Prognostic Score-based Clinical Factors and Metabolism-related Biomarkers for Predicting the Progression of Hepatocellular Carcinoma

Evolutionary Bioinformatics
Volume 16: 1–11
© The Author(s) 2020
Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/1176934320951571

Jia Yan[1,2,3], Ming Shu[1,2,3], Xiang Li[1,2,3], Hua Yu[1,2,3], Shuhuai Chen[1,2,3] and Shujie Xie[1,2,3]

[1]Department of Hepatobiliary Pancreatic Surgery, Hwa Mei Hospital, University of Chinese Academy of Sciences, Ningbo, Zhejiang, China. [2]Ningbo Institute of Life and Health Industry, University of Chinese Academy of Sciences, Ningbo, Zhejiang, China. [3]Key Laboratory of Diagnosis and Treatment of Digestive System Tumors of Zhejiang Province, Ningbo, Zhejiang, China.

ABSTRACT: Hepatocellular carcinoma (HCC) is a common malignant tumor representing more than 90% of primary liver cancer. This study aimed to identify metabolism-related biomarkers with prognostic value by developing the novel prognostic score (PS) model. Transcriptomic profiles derived from TCGA and EBIArray databases were analyzed to identify differentially expressed genes (DEGs) in HCC tumor samples compared with normal samples. The overlapped genes between DEGs and metabolism-related genes (crucial genes) were screened and functionally analyzed. A novel PS model was constructed to identify optimal signature genes. Cox regression analysis was performed to identify independent clinical factors related to prognosis. Nomogram model was constructed to estimate the predictability of clinical factors. Finally, protein expression of crucial genes was explored in different cancer tissues and cell types from the Human Protein Atlas (HPA). We screened a total of 305 overlapped genes (differentially expressed metabolism-related genes). These genes were mainly involved in "oxidation reduction," "steroid hormone biosynthesis," "fatty acid metabolic process," and "linoleic acid metabolism." Furthermore, we screened ten optimal DEGs (CYP2C9, CYP3A4, and TKT, among others) by using the PS model. Two clinical factors of pathologic stage ($P < .001$, HR: 1.512 [1.219-1.875]) and PS status ($P < .001$, HR: 2.259 [1.522-3.354]) were independent prognostic predictors by cox regression analysis. Nomogram model showed a high predicted probability of overall survival time, and the AUC value was 0.837. The expression status of 7 proteins was frequently altered in normal or differential tumor tissues, such as liver cancer and stomach cancer samples.We have identified several metabolism-related biomarkers for prognosis prediction of HCC based on the PS model. Two clinical factors were independent prognostic predictors of pathologic stage and PS status (high/low risk). The prognosis prediction model described in this study is a useful and stable method for novel biomarker identification.

KEYWORDS: Hepatocellular carcinoma, prognosis, biomarker, metabolism

RECEIVED: May 7, 2020. **ACCEPTED:** July 24, 2020.

TYPE: Original Research

FUNDING: The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This study was funded by Ningbo Clinical Research Center for Digestive System Tumors (No.2019A21003), Ningbo Medical Science and Technology Foundation (No.2018A67) and Research Foundation of Hwa Mei Hospital, University of Chinese Academy of Sciences (No.2018HMKY03).

DECLARATION OF CONFLICTING INTERESTS: The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

CORRESPONDING AUTHOR: Shujie Xie, MD, Department of Hepatobiliary Pancreatic Surgery, HwaMei Hospital, University of Chinese Academy of Sciences, No.41, Northwest Street, Ningbo, Zhejiang Province 315010, China. Email: siabc83@163.com

Introduction

Hepatocellular carcinoma (HCC) is the most common type of liver cancer in China and account for the second leading cause of cancer death.[1] An estimated 42,810 new cases were diagnosed as liver cancer in the United States and more than 30,160 patients died of this disease, according to cancer statistic in 2020.[2] Risk factors for HCC included infection of chronic hepatitis B or C (HBV/HCV) virus, alcohol abuse, diet of aflatoxin, and progress to cirrhosis induced by nonalcoholic steatohepatitis.[3,4] Treatment of HCC depends on the tumor stage, and different therapeutic options are available for early-stage patients, including orthotopic liver transplantation, surgical resection, radiofrequency ablation, and chemo- and radio-therapy.[5,6] Despite great improvement in surgical treatments over past decades, long-term outcome needs remain unmet, and patients with advanced stage still have limited therapeutic options. As for the greatly increased disease burden, it is important to identify useful and reliable tumor biomarkers for early stage detection and prognosis prediction of advanced HCC.

Changes in genomic and subsequent transcriptome expression are the most common characteristics driving tumorigenesis and have been extensively uncovered in liver cancer patients.[7-9] The recent development of sequential transcriptome analysis could potentially promote understanding of molecular mechanisms in human liver cancer. It facilitated the diagnosis and therapy of this disease. By using transcriptome analysis, researchers identified cohorts of genes as potential candidate biomarkers for prognosis in HCC patients, such as YWHAZ, ENAH, and HMGN4.[10] Furthermore, based on RNA-seq analyses, Jiang et al. identified several genes and miRNAs that might be pathogenic biomarkers of HCC.[11] According to integrative analyses of genome and transcriptome sequencing data, Miao et al. screened multiple prognostic biomarkers in hepatitis B virus-related HCC patients, including protein kinase TTK.[12] These findings provided novel insights and an important strategy to identify tumor biomarkers for early diagnosis and prognosis prediction of the disease.

In this study, we first analyzed the transcriptomic expression status of HCC derived from The Cancer Genome Atlas (TCGA) database and screened differentially expressed genes (DEGs) by integrated analysis of three genomic profiles related to the metabolism of amino acids, carbohydrates, and lipids. The prognostic score (PS) model was constructed to identify optimal prognosis-related genes. Finally, by analyzing the Human Protein Atlas (HPA) database, we analyzed the expression levels of critical genes in different cells, tissues, and cancer types. Our results identified novel prognostic biomarkers for HCC patients, which provides new insight on the understanding of HCC.

Materials and Methods

Microarray data resource

Microarray datasets of HCC downloaded from TCGA database (https://gdc-portal.nci.nih.gov/) on 10 February 2020 were tested on the Illumina HiSeq 2000 RNA Sequencing platform. The datasets comprised 423 samples, including 373 liver cancer samples and 50 normal samples. There were 367 tumor samples labeled with clinical prognostic information. This dataset was used as training data set.

Moreover, the genomic profiles of liver cancer under access number E-TABM-36[13] were derived from EBIArray database[14] (https://www.ebi.ac.uk/arrayexpress/) and tested on GPL96 [HG-U133A] Affymetrix Gene Chip Human Genome HG-U133A platform. E-TABM-36 dataset included a total of 65 samples, which comprised 11 normal samples and 44 liver cancer tissue samples with prognostic information. E-TABM-36 was considered as validation dataset.

Screening metabolism-related DEGs in liver cancer samples

Limma package[15] in Version 3.34.7 (https://bioconductor.org/packages/release/bioc/html/limma.html) was used to screen the DEGs between tumor samples and normal tissue samples in TCGA training sets under the criteria of $FDR < 0.05$and $| \log2FC | > 1$ (Parameters was set as 2 times).

