Journal of Clinical and Translational Hepatology **2021** vol. 9(6) | 898–908
DOI: 10.14218/JCTH.2020.00152

Original Article

Development of an Eight-gene Prognostic Model for Overall Survival Prediction in Patients with Hepatocellular Carcinoma

De-Zhen Guo[1#], Ao Huang[1#], Yu-Peng Wang[1], Ya Cao[2], Jia Fan[1,3,4], Xin-Rong Yang[1*] and Jian Zhou[1,3,4*]

[1]Department of Liver Surgery and Transplantation, Liver Cancer Institute, Zhongshan Hospital, Fudan University; Key Laboratory of Carcinogenesis and Cancer Invasion (Fudan University), Ministry of Education; Shanghai Key Laboratory of Organ Transplantation, Zhongshan Hospital, Fudan University, Shanghai, China; [2]Cancer Research Institute, Xiangya School of Medicine, Central South University; Key Laboratory of Carcinogenesis and Cancer Invasion, Ministry of Education, Xiangya Hospital, Central South University, Changsha, Hunan, China; [3]Institute of Biomedical Sciences, Fudan University, Shanghai, China; [4]State Key Laboratory of Genetic Engineering, Fudan University, Shanghai, China

Received: 6 December 2020 | **Revised:** 27 March 2021 | **Accepted:** 11 April 2021 | **Published:** 14 May 2021

Abstract

Background and Aims: The overall survival (OS) of hepatocellular carcinoma (HCC) remains dismal. Bioinformatic analysis of transcriptome data could identify patients with poor OS and may facilitate clinical decision. This study aimed to develop a prognostic gene model for HCC. ***Methods:*** GSE14520 was retrieved as a training set to identify differential expressed genes (DEGs) between tumor and adjacent liver tissues in HCC patients with different OS. A DEG-based prognostic model was then constructed and the TCGA-LIHC and ICGC-LIRI datasets were used to validate the model. The area under the receiver operating characteristic curve (AUC) and hazard ratio (HR) of the model for OS were calculated. A model-based nomogram was established and verified. ***Results:*** In the training set, differential expression analysis identified 80 genes dysregulated in oxidation-reduction and metabolism regulation. After univariate Cox and LASSO regression, eight genes (LPCAT1, DHRS1, SORBS2, ALDH5A1, SULT1C2, SPP1, HEY1 and GOLM1) were selected to build the prognostic model. The AUC for 1-, 3- and 5-year OS were 0.779, 0.736, 0.754 in training set and 0.693, 0.689, 0.693 in the TCGA-LIHC validation set, respectively. The AUC for 1- and 3-year OS were 0.767 and 0.705 in the ICGC-LIRI validation set. Multivariate analysis confirmed the model was an independent prognostic factor (training set: HR=4.422, $p<0.001$; TCGA-LIHC validation set: HR=2.561, $p<0.001$; ICGC-LIRI validation set: HR=3.931, $p<0.001$). Furthermore, a nomogram combining the model and AJCC stage was established and validated, showing increased OS predictive efficacy compared with the prognostic model ($p=0.035$) or AJCC stage ($p<0.001$). ***Conclusions:*** Our eight-gene prognostic model and the related nomogram represent as reliable prognostic tools for OS prediction in HCC patients.

Citation of this article: Guo DZ, Huang A, Wang YP, Cao Y, Fan J, Yang XR, *et al*. Development of an eight-gene prognostic model for overall survival prediction in patients with hepatocellular carcinoma. J Clin Transl Hepatol 2021;9(6):898–908. doi: 10.14218/JCTH.2020.00152.

Introduction

Hepatocellular carcinoma (HCC) is the sixth most frequent malignancy and the second leading cause of cancer-related mortality worldwide.[1] Globally, it was estimated that there were more than 840,000 new cases of HCC and nearly 780,000 deaths per year.[2] Despite the great progress in early diagnosis and early treatment, the overall survival (OS) remains unfavorable and approximately 70% of HCC patients would have tumor relapse within 5 years after curative resection or ablation.[3] Classification of HCC to guide prognostic stratification, clinical management and improve OS is thus of importance. However, currently available classification systems, such as Barcelona Clinic Liver Cancer (BCLC) and American Joint Committee on Cancer (AJCC) staging systems, focus on pretreatment classification rather than prognostication.[4] Therefore, it's necessary to develop a novel model for the prognosis prediction of HCC.

The advance of genome-sequencing technologies has resulted in large-scale tumor genome profiling and transcriptome analysis. Of note, mRNA profiling of tumors could identify potential biomarkers and gene signatures at the mRNA level have great potential in prognosis prediction.

Keywords: Prognostic model; Hepatocellular carcinoma; Bioinformatic analysis.

Abbreviations: AFP, alpha-fetoprotein; AJCC, American Joint Committee on Cancer; ALDH5A1, aldehyde dehydrogenase 5 family member A1; AUC, area under the receiver operating characteristic curve; BCLC, Barcelona Clinic Liver Cancer; C-index, concordance index; CI, confidence interval; DEG, differential expressed gene; DHRS1, dehydrogenase/reductase member 1; GEO, Gene Expression Omnibus; GO, gene ontology; GOLM1, Golgi membrane protein 1; HEY 1, Hes-related family bHLH transcription factor with YRPW motif 1; HCC, hepatocellular carcinoma; HR, hazard ratio; ICGC-LIRI, International Cancer Genome Consortium-Liver Cancer-RIKEN; KEGG, Kyoto Encyclopedia of Genes and Genomes; LPCAT1, lysophosphatidylcholine acyltransferase 1; LASSO, least absolute shrinkage and selection operator; OS, overall survival; ROC, receiver operating characteristic; SULT1Cs, 1C family of human cytosolic sulfotransferases; SPP1, secreted phosphoprotein-1; TCGA-LIHC, The Cancer Genome Atlas-Liver Hepatocellular Carcinoma.

[#]These authors contributed equally to this work.

[*]**Correspondence to:** Jian Zhou and Xin-Rong Yang, Liver Cancer Institute, Zhongshan Hospital, Fudan University, 136 Yi Xue Yuan Road, Shanghai 200032, China. ORCID: https://orcid.org/0000-0002-2118-1117 (JZ), https://orcid.org/0000-0002-2716-9338 (XRY). Tel: +86-21-64041990, Fax: +86-21-64037181, E-mail: zhou.jian@zs-hospital.sh.cn (JZ) or yang.xinrong@zs-hospital.sh.cn (XRY)

Currently, bioinformatics analysis of mRNA expression from publicly accessible databases has established gene signatures to predict the OS of HCC patients.[5–11] However, these studies had all unexceptionally employed differential analysis by comparing the mRNA expression between tumor tissues and non-tumor tissue, without considering the potential predictive value of differential expression genes (DEGs) between patients with different prognosis.

In this present study, we, for the first time, used a novel method to develop a prognostic model. We performed differential expression analysis not only on the mRNA expression of tumor and adjacent liver tissues but also on the tumor tissues between HCC patients with different prognosis. Then, the overlapped DEGs were retrieved to investigate survival-related biomarkers in the training set. An eight-gene risk model was then established and validated to be an independent index for OS. Moreover, we went a step further to construct a nomogram which had combined this eight-gene model with AJCC cancer staging system. This study has taken DEGs between different tissues and different patients into consideration simultaneously and developed a reliable and robust prognostic model.

Methods

Data collection

The GSE14520 dataset[12] (doi:10.1158/0008-5472) of the Gene Expression Omnibus database (GEO), The Cancer Genome Atlas-Liver Hepatocellular Carcinoma dataset (TCGA-LIHC, doi:10.1038/ng.2764)[13] and The International Cancer Genome Consortium-Liver Cancer-RIKEN dataset (ICGC-LIRI, doi: 10.1038/nature08987)[14] were used as data source. The data of 225 HCC patients (containing 225 tumor and 220 adjacent normal liver tissues) in the GSE14520 dataset was used as the training set, while the 355 HCC cases with prognosis information in the TCGA-LIHC dataset (containing 355 tumor and 50 adjacent normal liver tissues) and 243 HCC cases with prognosis information in the ICGC-LIRI dataset (containing 243 tumor and 202 adjacent normal liver tissues) were used to verify the predictive performance of the risk model and nomogram as external validation sets. The messenger RNA (mRNA) expression profiles of tumor tissues and the clinicopathological information of HCC patients from each dataset were obtained accordingly. The research was conducted in accordance with the Declaration of Helsinki (as revised in 2013). Approval from the ethics committee and patient consents were exempted since all data were derived from public databases. We strictly complied with the GEO, TCGA and ICGC publication guidelines and data access policies, and presented this article in accordance with the MDAR reporting checklist. The baseline characteristics of three datasets are shown in Table 1.

