

# Integrative analysis identifies key mRNA biomarkers for diagnosis, prognosis, and therapeutic targets of HCV-associated hepatocellular carcinoma

Yongqiang Zhang<sup>1,2,\*</sup>, Yuqin Tang<sup>3,\*</sup>, Chengbin Guo<sup>4</sup>, Gen Li<sup>4</sup>

<sup>1</sup>Molecular Medicine Center, West China Hospital, Sichuan University, Chengdu 610041, P.R. China

<sup>2</sup>West China School of Medicine, West China Hospital, Sichuan University, Chengdu 610041, P.R. China

<sup>3</sup>School of Basic Medical Sciences, Chengdu University of Traditional Chinese Medicine, Chengdu 611137, P.R. China

<sup>4</sup>Guangzhou Women and Children's Medical Center, Guangzhou Medical University, Guangzhou 510623, P.R. China

\*Equal contribution

**Correspondence to:** Gen Li; email: [superleegen@hotmail.com](mailto:superleegen@hotmail.com), <https://orcid.org/0000-0003-3625-2210>

**Keywords:** hepatocellular carcinomas, hepatitis C virus, differentially expressed genes, biomarkers, WGCNA

**Received:** December 30, 2020

**Accepted:** March 23, 2021

**Published:** May 4, 2021

**Copyright:** © 2021 Zhang et al. This is an open access article distributed under the terms of the [Creative Commons Attribution License](https://creativecommons.org/licenses/by/3.0/) (CC BY 3.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

## ABSTRACT

Hepatitis C virus-associated HCC (HCV-HCC) is a prevalent malignancy worldwide and the molecular mechanisms are still elusive. Here, we screened 240 differentially expressed genes (DEGs) of HCV-HCC from Gene expression omnibus (GEO) and the Cancer Genome Atlas (TCGA), followed by weighted gene coexpression network analysis (WGCNA) to identify the most significant module correlated with the overall survival. 10 hub genes (CCNB1, AURKA, TOP2A, NEK2, CENPF, NUF2, CDKN3, PRC1, ASPM, RACGAP1) were identified by four approaches (Protein-protein interaction networks of the DEGs and of the significant module by WGCNA, and diagnostic and prognostic values), and their abnormal expressions, diagnostic values, and prognostic values were successfully verified. A four hub gene-based prognostic signature was built using the least absolute shrinkage and selection operator (LASSO) algorithm and a multivariate Cox regression model with the ICGC-LIRI-JP cohort (N = 112). Kaplan-Meier survival plots ( $P = 0.0003$ ) and Receiver Operating Characteristic curves (ROC = 0.778) demonstrated the excellent predictive potential for the prognosis of HCV-HCC. Additionally, upstream regulators including transcription factors and miRNAs of hub genes were predicted, and candidate drugs or herbs were identified. These findings provide a firm basis for the exploration of the molecular mechanism and further clinical biomarkers development of HCV-HCC.

## INTRODUCTION

Globally, Hepatocellular carcinoma (HCC) represents the predominant histological form of liver cancer (accounting for 75%-85% of all cases), which is a commonly diagnosed malignancy with an increasing incidence rate and ranked fourth in mortality among all cancers [1]. In 2018, HCC leads to more than 781,000 deaths and about 841,000 newly diagnosed cases all over the world [1]. Hepatitis C virus (HCV) infection is one of the major causes of HCC, especially in western

countries and Japan [1]. According to a survey of 227,808 participants, the anti-HCV-positive rate was 3.0%, but more than 60% of the participants were not aware of their infection [2]. While the introduction of the vaccine has reduced the prevalence of Hepatitis B virus (HBV) infection with promise to decrease the incidence of HBV-associated HCC (HBV-HCC) in certain high-risk countries, there is no vaccine available for HCV infection [1]. On the other hand, although great advances have been achieved for the investigation of HCC in the last decades, its underlying mechanisms of different

etiologies vary dramatically, therefore extensive efforts are still needed to establish a better understanding of carcinogenesis and pathogenesis of HCV-associated HCC (HCV-HCC).

Recently, a growing number of candidate biomarkers for diagnosis or prognosis of HCC have been identified [3–12], among which the most commonly reported biomarkers are dysregulated genes [3, 6, 11], significant members of a certain gene family or gene set [4, 10], potential CpG methylation status [7, 9], and alternative splicing signatures [5, 12]. For example, a 24-mRNA-based risk signature has been developed as an independent risk classifier for the prediction of early recurrence in HCC patients [6]. Similarly, a nine immune-related mRNA signature was generated to predict the overall survival (OS) of HCC [10]. While most of the studies focused on HCC prognosis, its diagnosis has not yet been fully investigated. Besides, few studies characterized the stratified categorization by different risk factors (especially HCV infection), however, they may exert contrary outcomes even for the same risk group. Thus, additional markers are required for a more accurate risk prediction in HCV-HCC patients.

Of note, single cohort-based studies may result in false-positive outcomes because of the small sample size and limitation of technology platforms. Therefore, an integrated analysis combining multiple public databases such as The Cancer Genome Atlas (TCGA), The Gene Expression Omnibus (GEO), and International Cancer Genome Consortium (ICGC) could improve the accuracy and reliability of the results tremendously, providing an effective approach for the exploration of molecular landscape and the discovery of potential therapeutic targets or important biomarkers for diagnosis and prognosis of cancer. Thus, with the aim to identify the candidate crucial genes for diagnosis and prognosis of HCV-HCC from multiple public databases, which might also give a clue for seeking therapeutic targets in HCV-HCC, we enrolled eight gene expression datasets from TCGA, GEO, and ICGC, including a total of 304 HCV-HCC samples and 290 adjacent normal tissues in the present study. 240 differentially expressed genes (DEGs) were screened in the first step, followed by the identification of 10 hub genes with a combined analysis. Then, the diagnostic and prognostic values of these hub genes were verified. The least absolute shrinkage and selection operator (LASSO)-based penalized Cox regression (LASSO-COX) was performed to construct a prognostic risk signature, which was further evaluated by Kaplan-Meier curves and ROC plots. The relationships between the risk signature and tumor infiltration immune cells were also determined by Spearman correlation analysis. Moreover, Upstream regulations of the 10 hub

genes including miRNAs and transcription factors were also predicted. At last, network pharmacological analysis was conducted to seek possible small molecular drugs for HCV-HCC. Collectively, this study identified 10 hub genes concerning the crucial roles in the carcinogenesis of HCV-HCC, which may provide a firm basis for understanding the transcriptional regulatory mechanisms and advancing studies in clinical biomarker discovery of HCV-HCC. The flowchart summarizing the general process of this study was shown in Figure 1.

## RESULTS

### Screening of robust DEGs in HCV-HCC

By using GEO2R and the screening criteria of  $|\log \text{Fold change (FC)}| > 1$  and  $\text{FDR (adj.P.Val)} < 0.05$ , we extracted 1722 DEGs (842 upregulated and 880 downregulated) from GSE6764, 1459 DEGs (496 upregulated and 963 downregulated) from GSE41804, 1761 DEGs (1050 upregulated and 711 downregulated) from GSE62232, and 1163 DEGs (276 upregulated and 887 downregulated) from GSE107170. In the TCGA dataset, we fetched 3740 DEGs (1468 upregulated and 2272 downregulated) between HCV-HCC and normal tissues with the same threshold. As shown in Figure 2A, 2B, a total of 240 overlapping DEGs were identified, including 58 commonly upregulated genes, and 182 commonly downregulated genes. To increase the robustness of these common DEGs, we integrated the four microarray datasets into a combined dataset. The Combat function embedded in sva package was used to remove the batch effect. Plots of the Principal component analysis (PCA) indicated that after expression normalization, the batch effect was all removed successfully (Figure 2C, 2D). In addition, tumor samples and normal samples were clustered independently after batch removal (Figure 2E). Differential analysis by limma package revealed that all the 240 DEGs were still significant in the combined dataset (Figure 2F and Supplementary Table 2).

### Co-expression network construction and identification of the most important module

WGCNA is a useful approach to uncover gene expression patterns and to identify significant gene modules from multiple samples. We conducted WGCNA to disclose the most important module associated with HCV-HCC survival status. Briefly, 807 DEGs of the ICGC-LIRI-JP dataset were filtered (Supplementary Table 3), which were used to evaluate the outlier samples through sample hierarchical clustering using the average linkage method (Figure 3A). After the filtration, we obtained the adjacency matrix by using the appropriate soft threshold of 5 (scale-free  $R^2 = 0.87$ ), which was

transformed into the TOM, and transited into the dissTOM, followed by the accomplishment of the gene clustering dendrogram and module identification (Figure 3B). Highly similar modules were then merged by the cut line of 0.3. Seven modules were remained (Figure 3C). The Pearson correlation heatmap showed the turquoise module including 357 DEGs has the most significant correlation with survival status and thus was selected for further study (Figure 3D). Figure 3E presented the GS and MM for each gene in the turquoise module.

### PPI network construction

We constructed a PPI network with the 240 overlapping DEGs using the STRING online database and the Cytoscape software (Supplementary Figure 1). The network gave 129 nodes and 585 edges, and showed 41

upregulated genes and 88 downregulated genes. The average number of neighbors was 9.07 and the clustering coefficient was 0.461. Using the MCODE app, a significant sub-cluster was screened out with a cluster score of 29.5, comprising 30 nodes and 428 edges (Figure 4A). Interestingly, all of the 30 genes showed high degrees of connectivity by cytohuber analysis ( $>20$  for all cluster genes), indicating their potential hub roles for HCV-HCC tumorigenesis.

Besides, the 357 genes in the turquoise module by WGCNA were also used to construct a PPI network to identify candidate hub genes. The WGCNA-PPI network was composed of 245 nodes and 2581 edges (Figure 4B). There were 50 genes satisfied with the degree cutoff of  $\geq 50$  and defined as WGCNA-PPI-hub genes.

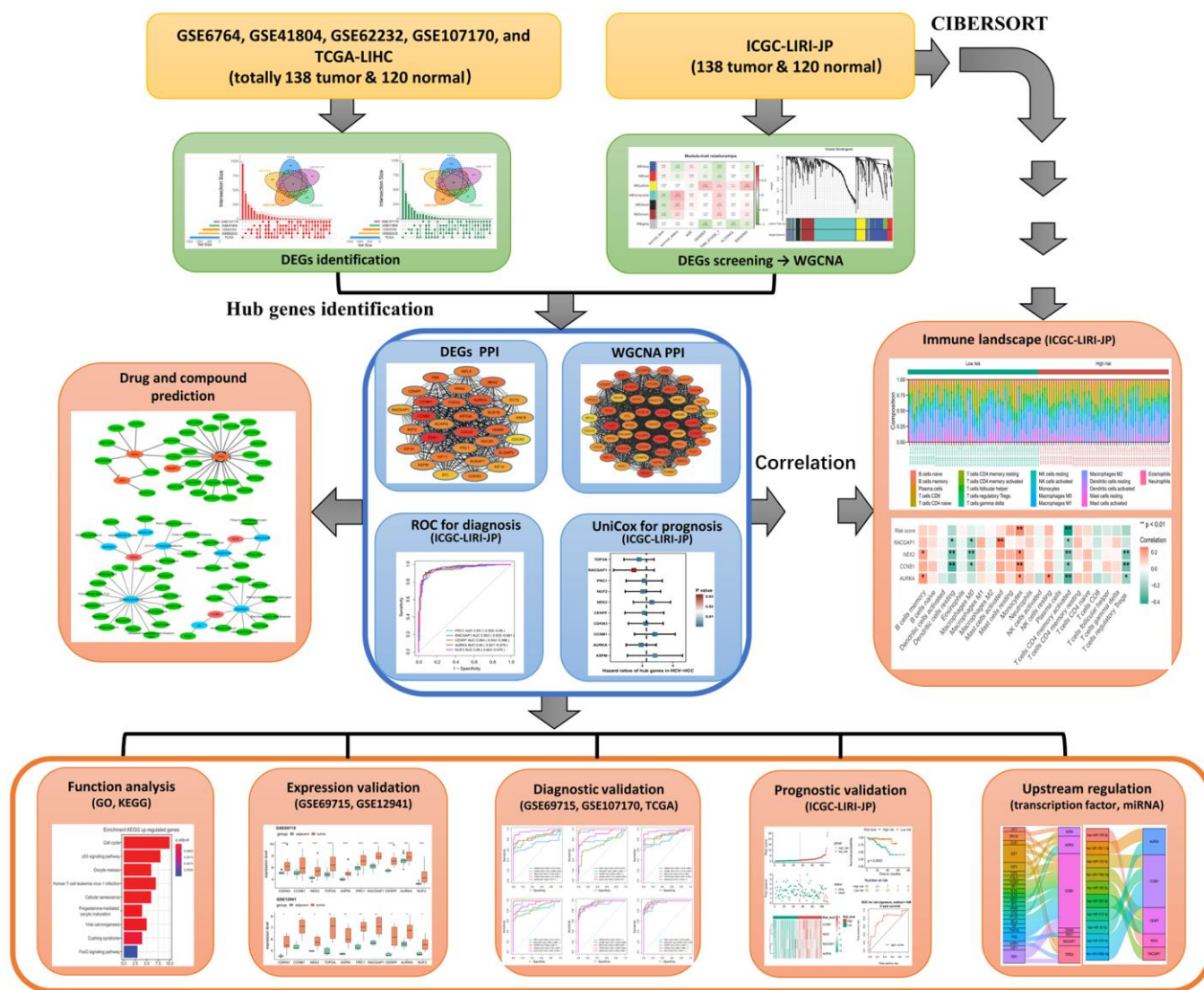
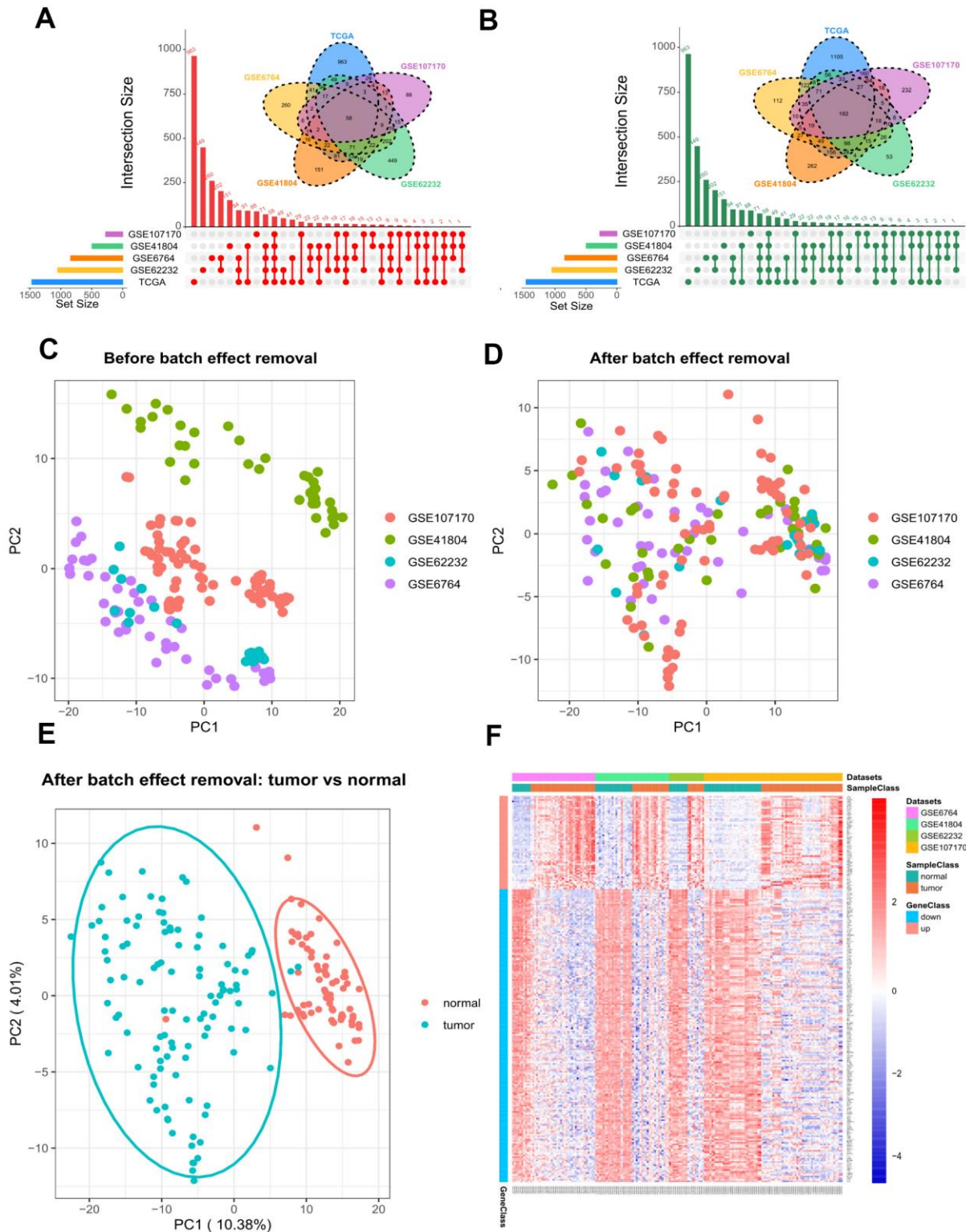
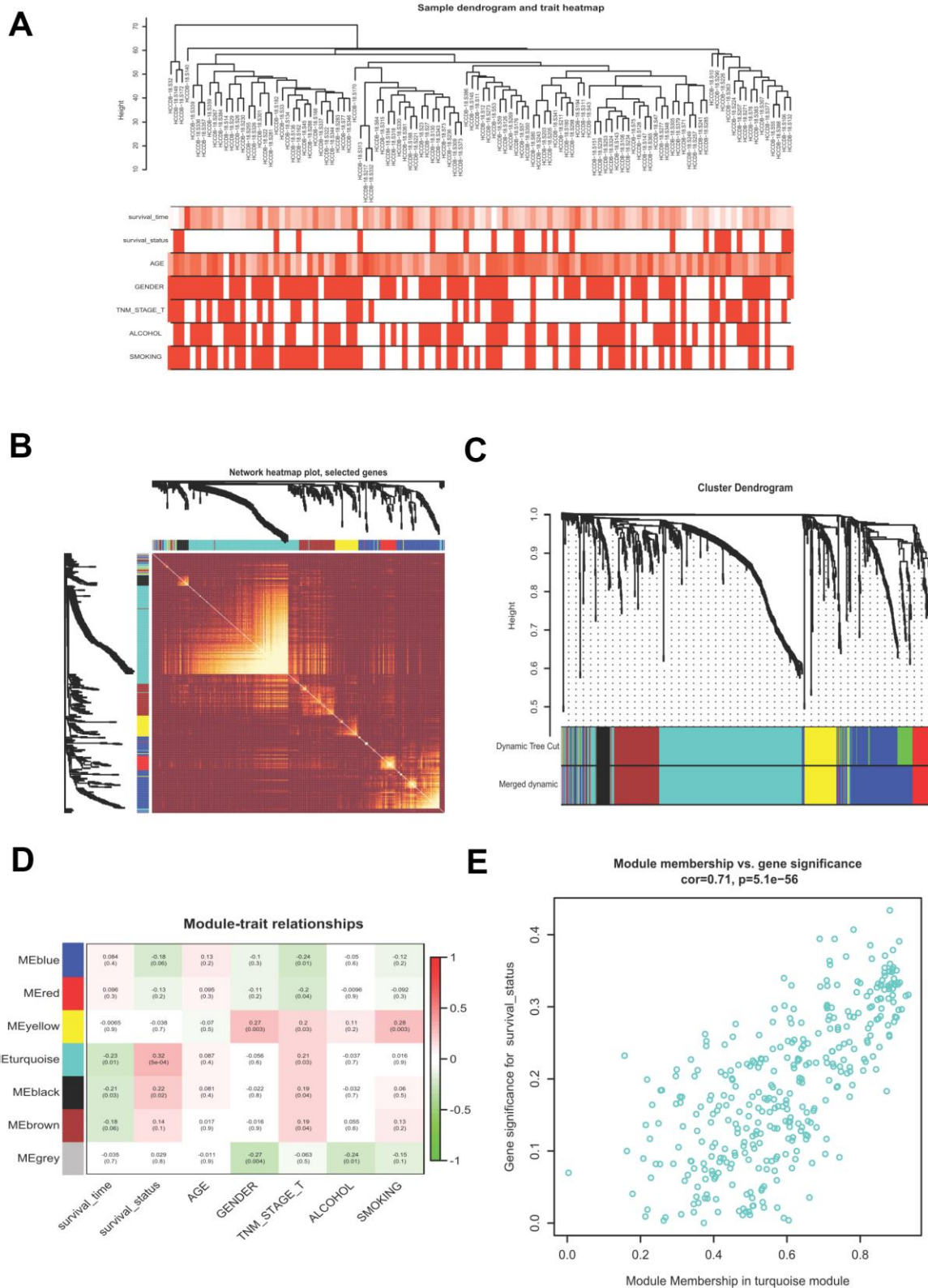