After exacting the expression value of DEGs from the training data set, we used pheatmap[16] (version1.0.8, https://cran.r-project.org/web/packages/pheatmap/index.html) in R3.4.1 software to perform hierarchical clustering analysis by using the algorithm of centered Pearson correlation.[17]

All genes related to the metabolism of amino acids, carbohydrates, and lipids were downloaded from the Gene Set Enrichment Analysis[18] (GSEA, http://software.broadinstitute.org/gsea/downloads.jsp) database. We analyzed the intersection between metabolism-related genes and DEGs of training datasets.

The overlapped genes were screened as candidate genes and by functional enrichment analysis, including GO biological process annotation and KEGG pathway analysis by using DAVID[19,20] (version 6.8, https://david.ncifcrf.gov/) tool. P values lower than .05 were considered to indicate significant difference.

Construction of PS model

As for the overlapped genes screening from TCGA training set and metabolism cohort gene, we conducted Cox regression analysis using the survival package (version 2.41-1, http://bioconductor.org/packages/survivalr/).[21] By the criteria of log-rank p value less than 0.05, we screened the prognosis related DEGs.

Penalized package[22] Version 0.9.50 (https://cran.r-project.org/web/packages/penalized/index.html) were used to construct the LASSO Cox regression model[23] for optimal DEGs screening. The parameter of "lambda" in the filter model was calculated by 1000 cross-validation likelihood (CVL) algorithm. Then, we developed a novel PS model by calculating the prognostic coefficient of each DEGs and the expression level of DEGs in training set samples. The formula for calculating PS was as follows:

$$\text{Prognostic score}\left(\text{PS}\right) = \sum \beta_{\text{DEGs}} \times \text{Exp}_{\text{DEGs}}$$

β_{DEGs} meant the prognostic coefficient of signature DEGs, and the Exp $_{\text{DEGs}}$ represented the expression levels of DEGs in the training dataset.

To assess the effectiveness and predictive ability of the prognostic model, we first calculated the PS values of whole samples in the TCGA data set. Then, these samples were divided into high- and low-risk groups, based on median PS value. Kaplan Meier curves were generated to evaluate overall survival times of patients between the 2 groups by using survival package Version2.41. Furthermore, the expression values of DEGs were extracted from the validation dataset, and all specimens were also divided into high- and low-risk groups according to the median value. Kaplan Meier curves were generated to evaluate the correlation between prognosis and differential risk groups.

Identifying clinical factors with independent prognostic value in HCC

In the TCGA training dataset, the survival package was used to analyzed HCC samples. Univariate and multivariate Cox regression analyses were conducted to identify independent prognostic factors associated with the survival of patients by the criteria of log-rank P value < .05.

Thus, liver cancer samples derived from the training dataset were divided into different groups according to pathological stage I-IV. Stratified analyses were conducted to explore the correlation between risk grouping and clinical factors.

Moreover, we developed a nomogram model to predict the 3 year- and 5 year- survival probability in liver cancer patients by using rms package[24] (version 5.1-2, https://cran.r-project.org/web/packages/rms/index.html). Nomogram is a logistic regression-based model that has been extensively applied in prognostic prediction of various cancers, such as colorectal cancer, renal cancer, and bladder cancer.[25-27] It formulates the scoring criteria according to regression coefficients of all independent variables

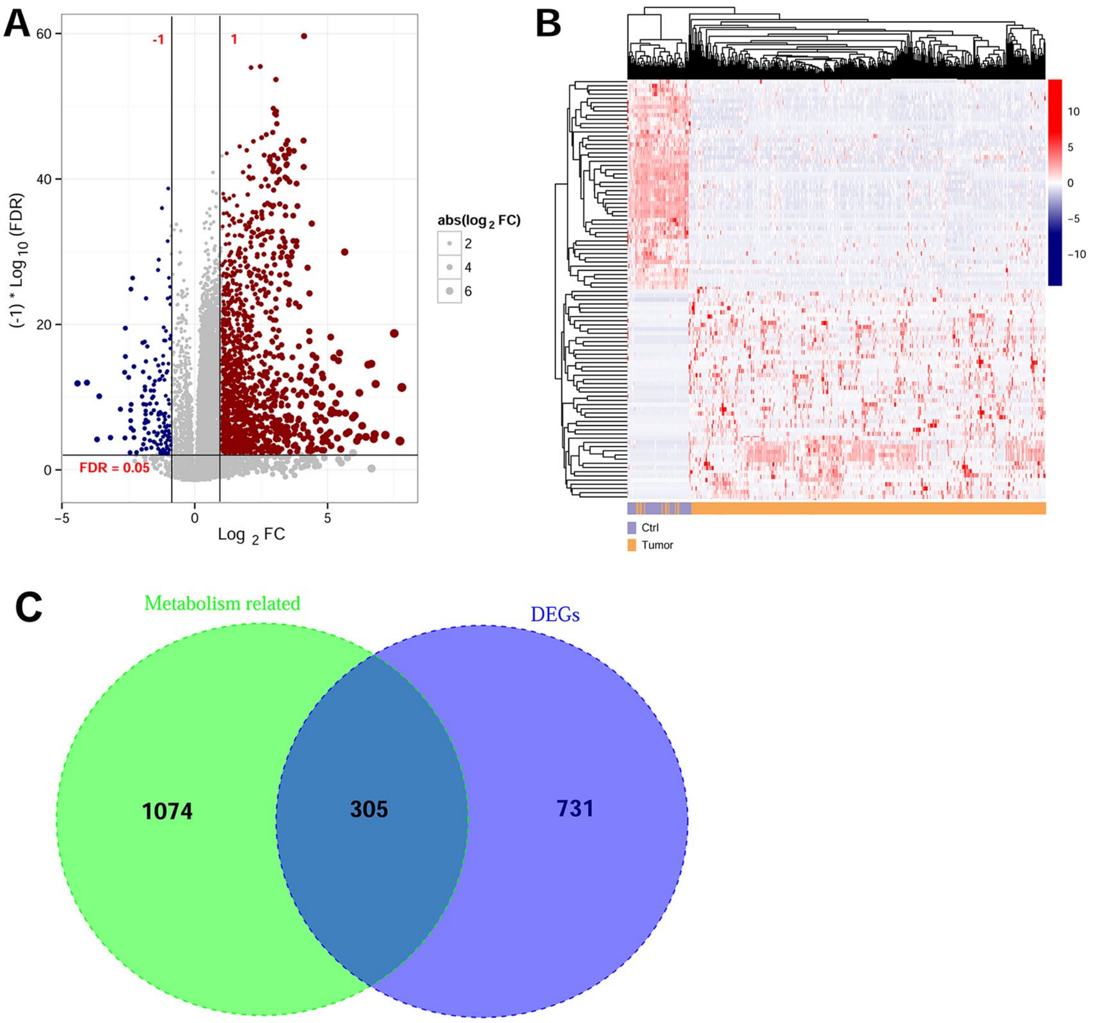

Figure1. Screening of differentially expressed genes (DEGs) associated with the metabolism of liver cancer patients. (A) Volcano diagram visualizing the DEGs screening from TCGA and EBI Array dataset under FDR $<$ 0.05 and | log2FC | $>$ 1 threshold. The red and blue dots represent upregulated and downregulated DEGs. The line in horizontal axis represents FDR <0.05; while the 2 vertical dotted lines represent |log2FC| $>$ 0.5. The dot size is consistent with absolute value of logFC. (B) Hierarchical clustering analysis results of the top 50 up- and down-regulated DEGs, sorted by increasing logFC values. (C) Venn diagram visualizing the crucial genes associated with liver cancer metabolism by taking the intersection of metabolism-related genes and DEGs of training datasets.

and assigns a score level for each independent variable. All of the selected samples entered the logistic regression model and the total score of each feature was calculated. The survival probability of each individual was calculated by the transfer function between the score and survival probability.