Gene differential expression analysis and gene set enrichment analysis

Genes with a very low expression level of "0" were precluded from bioinformatical analysis. The DEGs in the training set were investigated using the "limma" package.[15] DEGs between tumor and adjacent liver tissues were investigated and genes with an adjusted p-value <0.05 and log2 fold-change (logFC) >1 or <−1 were considered as up- and down-regulated respectively. Differential analysis was also applied between patients with different prognosis and those with an adjusted p-value <0.05 and logFC >0.5 or <−0.5 were identified as DEGs. Two groups of DEGs were consequently combined and those overlapped were retrieved for subsequent analysis. Gene ontology (GO) enrichment analyses and Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway analysis were performed by DAVID and visualized using the R packages "GOplot", with the intention to reveal the possible biological functions and pathways which might affect the prognosis of HCC. Adjusted p-value <0.05 was considered statistically significant.

Construction and validation of the prognostic model

Univariate Cox regression was performed for DEGs to identify the prognosis-related genes and those with $p<0.001$ were applied to least absolute shrinkage and selection operator (LASSO) regression analysis. After that, a prognostic model was established based on the genes derived from LASSO regression[16] and the risk score was calculated using the formula:

$$(\beta_{mRNA1} * expression_{mRNA1}) + (\beta_{mRNA2} * expression_{mRNA2}) + \ldots\ldots + (\beta_{mRNAn} * expression_{mRNAn}).$$

In the formula, the value of β was the regression coefficient derived from LASSO regression and the expression meant expression level of mRNA. The time-dependent predictive value of the prognostic model was evaluated using time-dependent receiver operating characteristic (ROC) curve via the R package "timeROC". The "survminer" package was applied to identify an optimal cut-off value of risk score and according to which patients were divided into high-risk and low-risk groups. Then, Kaplan-Meier analysis combined with log-rank test was used to compare the prognostic difference between the high-risk and low-risk groups using the "survival" package. Univariate and multivariate Cox regression analysis was applied to identify independent prognostic factors.

To validate the prognostic performance of the gene model, 355 HCC patients from TCGA-LIHC and 243 HCC patients from ICGC-LIRI were analyzed as external validation sets. The time-dependent ROC analysis, Kaplan-Meier analysis, subgroup analysis and multivariate Cox regression analysis were performed in two validation sets identically with that in training set.

Development and validation of the prognostic model-based nomogram

The nomogram was built using all independent prognostic factors identified by multivariate Cox regression in the training set. The concordance index (C-index) and calibration curve was applied to determine the discrimination and calibration of the nomogram respectively (by a bootstrap method with 1,000 resamples). The AJCC stage model, prognostic signature and the nomogram model were compared via time-dependent ROC curve and C-index. The prognostic model-based nomogram was also validated using time-dependent ROC curve, C-index and calibration curve in two validation sets.

Statistical analysis

Statistical analyses were performed using R software version 3.6.3 (R Foundation for Statistical Computing, Vienna, Austria, RRID:SCR_001905). Pearson $\chi 2$ test or Fisher's exact test were applied to analyze qualitative variables as appropriate. If not specified above, $p<0.05$ was considered

Table 1. Baseline characteristics of patients in the training and two validation datasets

	Level	Training dataset	Validation datasets	
		GSE14520, *n*=203	TCGA-LIHC, *n*=355	ICGC-LIRI, *n*=243
Sex	Male	175 (86.2)	240 (67.6)	182 (74.9)
	Female	28 (13.8)	115 (32.4)	61 (25.1)
Age in years	≤50	100 (49.3)	73 (20.6)	17 (7.0)
	>50	103 (50.7)	282 (79.4)	226 (93.0)
HBV infection	No	6 (3.0)	22 (6.2)	/
	Yes	195 (96.0)	134 (37.7)	/
	Unknown	2 (1.0)	199 (56.1)	/
HCV infection	No	/	57 (16.1)	/
	Yes	/	99 (27.9)	/
	Unknown	/	199 (56.1)	/
Alcohol consumption	No	/	232 (65.4)	/
	Yes	/	113 (31.8)	/
	Unknown	/	10 (2.8)	/
Cirrhosis	No	16 (7.9)	/	/
	Yes	187 (92.1)	/	/
Child-Pugh stage	A	/	211 (59.4)	/
	B/C	/	22 (6.2)	/
	Unknown	/	122 (34.4)	/
AFP in ng/mL	≤300	108 (53.2)	206 (58.0)	/
	>300	92 (45.3)	63 (17.7)	/
	Unknown	3 (1.5)	86 (24.2)	/
Tumor size in cm	≤5	134 (66.0)	/	/
	>5	69 (34.0)	/	/
Tumor number	Solitary	163 (80.3)	/	/
	Multiple	40 (19.7)	/	/
Edmondson grade	I/II	/	227 (63.9)	158 (65.0)
	III/IV	/	128 (36.1)	65 (26.7)
	Unknown	/	0	20 (8.2)
Vascular invasion	No	/	199 (56.1)	/
	Yes	/	102 (28.7)	/
	Unknown	/	54 (15.2)	/
AJCC stage	I/II	161 (79.3)	263 (74.1)	146 (60.1)
	III/IV	42 (20.7)	92 (25.9)	97 (39.9)
BCLC stage	A	157 (77.3)	/	/
	B/C	46 (22.7)	/	/

HBV, hepatitis B virus; HCV, hepatitis C virus.

statistically significant.

Ethics

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. The study was conducted in accordance with the Declaration of Helsinki (as revised in 2013). An approval by the ethics committee and patient consent were not required since all data were derived from public database. We strictly complied with the publication guidelines and data access policies of the TCGA, GEO and ICGC databases.

Results

DEG investigation and gene set enrichment analysis

The work flow of this study is illustrated in Supplementary Figure 1. First, mRNA expression profiles of tumor and non-tumor specimens were compared and 902 DEGs were identified, including 332 un-regulated DEGs and 570 down-regulated DEGs. Then, patients in GSE14520 dataset were classified into two groups according to OS: 136 patients with OS longer than three years were in the long-term survival group, while the other 67 patients died within three years were in the short-term survival group. mRNA expression profiles of tumor specimens were compared between the two groups and 126 DEGs were identified, including 47 un-regulated DEGs and 79 down-regulated DEGs. These DEGs were further overlapped with the DEGs between tumor and non-tumor specimens, and 80 DEGs were finally screened, including 19 up-regulated and 61 down-regulated genes respectively (Fig. 1A–C).

To elucidate the potential mechanism underlying the overlapping DEGs and prognosis, GO enrichment analysis was performed respectively through the online DAVID tool. In terms of biological processes, the overlapping DEGs were significantly enriched in oxidation-reduction process, steroid metabolic process and metabolic process (Fig. 1D). Enrichment analyses of cellular compartment and molecular functions are also shown in Figure 1D. Furthermore, we also applied KEGG pathway analysis and identified that these DEGs were mainly enriched in metabolic pathways, retinol metabolism and drug metabolism - cytochrome P450, which was concordant with the GO enrichment analysis (Fig. 1E).