Figure 1. Flowchart of this study.



**Figure 2. Differential gene expression between HCV-HCC tumor and adjacent normal tissues. (A, B)** The combination of Venn plot and Upset plot showing the common upregulated genes (A) and the common downregulated genes (B) in HCV-HCC according to five public datasets. The screening criteria was set as  $|\log \text{ Fold change (FC)}| > 1$  and  $\text{FDR (adj.P.Val)} < 0.05$ . (C, D) Principal component analysis (PCA) for the gene expression profiles from four microarray datasets before (C) and after (D) batch effect removal. The colors represent different datasets. (E) scatter plots visualizing the identified clusters of the tumor and normal samples based on the combined dataset. (F) Heatmap of the 240 DEGs showing their expression values for each patient. The scale bar indicates the gene expression value. Red indicates high expression level, and blue indicates low expression level. HCV-HCC, HCV-associated HCC. DEGs, differentially expressed genes.



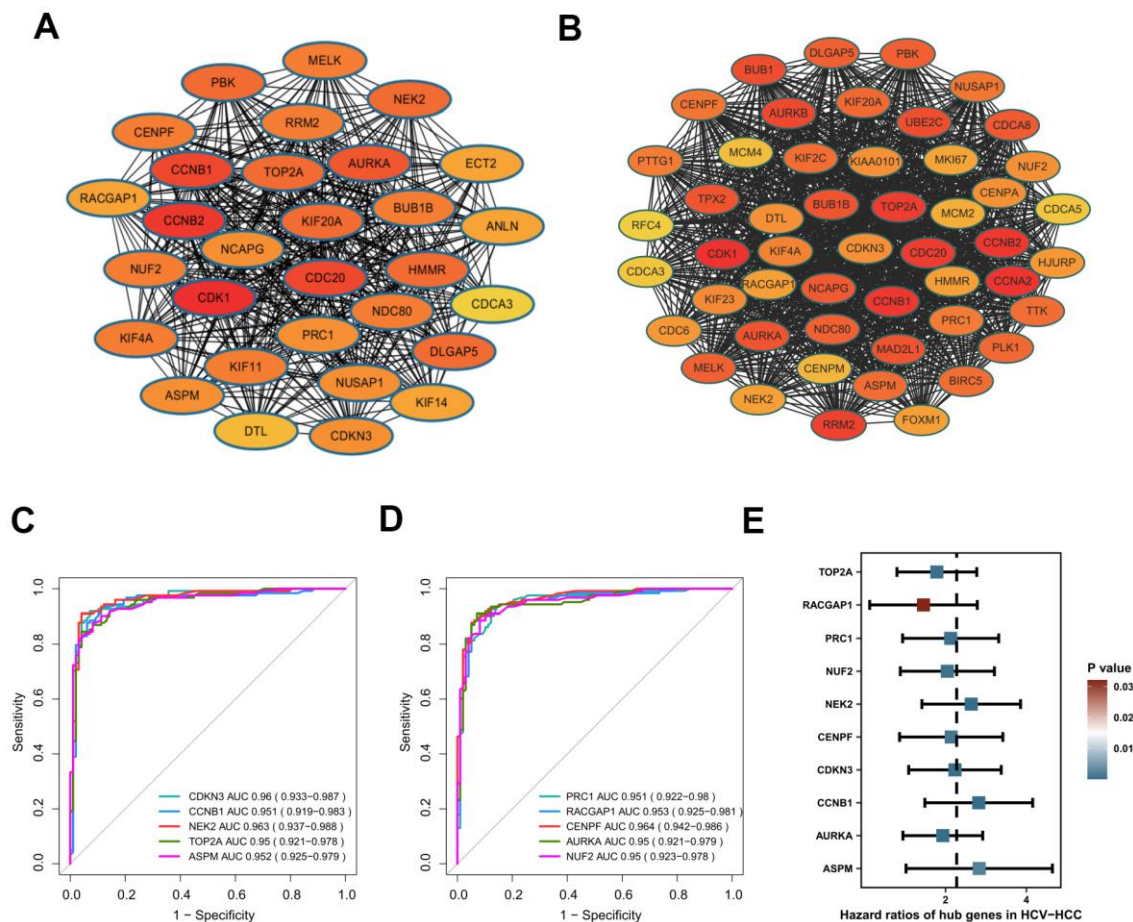
**Figure 3. Building a WGCNA network to identify the most significant module correlated with survival status.** (A) Sample clustering tree with clinical traits. (B) Heatmap showing the eigengene networks according to the topological overlap matrix (TOM) based dissimilarity. (C) Gene clustering dendrogram, with each color corresponding to an individual gene module. (D) Pearson correlation analysis between module eigengenes and clinical traits. (E) scatter plot showing the gene significance (GS) vs module membership (MM) for the turquoise module. WGCNA, Weight Gene Co-expression Network Analysis.

## Hub genes identification

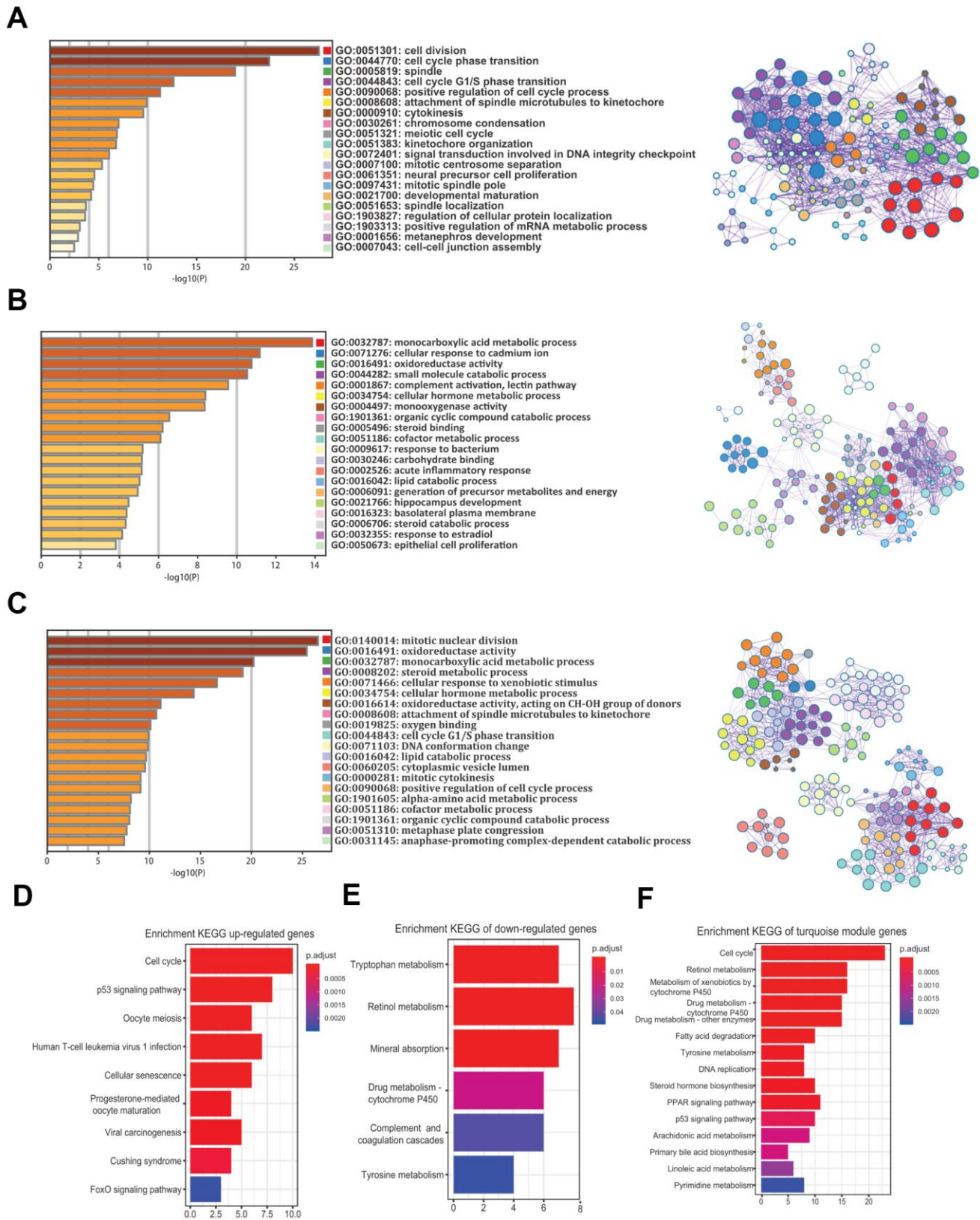
Based on the 30 DEGs-PPI-hub genes and the 50 WGCNA-PPI-hub genes, we preliminarily obtained a total of 26 overlapping genes (data not shown). Then we evaluated the AUROC scores of the 26 genes for discriminating HCV-HCC from normal tissue samples using the ICGC-LIRI-JP dataset. As a result, 10 genes (CCNB1, AURKA, TOP2A, NEK2, CENPF, NUF2, CDKN3, PRC1, ASPM, RACGAP1) showed superior discriminatory abilities with AUROC scores of  $\geq 0.95$  (Figure 4C, 4D), suggesting their excellent diagnostic values. More importantly, all of the 10 genes were also revealed significantly associated with the overall survival outcome of HCV-HCC patients by UniCox analysis, indicating their potential prognostic powers in clinical use (Figure 4E). Thus, we consider these 10 genes as hub genes in HCV-HCC.

## Functional enrichment analysis

To understand the biological functions of the robust DEGs and the turquoise module in HCV-HCC, we performed GO and KEGG analysis. GO enrichment analysis revealed that the commonly 58 upregulated genes were mostly involved in cell division, cell cycle phase transition, spindle, and other important GO terms, mainly related to cell proliferation (Figure 5A). The 182 commonly downregulated genes were mainly related to the monocarboxylic acid metabolic process, cellular response to cadmium ion, and oxidoreductase activity (Figure 5B). For the 357 genes in the turquoise module, mitotic nuclear division, oxidoreductase activity, and monocarboxylic acid metabolic process were the top GO terms (Figure 5C). On the other hand, KEGG analysis suggested that the most KEGG pathways associated with the upregulated genes were



**Figure 4. Identification of hub genes in HCV-HCC.** (A) The most significant cluster identified from the DEGs-PPI network. (B) The WGCNA-PPI network constructed by the turquoise module. (C, D) ROC curves showing the AUROC scores and AUC (95%CI) of the 10 hub genes for discriminating tumor from normal samples based on the ICGC-LIRI-JP dataset. Colored lines indicate the ROC curve for each hub gene, and the grey line indicates the reference line. (E) Forest plot presenting the results of the univariate Cox regression analysis for the 10 hub genes. HCV-HCC, HCV-associated HCC. DEGs, differentially expressed genes. PPI, protein-protein interaction. WGCNA, Weight Gene Co-expression Network Analysis. ROC, receiver operating characteristic. 95%CI, 95% confidence interval.

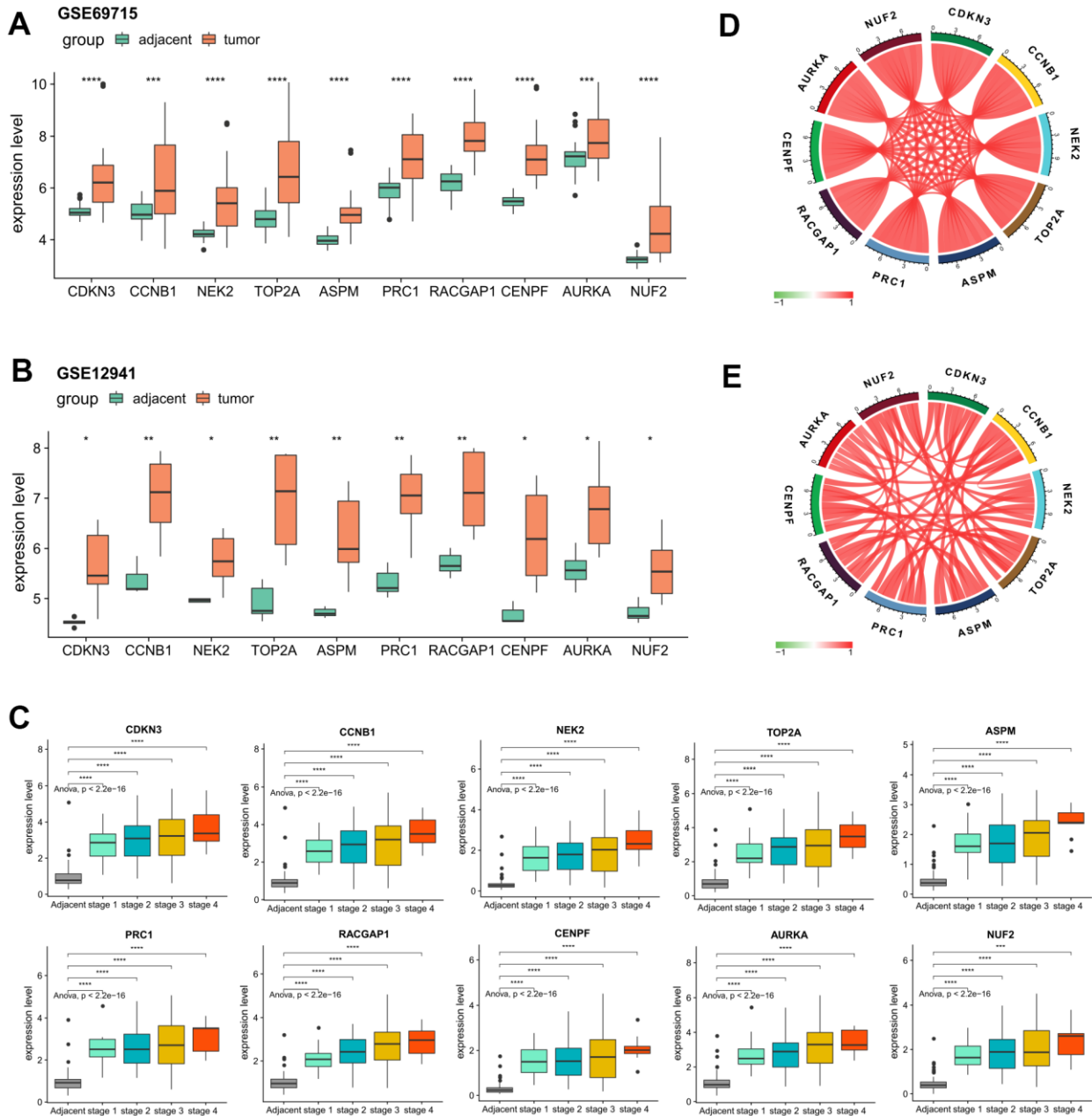


**Figure 5. GO and KEGG analysis of the 240 common DEGs and the turquoise module. (A–C)** GO enrichment analysis for the upregulated genes (A), downregulated genes (B), and the turquoise module (C) (Top 20 are shown). **(D–F)** Enrichment of KEGG pathways for the upregulated genes (D), downregulated genes (E), and the turquoise module (F). GO, gene ontology. KEGG, Kyoto Encyclopedia of Genes and Genomes. DEGs, differentially expressed genes.

cell cycle, p53 signaling pathway, and oocyte meiosis (Figure 5D), while the tryptophan metabolism, retinol metabolism, and mineral absorption were the top three pathways for the downregulated genes (Figure 5E). Moreover, the turquoise module was mostly associated with cell cycle, retinol metabolism, and metabolism of xenobiotics by cytochrome P450 (Figure 5F).