Expression analysis of target genes at the protein level (HPA database)

By using the Human Protein Atlas (HPA, https://www.proteinatlas.org/) Version 18 database, we analyzed the expression of target genes at the protein level in various cell types and different cancer tissues. HPA is an online database containing large amounts of antibody proteomics and transcriptomics data generated from normal and cancer tissues profiling and RNA sequencing technologies.[28-30] Although the HPA database provided immunohistochemistry images, proteomics, and transcriptomics data, it may not be optimal for dataset analysis and automatic retrieval of images. In this study, we performed data mining for critical genes by using HPA analyze packager[31] (version 1.4.3, http://www.bioconductor.org/packages/release/bioc/html/HPAanalyze.html), a software package designed for easy retrieval and analysis of HPA data.

Results

Screening DEGs associated with the metabolism of liver cancer

After data preprocessing, the specimens from TCGA database were divided into tumor and control groups according to clinical information of relapse. By the criteria of FDR $<$ 0.05 and | log2FC | $>$ 1, we screened a total of 1035 DEGs from tumor samples compared with normal specimens, including 196 downregulated and 840 upregulated DEGs (Figure 1A). These DEGs were sorted by increased logFC value, and the top 50 upregulated and downregulated genes were selected to conduct expression level-based hierarchical clustering analysis (Figure 1B). The results showed that clinical samples can be

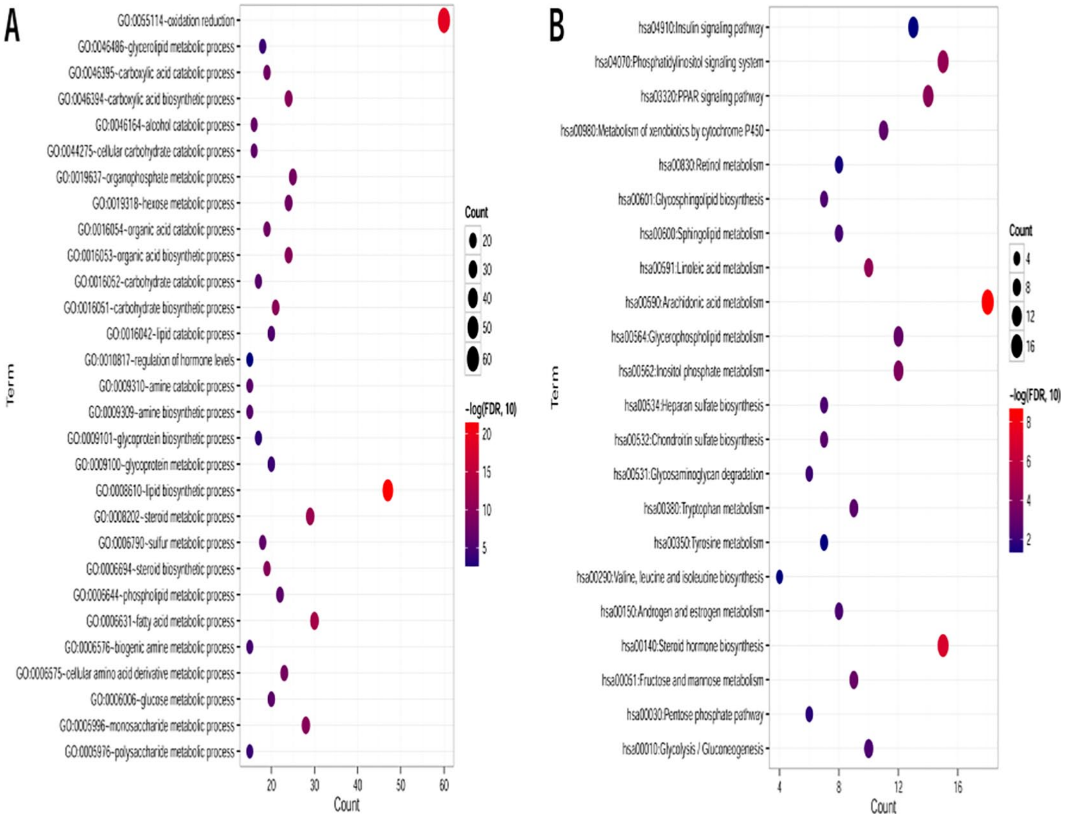

Figure 2. Functional enrichment analysis of Gene Ontology (A) and Kyoto Encyclopedia of Genes and Genomes pathway (B).
The horizontal axis shows gene number; vertical axis represents biological terms or pathways categories. The dot size is consistent with gene number, while color intensity of dots corresponds to P-value. A darker red color means higher statistical significance.

clustered into 2 different directions. Based on GSEA database screening, we obtained 372 amino acid metabolism related-genes, 293 carbohydrate metabolism-related genes, and 738 lipid metabolism-related genes. By comparison of DEGs, we identified 305 overlapped genes (Figure 1C).

Functional enrichment analysis revealed DEGs were mainly involved in 29 GO terms and 22 KEGG pathways (Figure 2, Table 1). GO terms included "oxidation reduction," "lipid biosynthetic process," "fatty acid metabolic process," "steroid metabolic process," "carbohydrate biosynthetic process," among others. The signaling pathway categories were associated with "steroid hormone biosynthesis," "glycerophospholipid metabolism," and "linoleic acid metabolism" pathway.

Prognostic model construction and assessment

Univariate analysis results showed 146 DEGs were associated with prognosis of liver cancer by criteria of log-rank P value < .05. Further multivariate cox regression analysis revealed 34 genes were identified as prognosis related genes. Based on the Cox-Proportional Hazards model, we screened 10 optimal DEGs for further analysis, such as CYP2C9, CYP3A4, CYP4A11, CYP7A1, G6PD, HMMR, LPCAT1, PYCR1, TAT, and TKT (Table 2).

In addition, we constructed the PS model according to the prognostic coefficient of each pre and expression level of DEGs in training set samples. All samples from the TCGA training set (n = 367) and E-TABM-36 validation set (n = 44) were divided into high- and low-risk based on median PS value. Significant survival differences were shown in Kaplan-Meier curves (Figure 3). Patients in low risk groups exhibit a longer overall survival time than those in high risk groups, and the result were consistent between both training datasets (*P* = 8.301e-08, HR:2.598 [1.809-3.730]) and validation datasets (*P* = 1.183e-02; HR:2.524 [1.197-5.324]). The results indicated that PS model-based risk grouping were significant correlated with actual prognosis for HCC patients. The AUC value of receiver operating characteristic curve (ROC) was 0.837 and 0.795 in the training set and validation set, which represented a high predictive probability.

