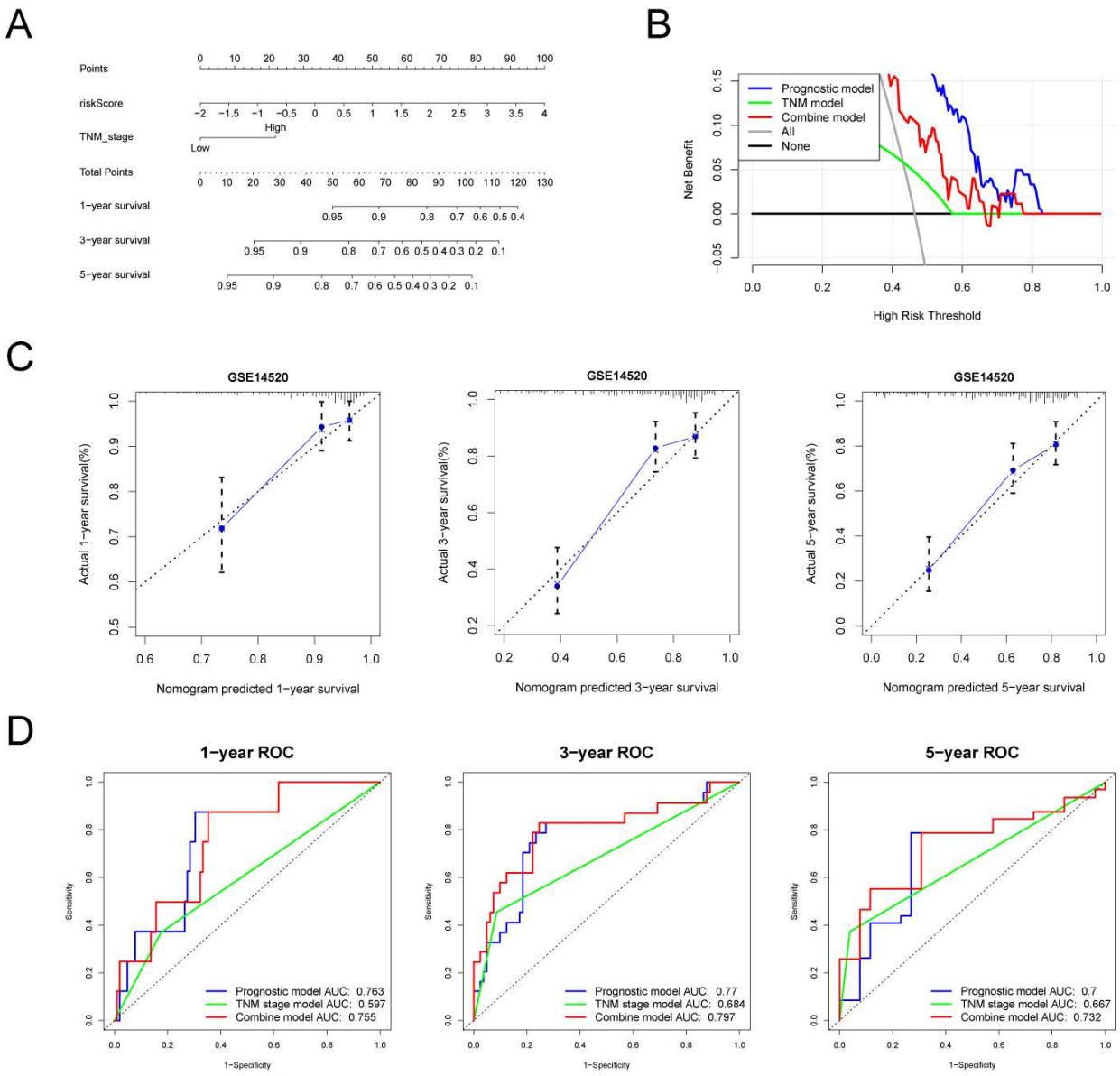

Figure 6

A predictive nomogram building and validating in GSE14520 set. (A) The nomogram was built by two independent prognostic factors. (B) The decision curve analysis (DCA) of prognostic, TNM stage and combined models for 5-year overall survival. (C) The calibration plots for internal validation of the nomogram for 1-year, 3-year and 5-year survival, respectively. (D) The time□dependent ROC curves of the nomograms compared for 1-year ,3-year and 5-year overall survival, respectively.