### Hub genes expression validation

For the validation of the expression patterns, based on the external validation datasets of GSE69715 and GSE12941, we observed significantly elevated gene expression levels of every hub genes in tumor samples compared with that of the adjacent normal samples (Figure 6A, 6B). A closer examination of the internal



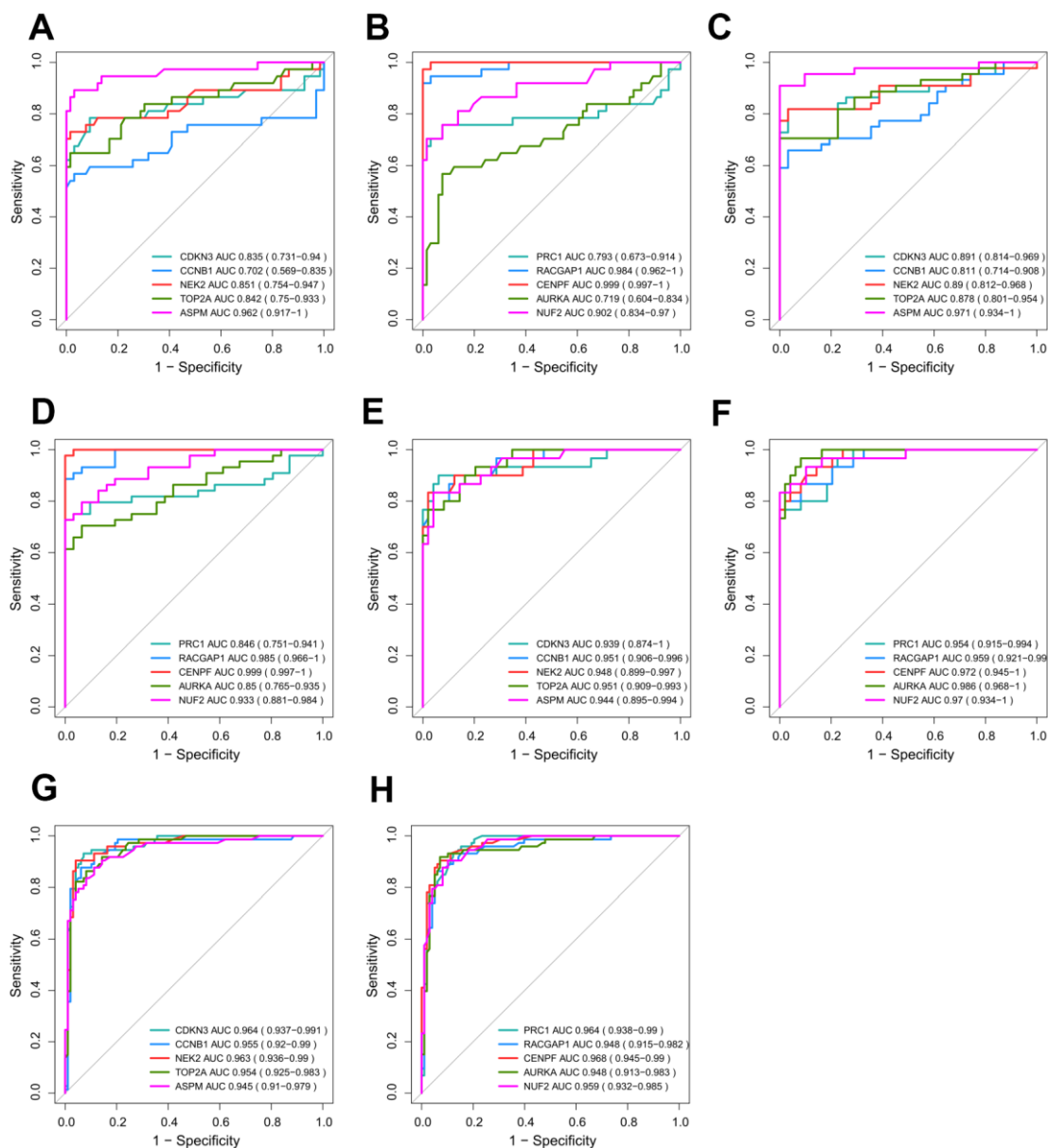
**Figure 6. Confirmation of the abnormal expression of the 10 selected hub genes and their expression correlations. (A, B)** Two external datasets (GSE69715 and GSE12941) to validate the increased expression levels of the hub genes in tumors compared with adjacent normal tissues. **(C)** Internal validation by ICGC-LIRI-JP dataset to verify the elevated levels of the hub genes concerning tumor stage. **(D, E)** Strong correlations among all of the hub genes according to the ICGC-LIRI-JP and TCGA-LIHC datasets. \* $P < 0.05$ , \*\* $P < 0.01$ , \*\*\* $P < 0.001$ , \*\*\*\* $P < 0.0001$ .



validation set of ICGC-LIRI-JP showed that the dysregulations of all the hub genes were statistically significant regardless of the tumor stage (Figure 6C), indicating the robustness of their crucial roles in tumor initiation of HCV-HCC. Strikingly, it was determined that in both ICGC-LIRI-JP and TCGA datasets, the relative expression levels of the hub genes were highly correlated with each other (Pearson correlation coefficient  $> 0.75$  for all gene pairs in both ICGC-LIRI-JP and TCGA-LIHC), suggesting their strong interactions and key roles in the development of HCV-HCC (Figure 6D, 6E).

## Validation of the diagnostic value

We presume that excellent discrimination capability may have great potential for cancer diagnosis to benefit HCV-HCC patients. Thus, we validated the performance of hub genes by plotting ROC curves of GSE69715, GSE107170, and TCGA-LIHC (Figure 7A–7F). Two hub genes (CENPF and RACGAP1) showed consistently high AUROC scores in all three datasets ( $>0.95$ ), indicating their potential utility as diagnostic biomarkers. Moreover, we used the internal validation set of ICGC-LIRI-JP to assess the distinguishing



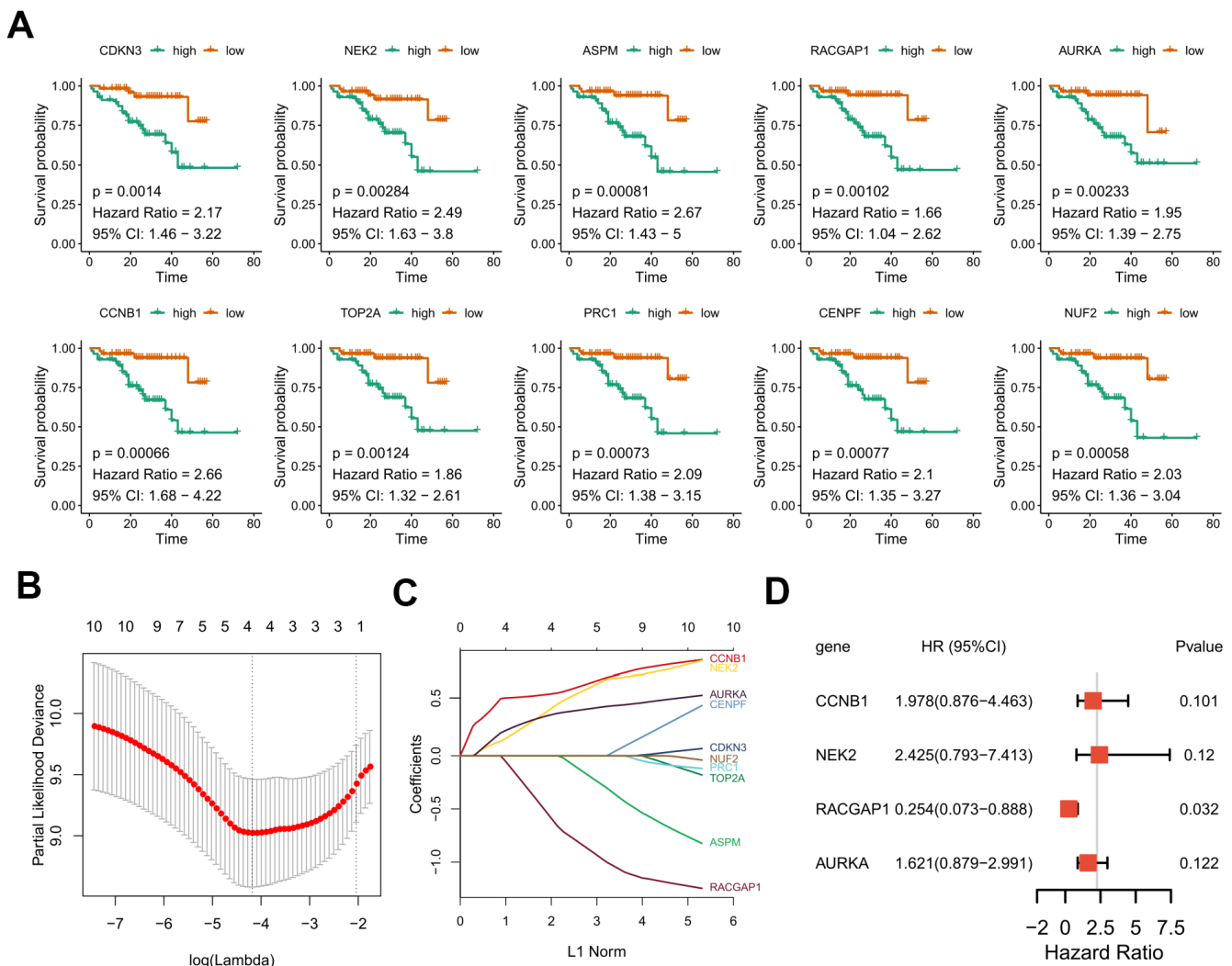
**Figure 7. Validation of the diagnostic efficiency for each of the 10 hub genes.** (A–F) Performance of the 10 hub genes in discriminating HCV-HCC from normal control based on GSE69715 (A, B), GSE107170 (C, D), and TCGA-LIHC (E, F). (G, H) Potential utilities of the hub genes for early tumor detection based on ICGC-LIRI-JP. HCV-HCC, HCV-associated HCC.

abilities of the hub genes for early phase tumor samples from adjacent normal tissue samples (Figure 7G, 7H). Surprisingly, ROC curves by all the hub genes revealed their great potential for early detection of HCV-HCC (AUROC score > 0.94 for each hub gene).

### Survival analysis

Due to the limited sample sizes of other datasets, we were only able to include the ICGC-LIRI-JP cohort that contained more than 100 HCV-HCC patients with adequate survival information to conduct the survival analysis (N = 112). Kaplan–Meier curves indicated that the overall survival of the high-risk group was significantly lower than that of the low-risk group

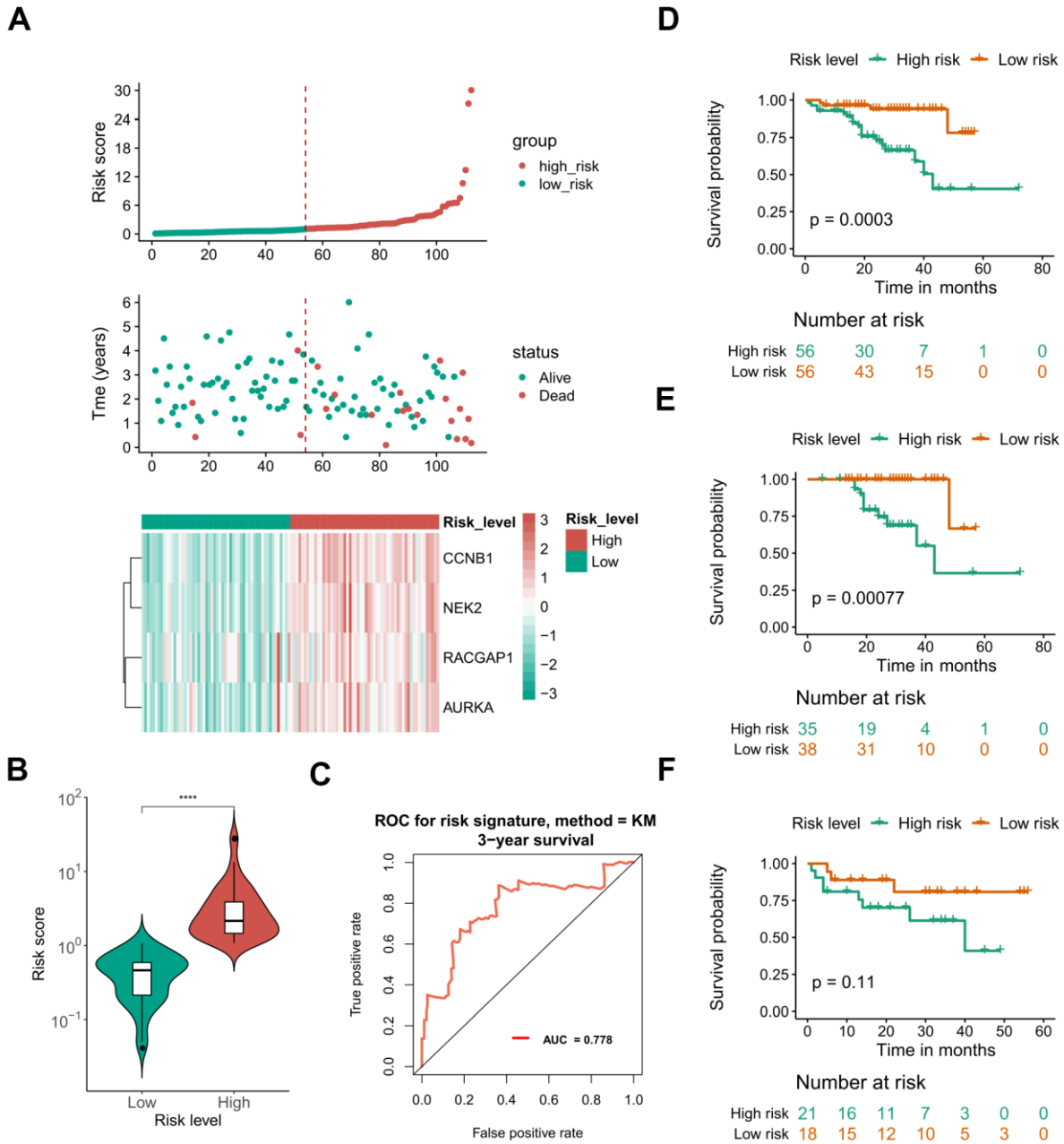
( $P < 0.01$  for all hub genes, Figure 8A). Furthermore, the LASSO-COX regression was used to reduce the variables with 10-fold cross-validation for the selection of the optimal turning parameter (Figure 8B). At the minimum lambda value, four hub genes were chosen with non-zero coefficients, including CCNB1, NEK2, RACGAP1, and AURKA (Figure 8C), which were next used to perform the multivariate Cox hazards regression analysis (Figure 8D). A risk signature was then generated to evaluate the risk score of HCV-HCC patients with the following formula: risk score =  $0.6819 \cdot \text{EXP}_{\text{CCNB1}} + 0.8859 \cdot \text{EXP}_{\text{NEK2}} - 1.3715 \cdot \text{EXP}_{\text{RACGAP1}} + 0.4831 \cdot \text{EXP}_{\text{AURKA}}$ . Patients were divided into the high- or low-risk groups according to the median risk score of 0.8822715 (Figure 9A). A significantly higher risk score



**Figure 8. Kaplan–Meier curves for overall survival of the 10 selected hub genes and construction of a prognostic signature using LASSO Cox regression.** (A) OS Kaplan–Meier curves of the 10 hub genes based on ICGC-LIRI-JP. (B) 10-fold cross-validation to select the optimal tuning parameter. The  $\lambda$  value of 0.015 was chosen with the lambda.min method. (C) LASSO coefficient profiles of the 10 hub genes. (D) Forest plot presenting the hazard ratio and 95% CI by multivariate Cox regression analysis for the four selected hub genes. OS, overall survival. LASSO, Least absolute shrinkage and selection operator. 95% CI, 95% confidence interval.

was observed in the high-risk group than that of the low-risk group (Figure 9B). The ROC curve at 3 years overall survival showed the area under the curve (AUC) value of 0.778 (Figure 9C), indicating a good predictive performance for the OS of HCV-HCC. Kaplan-Meier survival plots suggested the relatively poor survival in the high-risk group (Figure 9D). Besides, we carried out the stratified analysis using clinical parameters.