Independent prognostic factors for HCC patients by univariate and multivariate analysis

Univariate analysis of clinical factors showed that pathologic T(T1-T4/-, *P* = 7.867E-09), pathologic stage (I-IV /-, *P* = 4.468E-07) and PS status (High/Low, *P* = 8.301E-08) were significantly correlated with overall survival time (Table 3). A multivariate regression analysis was conducted and results indicated that only pathologic stage (I / II / III / IV /-, HR: 1.512 [1.219-1.875], *P* = 1.680E-04) and PS status (High/Low, HR: 2.259 [1.522-3.354], *P* = 5.250E-05) were

Table 1. Gene Ontology and Kyoto Encyclopedia of Genes and Genomes pathway analysis for intersection genes associated with metabolism of hepatocellular carcinoma.

CATEGORY	TERM	COUNT	P VALUE	FDR
Biology process	GO:0008610~lipid biosynthetic process	47	9.08E-26	1.53E-22
	GO:0055114~oxidation reduction	60	5.70E-23	9.58E-20
	GO:0006631~fatty acid metabolic process	30	7.19E-17	1.89E-13
	GO:0008202~steroid metabolic process	29	1.19E-15	2.05E-12
	GO:0016051~carbohydrate biosynthetic process	21	5.03E-14	8.46E-11
	GO:0006694~steroid biosynthetic process	19	9.78E-14	1.64E-10
	GO:0005996~monosaccharide metabolic process	28	1.06E-13	1.79E-10
	GO:0016053~organic acid biosynthetic process	24	1.08E-13	1.81E-10
	GO:0046394~carboxylic acid biosynthetic process	24	1.08E-13	1.81E-10
	GO:0019637~organophosphate metabolic process	25	3.54E-12	5.96E-09
	GO:0006575~cellular amino acid derivative metabolic process	23	3.92E-12	6.59E-09
	GO:0019318~hexose metabolic process	24	1.06E-11	1.77E-08
	GO:0016054~organic acid catabolic process	19	1.19E-11	2.00E-08
	GO:0046395~carboxylic acid catabolic process	19	1.19E-11	2.00E-08
	GO:0046164~alcohol catabolic process	16	9.03E-11	1.52E-07
	GO:0044275~cellular carbohydrate catabolic process	16	1.85E-10	3.11E-07
	GO:0006790~sulfur metabolic process	18	2.04E-10	3.44E-07
	GO:0006006~glucose metabolic process	20	3.76E-10	6.32E-07
	GO:0006644~phospholipid metabolic process	22	3.94E-10	6.62E-07
	GO:0009310~amine catabolic process	15	5.99E-10	1.01E-06
	GO:0016052~carbohydrate catabolic process	17	7.85E-10	1.32E-06
	GO:0009309~amine biosynthetic process	15	1.01E-09	1.70E-06
	GO:0016042~lipid catabolic process	20	3.08E-09	5.17E-06
	GO:0006576~biogenic amine metabolic process	15	1.15E-08	1.94E-05
	GO:0009100~glycoprotein metabolic process	20	3.98E-08	6.70E-05
	GO:0046486~glycerolipid metabolic process	18	4.21E-08	7.07E-05
	GO:0005976~polysaccharide metabolic process	15	6.68E-08	1.12E-04
	GO:0009101~glycoprotein biosynthetic process	17	1.74E-07	2.93E-04
	GO:0010817~regulation of hormone levels	15	2.99E-06	5.04E-03
KEGG Pathway	hsa00590:Arachidonic acid metabolism	18	2.29E-11	3.03E-09
	hsa00140:Steroid hormone biosynthesis	15	1.39E-09	9.14E-08
	hsa00591:Linoleic acid metabolism	10	8.03E-07	3.53E-05
	hsa04070:Phosphatidylinositol signaling system	15	9.72E-07	3.21E-05
	hsa03320:PPAR signaling pathway	14	2.53E-06	6.67E-05
	hsa00562:Inositol phosphate metabolism	12	6.95E-06	1.53E-04
	hsa00051:Fructose and mannose metabolism	9	4.31E-05	8.13E-04

(Continued)

Table 1. (Continued)

CATEGORY	TERM	COUNT	P VALUE	FDR
	hsa00564:Glycerophospholipid metabolism	12	6.79E-05	1.12E-03
	hsa00980:Metabolism of xenobiotics by cytochrome P450	11	1.13E-04	1.65E-03
	hsa00380:Tryptophan metabolism	9	1.48E-04	1.96E-03
	hsa00532:Chondroitin sulfate biosynthesis	7	1.69E-04	2.02E-03
	hsa00601:Glycosphingolipid biosynthesis	7	3.62E-04	3.97E-03
	hsa00534:Heparan sulfate biosynthesis	7	4.55E-04	4.61E-03
	hsa00150:Androgen and estrogen metabolism	8	5.57E-04	5.23E-03
	hsa00010:Glycolysis / Gluconeogenesis	10	5.63E-04	4.94E-03
	hsa00600:Sphingolipid metabolism	8	7.77E-04	6.39E-03
	hsa00531:Glycosaminoglycan degradation	6	1.19E-03	9.20E-03
	hsa00030:Pentose phosphate pathway	6	2.73E-03	1.98E-02
	hsa00830:Retinol metabolism	8	5.38E-03	3.68E-02
	hsa04910:Insulin signaling pathway	13	7.56E-03	4.88E-02
	hsa00350:Tyrosine metabolism	7	7.65E-03	4.71E-02
	hsa00290:Valine, leucine and isoleucine biosynthesis	4	8.27E-03	4.86E-02

Table 2. The ten optimal signature differential expressed genes related to prognosis of liver cancer according to multi-variate cox regression analysis.

SYMBOL	MULTI-VARIATE COX REGRESSION ANALYSIS			LASSO COEFFICIENT
	HR	95%CI	P VALUE	
CYP2C9	0.6841	0.533-0.878	2.830E-03	–0.09807
CYP3A4	1.2017	1.007-1.434	4.133E-02	0.01229
CYP4A11	1.7710	1.199-2.616	4.089E-03	0.04362
CYP7A1	0.7317	0.559-0.957	2.270E-02	–0.03181
G6PD	2.6743	1.199-5.966	1.627E-02	0.14709
HMMR	2.8779	1.263-6.555	1.184E-02	0.26357
LPCAT1	2.8097	1.414-5.583	3.190E-03	0.09797
PYCR1	1.6407	1.456-1.900	1.035E-02	0.01015
TAT	0.7211	0.531-0.979	3.622E-02	–0.00278
TKT	1.3329	1.133-1.833	1.878E-02	0.00218

independent prognostic predictors. KM curve in Figure 4A revealed patients with a lower pathologic stage could obtain a better prognosis, which was consistent with their actual prognosis.