Construction of the prognostic model

Univariate Cox regression was first performed for the 80 DEGs to identify genes of significant correlation with OS, and 49 genes with $p<0.001$ were selected into LASSO regression for further shrinkage. Upon the partial likelihood deviance reaching minimum in the LASSO regression, eight genes (LPCAT1, DHRS1, SORBS2, ALDH5A1, SULT1C2, SPP1, HEY1 and GOLM1) were identified and selected to construct the prognostic model (Supplementary Fig. 2). The formula for calculating the risk score was:

$$0.0326*\text{expression}_{LPCAT1} - 0.0483*\text{expression}_{DHRS1} - 0.1464*\text{expression}_{SORBS2} - 0.0005*\text{expression}_{ALDH5A1} + 0.0043*\text{expression}_{SULT1C2} + 0.0064*\text{expression}_{SPP1} + 0.0403*\text{expression}_{HEY1} + 0.0407*\text{expression}_{GOLM1}.$$

The optimal cut-off value for the risk score of the prognostic model was −0.6 and patients were classified into low- and high-risk groups accordingly (Fig. 2A, B). The Kaplan-Meier analysis demonstrated that high-risk group had a significant poorer OS compared with low-risk group (hazard ratio [HR]: 5.445, 95% confidence interval [CI]: 3.410–8.694, $P<0.0001$; Fig. 2A). Furthermore, we assessed the prognostic efficiency of the eight-gene model by operating a ROC curve and the AUCs for 1-, 3- and 5-year OS were 0.779, 0.736, 0.754, respectively (Fig. 2C). In univariate and multivariate Cox regression analysis, the AJCC stage and our model were both independent prognostic factors (Table 2).

Validation of the prognostic model

The predictive performance of the eight-gene prognostic model was then verified in the TCGA-LIHC validation set. Patients were grouped as low- and high-risk as well (Fig. 3A,B), at an optimal cut-off of −1.02. The OS was significantly shorter in the high-risk group than in the low-risk group (HR: 2.666, 95% CI: 1.862–3.818, $p<0.0001$; Fig. 3A). The AUCs for 1-, 3- and 5-year OS were 0.693, 0.689 and 0.693, respectively (Fig. 3C). In the Cox regression analysis, the eight-gene prognostic model was an independent factor for OS (Table 2). Besides, to assess the repeatability and reliability of the model, we further evaluated its performance in the ICGC-LIRI validation set. The optimal cut-off was -1.32 and the high-risk group showed significantly shorter OS than the low-risk group (HR: 5.889, 95% CI: 2.874–12.060, $p<0.0001$; Fig. 4A, B). The AUCs for 1- and 3-year OS were 0.767 and 0.705 (Fig. 4C). Considering the follow-up time of most patients in ICGC-LIRI did not reach 5 years, the AUC for 5-year OS was not assessed. Consistently, the model was also an independent factor for OS in ICGC-LIRI dataset (Table 2).

Additionally, we compared this prognostic model with those previously reported ones and found it had displayed comparable, or even better in certain condition, AUCs for OS prediction. Most importantly, this prognostic model demonstrated better reliability since its performance was satisfactory and consistent in two external validation sets (Table 3).

The prognostic gene model-related clinicopathological features

To further clarify the association between this model and prognosis, we applied correlation analysis between survival risk and clinicopathological features in the training and two validation sets. We found patients with high risk score generally had advanced tumor phenotype; in the training set, patients in the high-risk group were associated with higher alpha-fetoprotein (AFP) level, larger tumor size and advanced BCLC and AJCC stage (Supplementary Fig. 3A). Similarly, in the TCGA-LIHC validation set, patients with high-risk score were identified to be significantly associated with higher AFP level, advanced Edmondson grade, vascular invasion and advanced AJCC stage (Supplementary Fig. 3B). In the ICGC-LIRI validation set, high risk score was also correlated with advanced Edmondson grade and AJCC stage (Supplementary Fig. 3C).

Moreover, we found the eight-gene model could well differentiate patients into different prognostic groups for both AJCC stage I/II and AJCC stage III/IV patients, indicating that the signature could even distinguish the ones with poor survival among the early-stage patients (Supplementary Fig. 4A–F).

Development and validation of the prognostic model-based nomogram

Since the eight-gene risk score prognostic model and AJCC stage were found to be independent prognostic factors in the training set, we then tried to build a prognostic nomogram model combining the eight-gene risk score and AJCC stage (Fig. 5A). The calibration curves for 1-, 3- and 5-year OS are illustrated in Fig. 5B. The C-index was 0.740, 0.713 and 0.626 for the nomogram model, eight-gene risk score and AJCC stage, respectively. Correspondingly, the nomogram model demonstrated the largest AUC for OS compared with the prognostic model (3-year AUC: 0.774 vs. 0.736, $p=0.035$) and AJCC stage in training set (0.774 vs. 0.658; $p<0.001$) (Fig. 5E).

The nomogram was further tested in the TCGA-LIHC and

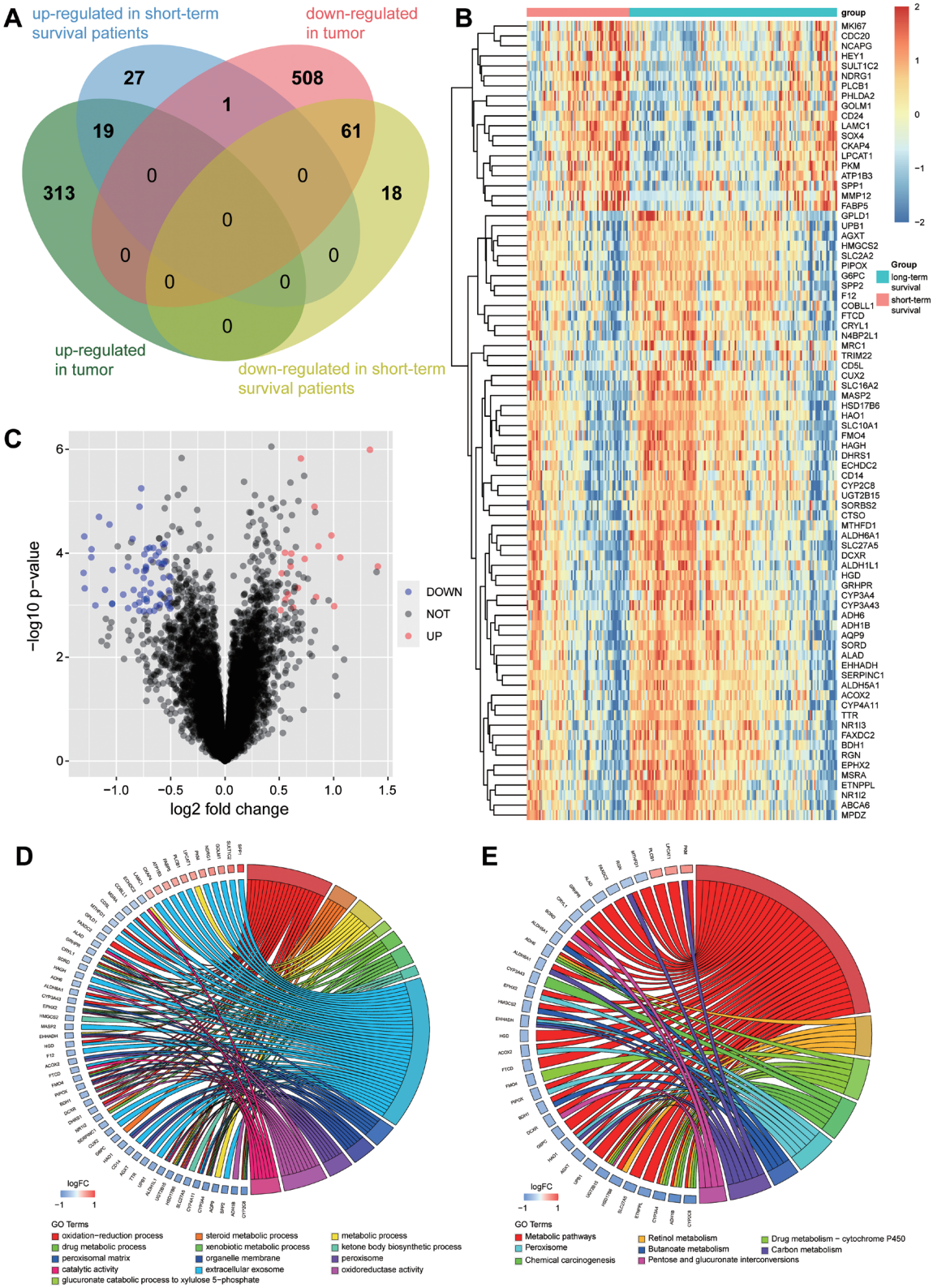

Fig. 1. Differential expression analysis and enrichment analysis identified genes associated with prognosis. (A) Venn plot of DEGs between tumor and adjacent liver tissues and between long-term (survived more than 3 years) and short-term (died within 3 years) survival patients. (B) Heatmap of overlapped genes in two group of DEGs. (C) Volcano plot showing the overlapped DEGs. (D, E) GO analysis (D) and KEGG pathway analysis (E) revealed the most significantly enriched biological functions or pathways of overlapped DEGs.