Consequently, in almost all subsets of patients with different age, gender, vein invasion status, alcohol consumption, and smoking status, the four-hub gene-based risk signature was still a significant prognostic factor (Supplementary Figure 2). Although the TNM staging system was considered as an important prognostic factor for HCC patients, conflict survival outcomes may exist for patients at the same stage. Therefore, we also

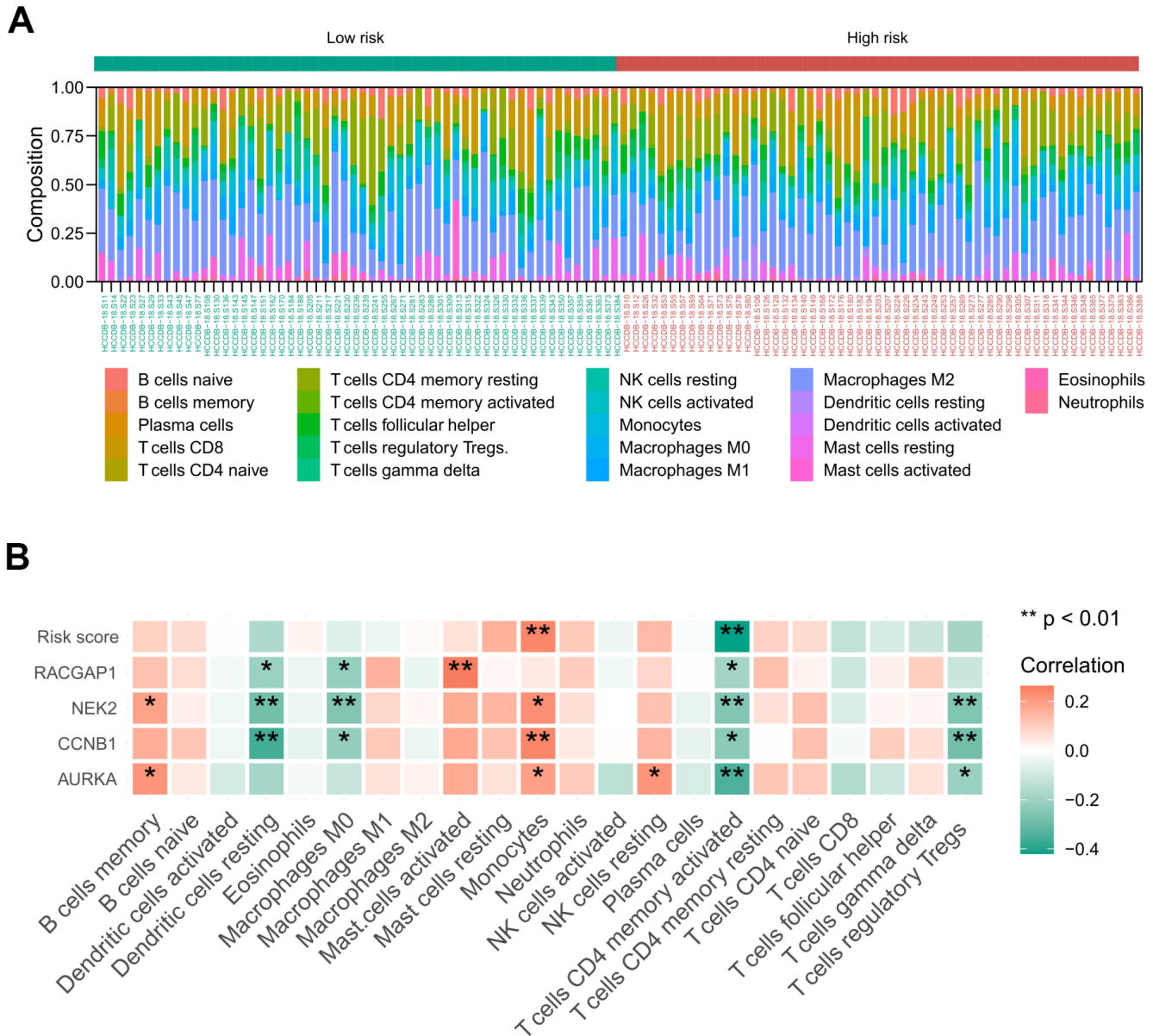


**Figure 9. Performance of the defined four mRNA-based risk signature with ICGC-LIRI-JP.** (A) Gene expression, risk score, and clinical outcome for all the patients in distinctive risk groups. (B) differential risk scores between high- and low-risk groups. (C) ROC plot at 3 years OS showing the AUROC score of 0.778. (D) OS Kaplan-Meier survival curves for high- and low-risk patients. (E, F) OS Kaplan-Meier survival curves for different risk groups of early stage (E) and advanced stage patients (F). \*\*\*\*,  $P < 0.0001$ . OS, overall survival. ROC, receiver operating characteristic. AUROC, the area under the receiver operating characteristic curve.

performed the stratification survival analysis based on the TNM stage. Notably, patients in the low-risk group possessed a better OS compared with the high-risk group in the early stage subset ( $N = 73, P < 0.01$ ) (Figure 9F), while no significant difference was observed for the advanced stage of HCV-HCC ( $N = 39, P = 0.11$ ) (Figure 9F). Besides, we also conducted the univariate Cox analysis to evaluate the other underlying risk factors, however, no significant associations were observed at a statistical level of 0.05, which might partly due to the small sample size.

### The risk signature was associated with the abundance of immune infiltration cells

Based on the ICGC-LIRI-JP cohort, we achieved the landscape of the 22 tumor immune infiltration cells for HCV-HCC via the CIBERSORT algorithm (Figure 10A). Then the Spearman correlation coefficient and corresponding  $P$  value between risk score and infiltration level of each immune cell were calculated. As a result, monocytes were positively associated with the risk score and the expression of *NEK2*, *CCNB1*, and

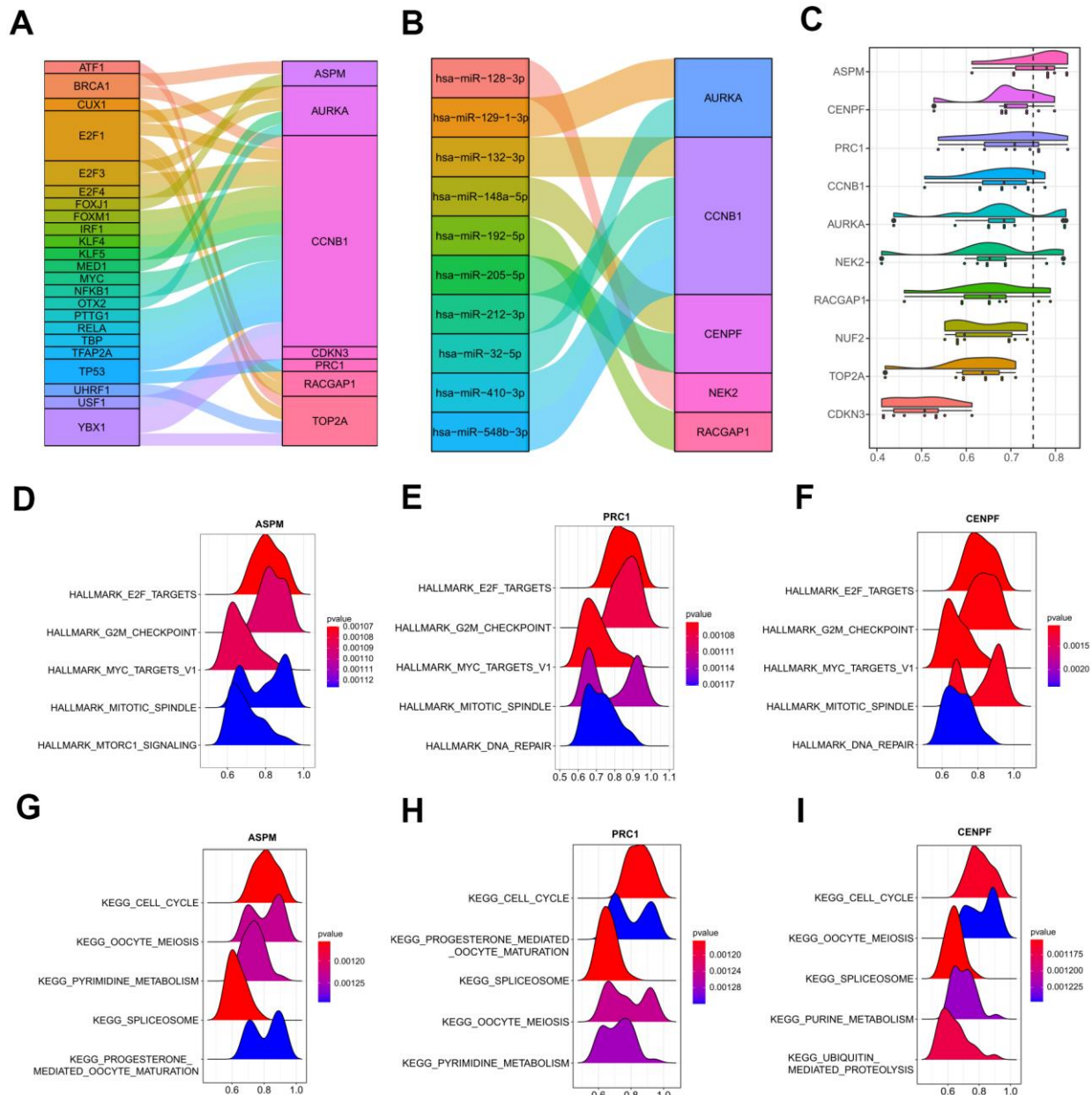


**Figure 10. Relationship between the identified risk signature and tumor immune cell infiltration based on the ICGC-LIRI-JP cohort.** (A) The landscape of immune infiltration in each of the tumor samples of low- and high-risk groups. (B) Heatmap representing the correlation matrix of the four signature genes, risk score, and relative abundance of 22 immune cell types. Red indicates the positive correlation, while green indicates the negative correlation. \*  $P < 0.05$ , \*\*  $P < 0.01$ .

AURKA. Activated CD4 memory T cells displayed negative correlations with the risk score and all of the four signature hub genes. Other immune cells manifested no significant correlation with the risk score, except resting dendritic cells and M0 macrophages, which were negatively associated with the expression of RACGAP2, NEK2, and CCNB1. T cells regulatory Tregs were negatively associated with the expression of NEK2, CCNB1, and AURKA (Figure 10B).

### Prediction of upstream regulations

Next, crucial transcription factors in the upstream of the 10 hub genes were determined by the TRRUST database that was integrated into the web-based application of miRNet (Supplementary Table 4). A transcription factor-hub gene network was then constructed and visualized by a Sankey diagram. 23 transcription factors and 7 hub genes were found in this network (Figure 11A). Among



**Figure 11. Upstream regulations of the ten hub genes and GO semantic similarities analysis.** (A) The transcription factor-hub gene network predicted by miRNet. (B) 10 function MTIs predicted through miRTarBase 8.0. (C) Raincloud plot showing the ranking list of function semantic similarities for the 10 hub genes using the ICGC-LIRI-JP dataset. ASPM, CENPF, and PRC1 were the top three hub genes with the highest scores. (D–F) GSEA results of ASPM, CENPF, and PRC1 based on the hallmark gene set. (G–I) GSEA results of ASPM, CENPF, and PRC1 based on the KEGG database. GO, gene ontology. MTIs, miRNA-target interactions. KEGG, Kyoto Encyclopedia of Genes and Genomes.

all the genes, CCNB1 was the most important node with the highest degree, which was regulated by a total of 19 transcription factors, including several important tumor-associated genes, such as BRCA1, MYC, TP53, and three E2Fs family members (E2F1, E2F3, and E2F4). Moreover, miRNA-hub gene interactions were predicted by employing miRTarBase 8.0 to explore the underlying regulatory mechanism by miRNAs in human. A total of 428 miRNA-hub gene interaction pairs were obtained, including 10 function MTIs, 417 weak function MTIs, and one none function MTI (Supplementary Table 5). As illustrated in Figure 11B, among the 10 function MTIs, CCNB1 gave the highest connection degree, which was targeted by four miRNAs (hsa-miR-132-3p, hsa-miR-212-3p, hsa-miR-548b-3p, hsa-miR-410-3p) in experiment validation, suggesting its increasing correlation with human cancer. Other predictive weak function MTIs were still needed to be investigated in the future. Hence, the transcription factor-hub gene interaction network and the identified miRNA-hub gene pairs could provide insight into the further exploration of the molecular mechanisms of HCV-HCC.

### Semantic similarities and GSEA

In order to achieve a deeper and better understanding of the molecular mechanism of the 10 hub genes, we adopted the geometric mean of CCs and MFs with GO semantic similarity analysis. ASPM, CENPF, and PRC1 were the top three hub genes with the highest-ranking scores (Figure 11C), which were verified by GSEA analysis with the ICGC-LIRI-JP dataset. The top 5000 genes that correlated with the ASPM, CENPF, and PRC1 were used to operate the GSEA, and the most significant terms that enriched by the hallmark gene set (h.all.v7.0) were similar (HALLMARK\_E2F\_TARGETS, HALLMARK\_G2M\_CHECKPOINT, HALLMARK\_MYC\_TARGETS\_V1, and HALLMARK\_MITOTIC\_SPINDLE were the top four gene sets for all of the three hub genes) (Figure 11D–11F). For the pathways distribution from the KEGG database (c2.cp.kegg.v7.0), ASPM, CENPF, and PRC1 shared three important tumor-associated terms (KEGG\_CELL\_CYCLE, KEGG\_OOCYTE\_MEIOSIS, and KEGG\_SPLICEOSOME) in the top five significant pathways (Figure 11G–11I). These findings were in agreement with the above-mentioned results that hub genes may strongly interact with each other and strengthened the pivotal roles of the hub genes in the tumor initiation and development in HCV-HCC.

### Drug-hub gene network and candidate compounds identification

The aforementioned results prompted us to focus on the therapeutic utility of the hub genes. Thus, we searched

the DGIdb to obtain the potential drug-gene interactions. As a result, four (TOP2A, AURKA, NEK2, and RACGAP1) of the 10 hub genes were targeted by therapeutic drugs that were approved by FDA. Then a drug-hub gene network was established including the four hub genes and 29 anti-neoplastic drugs (Figure 12A and Supplementary Table 6). As presented in Figure 12A, TOP2A might be inhibited by most of the drugs, followed by AURKA. Among these drugs, Etoposide was supported by the largest number of literature as the cancer chemotherapeutic drug including HCC [13]. Additionally, TCM-MESH and TCM-ID were used to investigate candidate active ingredients and herbs that may target these hub genes. A network consisting of 3 hub genes (TOP2A, CCNB1, and NUF2), 9 effective compounds, and 40 related herbs was constructed (Figure 12B). Similarly, TOP2A was putatively targeted by most compounds (3,3',4',5,5',7-hexahydroxyflavone, proanthocyanidin b2, epigallocatechin 3-gallate, howiinol a, and betulic acid). Among all the compounds, proanthocyanidin b2 and plumbagin showed the top two nodes with the highest degrees (proanthocyanidin b2 was contained in 17 herbs and plumbagin was contained in 9 herbs). Proanthocyanidin b2 was well-documented with anticarcinogenic properties via anti-inflammatory and antioxidant potential, and was demonstrated to exert anti-tumor efficacy for HCC *in vitro* and *in vivo* [14]. Plumbagin was also indicated to suppress HCC carcinogenesis through induction of cell arrest and cellular apoptosis [15]. These data may shed light upon target therapy for HCV-HCC patients.

### Comparison of the hub genes and pathways between HCV-HCC and HBV-HCC

In a previous study, we reported 17 hub genes with diagnostic and prognostic values in HBV-HCC [16]. Interestingly, three of them (CCNB1, TOP2A, NEK2) were also identified as crucial genes in HCV-HCC, which might to some extent reflect the common transcriptome regulatory mechanisms in liver cancer induced by viral hepatitis. We also compared the robust DEGs between HCV-HCC and HBV-HCC, as the result, we found 38 common upregulated DEGs and 95 common downregulated DEGs. Notably, commonly important KEGG pathways enriched by robust DEGs were identified between HCV-HCC and HBV-HCC including cell cycle, p53 signaling pathway, oocyte meiosis, progesterone-mediated oocyte maturation, Human T-cell leukemia virus 1 infection, cellular senescence, retinol metabolism, tryptophan metabolism, complement and coagulation cascades, drug metabolism - cytochrome P450, tyrosine metabolism (Supplementary Figure 3). Knowledge like that may reveal indispensable and key pathways for the complete transition from hepatitis to HCC, and therefore would throw light on the

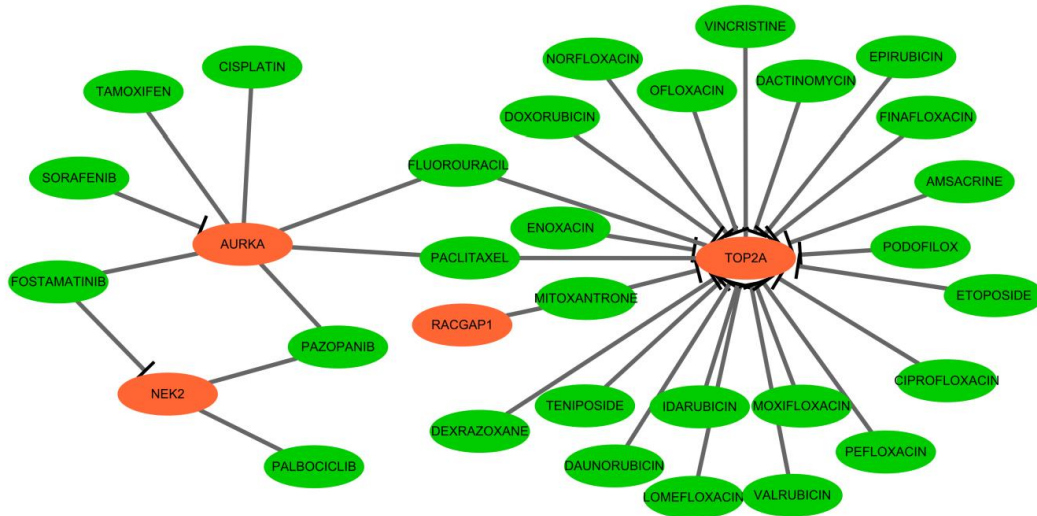
yielding of possible predictors or biomarkers during the process.