Moreover, stratified analysis was conducted for patients in differential pathologic stage, stage I-II, and stage III-IV. After risk grouping, KM curve analysis showed (Figure 5) patients in low-risk group exhibited a better prognosis than high-risk group individuals (Stage I-II, P = 9.697e-05, HR:2.560 [1.569-4.177]; Stage III-IV, P = 1.582e-02, HR:2.168 [1.139-4.125])

In the nomogram model, 4 clinical factors were included, pathologic stage I -IV, PS status, and 3- and 5-year survival times. As shown in Figure 4A, total points in nomograms integrated clinical factors to predict survival probability of each individual at 3-and 5-year times. The diagnostic value of the nomogram model was performed by comparison of the nomogram-predicted probability of OS and actual OS. The C-index

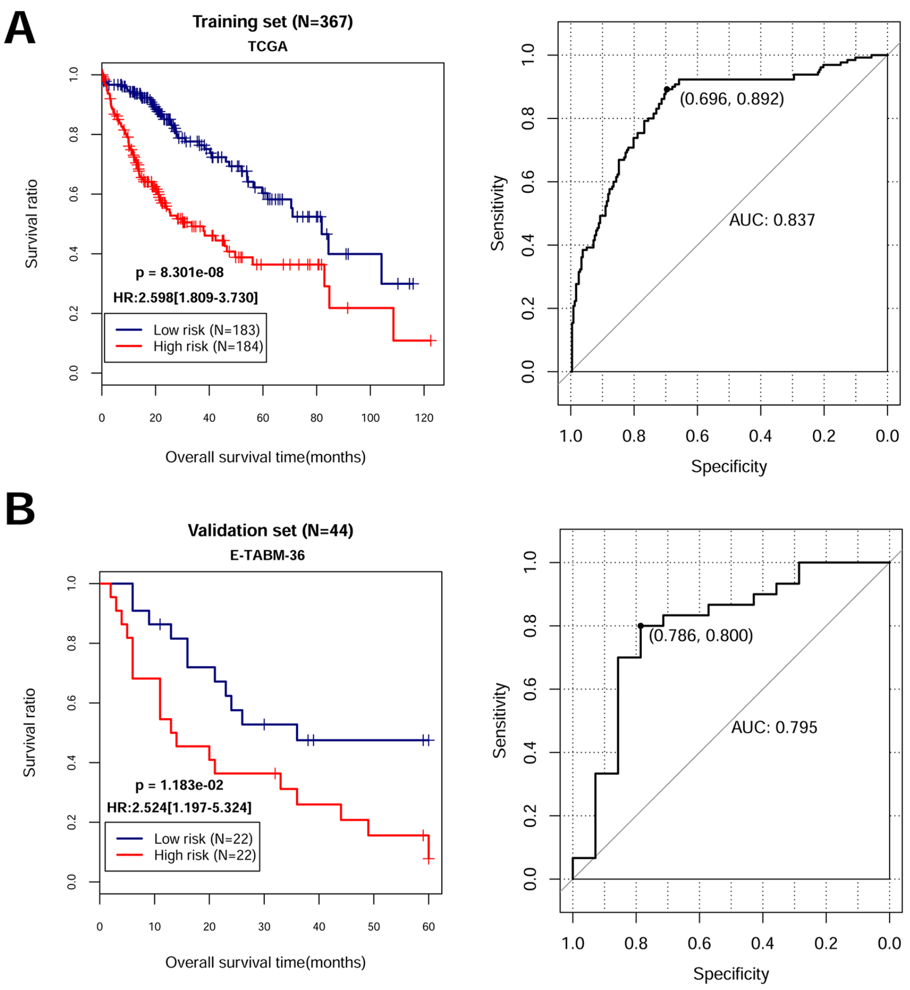

Figure 3. Kaplan-Meier curves for liver cancer patients in TCGA (A) and E-TABM-36 dataset (B).
A-B left: Overall survival rates were analyzed based using the prognostic score prediction model. Blue and red color curves indicate low- and high-risk group patients, respectively.
A-B right: Receiver operating characteristic (ROC) curve analysis for the prognostic score model. The data in brackets represents the specificity and sensitivity of the ROC curve.

Table 3. Univariate and multivariate analysis result to identify independent prognostic clinical factors for liver cancer patients.

CLINICAL CHARACTERISTICS	TCGA (N = 367)	UNI-VARIABLE COX		MULTI-VARIABLE COX	
		HR (95% CI)	P VALUE	HR (95% CI)	P VALUE
Age (years, mean ± sd)	59.67 ± 13.33	1.012 [0.998-1.026]	8.662E-02	–	–
Gender (Male/Female)	248/119	0.801 [0.562-1.142]	2.193E-01	–	–
Pathologic M (M0/M1/-)	264/3/100	4.054 [0.974-12.90]	5.122E-02	–	–
Pathologic N (N0/N1/-)	249/4/114	2.017 [0.494-8.231]	3.185E-01	–	–
Pathologic T (T1/T2/T3/T4/-)	181/92/78/13/3	1.683 [1.404-2.018]	7.867E-09	1.352 [0.640-2.855]	4.290E-01
Pathologic stage (I / II / III / IV /-)	171/85/83/4/24	1.670 [1.362-2.048]	4.468E-07	1.512 [1.219-1.875]	1.680E-04
Histologic grade (G1/G2/G3/G4/-)	55/176/119/12/5	1.114 [0.881-1.408]	3.687E-01	–	–
Vascular invasion (Yes/No/-)	107/206/54	1.333 [0.880-2.019]	1.733E-01	–	–
Radiation therapy (Yes/No/-)	8/338/21	0.979 [0.311-3.086]	9.717E-01	–	–
Recurrence (Yes/No/-)	141/179/47	1.342 [0.892-2.018]	1.572E-01	–	–
PS status (High/Low)	183/184	2.598 [1.809-3.730]	8.301E-08	2.259 [1.522-3.354]	5.250E-05
Death (Dead/Alive)	130/237	–	–	–	–
Overall survival time (months, mean ± sd)	27.37 ± 24.42	–	–	–	–

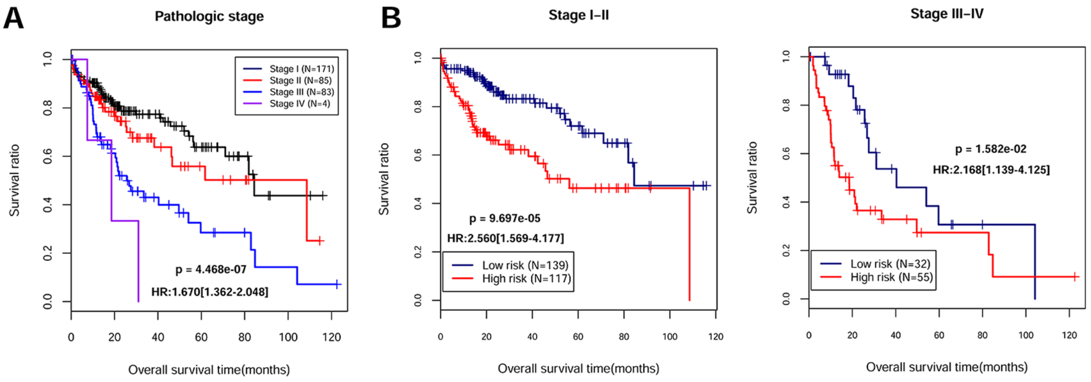

Figure 4. Construction and assessment of nomogram model for 2 independent clinical factors, pathologic stage, and prognostic model status. (A) Nomogram model estimating the 3- and 5-year survival probability of independent prognostic factors. (B) The calibration curve for comparison of the nomogram-predicted probability of overall survival time (OS) and actual OS. The horizontal axis represents the predicted OS survival rate and the vertical axis referred to the actual OS survival rate. Red and black represent the 3- and 5-year predicted line charts, respectively. (C) Receiver operating characteristic (ROC) curve for the accuracy assessment of the differential prognostic models. Black, red and blue, respectively, represent the mRNA model, Stage model, and the mRNA-stage model.