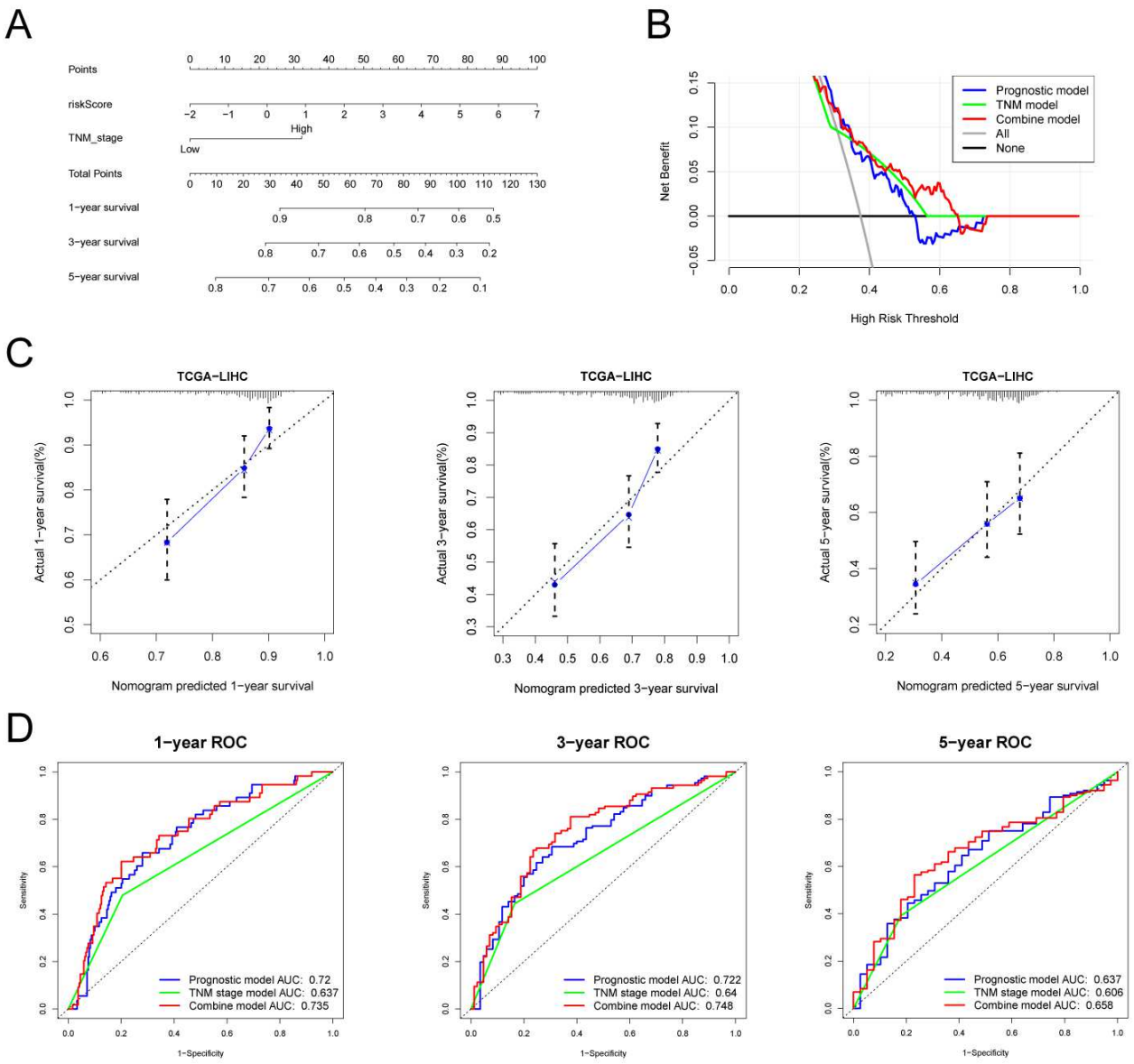

Figure 7

A predictive nomogram building and validating in TCGA-LIHC set. (A) The nomogram was built based on two independent prognostic factors. (B) The decision curve analysis (DCA) of prognostic, TNM stage and combined models for 5-year overall survival. (C) The calibration plots for internal validation of the nomogram for 1-year, 3-year and 5-year survival, respectively. (D) The time□dependent ROC curves of the nomograms compared for 1-year ,3-year and 5-year overall survival, respectively.