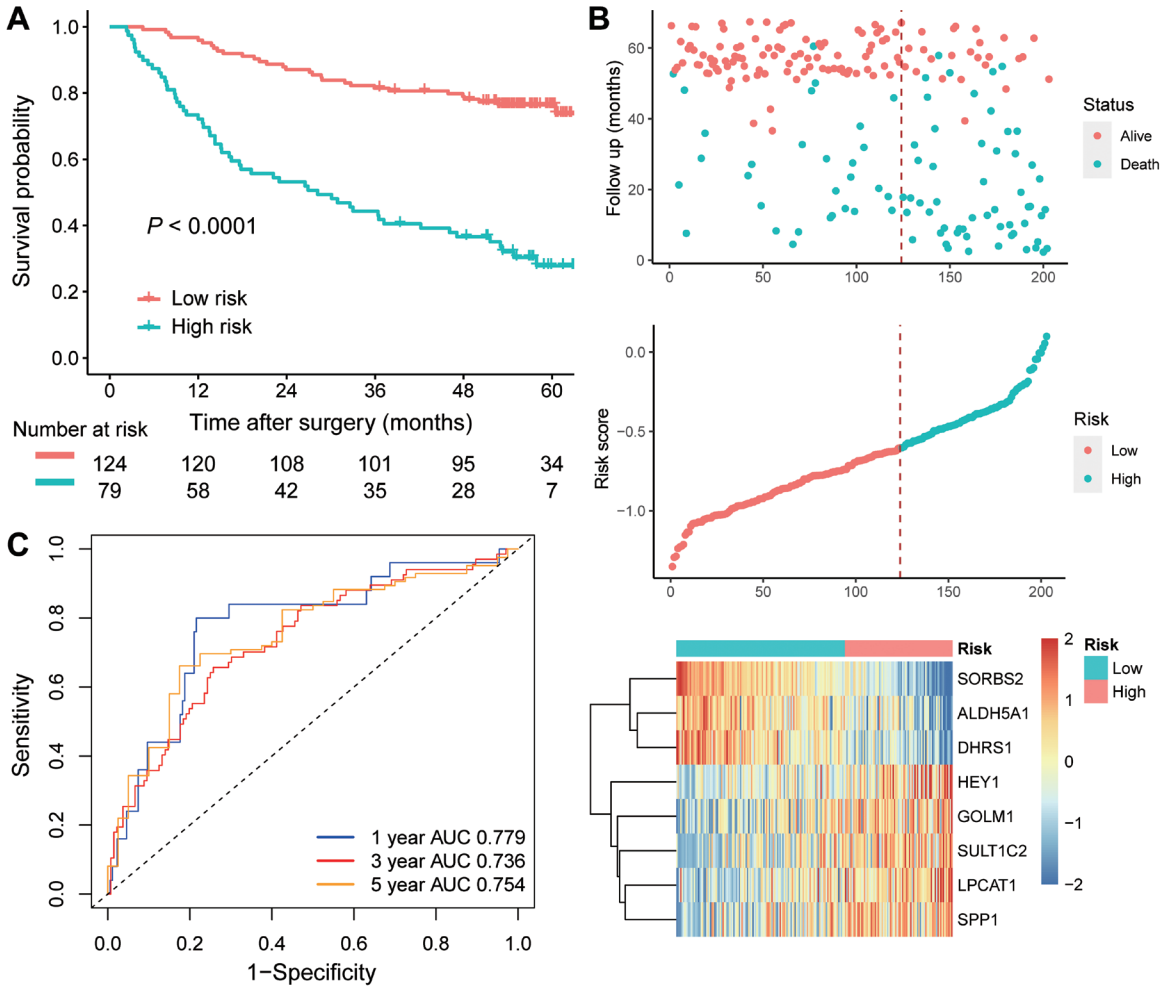

Fig. 2. Kaplan-Meier curve, risk score analysis and time-dependent ROC analysis for the eight-gene model in the GSE14520 dataset. (A) Kaplan-Meier curve for the eight-gene model in the GSE14520 dataset. (B) Risk score distribution and heatmap of the eight genes in model in the GSE14520 dataset. (C) Time-dependent ROC analysis of the eight-gene model for 1-, 3- and 5-year OS in the GSE14520 dataset.

ICGC-LIRI validation sets. The calibration curves for 1-, 3- and 5-year OS are illustrated in Figure 5C and 5D. In the TCGA-LIHC dataset, the C-index was 0.677, 0.654 and 0.598 for the nomogram model, signature and AJCC stage, respectively; meanwhile, the C-index in the ICGC-LIRI dataset was 0.723, 0.688 and 0.645, respectively. The AUCs of the nomogram were also the largest compared with the prognostic model (3-year AUC: 0.727 vs. 0.689, p=0.112) and AJCC stage (0.727 vs. 0.640, p=0.005) in the TCGA-LIHC dataset (Fig. 5F). Consistently, in the ICGC-LIRI dataset, the nomogram showed the best predictive performance compared with the prognostic model (0.715 vs. 0.705, p=0.821) and AJCC stage (0.715 vs. 0.602, p=0.008) (Fig. 5G). Taken together, the nomogram which was built from the eight-gene risk score prognostic model and AJCC stage showed improved sensitivity and specificity for prognosis prediction compared with the eight-gene prognostic model or the AJCC stage alone.

Availability of data and materials

The data we used in this study are available in the TCGA repository (http://cancergenome.nih.gov/), GEO (https://www.ncbi.nlm.nih.gov/) and ICGC Data Portal (https://dcc.icgc.org/).

Discussion

Despite the progression of early diagnosis and the advent of novel treatment modalities, the OS of HCC patients still remains poor and much effort has been made to generate a prognosis predicting model to identify patients with high risk of poor survival. Recently, with the development of genomic sequencing technology, the aberrant mRNA expression-based gene signature has attracted much attention and showed potential in prognostication.[5–11] Although several gene models had been established, they were developed unequivocally using the DEGs between tumor and non-tumor tissues, which might overlook the predictive significance of the DEGs between patients with different prognoses. Thus, a novel prognostication model using both the mRNA profiling of tumor and adjacent liver tissues and tumor tissues between HCC patients with different prognosis might be more comprehensive.

In the present study, we have adopted a strategy different from conventional bioinformatics analysis process;

Table 2. Univariate and multivariate Cox regression analysis for overall survival in the GSE14520 dataset, TCGA-LIHC dataset and ICGC-LIHC dataset

	Univariate analysis		Multivariate analysis	
	HR (95% CI)	***p***	**HR (95% CI)**	***p***
GSE14520				
Sex (male: female)	0.571 (0.276–1.182)	0.131	/	/
Age in years (≤50: >50)	1.009 (0.660–1.545)	0.966	/	/
Tumor size in cm (≤5: >5)	2.073 (1.350–3.184)	0.001	1.068 (0.642–1.778)	0.801
Tumor number (solitary: multiple)	1.634 (1.012–2.638)	0.045	0.789 (0.443–1.406)	0.421
Cirrhosis (no: yes)	4.589 (1.128–18.664)	0.033	2.743 (0.665–11.314)	0.163
AJCC stage (I/II: III/IV)	3.849 (2.460–6.024)	<0.001	2.936 (1.587–5.432)	0.001
Risk score (low: high)	4.422 (2.826–6.919)	<0.001	3.429 (2.157–5.452)	<0.001
TCGA-LIHC				
Sex (male: female)	0.823 (0.576–1.175)	0.284	/	/
Age in years (≤50: >50)	1.086 (0.705–1.673)	0.708	/	/
Edmondson grade (I/II: III/IV)	1.056 (0.737–1.513)	0.765	/	/
AJCC stage (I/II: III/IV)	2.477 (1.743–3.521)	<0.001	2.14 (1.497–3.060)	<0.001
Risk score (low/ high)	2.561 (1.793–3.656)	<0.001	2.277 (1.585–3.272)	<0.001
ICGC-LIRI				
Sex (male: female)	2.013 (1.084–3.738)	0.027	2.21 (1.155–4.227)	0.017
Age in years (≤50: >50)	3.141 (0.432–22.817)	0.258	/	/
AJCC stage (I/II: III/IV)	2.119 (1.169–3.842)	0.013	2.059 (1.095–3.869)	0.025
Risk score (low/ high)	3.931 (2.167–7.132)	<0.001	3.27 (1.777–6.017)	<0.001