## DISCUSSION

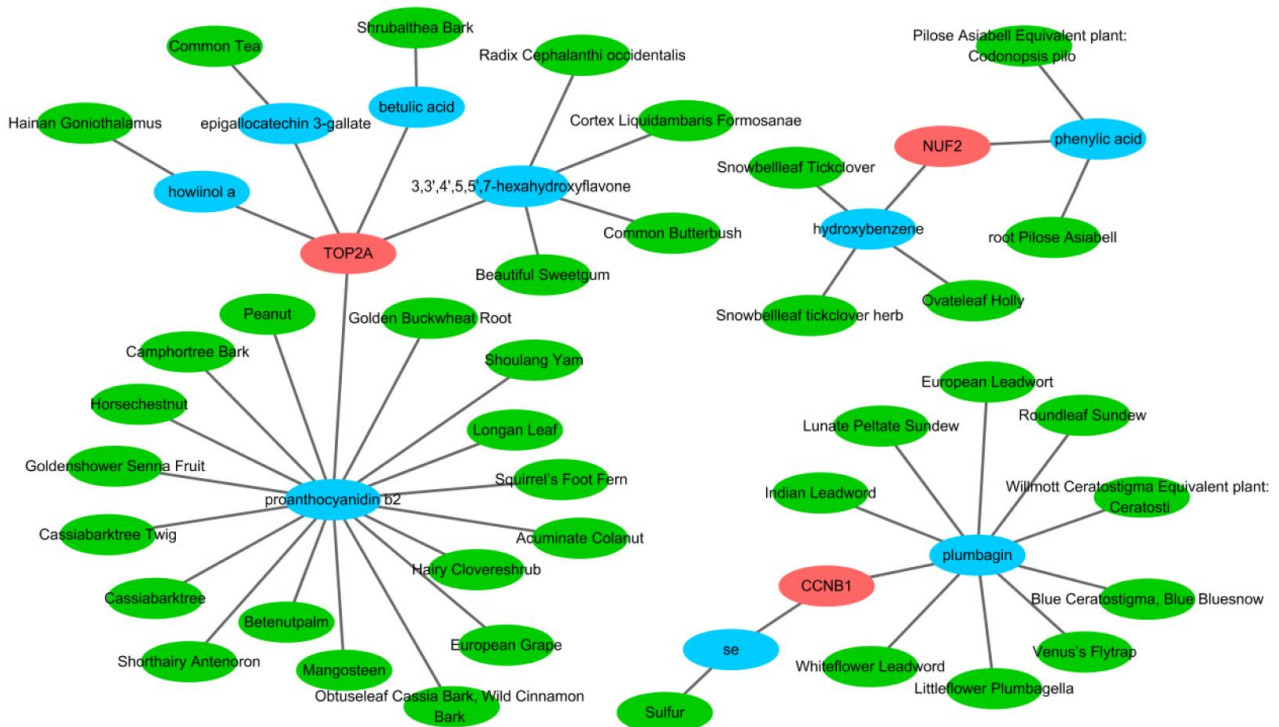
Despite intense efforts that have been made for the investigation of HCC pathogenesis and its candidate

biomarker searching, the overall prognosis for HCC patients was still unfavorable, and the comprehensive explanation for its transcriptional and genetic mechanisms remained elusive, especially for HCV associated HCC. In the current study, 240 robust DEGs of HCV-HCC were, for the first time, screened based on five public datasets, including 58 upregulated genes and

**A**



**B**



**Figure 12. Network pharmacological analysis to identify candidate drugs and effective compounds for therapeutic targets of HCV-HCC. (A) Drug-hub gene network identified from the DGIdb. Green nodes indicate the predictive miRNAs and red nodes indicate the targeted hub genes. (B) Herb-compounds-hub gene network predicted by TCM-MESH and TCM-ID. red nodes indicate hub genes, blue nodes indicate the active compounds and green nodes indicate the putative herbs containing these compounds.**

182 downregulated genes. The upregulated genes mainly participated in cell cycle-associated GO terms, such as cell division, cell cycle phase transition, and spindle. Cell-cycle aberration was considered a hallmark of cancer [17]. In the present study, those cell cycle-related genes such as CDK1, CCNB1, CDC20, NEK2, AURKA, RACGAP1, CDKN2A, CDKN2B, CDKN3, RRM2, and ASPM were significantly upregulated in HCV-HCC. Downregulated genes were mostly involved in the monocarboxylic acid metabolic process, cellular response to cadmium ion, and oxidoreductase activity. KEGG analysis revealed that, while the upregulated genes were significantly related to cell cycle, p53 signaling pathway, and oocyte meiosis, the downregulated genes were associated with tryptophan metabolism, retinol metabolism, and mineral absorption. The common pathways identified via a deeper examination between HCV-HCC and HBV-HCC may help reveal the generality of the transition from viral hepatitis to HCC.

Subsequently, 10 hub genes were identified through multiple approaches by overlapping a panel of 30 closely correlated DEGs-PPI-hub genes and 50 WGCNA-PPI-hub genes, as well as assessing their diagnostic and prognostic power. Interestingly, in comparison with the adjacent normal tissues, all of the hub genes were overexpressed in the tumor tissue samples, suggesting their tumor-driven function in oncogenesis. Next, the consistent and significant expression trend of the hub genes was validated internally and externally. Notably, the expression levels of all the hub genes were associated with each other, indicating their tight connections and pivotal roles during HCV-HCC. Moreover, ROC curves revealed that CENPF and RACGAP1 exhibited robust discrimination between tumor and normal tissue samples. More importantly, all of the hub genes displayed great potential for early diagnosis of HCV-HCC according to the ICGC-LIRI-JP dataset.

To improve the reliability of their prognostic values, we depicted the Kaplan–Meier OS survival plot for each of the hub genes, combined with a log-rank test. As a result, all the hub genes showed a negative impact on patients' OS. Furthermore, we generated a four-hub gene-based risk signature (CCNB1, NEK2, RACGAP1, and AURKA) via LASSO-COX proportional regression analysis, which was demonstrated to have excellent prognostic accuracy for OS of HCV-HCC patients. It was worth noting that even though the TNM staging system was most frequently used to predict the outcomes of HCC, our model could still offer additional value for subgroups with different TNM stages and other clinicopathological variables.

In previous studies, all of the four risk signature genes had been reported to play oncogenic roles in HCC and

their elevated expression levels were closely related to the reduced overall survival of HCC patients [18–27], which is in agreement with this study. For example, CCNB1 and AURKA were proved to be significantly associated with the prognosis of HCC and HBV-HCC and proposed as hub genes in these cancers [16, 19, 25–28]. Weng, L. et al. demonstrated that the elevated expression of CCNB1 was an independent prognostic indicator for the recurrence in HBV-HCC patients [27]. Recently, CCNB1 was used to build an mRNA risk signature to predict the prognosis of HCC [28], and AURKA was also involved in a 24 mRNA-based signature for early relapse prediction of HCC [6]. NEK2 was showed to promote tumor growth, angiogenesis, and metastasis *in vivo*. Its overexpression was found in liver tumor tissues on both mRNA and proteomic levels, and was associated with the poor survival of HCC [20]. Another study revealed RACGAP1 as a prognostic factor for early recurrence of HCC. Silencing of RACGAP1 could significantly inhibit Hep3B and MHCC97-H cell invasion and migration [22]. Other hub genes were also reported to be predictors for HCC prognosis. It is notable that TOP2A was a widely accepted hub gene in both HCC and HBV-HCC [16, 19, 28, 29], and it was also used to establish an mRNA-based signature for prognosis prediction in HBV-HCC previously [16]. Moreover, CENPF was correlated with higher cumulative recurrence rates and shorter overall survival of HCC [30, 31]. Besides, high expression levels of NUF2 [32], CDKN3 [33], ASPM [33], PRC1 [34] were also related to poor prognosis of HCC. What's important, CDKN3 was linked to the activated or inhibited cell cycle modules for the transformation of non-malignancy-associated hepatitis/cirrhosis to HCC.

Given that dynamic crosstalk among tumor immune infiltration cells in the tumor microenvironment may trigger and accelerate tumor growth or progression [35], we put emphasis on the correlations between risk score and relative abundance of immune cells. Eight putative immune cell types were found to have significant correlations with risk score or signature hub genes.

To gain insights into the upstream regulatory mechanisms of these hub genes, a transcription factor-hub gene network and 10 function MTIs were identified. Remarkably, most of the transcription factors were well-established oncogenes or tumor suppressor genes. Among the 10 experimentally verified miRNAs, most of them, including miR-128-3p [36], miR-132-3p [37], miR-148a-5p [38], miR-192-5p [39], miR-205-5p [40], miR-212-3p [41], miR-32-5p [42] were previously linked to tumor cell proliferation, invasion, migration and prognosis of HCC. Meanwhile, further researches are still required to elucidate the biological functions of the remaining three miRNAs (miR-129-1-3p, miR-410-3p,



and miR-548b-3p) on tumorigenesis and progression of HCC. Moreover, GO semantic similarity analysis and GSEA suggested that ASPM, CENPF, and PRC1 may share common molecular mechanisms during the pathogenesis of HCV-HCC.

For the evaluation of therapeutic implications of the hub genes, we carried out network pharmacological analysis. We found that four of the hub genes (TOP2A, AURKA, NEK2, and RACGAP1) can serve as tumor therapeutic targets for drugs approved by FDA. Specifically, a set of TOP2A inhibitors were determined as potential chemoprotective drugs in various types of cancer, such as doxorubicin in solid tumors, leukemias and lymphomas [43], Idarubicin in HCC [44], acute myelogenous leukemia, advanced breast cancer, multiple myeloma, non-Hodgkin's lymphoma, and other malignancies [45], and etoposide in several malignant tumors [46–50] and metastatic tumors (such as brain metastasis of breast cancer) [51, 52]. Next, we identified candidate herbs and their effective components that may have an inhibitory impact on tumor progression via three hub genes (TOP2A, NUF2, and CCNB2). Proanthocyanidin b2 and plumbagin were the most common compounds in herbs related to TOP2A and CCNB1, showing good potential for cancer treatment including HCC [14, 15].

Compared with previous studies, the present study has at least several strengths: first, most studies only enrolled one cohort or single method to screen DEGs in cancer, while a total of eight high-quality gene expression profile datasets with stringent approaches (combining the overlapping strategy and the integrating strategy) improved the robustness. Second, we performed four approaches to identify potential key genes in HCV-HCC, which was different from those derived from only one algorithm (such as PPI network or WGCNA). Third, unlike previous studies that neglected population stratification while constructing a gene signature, we focused on a specific cohort of HCC that was influenced by HCV. Furthermore, the comparison between HCV-HCC and HBV-HCC may help understand the generality and specificity of the transformation from hepatitis B or hepatitis C to HCC. Additionally, the hub gene-based drugs or effective compounds may provide new insight for targeted therapy in HCV-HCC.

Several limitations, however, should be addressed in this study. First, due to the strict patient inclusion criteria applied in this study, only one available cohort (ICGC-LIRI-JP) was included for survival analysis, which may introduce imprecision or potential bias in the evaluation of risk factors, and increase the risk of overfitting during the construction of the prognostic gene signature. Therefore, more external validation cohorts with larger sample sizes are required to validate our prognostic

signature and their relevance to immune cell infiltration. Second, more *in vitro* and *in vivo* experiments should be performed to uncover the molecular mechanisms of the predicted transcription factor-hub gene pairs and putative miRNAs that may target the hub genes during HCC tumorigenesis and cancer progression. Third, it should be noted that the candidate drugs and potential active components targeting the hub genes should be further investigated, from structural analysis (such as molecular docking) to in-depth experimental studies for functional exploration, which may help accelerate the development of novel promising drugs for target therapy of HCC.

In summary, we identified 10 hub genes, which may play crucial roles in the carcinogenesis and pathogenesis of HCV-HCC, from multiple datasets with comprehensive bioinformatics approaches. The dysregulation of the hub genes was linked to tumor diagnosis and prognosis and might serve as potential therapeutic targets of HCV-HCC patients. A risk signature was constructed for OS survival classification. A transcription factor-hub gene network and a series of targeted miRNAs were predicted. Potential drugs and candidate compounds for these hub genes were identified. All these results from the multi-dimension analysis provide a strong foundation for a better understanding of the complex transcriptional regulatory mechanisms underlying HCV-HCC, which might shed light on the discovery of potential biomarkers for early diagnosis, prognosis, and treatment for HCV-HCC patients.

## MATERIALS AND METHODS

### Data acquisition

Six gene expression profiles of HCC were selected from the GEO (<https://www.ncbi.nlm.nih.gov/geo/>) database with the GSE number of GSE6764 [53], GSE41804 [54], GSE62232 [55], GSE107170 [56], GSE12941 [57], and GSE69715 [58]. These datasets met the following strict criteria: (1) including both tumor and normal human tissues; (2) with information of HCV infection; (3) containing at least six HCC-HCV samples. HCV-HCC cases were carefully examined and picked out. Five datasets (GSE6764, GSE41804, GSE62232, GSE107170, GSE69715) were based on GPL570 (Affymetrix Human Genome U133 Plus 2.0 Array) and GSE12941 was based on GPL5175 (Affymetrix Human Exon 1.0 ST Array). We also collected the pretreated data of HCV-HCC samples and the corresponding clinical information of TCGA-LIHC (<http://www.tcg.org/>) and ICGC-LIRI-JP (<https://icgc.org/>) from the HCCD database [59]. Five of them (GSE6764, GSE41804, GSE62232, GSE107170, and TCGA-LIHC) were served to screen DEGs, and the remaining three sets were used for further analysis. All of

the above studies comprised a total of 304 HCV-HCC and 290 adjacent normal, and detailed information was summarized in Supplementary Table 1.

### Screening of differentially expressed genes (DEGs)

Differential analysis for each of the above-mentioned microarray datasets was performed by GEO2R (<https://www.ncbi.nlm.nih.gov/geo/geo2r/>) with default settings. For the TCGA-LIHC dataset, the level three normalized mRNA expression profile was downloaded from the HCCD database, and the limma package [60] was adopted to pick out DEGs between HCV-HCC and normal samples. Statistical significance was set as  $|\log \text{Fold change (FC)}| > 1$  and  $\text{FDR (adj.P.Val)} < 0.05$ . Thereafter, the intersected DEGs were obtained and visualized by the UpSetR [61] and VennDiagram [62] packages. In order to further validate the robustness of the DEGs, we conducted the integrated analysis and differential analysis of the four microarray datasets with the aid of sva and limma packages [63].

### Weight Gene Co-expression Network Analysis (WGCNA) and module identification

The WGCNA network was constructed by the WGCNA package [64] based on the gene expression data of ICGC-LIRI-JP. In the beginning, the DEGs from ICGC-LIRI-JP dataset were screened by limma package at the cutoff of  $|\log \text{Fold change (FC)}| > 1$  and  $\text{FDR} < 0.05$ , which were used to detect and eliminate outlier samples through the sample clustering tree. Next, an appropriate soft threshold was used to obtain scale-free networks. Then topological overlap matrix (TOM) and the dissimilarity (dissTOM) were computed and used to implement the gene dendrogram and module recognition ( $\text{minClusterSize} = 30$ ). Similar modules were merged into larger ones at a cutline of 0.3. To determine their relevance to clinical traits, Pearson correlations between module eigengenes and clinical phenotypes including age, gender, TNM stage, alcohol consumption, smoking status, survival time, and survival status were calculated and shown with a correlation heatmap. In this study, we chose the most significant module that correlated with survival status for further analysis, and gene significance (GS) and module membership (MM) were also calculated.

### Protein-protein interaction (PPI) network construction

PPI network is a useful approach to explore molecular interactions related to tumorigenesis and progression. In this study, a PPI network comprising the overlapping

DEGs was constructed by the Search Tool for the Retrieval of Interacting Genes (STRING) database (version 11.0; <http://string-db.org/>). A comprehensive interaction score of  $\geq 0.7$  was set as the threshold (high confidence). Visualization of the PPI network was done by Cytoscape (version 3.2.1; <http://www.cytoscape.org>) [65]. The MCODE plugin of Cytoscape was used to obtain the most significant cluster in the network. Topological parameters were calculated by cytohuber app [66] and we chose the top 30 nodes that had a degree of  $> 20$  as DEGs-PPI hub genes. Besides, to fetch the hub genes in the significant module that correlated with survival status, we also uploaded the corresponding genes in the selected module to the STRING database to establish the WGCNA-PPI network, which was used to identify WGCNA hub genes according to the node degree threshold ( $\geq 50$ ).