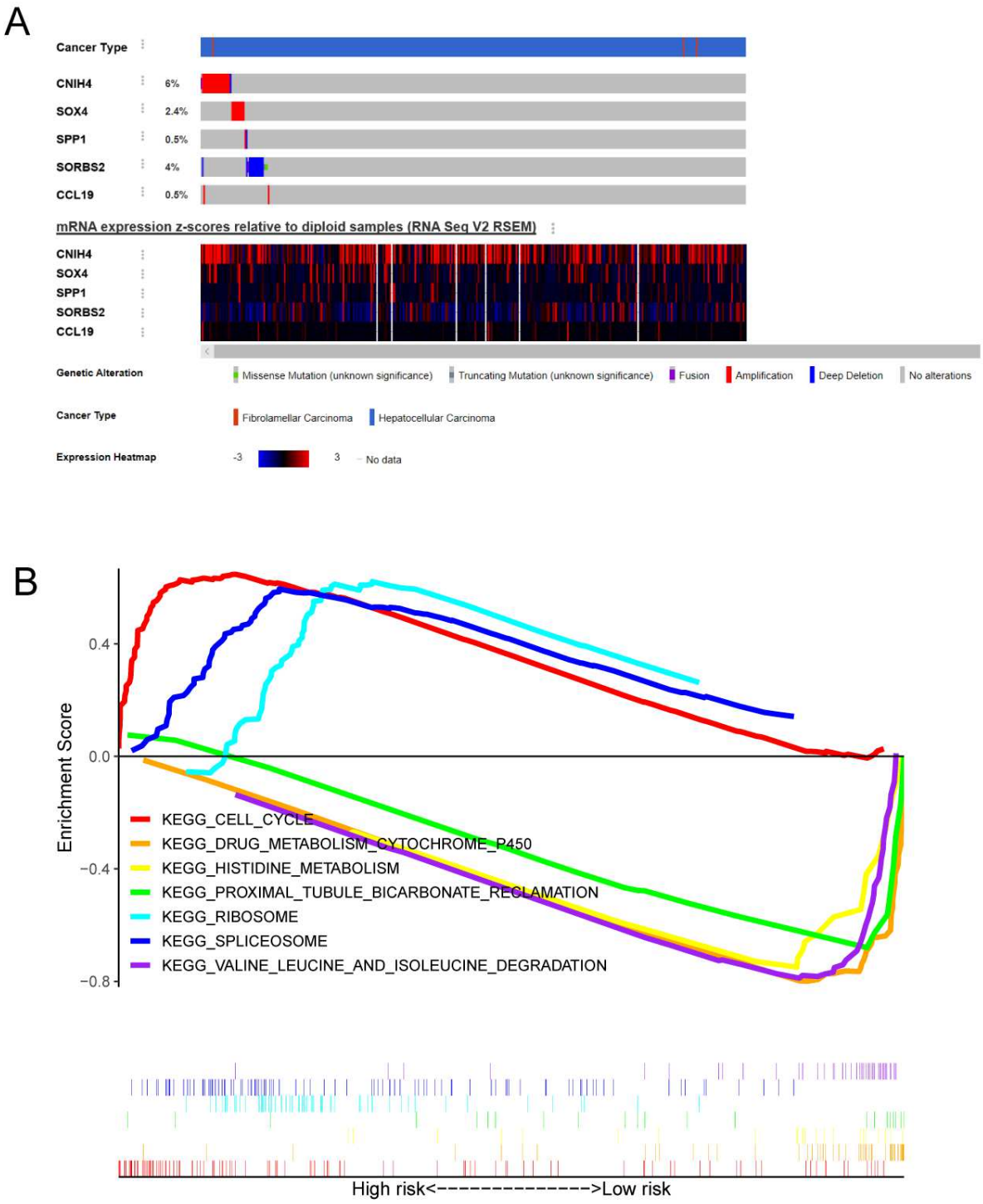

Figure 8

Mutation information and gene set enrichment analysis (GSEA). (A)the mutation information of five prognostic genes in cBioPortal online website. (B) seven obvious Kyoto Encyclopedia of Genes and Genomes (KEGG) pathways enriched in high and low risk group in GSE14520 set.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- Supplementaryfile2.txt
- Supplementaryfile3high.xlsx
- SuplementaryFigureLegend.docx
- SupplementaryTable1.docx
- SupplementaryFig1.tif
- Supplementaryfile3low.xlsx
- Supplementaryfile1test.txt
- Supplementaryfile1train.txt
- SupplementaryFig2.tif