instead of restriction to the DEGs between tumor and adjacent liver tissues, we integrated the differences of mRNA expression between tumor and adjacent liver tissues and between the tumors of HCC patients with long and short OS. The DEGs were first mined in the training set of GSE14520 and narrowed down, using univariate Cox regression analysis and LASSO regression, until eight genes were identified to construct the prognostic model. This eight-gene prognostic model worked well in two validation sets, including TCGA-LIHC and ICGC-LIRI datasets, and multivariate Cox regression analysis demonstrated that this model was an independent prognostic factor superior to clinicopathological factors. Moreover, subgroup analysis found that in early-stage patients, who were generally eligible for curative therapy and thus might obtain a better survival, this eight-gene prognostic model could also identify patients at high risk for poorer survival. Thus, the prognostic model in this study has important clinical significance.

Most of the genes in our model had been reported to be involved in the development and progression of HCC. Lysophosphatidylcholine acyltransferase 1 (LPCAT1) can catalyze lysophosphatidylcholine into phosphatidylcholine, modulate phospholipid composition to create favorable conditions for HCC cells, and promote cell proliferation, migration and invasion.[17] As a member of the short-chain dehydrogenase/reductase superfamily, dehydrogenase/reductase member 1 (DHRS1) interacts with the membrane of the endoplasmic reticulum and catalyzes the reductive conversion of some steroids *in vitro*, as well as of other endogenous substances and xenobiotics.[18] It was identified that decreased DHRS1 expression may be involved in the carcinogenesis of IDH1-mutated melanoma.[19] Sorbin and SH3 domain-containing 2 (SORBS2) is critical for regulating cell adhesion and actin/cytoskeletal organization.[20,21] SORBS2 was down-regulated in HCC tissues and down-regulation of SORBS2 significantly correlated with poor survival of HCC patients.[22] Aldehyde dehydrogenase 5 family member A1 (ALDH5A1) belongs to the superfamily of aldehyde dehydrogenases (ALDHs), which is essential for the synthesis of various molecules such as retinoic acid, bine, and g-aminobutyric acid (GABA).[23] Expression of ALDH5A1 was down-regulated in certain type of tumors, such as ovarian cancer.[24] Down-regulation of ALDH5A1 could reprogram GABA metabolism and lead to stem-like cell differentiation in the tumor.[25] SULT1C2 is a member of the 1C family of human cytosolic sulfotransferases (SULT1Cs), which catalyzes the conjugation of myriad drugs, environmental chemicals, hormones and sterols.[26] SULT1Cs are most noted for their ability to bioactivate potent procarcinogens, such as N-hydroxy-2-acetylaminofluorene, and it has been reported that SULT1C2 expression was up-regulated in malignant breast tissue.[27] Secreted phosphoprotein-1 (SPP1), also called osteopontin, is a secreted arginine-glycine-aspartate-containing phosphoprotein, which has been demonstrated as being overexpressed and serving as an prognostic biomarker in many cancers, including lung adenocarcinoma,[28] upper tract urothelial carcinomas[29] and HCC.[30] Recently, SPP1 has been found to be involved in tumor immunosuppression and to influence the tumor microenvironment.[31] Golgi membrane protein 1 (GOLM1), a type II cis-Golgi-localized transmembrane protein, is associated with tumor progression,[32] metastasis[33] and immunosuppression.[34] Increased expression level of GOLM1 has been reported in several types of cancer, such as HCC,[33] lung adenocarcinoma[35] and prostate cancer.[36] Hes-related family bHLH transcription factor with YRPW motif 1 (HEY 1) is a transcriptional repressor in the NOTCH pathway, which is consistently induced by hypoxia.[37,38] Elevated expression of HEY1 was identified in HCC tissues and correlated with unfavorable outcomes.[38] In general, most of genes play an important role in metabolism, which coincides

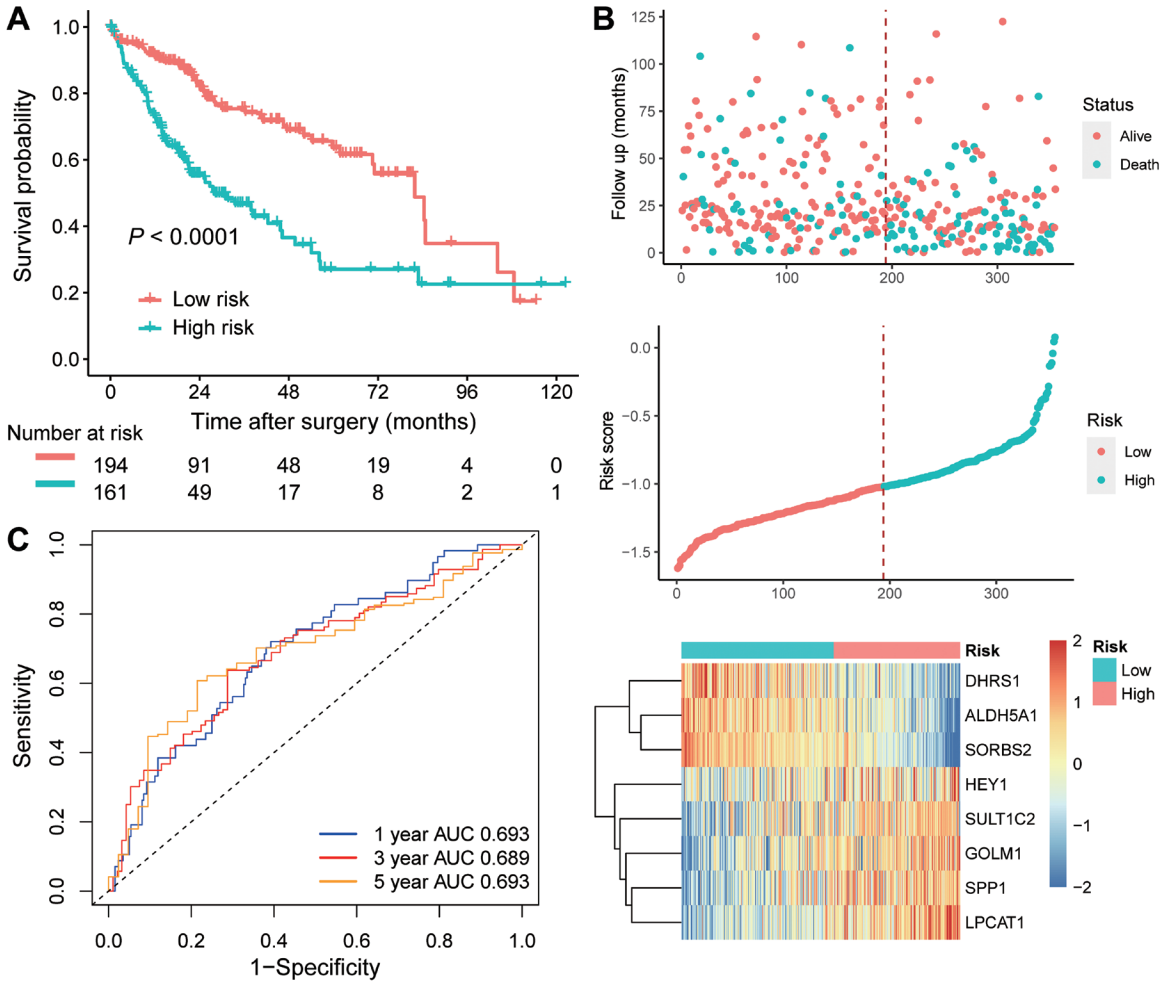

Fig. 3. Kaplan-Meier curve, risk score analysis and time-dependent ROC analysis for the eight-gene model in the TCGA-LIHC dataset. (A) Kaplan-Meier curve for the eight-gene model in the TCGA-LIHC dataset. (B) Risk score distribution and heatmap of the eight genes in model in the TCGA-LIHC dataset. (C) Time-dependent ROC analysis of the eight-gene model for 1-, 3- and 5-year OS in the TCGA-LIHC dataset.

with the GO and KEGG enrichment analyses; poor survival HCCs are associated with dysregulated metabolism.