### Hub genes identification

In addition to their putative pivotal role in fostering tumorigenesis of cancer, we envisaged that hub genes would provide diagnostic and prognostic values in HCV-HCC patients. So, we picked out the overlapping genes in the PPI hub genes and the WGCNA hub genes and assessed their predictive capabilities for diagnosis and prognosis based on the expression profile of the ICGC-LIRI-JP dataset. For the assessment of their diagnostic powers, we depicted the ROC curves of the overlapping genes by the pROC package [67] to rank their area under the receiver operating characteristic curve (AUROC) scores from high to low, and an AUROC score of  $> 0.95$  was used set as the criterion for selection. To evaluate their prognostic values, only 112 HCV-HCC patients with complete clinicopathologic characteristics (age, gender, TNM stage, vein invasion, alcohol consumption, and smoking status) and available follow-up information (overall survival outcome) were included. The prognostic powers of overlapping genes were estimated by univariate Cox regression (UniCox) with a  $P$ -value threshold of less than 0.05. A forest plot was drawn to present the hazard ratio (HR) and  $P$ -value obtained from UniCox analysis. Only genes that satisfied all these conditions were regarded as hub genes in this study.

### Function enrichment

Metascape database [68] was used to perform the gene ontology (GO) analysis of the upregulated genes, the downregulated genes and of the most significant module in the WGCNA network. Significant terms were defined with a  $P < 0.01$  and count  $> 3$ . For the Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway analysis, the “clusterProfiler” package [69] was utilized and  $\text{FDR} < 0.05$  was set as a cutoff.

## Validation of the hub genes' dysregulation patterns

Three gene expression datasets including ICGC-LIRI-JP, GSE69715, and GSE12941 were used for the validation of the expression patterns of the identified hub genes. We firstly used GSE69715 and GSE12941 as the external datasets to compare the expression levels of the hub genes in tumor vs normal by *t*-test, followed by the investigation of the comparison of that according to different TNM stages, which was conducted through the internal validation set of ICGC-LIRI-JP. Moreover, Pearson correlations of the hub genes' expression values were also carried out with ICGC-LIRI-JP and TCGA datasets.

## Validation of the hub genes' diagnostic abilities

For the evaluation of the hub genes' diagnostic efficiencies, we depicted the ROC curves of GSE69715, GSE107170, and TCGA-LIHC with the pROC package, using the corresponding gene expression profiles. To explore their performance in differentiating the early phase of HCV-HCC from normal liver tissues for early detection possibilities, we used the ICGC-LIRI-JP dataset to generate ROC curves to quantify their predictive powers to discriminate between early stage tumor samples (N =73) and normal control (N =98).

## Survival analysis and construction of the risk signature

To further validate the prognostic values of the identified hub genes, Kaplan–Meier overall survival (OS) curves and log-rank tests were carried out based on the gene expression profile of the ICGC-LIRI-JP dataset. Patients were assigned to different risk groups according to the median expression value, then the survival package [70] was used to measure HR and clinical significance. To construct a prognostic risk signature, we conducted LASSO-COX analysis with the glmnet package [71] to identify the most important hub genes for OS of HCV-HCC. Genes retained by the LASSO regression algorithm were used to build a linear combination to generate the risk score for each patient with the following formula: risk score =  $\sum$  (coef (hub gene)  $\times$  expression value of hub gene). Patients were separated into high- or low-risk groups according to the median risk score, and survival difference between the two groups were compared by Kaplan–Meier curves. The predictive power of the risk signature was investigated by the ROC curve at 3 years, which was accomplished with the survivalROC package [72]. Besides, Stratified analysis was also performed for clinicopathological features including TNM stage, age, gender, vein invasion status, alcohol consumption, and smoking status. For all statistical tests,  $P < 0.05$  was set as the significant cutoff.

## Correlations between immune response and the risk signature

To explore the relationship between our risk signature and immune response, we utilized the CIBERSORT algorithm [73] to obtain the estimation of the percentage for 22 immune cell types in each of the HCV-HCC patients based on the ICGC-LIRI-JP cohort. The relative abundance of immune cells in high- and low-risk groups was computed and presented by a heatmap plot. Spearman correlation analysis was applied to determine the relevance of risk score and immune cell infiltration. Besides, the correlation between each of the risk signature genes and the immune cell was also investigated and visualized by a correlation heatmap.

## Prediction of upstream regulators for the hub genes

The upstream transcription factors of hub genes were explored by miRNet 2.0 [74], an up to date comprehensive platform illustrating “multiple-to-multiple” relationships and functional interpretation via network-based visualization. We chose the TRRUST database [75] to capture the possible transcription factor-hub gene interactions in the current study, which offered 8,444 TF-target regulatory relationships of 800 human TFs derived from PubMed articles. Besides, we predicted the putative targeted miRNAs of all the hub genes with miRTarBase 8.0, which collected the most updated experimentally validated miRNA-target interactions (MTIs) that were validated by reporter assay, western blot, microarray, and next-generation sequencing experiments [76]. The transcription factor-hub gene interactions and predictive miRNA-hub genes interactions were visualized in Sankey diagrams.

## Semantic similarities and GSEA

The GOSemSim package [77] was used to compute the semantic similarities of the identified hub genes among GO terms with Wang's methods. The geometric mean in cellular components (CCs) and molecular functions (MFs) was adopted to evaluate the similarity index for each hub gene. The top three hub genes with the highest similarity index were further verified by GSEA using the gene expression profile of ICGC-LIRI-JP. Briefly, we calculated the Spearman correlation coefficients and corresponding *P* values between each of the genes in the profile and all of the three hub genes and selected the top 5000 of the most significantly correlated genes that had the largest correlation coefficients at a *P*-value of  $< 0.01$ . GSEA was performed using the clusterProfiler package [69] with the pre-ranked gene lists by correlation coefficients, which was based on the database of h.all.v7.0.symbols and c2.cp.kegg.v7.0.symbols, respectively.

## Network pharmacology analysis for HCV-HCC

The drug-gene interaction database (DGIdb) (<https://dgidb.genome.wustl.edu/>; version: 4.2.0 - sha1 afd9f30b) [78] was used to determine the promising drugs that target the identified hub genes. It contains more than 100,000 drug-gene interactions mined from PharmGKB, DrugBank, ChEMBL, TTD, Drug Target Commons, and others. Possible drugs were prefiltered by FDA approval and anti-neoplastic function. TCM-MESH [79] and TCM-ID [80] were used to predict the potential herbs and active components that targeting the hub genes. Only compounds supported by both of the two databases were regarded as potential effective herbal ingredients.

## Abbreviations

HCV-HCC: Hepatitis C virus-associated HCC; DEGs: differently expressed genes; GEO: gene expression omnibus; TCGA: the Cancer Genome Atlas; ICGC: International Cancer Genome Consortium; WGCNA: weighted gene coexpression network analysis; LASSO: the least absolute shrinkage and selection Operator; ROC: Receiver Operating Characteristic; OS: overall survival; TOM: topological overlap matrix; PPI: protein-protein interaction; STRING: the Search Tool for the Retrieval of Interacting Genes; AUROC: area under the receiver operating characteristic curve; UniCox: univariate Cox regression; GO: gene ontology; KEGG: Kyoto Encyclopedia of Genes and Genomes; CCs: cellular components; MFs: molecular functions; DGIdb: the drug-gene interaction database.

## AUTHOR CONTRIBUTIONS

Y. Z. and Y. T. performed the research and drafted the manuscript. C.G. participated in data collecting and preliminary analysis. G. L. conceived and supervised the research, and revised the manuscript. All authors read and approved the final version of the manuscript.

## ACKNOWLEDGMENTS

We acknowledge the Gene Expression Omnibus (GEO) database, Cancer Genome Atlas (TCGA), and International Cancer Genome Consortium for data sharing.

## CONFLICTS OF INTEREST

The authors declare that they have no conflicts of interest.

## REFERENCES

1. Bray F, Ferlay J, Soerjomataram I, Siegel RL, Torre LA, Jemal A. Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA Cancer J Clin.* 2018; 68:394–424. <https://doi.org/10.3322/caac.21492> PMID:30207593
2. Zhang Q, Qi W, Wang X, Zhang Y, Xu Y, Qin S, Zhao P, Guo H, Jiao J, Zhou C, Ji S, Wang J. Epidemiology of Hepatitis B and Hepatitis C Infections and Benefits of Programs for Hepatitis Prevention in Northeastern China: A Cross-Sectional Study. *Clin Infect Dis.* 2016; 62:305–12. <https://doi.org/10.1093/cid/civ859> PMID:26433720
3. 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. <https://doi.org/10.1186/s12935-019-0858-2> PMID:31139015
4. Huang YL, Ning G, Chen LB, Lian YF, Gu YR, Wang JL, Chen DM, Wei H, Huang YH. Promising diagnostic and prognostic value of E2Fs in human hepatocellular carcinoma. *Cancer Manag Res.* 2019; 11:1725–40. <https://doi.org/10.2147/CMAR.S182001> PMID:30863181
5. Wu F, Chen Q, Liu C, Duan X, Hu J, Liu J, Cao H, Li W, Li H. Profiles of prognostic alternative splicing signature in hepatocellular carcinoma. *Cancer Med.* 2020; 9:2171–80. <https://doi.org/10.1002/cam4.2875> PMID:31975560
6. Cai J, Tong Y, Huang L, Xia L, Guo H, Wu H, Kong X, Xia Q. Identification and validation of a potent multi-mRNA signature for the prediction of early relapse in hepatocellular carcinoma. *Carcinogenesis.* 2019; 40:840–52. <https://doi.org/10.1093/carcin/bgz018> PMID:31059567
7. Li G, Xu W, Zhang L, Liu T, Jin G, Song J, Wu J, Wang Y, Chen W, Zhang C, Chen X, Ding Z, Zhu P, Zhang B. Development and validation of a CIMP-associated prognostic model for hepatocellular carcinoma. *EBioMedicine.* 2019; 47:128–41. <https://doi.org/10.1016/j.ebiom.2019.08.064> PMID:31492561
8. Gu Y, Li J, Guo D, Chen B, Liu P, Xiao Y, Yang K, Liu Z, Liu Q. Identification of 13 Key Genes Correlated With Progression and Prognosis in Hepatocellular Carcinoma by Weighted Gene Co-expression Network Analysis. *Front Genet.* 2020; 11:153. <https://doi.org/10.3389/fgene.2020.00153> PMID:32180800

9. Qiu J, Peng B, Tang Y, Qian Y, Guo P, Li M, Luo J, Chen B, Tang H, Lu C, Cai M, Ke Z, He W, et al. CpG Methylation Signature Predicts Recurrence in Early-Stage Hepatocellular Carcinoma: Results From a Multicenter Study. *J Clin Oncol*. 2017; 35:734–42. <https://doi.org/10.1200/JCO.2016.68.2153> PMID:[28068175](https://pubmed.ncbi.nlm.nih.gov/28068175/)
10. Wang Z, Zhu J, Liu Y, Liu C, Wang W, Chen F, Ma L. Development and validation of a novel immune-related prognostic model in hepatocellular carcinoma. *J Transl Med*. 2020; 18:67. <https://doi.org/10.1186/s12967-020-02255-6> PMID:[32046766](https://pubmed.ncbi.nlm.nih.gov/32046766/)
11. Long J, Chen P, Lin J, Bai Y, Yang X, Bian J, Lin Y, Wang D, Yang X, Zheng Y, Sang X, Zhao H. DNA methylation-driven genes for constructing diagnostic, prognostic, and recurrence models for hepatocellular carcinoma. *Theranostics*. 2019; 9:7251–67. <https://doi.org/10.7150/thno.31155> PMID:[31695766](https://pubmed.ncbi.nlm.nih.gov/31695766/)
12. Zhu GQ, Zhou YJ, Qiu LX, Wang B, Yang Y, Liao WT, Luo YH, Shi YH, Zhou J, Fan J, Dai Z. Prognostic alternative mRNA splicing signature in hepatocellular carcinoma: a study based on large-scale sequencing data. *Carcinogenesis*. 2019. [Epub ahead of print]. <https://doi.org/10.1093/carcin/bgz073> PMID:[31099827](https://pubmed.ncbi.nlm.nih.gov/31099827/)
13. Yuan JN, Chao Y, Lee WP, Li CP, Lee RC, Chang FY, Yen SH, Lee SD, Whang-Peng J. Chemotherapy with etoposide, doxorubicin, cisplatin, 5-fluorouracil, and leucovorin for patients with advanced hepatocellular carcinoma. *Med Oncol*. 2008; 25:201–06. <https://doi.org/10.1007/s12032-007-9013-3> PMID:[18488159](https://pubmed.ncbi.nlm.nih.gov/18488159/)
14. Liu G, Shi A, Wang N, Li M, He X, Yin C, Tu Q, Shen X, Tao Y, Wang Q, Yin H. Polyphenolic Proanthocyanidin-B2 suppresses proliferation of liver cancer cells and hepatocellular carcinogenesis through directly binding and inhibiting AKT activity. *Redox Biol*. 2020; 37:101701. <https://doi.org/10.1016/j.redox.2020.101701> PMID:[32863234](https://pubmed.ncbi.nlm.nih.gov/32863234/)
15. Zhou R, Wu K, Su M, Li R. Bioinformatic and experimental data decipher the pharmacological targets and mechanisms of plumbagin against hepatocellular carcinoma. *Environ Toxicol Pharmacol*. 2019; 70:103200. <https://doi.org/10.1016/j.etap.2019.103200> PMID:[31158732](https://pubmed.ncbi.nlm.nih.gov/31158732/)
16. Tang Y, Zhang Y, Hu X. Identification of Potential Hub Genes Related to Diagnosis and Prognosis of Hepatitis B Virus-Related Hepatocellular Carcinoma via Integrated Bioinformatics Analysis. *Biomed Res Int*. 2020; 2020:4251761. <https://doi.org/10.1155/2020/4251761> PMID:[33376723](https://pubmed.ncbi.nlm.nih.gov/33376723/)
17. Dominguez-Brauer C, Thu KL, Mason JM, Blaser H, Bray MR, Mak TW. Targeting Mitosis in Cancer: Emerging Strategies. *Mol Cell*. 2015; 60:524–36. <https://doi.org/10.1016/j.molcel.2015.11.006> PMID:[26590712](https://pubmed.ncbi.nlm.nih.gov/26590712/)
18. Gu J, Liu X, Li J, He Y. MicroRNA-144 inhibits cell proliferation, migration and invasion in human hepatocellular carcinoma by targeting CCNB1. *Cancer Cell Int*. 2019; 19:15. <https://doi.org/10.1186/s12935-019-0729-x> PMID:[30651720](https://pubmed.ncbi.nlm.nih.gov/30651720/)
19. Li N, Li L, Chen Y. The Identification of Core Gene Expression Signature in Hepatocellular Carcinoma. *Oxid Med Cell Longev*. 2018; 2018:3478305. <https://doi.org/10.1155/2018/3478305> PMID:[29977454](https://pubmed.ncbi.nlm.nih.gov/29977454/)
20. Wu SM, Lin SL, Lee KY, Chuang HC, Feng PH, Cheng WL, Liao CJ, Chi HC, Lin YH, Tsai CY, Chen WJ, Yeh CT, Lin KH. Hepatoma cell functions modulated by NEK2 are associated with liver cancer progression. *Int J Cancer*. 2017; 140:1581–96. <https://doi.org/10.1002/ijc.30559> PMID:[27925179](https://pubmed.ncbi.nlm.nih.gov/27925179/)
21. Lin S, Zhou S, Jiang S, Liu X, Wang Y, Zheng X, Zhou H, Li X, Cai X. NEK2 regulates stem-like properties and predicts poor prognosis in hepatocellular carcinoma. *Oncol Rep*. 2016; 36:853–62. <https://doi.org/10.3892/or.2016.4896> PMID:[27349376](https://pubmed.ncbi.nlm.nih.gov/27349376/)
22. Wang SM, Ooi LL, Hui KM. Upregulation of Rac GTPase-activating protein 1 is significantly associated with the early recurrence of human hepatocellular carcinoma. *Clin Cancer Res*. 2011; 17:6040–51. <https://doi.org/10.1158/1078-0432.CCR-11-0557> PMID:[21825042](https://pubmed.ncbi.nlm.nih.gov/21825042/)
23. Yang XM, Cao XY, He P, Li J, Feng MX, Zhang YL, Zhang XL, Wang YH, Yang Q, Zhu L, Nie HZ, Jiang SH, Tian GA, et al. Overexpression of Rac GTPase Activating Protein 1 Contributes to Proliferation of Cancer Cells by Reducing Hippo Signaling to Promote Cytokinesis. *Gastroenterology*. 2018; 155:1233–1249.e22. <https://doi.org/10.1053/j.gastro.2018.07.010> PMID:[30009820](https://pubmed.ncbi.nlm.nih.gov/30009820/)
24. Xie W, Wang B, Wang X, Hou D, Su H, Huang H. Nine hub genes related to the prognosis of HBV-positive hepatocellular carcinoma identified by protein interaction analysis. *Ann Transl Med*. 2020; 8:478. <https://doi.org/10.21037/atm.2020.03.94> PMID:[32395522](https://pubmed.ncbi.nlm.nih.gov/32395522/)
25. Chai N, Xie HH, Yin JP, Sa KD, Guo Y, Wang M, Liu J, Zhang XF, Zhang X, Yin H, Nie YZ, Wu KC, Yang AG, Zhang R. FOXM1 promotes proliferation in human