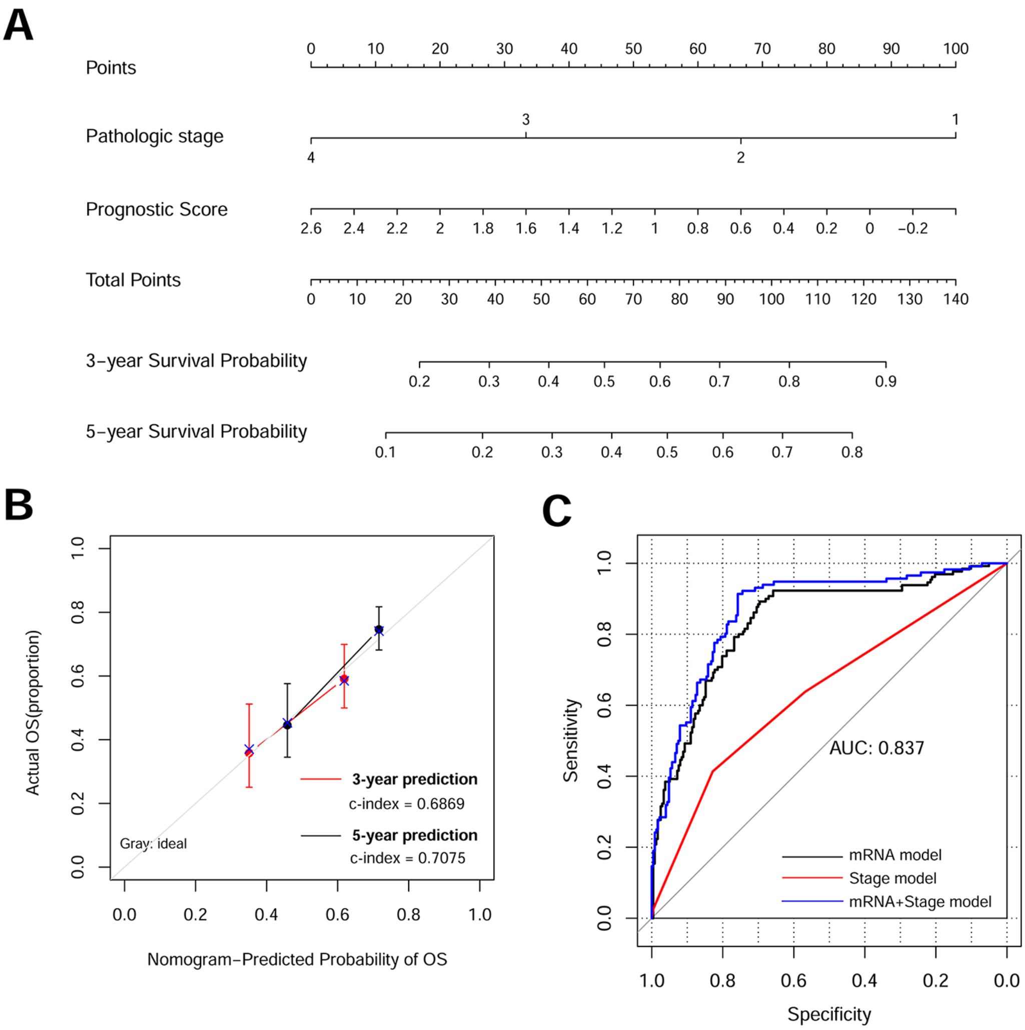

Figure 5. Stratified analysis results of liver cancer patients in different pathologic stage, including Stage I-IV patients. (A) Kaplan Meier curve shows the overall survival time of TCGA samples in different pathology stages. Black, red, blue, and purple color represent Stage I-IV group samples, respectively. (B) Kaplan Meier curve shows the overall survival time of differential pathologic stage patients sorted by prognostic score model. Blue and red color indicate low- and high-risk samples.

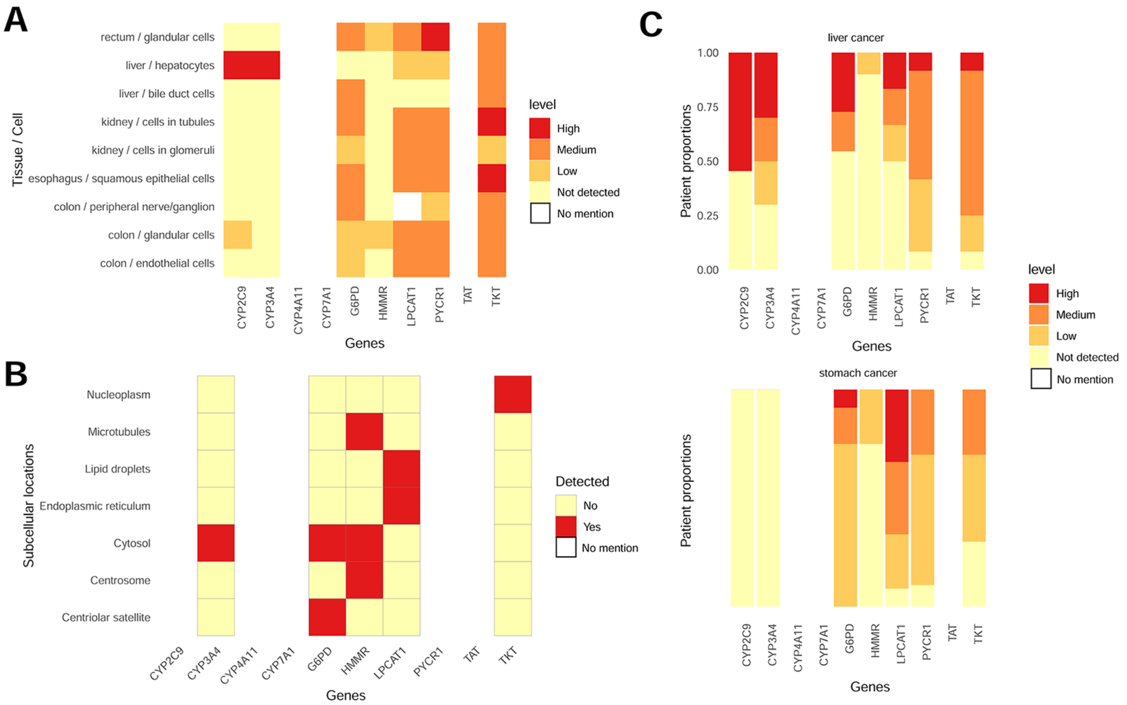

Figure 6. Data mining to explore the protein expression of target genes in various cancer types and cell types.
(A) Heatmap of protein expression levels in differential tissues and cells. (B) Subcellular location data from HPA analysis showing protein expression of ten target genes. (C) Protein expression data of genes in liver cancer and stomach cancer based on HPA data analysis.

value of 3- and 5-year survival time was 0.6869 and 0.7075, respectively (Figure 4B). ROC curve showed the predictive performance for liver cancer by using 2 models, stage model and mRNA model (Figure 4C). The AUC value of ROC curve was 0.837, indicating high accuracy of the prognostic model.

Protein expression of crucial target genes associated with liver cancer

We investigated the protein expression data of target genes in different cancer tissues and cell types from the HPA analysis. The expression status of 7 proteins was frequently altered in normal compared with tumor tissues, such as liver cancer and stomach cancer samples (Figure 6A and C). Three genes (CYP4A11, CYP7A1, and TAT) were not identified in the HPA database. Subcellular location data from the HPA analysis is also shown (Figure 6B) to visualize the expression status of these target genes.