We have built a nomogram that combined the prognostic model and AJCC cancer staging system. This nomogram showed better efficacy in predicting OS than the prognostic model or AJCC staging system alone. This indicated that when used alone, a pure bioinformatics analysis model or clinical staging system could only reflect biological or clinical features of a certain disease and thus might not predict clinical prognosis well. We herein provided a new insight for future bioinformatics study that the incorporation of clinicopathological parameters into a mathematical model might improve prediction outcome.

To our knowledge, this is the first prognostic model that was constructed by integrating the DEGs between tumors and adjacent liver tissues with those between patients with short and long survival. Compared with previously reported ones, the prognosis prediction performance of this model was comparable or even better in certain conditions, especially when there were multiple validation sets. Although the underlying mechanism is not clear, we proposed that some survival-associated DEGs might not be differentially expressed between tumors and adjacent liver tissues. Former lines of research which focused solely on DEGs between tumors and adjacent liver tissues might thus neglect other potential prognostic factors.

There are several limitations in the present study. First, compared with that of the training set, AUCs of the prognostic model in two validation datasets have slightly decreased. One possible explanation might be the difference of patient ethnicities and underlying etiology. Most patients in the TCGA-LIHC dataset were Asian or White and the major causes of HCC were hepatitis C virus infection and alcohol consumption, while the majority of patients in GSE14520 came from China and the predominant etiology of HCC was hepatitis B virus infection. In addition, most patients in the ICGC-LIRI dataset were from Japan but the ethnicity and etiology were unclarified. Secondly, external tests in other datasets or clinical cohorts are necessary to further verify the prognostic value of the eight-gene model. Thirdly, the mechanical investigation of the enrolled genes in the model was largely descriptive and future research is needed to clarify the functions of certain genes that have not been widely investigated in cancers, such as SULT1C2.

Conclusions

In conclusion, this study provided, for the first time, an

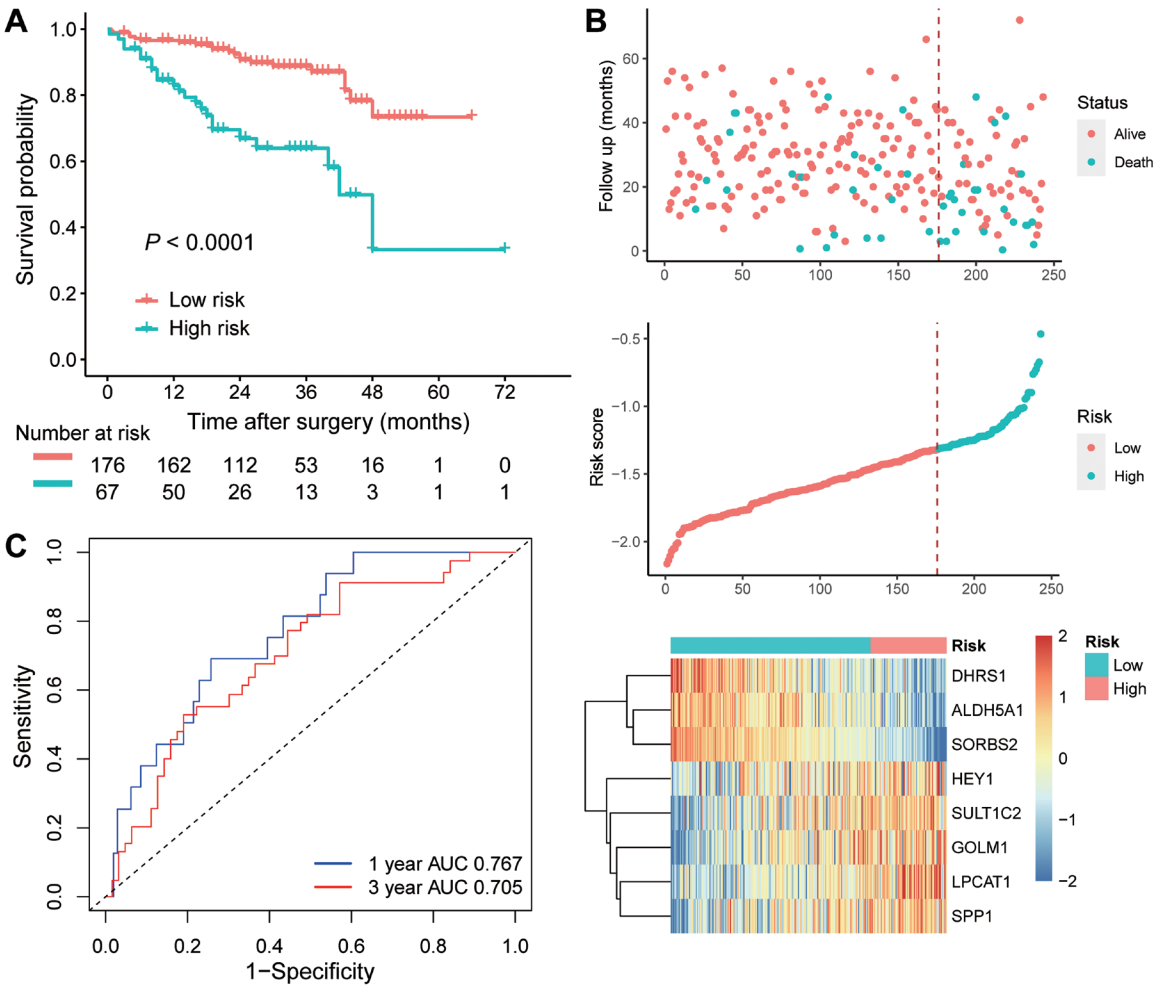

Fig. 4. Kaplan-Meier curve, risk score analysis and time-dependent ROC analysis for the eight-gene model in the ICGC-LIRI dataset. (A) Kaplan-Meier curve for the eight-gene model in the ICGC-LIRI dataset. (B) Risk score distribution and heatmap of the eight genes in model in the ICGC-LIRI dataset. (C) Time-dependent ROC analysis of the eight-gene model for 1- and 3-year OS in the ICGC-LIRI dataset.

Table 3. Comparison with previously reported prognostic models

Ref.	Model	Training set	AUC (1-, 3-, 5-year OS)	Validation set one	AUC (1-, 3-, 5-year OS)	Validation set two	AUC (1-, 3-, 5-year OS)
Our model	8-gene	GSE14520 (*n*=203)	0.78, 0.74, 0.75	TCGA (*n*=355)	0.69, 0.69, 0.69	ICGC (*n*=243)	0.77, 0.71, –
Long *et al*. 2018[9]	4-gene	TCGA (*n*=365)	0.77, 0.70,0.70	GSE54236 (*n*=78)	0.78, 0.59, –	/	/
Qiao *et al*. 2019[10]	8-gene	TCGA (*n*=332)	–, 0.78, 0.77	GSE14520 (*n*=221)	–, 0.71, 0.69	/	/
Liu *et al*. 2019[6]	6-gene	TCGA (*n*=172)	0.83, 0.85, 0.77	TCGA (*n*=171)	0.71, 0.59, 0.60	GSE14520 (*n*=215)	0.68, 0.64, 0.63
Yan *et al*. 2019[11]	4-gene	TCGA (*n*=236)	0.72, 0.71,0.61	TCGA (*n*=118)	0.71, 0.57, 0.55	GSE76427 (*n*=115)	0.63, 0.66, 0.72
Li *et al*. 2020[5]	6-gene	TCGA (*n*=365)	0.76, 0.68, 0.69	ICGC (*n*=243)	0.68, 0.7, 0.68	/	/
Zhang *et al*. 2020[7]	14-gene	TCGA (*n*=312)	0.71, 0.74, 0.64	GSE14520 (*n*=225)	0.64, 0.59, 0.65	GSE76427 (*n*=114)	0.60, 0.61, 0.60
Liu *et al*. 2020[8]	4-gene	TCGA (*n*=343)	0.70, 0.71, 0.68	GSE14520 (*n*=215)	0.72, 0.70, 0.68	/	/