- hepatocellular carcinoma cells by transcriptional activation of CCNB1. *Biochem Biophys Res Commun.* 2018; 500:924–29.  
<https://doi.org/10.1016/j.bbrc.2018.04.201>  
PMID:[29705704](https://pubmed.ncbi.nlm.nih.gov/29705704/)
26. Ji Y, Yin Y, Zhang W. Integrated Bioinformatic Analysis Identifies Networks and Promising Biomarkers for Hepatitis B Virus-Related Hepatocellular Carcinoma. *Int J Genomics.* 2020; 2020:2061024.  
<https://doi.org/10.1155/2020/2061024>  
PMID:[32775402](https://pubmed.ncbi.nlm.nih.gov/32775402/)
  27. Weng L, Du J, Zhou Q, Cheng B, Li J, Zhang D, Ling C. Identification of cyclin B1 and Sec62 as biomarkers for recurrence in patients with HBV-related hepatocellular carcinoma after surgical resection. *Mol Cancer.* 2012; 11:39.  
<https://doi.org/10.1186/1476-4598-11-39>  
PMID:[22682366](https://pubmed.ncbi.nlm.nih.gov/22682366/)
  28. Shi Q, Meng Z, Tian XX, Wang YF, Wang WH. Identification and validation of a hub gene prognostic index for hepatocellular carcinoma. *Future Oncol.* 2021. [Epub ahead of print].  
<https://doi.org/10.2217/fon-2020-1112>  
PMID:[33620260](https://pubmed.ncbi.nlm.nih.gov/33620260/)
  29. Deng JL, Xu YH, Wang G. Identification of Potential Crucial Genes and Key Pathways in Breast Cancer Using Bioinformatic Analysis. *Front Genet.* 2019; 10:695.  
<https://doi.org/10.3389/fgene.2019.00695>  
PMID:[31428132](https://pubmed.ncbi.nlm.nih.gov/31428132/)
  30. Yang X, Miao BS, Wei CY, Dong RZ, Gao PT, Zhang XY, Lu JC, Gao C, Wang XY, Sun HC, Zhou J, Fan J, Ke AW, et al. Lymphoid-specific helicase promotes the growth and invasion of hepatocellular carcinoma by transcriptional regulation of centromere protein F expression. *Cancer Sci.* 2019; 110:2133–44.  
<https://doi.org/10.1111/cas.14037> PMID:[31066149](https://pubmed.ncbi.nlm.nih.gov/31066149/)
  31. Dai Y, Liu L, Zeng T, Zhu YH, Li J, Chen L, Li Y, Yuan YF, Ma S, Guan XY. Characterization of the oncogenic function of centromere protein F in hepatocellular carcinoma. *Biochem Biophys Res Commun.* 2013; 436:711–18.  
<https://doi.org/10.1016/j.bbrc.2013.06.021>  
PMID:[23791740](https://pubmed.ncbi.nlm.nih.gov/23791740/)
  32. Guo L, Wang Z, Du Y, Mao J, Zhang J, Yu Z, Guo J, Zhao J, Zhou H, Wang H, Gu Y, Li Y. Random-forest algorithm based biomarkers in predicting prognosis in the patients with hepatocellular carcinoma. *Cancer Cell Int.* 2020; 20:251.  
<https://doi.org/10.1186/s12935-020-01274-z>  
PMID:[32565735](https://pubmed.ncbi.nlm.nih.gov/32565735/)
  33. Zhou Z, Li Y, Hao H, Wang Y, Zhou Z, Wang Z, Chu X. Screening Hub Genes as Prognostic Biomarkers of Hepatocellular Carcinoma by Bioinformatics Analysis. *Cell Transplant.* 2019; 28:765–86S.  
<https://doi.org/10.1177/0963689719893950>  
PMID:[31822116](https://pubmed.ncbi.nlm.nih.gov/31822116/)
  34. Chen J, Rajasekaran M, Xia H, Zhang X, Kong SN, Sekar K, Seshachalam VP, Deivasigamani A, Goh BK, Ooi LL, Hong W, Hui KM. The microtubule-associated protein PRC1 promotes early recurrence of hepatocellular carcinoma in association with the Wnt/ $\beta$ -catenin signalling pathway. *Gut.* 2016; 65:1522–34.  
<https://doi.org/10.1136/gutjnl-2015-310625>  
PMID:[26941395](https://pubmed.ncbi.nlm.nih.gov/26941395/)
  35. Diakos CI, Charles KA, McMillan DC, Clarke SJ. Cancer-related inflammation and treatment effectiveness. *Lancet Oncol.* 2014; 15:e493–503.  
[https://doi.org/10.1016/S1470-2045\(14\)70263-3](https://doi.org/10.1016/S1470-2045(14)70263-3)  
PMID:[25281468](https://pubmed.ncbi.nlm.nih.gov/25281468/)
  36. Huang CY, Huang XP, Zhu JY, Chen ZG, Li XJ, Zhang XH, Huang S, He JB, Lian F, Zhao YN, Wu GB. miR-128-3p suppresses hepatocellular carcinoma proliferation by regulating PIK3R1 and is correlated with the prognosis of HCC patients. *Oncol Rep.* 2015; 33:2889–98.  
<https://doi.org/10.3892/or.2015.3936> PMID:[25962360](https://pubmed.ncbi.nlm.nih.gov/25962360/)
  37. Liu K, Li X, Cao Y, Ge Y, Wang J, Shi B. MiR-132 inhibits cell proliferation, invasion and migration of hepatocellular carcinoma by targeting PIK3R3. *Int J Oncol.* 2015; 47:1585–93.  
<https://doi.org/10.3892/ijo.2015.3112> PMID:[26252738](https://pubmed.ncbi.nlm.nih.gov/26252738/)
  38. Han H, Sun D, Li W, Shen H, Zhu Y, Li C, Chen Y, Lu L, Li W, Zhang J, Tian Y, Li Y. A c-Myc-MicroRNA functional feedback loop affects hepatocarcinogenesis. *Hepatology.* 2013; 57:2378–89.  
<https://doi.org/10.1002/hep.26302> PMID:[23389829](https://pubmed.ncbi.nlm.nih.gov/23389829/)
  39. Zhu MX, Wei CY, Zhang PF, Gao DM, Chen J, Zhao Y, Dong SS, Liu BB. Elevated TRIP13 drives the AKT/mTOR pathway to induce the progression of hepatocellular carcinoma via interacting with ACTN4. *J Exp Clin Cancer Res.* 2019; 38:409.  
<https://doi.org/10.1186/s13046-019-1401-y>  
PMID:[31533816](https://pubmed.ncbi.nlm.nih.gov/31533816/)
  40. Lu J, Lin Y, Li F, Ye H, Zhou R, Jin Y, Li B, Xiong X, Cheng N. MiR-205 suppresses tumor growth, invasion, and epithelial-mesenchymal transition by targeting SEMA4C in hepatocellular carcinoma. *FASEB J.* 2018. [Epub ahead of print].  
<https://doi.org/10.1096/fj.201800113R>  
PMID:[29799789](https://pubmed.ncbi.nlm.nih.gov/29799789/)
  41. Chen JQ, Ou YL, Huang ZP, Hong YG, Tao YP, Wang ZG, Ni JS, Hao LQ, Lin H. MicroRNA-212-3p inhibits the Proliferation and Invasion of Human Hepatocellular Carcinoma Cells by Suppressing CTGF expression. *Sci Rep.* 2019; 9:9820.

- <https://doi.org/10.1038/s41598-019-46088-w>  
PMID:[31285444](https://pubmed.ncbi.nlm.nih.gov/31285444/)
42. Fu X, Liu M, Qu S, Ma J, Zhang Y, Shi T, Wen H, Yang Y, Wang S, Wang J, Nan K, Yao Y, Tian T. Exosomal microRNA-32-5p induces multidrug resistance in hepatocellular carcinoma via the PI3K/Akt pathway. *J Exp Clin Cancer Res.* 2018; 37:52.  
<https://doi.org/10.1186/s13046-018-0677-7>  
PMID:[29530052](https://pubmed.ncbi.nlm.nih.gov/29530052/)
43. Rivankar S. An overview of doxorubicin formulations in cancer therapy. *J Cancer Res Ther.* 2014; 10:853–58.  
<https://doi.org/10.4103/0973-1482.139267>  
PMID:[25579518](https://pubmed.ncbi.nlm.nih.gov/25579518/)
44. Guiu B, Chevallier P, Assenat E, Barbier E, Merle P, Bouvier A, Dumortier J, Nguyen-Khac E, Gugenheim J, Rode A, Oberti F, Valette PJ, Yzet T, et al. Idarubicin-loaded Beads for Chemoembolization of Hepatocellular Carcinoma: The IDASPHERE II Single-Arm Phase II Trial. *Radiology.* 2019; 291:801–08.  
<https://doi.org/10.1148/radiol.2019182399>  
PMID:[31038408](https://pubmed.ncbi.nlm.nih.gov/31038408/)
45. Hollingshead LM, Faulds D. Idarubicin. A review of its pharmacodynamic and pharmacokinetic properties, and therapeutic potential in the chemotherapy of cancer. *Drugs.* 1991; 42:690–719.  
<https://doi.org/10.2165/00003495-199142040-00010>  
PMID:[1723369](https://pubmed.ncbi.nlm.nih.gov/1723369/)
46. Paz-Ares L, Dvorkin M, Chen Y, Reinmuth N, Hotta K, Trukhin D, Statsenko G, Hochmair MJ, Özgüroğlu M, Ji JH, Voitko O, Poltoratskiy A, Ponce S, et al, and CASPIAN investigators. Durvalumab plus platinum-etoposide versus platinum-etoposide in first-line treatment of extensive-stage small-cell lung cancer (CASPIAN): a randomised, controlled, open-label, phase 3 trial. *Lancet.* 2019; 394:1929–39.  
[https://doi.org/10.1016/S0140-6736\(19\)32222-6](https://doi.org/10.1016/S0140-6736(19)32222-6)  
PMID:[31590988](https://pubmed.ncbi.nlm.nih.gov/31590988/)
47. Ladenstein R, Pötschger U, Pearson AD, Brock P, Luksch R, Castel V, Yaniv I, Papadakis V, Laureys G, Malis J, Balwierz W, Ruud E, Kogner P, et al, and SIOP Europe Neuroblastoma Group (SIOPEN). Busulfan and melphalan versus carboplatin, etoposide, and melphalan as high-dose chemotherapy for high-risk neuroblastoma (HR-NBL1/SIOPEN): an international, randomised, multi-arm, open-label, phase 3 trial. *Lancet Oncol.* 2017; 18:500–14.  
[https://doi.org/10.1016/S1470-2045\(17\)30070-0](https://doi.org/10.1016/S1470-2045(17)30070-0)  
PMID:[28259608](https://pubmed.ncbi.nlm.nih.gov/28259608/)
48. Lan CY, Wang Y, Xiong Y, Li JD, Shen JX, Li YF, Zheng M, Zhang YN, Feng YL, Liu Q, Huang HQ, Huang X. Apatinib combined with oral etoposide in patients with platinum-resistant or platinum-refractory ovarian cancer (AERO): a phase 2, single-arm, prospective study. *Lancet Oncol.* 2018; 19:1239–46.  
[https://doi.org/10.1016/S1470-2045\(18\)30349-8](https://doi.org/10.1016/S1470-2045(18)30349-8)  
PMID:[30082170](https://pubmed.ncbi.nlm.nih.gov/30082170/)
49. Budde LE, Wu D, Martin DB, Philip M, Shustov AR, Smith SD, Gooley TA, Chen TL, Libby EN, Chen EY, Kojouri K, Langerak A, Roden JE, et al. Bendamustine with rituximab, etoposide and carboplatin (T(R)EC) in relapsed or refractory aggressive lymphoma: a prospective multicentre phase 1/2 clinical trial. *Br J Haematol.* 2018; 183:601–07.  
<https://doi.org/10.1111/bjh.15585> PMID:[30596402](https://pubmed.ncbi.nlm.nih.gov/30596402/)
50. Sato S, Yamamoto E, Niimi K, Ino K, Nishino K, Suzuki S, Kotani T, Kajiyama H, Kikkawa F. The efficacy and toxicity of 4-day chemotherapy with methotrexate, etoposide and actinomycin D in patients with choriocarcinoma and high-risk gestational trophoblastic neoplasia. *Int J Clin Oncol.* 2020; 25:203–09.  
<https://doi.org/10.1007/s10147-019-01540-9>  
PMID:[31520175](https://pubmed.ncbi.nlm.nih.gov/31520175/)
51. Lu YS, Chen TW, Lin CH, Yeh DC, Tseng LM, Wu PF, Rau KM, Chen BB, Chao TC, Huang SM, Huang CS, Shih TT, Cheng AL, and Taiwan Breast Cancer Consortium. Bevacizumab preconditioning followed by Etoposide and Cisplatin is highly effective in treating brain metastases of breast cancer progressing from whole-brain radiotherapy. *Clin Cancer Res.* 2015; 21:1851–58.  
<https://doi.org/10.1158/1078-0432.CCR-14-2075>  
PMID:[25700303](https://pubmed.ncbi.nlm.nih.gov/25700303/)
52. Franciosi V, Cocconi G, Michiara M, Di Costanzo F, Fossier V, Tonato M, Carlini P, Boni C, Di Sarra S. Front-line chemotherapy with cisplatin and etoposide for patients with brain metastases from breast carcinoma, nonsmall cell lung carcinoma, or malignant melanoma: a prospective study. *Cancer.* 1999; 85:1599–605.  
PMID:[10193952](https://pubmed.ncbi.nlm.nih.gov/10193952/)
53. Wurmbach E, Chen YB, Khitrov G, Zhang W, Roayaie S, Schwartz M, Fiel I, Thung S, Mazzaferro V, Bruix J, Bottinger E, Friedman S, Waxman S, Llovet JM. Genome-wide molecular profiles of HCV-induced dysplasia and hepatocellular carcinoma. *Hepatology.* 2007; 45:938–47.  
<https://doi.org/10.1002/hep.21622>  
PMID:[17393520](https://pubmed.ncbi.nlm.nih.gov/17393520/)
54. Hodo Y, Honda M, Tanaka A, Nomura Y, Arai K, Yamashita T, Sakai Y, Yamashita T, Mizukoshi E, Sakai A, Sasaki M, Nakanuma Y, Moriyama M, Kaneko S. Association of interleukin-28B genotype and hepatocellular carcinoma recurrence in patients with chronic hepatitis C. *Clin Cancer Res.* 2013; 19:1827–37.  
<https://doi.org/10.1158/1078-0432.CCR-12-1641>  
PMID:[23426277](https://pubmed.ncbi.nlm.nih.gov/23426277/)