Discussion

We have screened numerous DEGs between tumor tissue samples and normal specimens from TCGA dataset. After integration with metabolism-related genes from GSEA, we finally identified 305 overlapped genes associated with liver cancer metabolism. Functional analysis showed that these genes were enriched in several biological processes and signaling pathways, such as "steroid hormone biosynthesis," "arachidonic acid metabolism," and "linoleic acid metabolism." Based on the PS model, we obtained ten optimal signature DEGs with prognostic value, including CYP2C9, CYP3A4, and TKT.

Among these genes, CYP2C9, CYP3A4, CYP4A11, and CYP7A1 belong to the Cytochrome P450 (CYPs) superfamily. CYPs enzymes metabolize nearly 60% of prescribed drugs; CYP3A4 is responsible for half of these drug interactions.[32] Although CYP3A4 is mainly expressed in the liver (adult hepatocytes) and small intestine, it also could be transcriptionally induced by multiple xenochemicals.[33-35] Regulation of CYP3A4 is complex since numerous transcription factors could interact with the promoter region, and contribute to hepatic-specific expression of this gene, such as C/EBPα-β, CAR and PXR.[35,36] Recently, estrogen receptor alpha (ESR1) was also identified as a major transcription factor of CYP3A4, and involved in the regulation of xenobiotics metabolism in human liver.[37] By using immunohistochemistry methods, Fanni et al revealed that CYP3A4 serves as a major metabolic factor of sorafenib, and was present in surrounding hepatocytes in most cases of clinical samples, indicating the potential prognostic predictor roles for HCC management.[38] Furthermore, CYP3A4 was reported as a novel tumor suppressor gene predicted poor prognosis of HCC.[39] Similarly, in this study, our results found CYP3A4 expressed at mostly medium to high level in liver hepatocytes. Based on the PS model, we also identified it as a crucial gene with prognostic prediction value for HCC patients, which is consistent with previous studies. In addition to liver cancer, the dysregulation of CYP3A4 was also found in gastric cancer samples compared with chronic atrophic gastritis cases, which indicates a correlation between CYP3A4 expression and carcinogenic transformation of gastric cancer.[40] Moreover, a similar trend was explored in the influence of CYP2C9 expression. The hepatic CYP2C9 proteins account for nearly 20% of CYP content, and CYP2C9 functions as a major enzyme that participates in the metabolism of various drugs,[41] including warfarin, non-steroidal anti-inflammatory drugs, phenytoin, etc.. This protein was also associated with the bioactivation of carcinogens.[41] A recent study elucidated the mechanism of CYP2C9 regulation; suppression of CYP2C9 by hsa-miR-128-3p was significantly correlated with HCC pathogenesis.[42] By using bioinformatics analysis, another study revealed that CYP2C9 promoted the development of HCC and might serve as a potential prognostic marker for liver cancer.[43] Taken together, these findings suggested CYP3A4 and CYP2C9 protein might be useful and reliable biomarkers for prognosis prediction of HCC.

In addition, metabolism-related genes of TKT were also identified as candidate genes associated with HCC prognosis. The molecular mechanism of TKT remains unclear in liver cancer. It is well known that cancer cells experience increasing oxidative stress and metabolic reprogramming; a previous study reported that TKT-encoded transketolase protein counteracted oxidative stress to drive cancer development by regulating NAPDH production.[44] Qin et al. revealed that TKT can promote HCC progression both in metabolic and a non-metabolic manner through nuclear localization and EGFR signaling pathway.[45] Moreover, in HBV-related HCC, the virus induced higher levels of SH2D5 that are capable of binding to TKT and leading to the promotion of cancer cell proliferation.[46]

Univariate and multivariate analysis showed that 2 clinical characteristics (pathologic stage and PS model status) were associated with the prognosis of HCC. To assess the predictive ability of clinical factors, a nomogram model was constructed with 4 variables: pathologic stage, PS model status, and 3- and 5-year survival probability. A previous study reported that a multiple factor-based nomogram combined with prognostic score could predict the survival of gastric cancer patients with adjuvant chemotherapy.[47] The factors included inflammatory, nutritional, and preoperative prognostic markers. Recently, Huang et al. designed a nomogram model consistent with factors of molecular markers and TNM staging, and this model could predict the overall survival of HCC patients.[48] However, the present nomogram in our study showed the AUC value of 2 factor-based ROC was 0.837, which is high specificity and sensitivity. Therefore, the development of a nomogram with 2 clinical factors may facilitate the prediction of overall survival for HCC patients.

There are some limitations in this study. For example, the number of individuals was small, and more patients derived from multiple clinical centers should be used to validate the performance of the prediction model. Second, the molecular mechanisms of optimum DEGs also need to be further explored and experimentally validated.

In summary, this study developed a novel useful prognostic model for the identification of biomarkers and independent clinical factors with prognostic values in HCC. CYP2C9, CYP3A4, and TKT were correlated to the prognosis of liver cancer patients; these genes might be reliable tumor biomarkers for HCC prognosis prediction.

Author contribution

SX, JY and MS designed the research; HY and SC acquired the data; XL, HY and MS analyzed and interpreted data. SX obtained the funding. JY drafted the manuscript. SX revised the manuscript for important content. All authors read and approved the final version of the manuscript.

REFERENCES

1. He J, Gu D, Wu X, et al. Major causes of death among men and women in China. *N Engl J Med*. 2005;353:1124-1134.
2. Siegel RL, Miller KD, Jemal A. Cancer statistics, 2020. *CA Cancer J Clin*. 2020;70:7-30.
3. Alter MJ. Epidemiology of hepatitis C virus infection. *World J Gastroenterol*. 2007;13:2436-2441.
4. White DL, Kanwal F, El–Serag HB. Association between nonalcoholic fatty liver disease and risk for hepatocellular cancer, based on systematic review. *Clin Gastroenterol Hepatol*. 2012;10:1342-135900.
5. Marrero JA. Current treatment approaches in HCC. *Clin Adv Hematol Oncol*. 2013;11(suppl 5):15-18.
6. Raza A, Sood GK. Hepatocellular carcinoma review: current treatment, and evidence-based medicine. *World J Gastroenterol*. 2014;20:4115-4127.
7. Skawran B, Steinemann D, Weigmann A, et al. Gene expression profiling in hepatocellular carcinoma: upregulation of genes in amplified chromosome regions. *Mod Pathol*. 2008;21:505-516.
8. Chow EY-C, Zhang J, Qin B, Chan T. Characterization of hepatocellular carcinoma cell lines using a fractionation-then-sequencing approach reveals nuclear-enriched HCC-associated lncRNAs. *Front Genet*. 2019;10:1081.
9. Marquardt JU, Seo D, Andersen JB, Gillen MC, Thorgeirsson SS. Sequential transcriptome analysis of human liver cancer indicates late stage acquisition of malignant traits. *J Hepatol*. 2014;60:346-353.
10. Xia Q, Li Z, Zheng J, et al. Identification of novel biomarkers for hepatocellular carcinoma using transcriptome analysis. *J Cell Physiol*. 2019;234:4851-4863.
11. Jiang W, Zhang L, Guo Q, et al. Identification of the pathogenic biomarkers for hepatocellular carcinoma based on RNA-seq analyses. *Pathol Oncol Res*. 2019;25:1207-1213.
12. Miao R, Luo H, Zhou H, et al. Identification of prognostic biomarkers in hepatitis B virus-related hepatocellular carcinoma and stratification by integrative multi-omics analysis. *J Hepatol*. 2014;61:840-849.
13. Kim SM, Leem SH, Chu IS, et al. Sixty-five gene-based risk score classifier predicts overall survival in hepatocellular carcinoma. *Hepatology*. 2012;55:1443-1452.
14. Helen P, Misha K, Nikolay K, et al. ArrayExpress update—from an archive of functional genomics experiments to the atlas of gene expression. *Nucleic Acids Res*. 2009;37:D868–D872.
15. Ritchie ME, Belinda P, Di W, et al. limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res*. 2015;7:247.
16. Wang L, Cao C, Ma Q, et al. RNA-seq analyses of multiple meristems of soybean: novel and alternative transcripts, evolutionary and functional implications. *BMC Plant Biol*.2014;14:169.
17. Eisen MB, Spellman PT, Brown PO, Botstein D. Cluster analysis and display of genome-wide expression patterns. *Proceedings of the National Academy of Sciences*. 1998;95:14863-14868.
18. 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 USA*. 2005;102:15545-15550.
19. Da WH, Sherman BT, Lempicki RA. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat Protoc*. 2008;4:44-57.
20. Huang DW, Sherman BT, Lempicki RA. Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. *Nucleic Acids Res*. 2009;37:1–13.