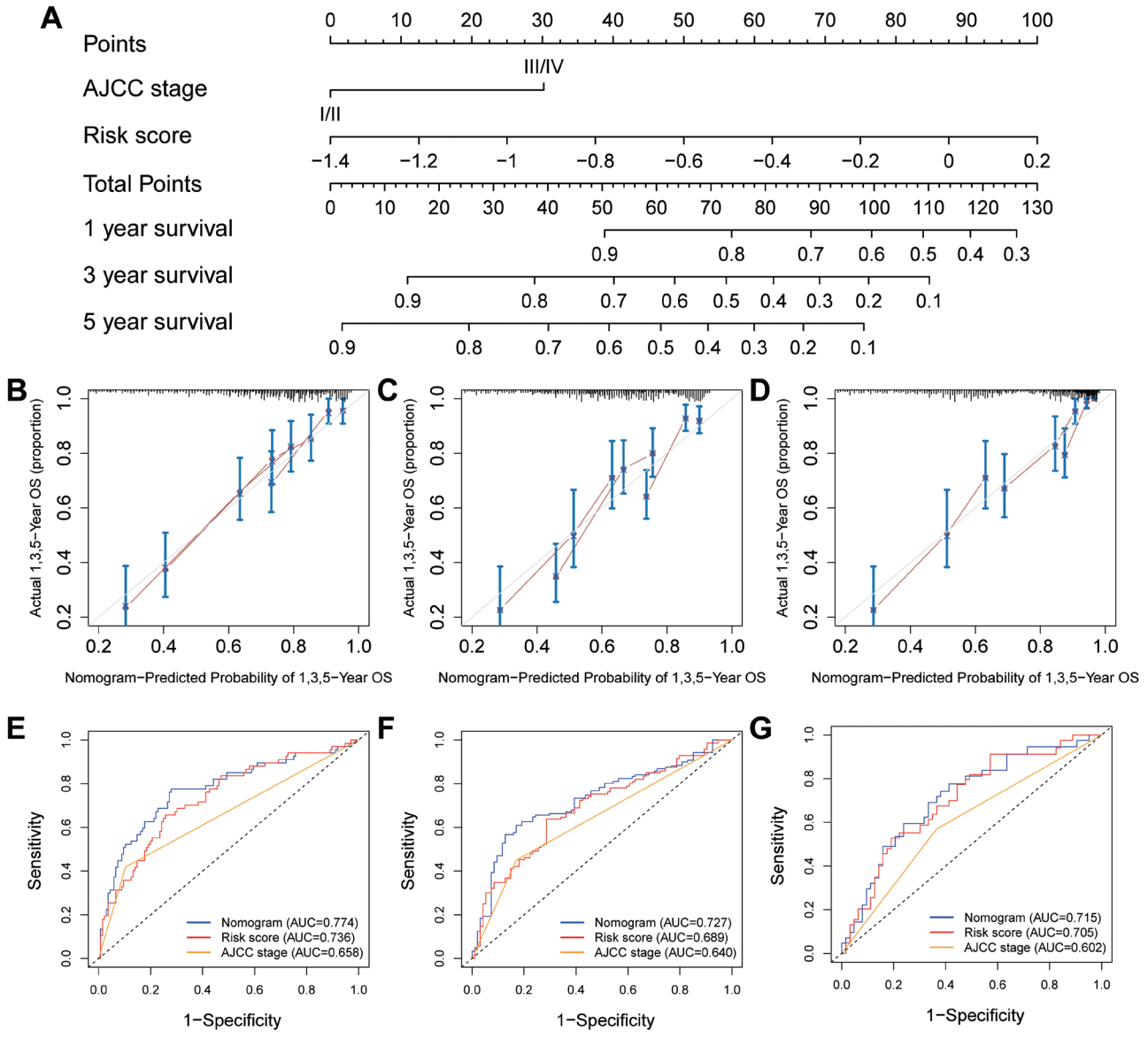

Fig. 5. Establishment and validation of the eight-gene prognostic model-based nomogram predicting OS for HCC patients. (A) The nomogram combined the prognostic model and AJCC stage was built in the GSE14520 dataset. (B–D) The calibration curve of the nomogram in the GSE14520 dataset (B), TCGA-LIHC dataset (C) and ICGC-LIRI dataset (D). (E–G) The time-dependent ROC curves of the nomogram for 3-year OS in the GSE14520 dataset (E), TCGA-LIHC dataset (F) and ICGC-LIRI dataset (G).

eight-gene model to predict OS for HCC by comprehensively comparing the transcriptome profiling of tumor and adjacent liver tissues and tumor tissues from patients with different OS. The enrolled genes in the model suggested that metabolism played important role in the development and progression of HCC. Future work of these hub genes may facilitate our gaining a greater understanding of and treatment for HCC. Moreover, bioinformatics modeling, if combined with clinicopathological features, could produce improved prediction performance.

Funding

This study was jointly supported by the National Key R&D Program of China (Nos. 2019YFC1315800, 2019YFC1315802), National Natural Science Foundation of China (Nos. 81830102, 81772578, 81802991), STCSM (No. 18YF1403600), and Shanghai Municipal Key Clinical Specialty.

Conflict of interest

The authors have no conflict of interests related to this publication.

Author contributions

Study concept and design (JZ, DZG, AH), collection and assembly of data (JZ, DZG, AH, YPW, XRY), data analysis and

interpretation (DZG, AH, YPW, YC, JF), drafting of the manuscript (DZG, AH, XRY), critical revision of the manuscript for important intellectual content (JZ, XRY).

Data sharing statement

All data are available upon request.