21. Wang P, Wang Y, Hang B, Zou X, Mao JH. A novel gene expression-based prognostic scoring system to predict survival in gastric cancer. *Oncotarget*. 2016;7:55343-55351.
22. Goeman JJ. Penalized estimation in the cox proportional hazards model. *Biom J*. 2010;52:70-84.
23. Tibshirani R. The lasso method for variable selection in the cox model. *Stat Med*. 1997;16:385-395.
24. Eng KH, Schiller E, Morrel K. On representing the prognostic value of continuous gene expression biomarkers with the restricted mean survival curve. *Oncotarget*. 2015;6:36308-36318.
25. Chen Y, Jiang S, Lu Z, et al. Development and verification of a nomogram for prediction of recurrence-free survival in clear cell renal cell carcinoma. *J Cell Mol Med*. 2020;24:1245-1255.
26. Yao Z, Zheng Z, Ke W, et al. Prognostic nomogram for bladder cancer with brain metastases: a National Cancer Database analysis. *J Transl Med*. 2019;17:411.
27. Xie W, Liu J, Huang X, et al. A nomogram to predict vascular invasion before resection of colorectal cancer. *Oncol Lett*. 2019;18:5785-5792.
28. Pontén F, Jirström K, Uhlen M. The human protein atlas – A tool for pathology. *J Pathol*. 2008;216:387-393.
29. Uhlén M, Björling E, Agaton C, et al. A human protein atlas for normal and cancer tissues based on antibody proteomics. *Mol Cell Proteomics*. 2005;4(12): 1920-1932.
30. Uhlén M, Fagerberg L, Hallström BM, et al. Tissue-based map of the human proteome. *Science*. 2015;347:1260419.
31. Tran AN, Dussaq AM, Kennell T, Willey CD, Hjelmeland AB. HPAanalyze: an R package that facilitates the retrieval and analysis of the Human Protein Atlas data. *BMC Bioinformatics*. 2019;20:463.
32. Schwab M. Cytochrome P450 enzymes in drug metabolism: regulation of gene expression, enzyme activities, and impact of genetic variation. *Pharmacol Ther*. 2013;138:103-141.
33. Kamdem LK, Meineke I, Gödtel-Armbrust U, Brockmöller J, Wojnowski L. Dominant contribution of P450 3A4 to the hepatic carcinogenic activation of aflatoxin B_1. *Chem Res Toxicol*.2006; 19:577-586.
34. Watanabe M, Kumai T, Matsumoto N, et al. Expression of CYP3A4 mRNA is correlated with CYP3A4 protein level and metabolic activity in human liver. *J Pharmacol Sci*. 2004;94:459-462.
35. Martinez-Jimenez C, Jover R, Teresa Donato M, Castell J, Jose Gomez-Lechon M. Transcriptional regulation and expression of CYP3A4 in hepatocytes. *Curr Drug Metab*. 2007;8:185-194.
36. RodríguezAntona C, Bort R, Jover R, et al. Transcriptional regulation of human CYP3A4 basal expression by CCAAT enhancer-binding protein α and hepatocyte nuclear factor-3γ. *Mol Pharmacol*. 2003;63:1180-1189.
37. Wang D, Lu R, Rempala G, Sadee W. Ligand-free estrogen receptor alpha (ESR1) as master regulator for the expression of CYP3A4 and other cytochrome P450 enzymes in the human liver. *Mol Pharmacol*. 2019;96(4): 430-440.
38. Fanni D, Manchia M, Lai F, Gerosa C, Ambu R, Faa G. Immunohistochemical markers of CYP3A4 and CYP3A7: a new tool towards personalized pharmacotherapy of hepatocellular carcinoma. *Eur J Histochem*. 2016;60:2614.
39. Ashida R, Okamura Y, Ohshima K, Kakuda Y, Yamaguchi K. CYP3A4 gene is a novel biomarker for predicting a poor prognosis in hepatocellular carcinoma. *Cancer Genomics Proteomics*. 2017;14:445-453.
40. Zhang F, Wang F, Chen C, et al. Prediction of progression of chronic atrophic gastritis with Helicobacter pylori and poor prognosis of gastric cancer by CYP3A4. *J Gastroenterol Hepatol*. 2020;35:425-432.
41. Miners JO, Birkett DJ. Cytochrome P4502C9: an enzyme of major importance in human drug metabolism. *Br J Clin Pharmacol*. 1998;45:525-538.
42. 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. *Scie Rep*. 2015;5:8534.
43. Shuaichen L, Guangyi W. Bioinformatic analysis reveals CYP2C9 as a potential prognostic marker for HCC and liver cancer cell lines suitable for its mechanism study. *Cell Mol Biol (Noisy-le-Grand, France)*. 2018;64:70-74.
44. Xu IM, Lai RK, Lin SH, et al. Transketolase counteracts oxidative stress to drive cancer development. *Proc Natl Acad Sci USA*. 2016;113:25.
45. Qin Z, Xiang C, Zhong F, et al. Transketolase (TKT) activity and nuclear localization promote hepatocellular carcinoma in a metabolic and a non-metabolic manner. *J Exp Clin Cancer Res*. 2019;38:154.
46. Zheng Y, Ming P, Zhu C, et al. Hepatitis B virus X protein-induced SH2 domain-containing 5 (SH2D5) expression promotes hepatoma cell growth via an SH2D5-transketolase interaction. *J Biol Chem*. 2019;294:4815-4827.
47. Liu X, Wu Z, Lin E, et al. Systemic prognostic score and nomogram based on inflammatory, nutritional and tumor markers predict cancer-specific survival in stage II–III gastric cancer patients with adjuvant chemotherapy. *Clin Nutrit*. 2019;38:1853-1860.
48. Xi-Tai Huang, Liu-Hua, et al. Establishment of a nomogram by integrating molecular markers and tumor-node-metastasis staging system for predicting the prognosis of hepatocellular carcinoma. *Dig Surg*. 2019;36:426-432.