References

[1] EASL Clinical Practice Guidelines: Management of hepatocellular carcinoma. J Hepatol 2018;69(1):182–236. doi:10.1016/j.jhep.2018.03.019.
[2] Bray F, Ferlay J, Soerjomataram I, Siegel RL, Torre LA, Jemal A. Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA Cancer J Clin 2018;68(6):394–424. doi:10.3322/caac.21492.
[3] Comprehensive and integrative genomic characterization of hepatocellular carcinoma. Cell 2017;169(7):1327–1341.e23. doi:10.1016/j.cell.2017.05.046.
[4] Kamarajah SK, Frankel TL, Sonnenday C, Cho CS, Nathan H. Critical evaluation of the American Joint Commission on Cancer (AJCC) 8th edition staging system for patients with Hepatocellular Carcinoma (HCC): A Surveillance, Epidemiology, End Results (SEER) analysis. J Surg Oncol 2018;117(4):644–650. doi:10.1002/jso.24908.
[5] Li W, Lu J, Ma Z, Zhao J, Liu J. An integrated model based on a six-gene signature predicts overall survival in patients with hepatocellular carcinoma. Front Genet 2020;10:1323. doi:10.3389/fgene.2019.01323.
[6] Liu GM, Zeng HD, Zhang CY, Xu JW. Identification of a six-gene signature predicting overall survival for hepatocellular carcinoma. Cancer Cell Int 2019;19:138. doi:10.1186/s12935-019-0858-2.
[7] Zhang BH, Yang J, Jiang L, Lyu T, Kong LX, Tan YF, *et al*. Development and validation of a 14-gene signature for prognosis prediction in hepatocellular carcinoma. Genomics 2020;112(4):2763–2771. doi:10.1016/j.ygeno.2020.03.013.
[8] Liu GM, Xie WX, Zhang CY, Xu JW. Identification of a four-gene metabolic signature predicting overall survival for hepatocellular carcinoma. J Cell Physiol 2020;235(2):1624–1636. doi:10.1002/jcp.29081.
[9] Long J, Zhang L, Wan X, Lin J, Bai Y, Xu W, *et al*. A four-gene-based prognostic model predicts overall survival in patients with hepatocellular carcinoma. J Cell Mol Med 2018;22(12):5928–5938. doi:10.1111/jcmm.13863.
[10] Qiao GJ, Chen L, Wu JC, Li ZR. Identification of an eight-gene signature for survival prediction for patients with hepatocellular carcinoma based on integrated bioinformatics analysis. PeerJ 2019;7:e6548. doi:10.7717/peerj.6548.
[11] Yan Y, Lu Y, Mao K, Zhang M, Liu H, Zhou Q, *et al*. Identification and validation of a prognostic four-genes signature for hepatocellular carcinoma: integrated ceRNA network analysis. Hepatol Int 2019;13(5):618–630. doi:10.1007/s12072-019-09962-3.
[12] Roessler S, Jia HL, Budhu A, Forgues M, Ye QH, Lee JS, *et al*. A unique metastasis gene signature enables prediction of tumor relapse in early-stage hepatocellular carcinoma patients. Cancer Res 2010;70(24):10202–10212. doi:10.1158/0008-5472.CAN-10-2607.
[13] Weinstein JN, Collisson EA, Mills GB, Shaw KR, Ozenberger BA, Ellrott K, *et al*. The Cancer Genome Atlas Pan-Cancer analysis project. Nat Genet 2013;45(10):1113–1120. doi:10.1038/ng.2764.
[14] Hudson TJ, Anderson W, Artez A, Barker AD, Bell C, Bernabé RR, *et al*. International network of cancer genome projects. Nature 2010;464(7291):993–998. doi:10.1038/nature08987.
[15] Diboun I, Wernisch L, Orengo CA, Koltzenburg M. Microarray analysis after RNA amplification can detect pronounced differences in gene expression using limma. BMC Genomics 2006;7:252. doi:10.1186/1471-2164-7-252.
[16] Tibshirani R. The lasso method for variable selection in the Cox model. Stat Med 1997;16(4):385–395.
[17] Morita Y, Sakaguchi T, Ikegami K, Goto-Inoue N, Hayasaka T, Hang VT, *et al*. Lysophosphatidylcholine acyltransferase 1 altered phospholipid composition and regulated hepatoma progression. J Hepatol 2013;59(2):292–299. doi:10.1016/j.jhep.2013.02.030.
[18] Zemanová L, Navrátilová H, Andrýs R, Šperková K, Andrejs J, Kozáková K, *et al*. Initial characterization of human DHRS1 (SDR19C1), a member of the short-chain dehydrogenase/reductase superfamily. J Steroid Biochem Mol Biol 2019;185:80–89. doi:10.1016/j.jsbmb.2018.07.013.
[19] Shibata T, Kokubu A, Miyamoto M, Sasajima Y, Yamazaki N. Mutant IDH1 confers an in vivo growth in a melanoma cell line with BRAF mutation. Am J Pathol 2011;178(3):1395–1402. doi:10.1016/j.ajpath.2010.12.011.
[20] Kawabe H, Hata Y, Takeuchi M, Ide N, Mizoguchi A, Takai Y. nArgBP2, a novel neural member of ponsin/ArgBP2/vinexin family that interacts with synapse-associated protein 90/postsynaptic density-95-associated protein (SAPAP). J Biol Chem 1999;274(43):30914–30918. doi:10.1074/jbc.274.43.30914.
[21] Taieb D, Roignot J, André F, Garcia S, Masson B, Pierres A, *et al*. ArgBP2-dependent signaling regulates pancreatic cell migration, adhesion, and tumorigenicity. Cancer Res 2008;68(12):4588–4596. doi:10.1158/0008-5472.CAN-08-0958.
[22] Han L, Huang C, Zhang S. The RNA-binding protein SORBS2 suppresses hepatocellular carcinoma tumourigenesis and metastasis by stabilizing RORA mRNA. Liver Int 2019;39(11):2190–2203. doi:10.1111/liv.14202.
[23] Jackson B, Brocker C, Thompson DC, Black W, Vasiliou K, Nebert DW, *et al*. Update on the aldehyde dehydrogenase gene (ALDH) superfamily. Hum Genomics 2011;5(4):283–303. doi:10.1186/1479-7364-5-4-283.
[24] Tian X, Han Y, Yu L, Luo B, Hu Z, Li X, *et al*. Decreased expression of ALDH5A1 predicts prognosis in patients with ovarian cancer. Cancer Biol Ther 2017;18(4):245–251. doi:10.1080/15384047.2017.1295175.
[25] El-Habr EA, Dubois LG, Burel-Vandenbos F, Bogeas A, Lipecka J, Turchi L, *et al*. A driver role for GABA metabolism in controlling stem and proliferative cell state through GHB production in glioma. Acta Neuropathol 2017;133(4):645–660. doi:10.1007/s00401-016-1659-5.
[26] Runge-Morris M, Kocarek TA. Expression of the sulfotransferase 1C family: implications for xenobiotic toxicity. Drug Metab Rev 2013;45(4):450–459. doi:10.3109/03602532.2013.835634.
[27] Aust S, Obrist P, Klimpfinger M, Tucek G, Jäger W, Thalhammer T. Altered expression of the hormone- and xenobiotic-metabolizing sulfotransferase enzymes 1A2 and 1C1 in malignant breast tissue. Int J Oncol 2005;26(4):1079–1085. doi:10.3892/ijo.26.4.1079.
[28] Shen XY, Liu XP, Song CK, Wang YJ, Li S, Hu WD. Genome-wide analysis reveals alcohol dehydrogenase 1C and secreted phosphoprotein 1 for prognostic biomarkers in lung adenocarcinoma. J Cell Physiol 2019;234(12):22311–22320. doi:10.1002/jcp.28797.
[29] Li Y, He S, He A, Guan B, Ge G, Zhan Y, *et al*. Identification of plasma secreted phosphoprotein 1 as a novel biomarker for upper tract urothelial carcinomas. Biomed Pharmacother 2019;113:108744. doi:10.1016/j.biopha.2019.108744.
[30] Shin HD, Park BL, Cheong HS, Yoon JH, Kim YJ, Lee HS. SPP1 polymorphisms associated with HBV clearance and HCC occurrence. Int J Epidemiol 2007;36(5):1001–1008. doi:10.1093/ije/dym093.
[31] Shurin MR. Osteopontin controls immunosuppression in the tumor microenvironment. J Clin Invest 2018;128(12):5209–5212. doi:10.1172/JCI124918.
[32] Liu G, Zhang Y, He F, Li J, Wei X, Li Y, *et al*. Expression of GOLPH2 is associated with the progression of and poor prognosis in gastric cancer. Oncol Rep 2014;32(5):2077–2085. doi:10.3892/or.2014.3404.
[33] Ye QH, Zhu WW, Zhang JB, Qin Y, Lu M, Lin GL, *et al*. GOLM1 modulates EGFR/RTK cell-surface recycling to drive hepatocellular carcinoma metastasis. Cancer Cell 2016;30(3):444–458. doi:10.1016/j.ccell.2016.07.017.
[34] Zhang X, Zhu C, Wang T, Jiang H, Ren Y, Zhang Q, *et al*. GP73 represses host innate immune response to promote virus replication by facilitating MAVS and TRAF6 degradation. PLoS Pathog 2017;13(4):e1006321. doi:10.1371/journal.ppat.1006321.
[35] Zhang F, Gu Y, Li X, Wang W, He J, Peng T. Up-regulated Golgi phosphoprotein 2 (GOLPH2) expression in lung adenocarcinoma tissue. Clin Biochem 2010;43(12):983–991. doi:10.1016/j.clinbiochem.2010.05.010.
[36] Kristiansen G, Fritzsche FR, Wassermann K, Jäger C, Tölls A, Lein M, *et al*. GOLPH2 protein expression as a novel tissue biomarker for prostate cancer: implications for tissue-based diagnostics. Br J Cancer 2008;99(6):939–948. doi:10.1038/sj.bjc.6604614.
[37] Bray SJ. Notch signalling: a simple pathway becomes complex. Nat Rev Mol Cell Biol 2006;7(9):678–689. doi:10.1038/nrm2009.
[38] Kung-Chun Chiu D, Pui-Wah Tse A, Law CT, Ming-Jing Xu I, Lee D, Chen M, *et al*. Hypoxia regulates the mitochondrial activity of hepatocellular carcinoma cells through HIF/HEY1/PINK1 pathway. Cell Death Dis 2019;10(12):934. doi:10.1038/s41419-019-2155-3.