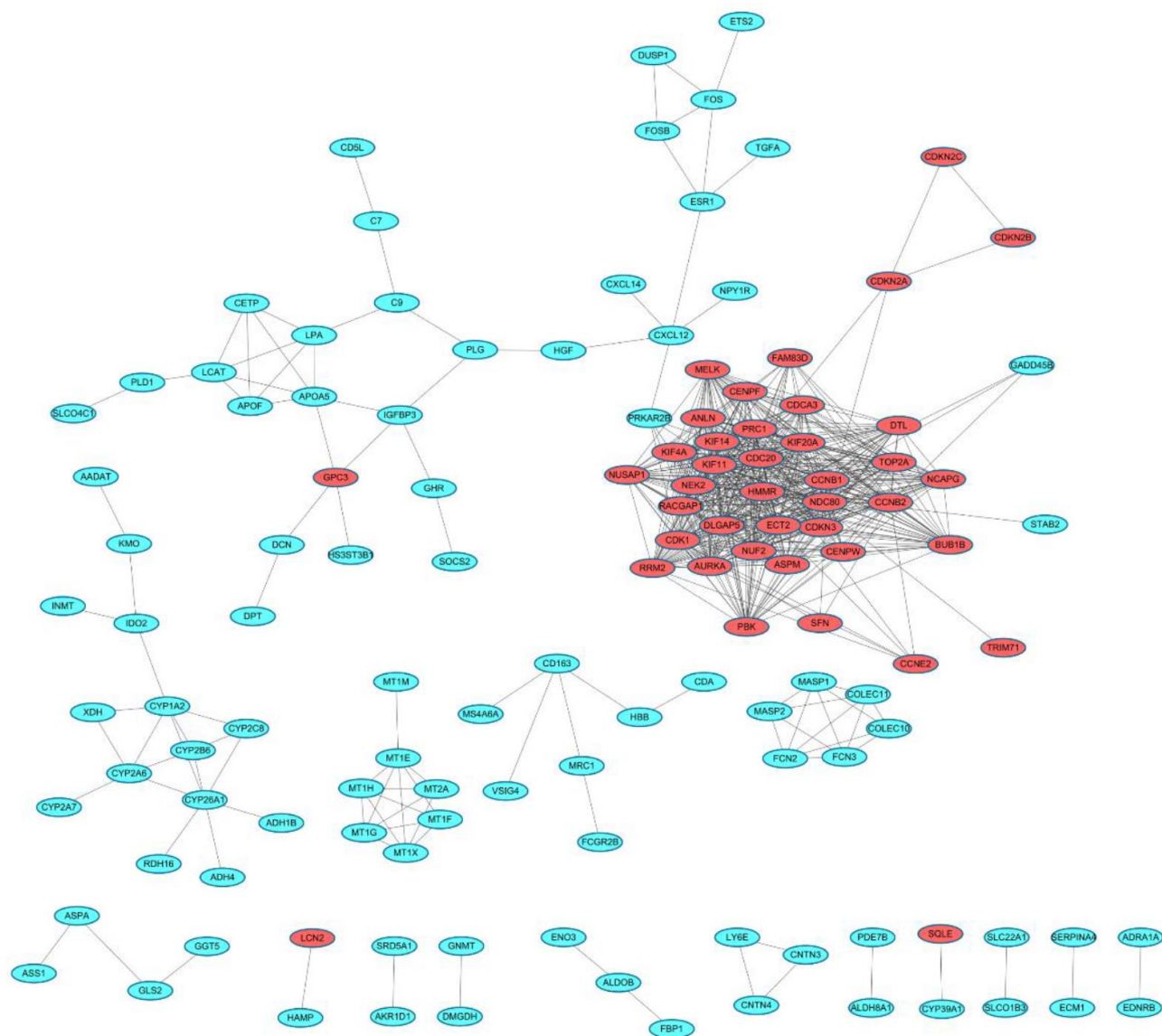
55. Schulze K, Imbeaud S, Letouzé E, Alexandrov LB, Calderaro J, Rebouissou S, Couchy G, Meiller C, Shinde J, Soysouvanh F, Calatayud AL, Pinyol R, Pelletier L, et al. Exome sequencing of hepatocellular carcinomas identifies new mutational signatures and potential therapeutic targets. *Nat Genet.* 2015; 47:505–11. <https://doi.org/10.1038/ng.3252> PMID:25822088
56. Diaz G, Engle RE, Tice A, Melis M, Montenegro S, Rodriguez-Canales J, Hanson J, Emmert-Buck MR, Bock KW, Moore IN, Zamboni F, Govindarajan S, Kleiner DE, Farci P. Molecular Signature and Mechanisms of Hepatitis D Virus-Associated Hepatocellular Carcinoma. *Mol Cancer Res.* 2018; 16:1406–19. <https://doi.org/10.1158/1541-7786.MCR-18-0012> PMID:29858376
57. Satow R, Shitashige M, Kanai Y, Takeshita F, Ojima H, Jigami T, Honda K, Kosuge T, Ochiya T, Hirohashi S, Yamada T. Combined functional genome survey of therapeutic targets for hepatocellular carcinoma. *Clin Cancer Res.* 2010; 16:2518–28. <https://doi.org/10.1158/1078-0432.CCR-09-2214> PMID:20388846
58. Sekhar V, Pollicino T, Diaz G, Engle RE, Alayli F, Melis M, Kabat J, Tice A, Pomeranke A, Altan-Bonnet N, Zamboni F, Lusso P, Emerson SU, Farci P. Infection with hepatitis C virus depends on TACSTD2, a regulator of claudin-1 and occludin highly downregulated in hepatocellular carcinoma. *PLoS Pathog.* 2018; 14:e1006916. <https://doi.org/10.1371/journal.ppat.1006916> PMID:29538454
59. Lian Q, Wang S, Zhang G, Wang D, Luo G, Tang J, Chen L, Gu J. HCCDB: A Database of Hepatocellular Carcinoma Expression Atlas. *Genomics Proteomics Bioinformatics.* 2018; 16:269–75. <https://doi.org/10.1016/j.gpb.2018.07.003> PMID:30266410
60. Ritchie ME, Phipson B, Wu D, Hu Y, Law CW, Shi W, Smyth GK. limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res.* 2015; 43:e47. <https://doi.org/10.1093/nar/gkv007> PMID:25605792
61. Lex A, Gehlenborg N, Strobel H, Vuillemot R, Pfister H. UpSet: Visualization of Intersecting Sets. *IEEE Trans Vis Comput Graph.* 2014; 20:1983–92. <https://doi.org/10.1109/TVCG.2014.2346248> PMID:26356912
62. Chen H, Boutros PC. VennDiagram: a package for the generation of highly-customizable Venn and Euler diagrams in R. *BMC Bioinformatics.* 2011; 12:35. <https://doi.org/10.1186/1471-2105-12-35> PMID:21269502
63. Leek JT, Johnson WE, Parker HS, Jaffe AE, Storey JD. The sva package for removing batch effects and other unwanted variation in high-throughput experiments. *Bioinformatics.* 2012; 28:882–83. <https://doi.org/10.1093/bioinformatics/bts034> PMID:22257669
64. Langfelder P, Horvath S. WGCNA: an R package for weighted correlation network analysis. *BMC Bioinformatics.* 2008; 9:559. <https://doi.org/10.1186/1471-2105-9-559> PMID:19114008
65. Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, Amin N, Schwikowski B, Ideker T. Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res.* 2003; 13:2498–504. <https://doi.org/10.1101/gr.1239303> PMID:14597658
66. Chin CH, Chen SH, Wu HH, Ho CW, Ko MT, Lin CY. cytoHubba: identifying hub objects and sub-networks from complex interactome. *BMC Syst Biol.* 2014 (Suppl 4); 8:S11. <https://doi.org/10.1186/1752-0509-8-S4-S11> PMID:25521941
67. Robin X, Turck N, Hainard A, Tiberti N, Lisacek F, Sanchez JC, Müller M. pROC: an open-source package for R and S+ to analyze and compare ROC curves. *BMC Bioinformatics.* 2011; 12:77. <https://doi.org/10.1186/1471-2105-12-77> PMID:21414208
68. Zhou Y, Zhou B, Pache L, Chang M, Khodabakhshi AH, Tanaseichuk O, Benner C, Chanda SK. Metascape provides a biologist-oriented resource for the analysis of systems-level datasets. *Nat Commun.* 2019; 10:1523. <https://doi.org/10.1038/s41467-019-09234-6> PMID:30944313
69. Yu G, Wang LG, Han Y, He QY. clusterProfiler: an R package for comparing biological themes among gene clusters. *OMICS.* 2012; 16:284–87. <https://doi.org/10.1089/omi.2011.0118> PMID:22455463
70. Therneau TM, Lumley TL. Package ‘survival’. *R Top Doc.* 2015; 128.
71. Friedman J, Hastie T, Tibshirani R. Regularization Paths for Generalized Linear Models via Coordinate Descent. *J Stat Softw.* 2010; 33:1–22. PMID:20808728
72. Heagerty PJ, Lumley T, Pepe MS. Time-dependent ROC curves for censored survival data and a diagnostic marker. *Biometrics.* 2000; 56:337–44. <https://doi.org/10.1111/j.0006-341x.2000.00337.x> PMID:10877287



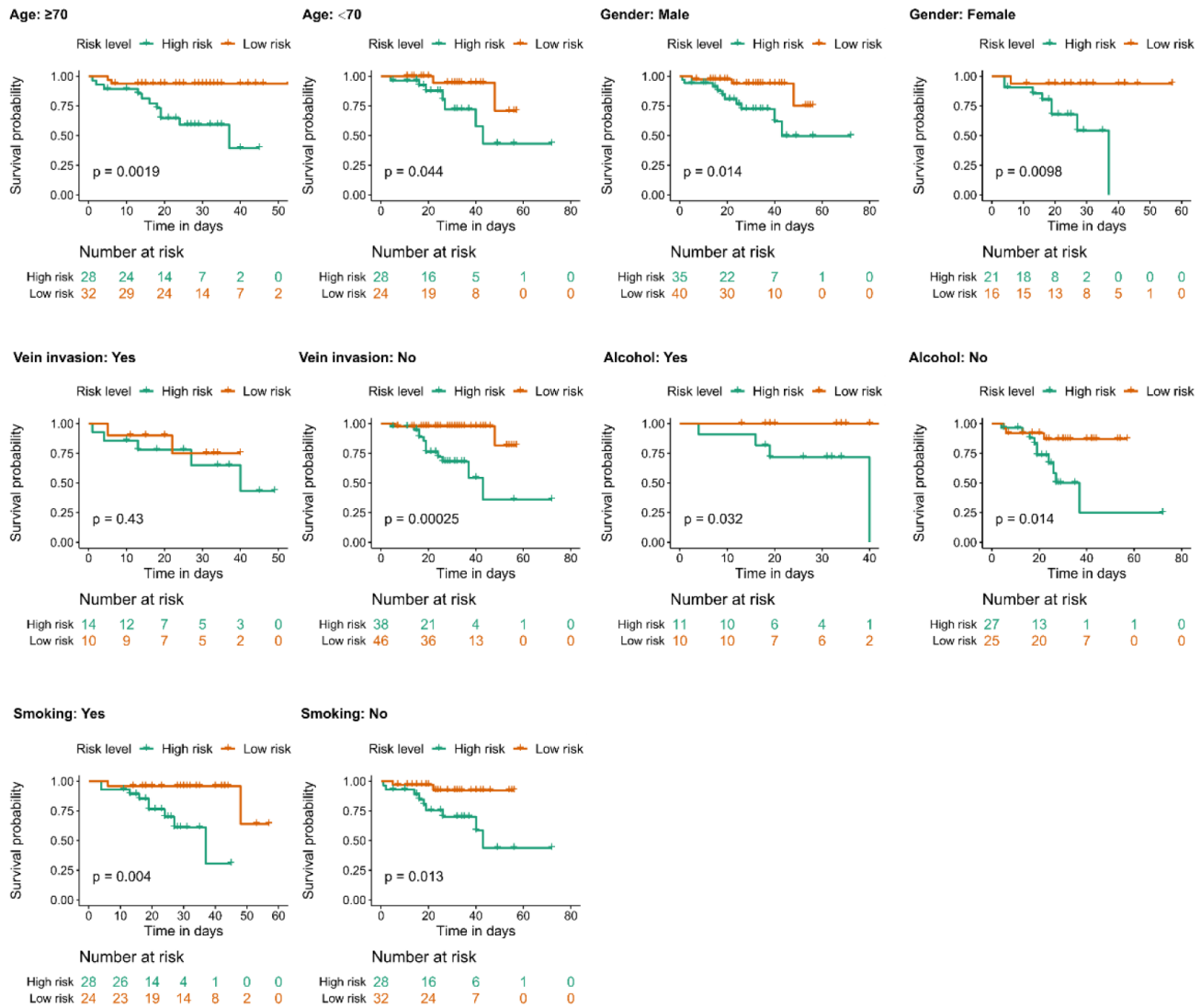
73. Newman AM, Liu CL, Green MR, Gentles AJ, Feng W, Xu Y, Hoang CD, Diehn M, Alizadeh AA. Robust enumeration of cell subsets from tissue expression profiles. *Nat Methods*. 2015; 12:453–57.  
<https://doi.org/10.1038/nmeth.3337>  
PMID:[25822800](https://pubmed.ncbi.nlm.nih.gov/25822800/)
74. Chang L, Zhou G, Soufan O, Xia J. miRNet 2.0: network-based visual analytics for miRNA functional analysis and systems biology. *Nucleic Acids Res*. 2020; 48:W244–51.  
<https://doi.org/10.1093/nar/gkaa467>  
PMID:[32484539](https://pubmed.ncbi.nlm.nih.gov/32484539/)
75. Han H, Cho JW, Lee S, Yun A, Kim H, Bae D, Yang S, Kim CY, Lee M, Kim E, Lee S, Kang B, Jeong D, et al. TRRUST v2: an expanded reference database of human and mouse transcriptional regulatory interactions. *Nucleic Acids Res*. 2018; 46:D380–86.  
<https://doi.org/10.1093/nar/gkx1013>  
PMID:[29087512](https://pubmed.ncbi.nlm.nih.gov/29087512/)
76. Chou CH, Shrestha S, Yang CD, Chang NW, Lin YL, Liao KW, Huang WC, Sun TH, Tu SJ, Lee WH, Chiew MY, Tai CS, Wei TY, et al. miRTarBase update 2018: a resource for experimentally validated microRNA-target interactions. *Nucleic Acids Res*. 2018; 46:D296–302.  
<https://doi.org/10.1093/nar/gkx1067>  
PMID:[29126174](https://pubmed.ncbi.nlm.nih.gov/29126174/)
77. Yu G, Li F, Qin Y, Bo X, Wu Y, Wang S. GOSemSim: an R package for measuring semantic similarity among GO terms and gene products. *Bioinformatics*. 2010; 26:976–78.  
<https://doi.org/10.1093/bioinformatics/btq064>  
PMID:[20179076](https://pubmed.ncbi.nlm.nih.gov/20179076/)
78. Cotto KC, Wagner AH, Feng YY, Kiwala S, Coffman AC, Spies G, Wollam A, Spies NC, Griffith OL, Griffith M. DGIdb 3.0: a redesign and expansion of the drug-gene interaction database. *Nucleic Acids Res*. 2018; 46:D1068–73.  
<https://doi.org/10.1093/nar/gkx1143>  
PMID:[29156001](https://pubmed.ncbi.nlm.nih.gov/29156001/)
79. Zhang RZ, Yu SJ, Bai H, Ning K. TCM-Mesh: The database and analytical system for network pharmacology analysis for TCM preparations. *Sci Rep*. 2017; 7:2821.  
<https://doi.org/10.1038/s41598-017-03039-7>  
PMID:[28588237](https://pubmed.ncbi.nlm.nih.gov/28588237/)
80. Xue R, Fang Z, Zhang M, Yi Z, Wen C, Shi T. TCMID: Traditional Chinese Medicine integrative database for herb molecular mechanism analysis. *Nucleic Acids Res*. 2013; 41:D1089–95.  
<https://doi.org/10.1093/nar/gks1100>  
PMID:[23203875](https://pubmed.ncbi.nlm.nih.gov/23203875/)

## SUPPLEMENTARY MATERIALS

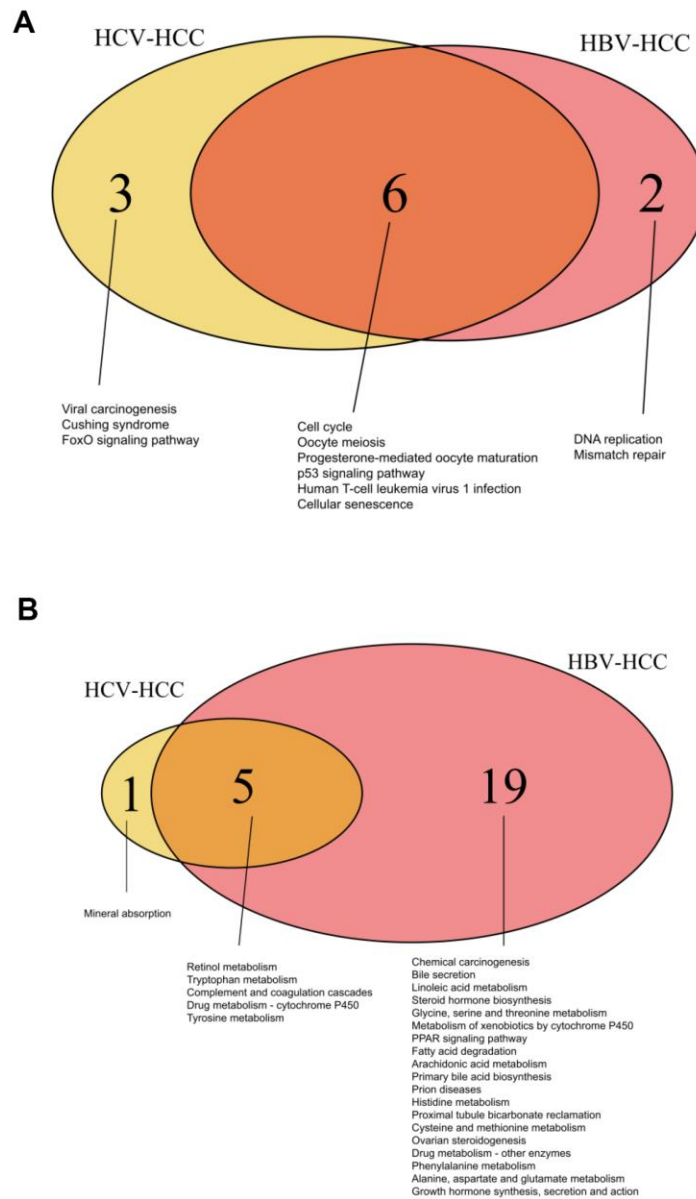
### Supplementary Figures



**Supplementary Figure 1. PPI network of 240 DEGs based on the STRING database.** Red nodes denote the upregulated genes, while blue nodes denote the downregulated genes. PPI, protein-protein interaction. DEGs, differentially expressed genes. PPI, protein-protein interaction. STRING, Search Tool for the Retrieval of Interacting Genes.



Supplementary Figure 2. Stratified survival analysis according to clinicopathological features.



**Supplementary Figure 3. Common KEGG pathways enriched by robust DEGs of HBV-HCC and HCV-HCC. (A)** Venn plot showing the common or individual pathways enriched by the upregulated genes of HCV-HCC and HBV-HCC. **(B)** Venn plot showing the common or individual pathways enriched by the downregulated genes of HCV-HCC and HBV-HCC. HBV-HCC, HBV - associated HCC. HCV-HCC, HCV-associated HCC. KEGG, Kyoto Encyclopedia of Genes and Genomes. DEGs, differentially expressed genes.

## Supplementary Tables

Please browse Full Text version to see the data of Supplementary Tables 2, 3, 5, 6.

**Supplementary Table 1. Characteristics of all datasets enrolled in this study.**

Study	Platforms	Etiology	Sample number			Technology
			Tumor	Normal	Total	
GSE6764	GPL570	HCV	35	10	45	microarray
GSE41804	GPL570	HCV	20	20	40	microarray
GSE62232	GPL570	HCV	9	10	19	microarray
GSE107170	GPL570	HCV	44	31	75	microarray
TCGA-LIHC	Illumina Hiseq	HCV	30	49	79	RNA-seq
HCCD-ICGC	Illumina Hiseq	HCV	123	98	221	RNA-seq
GSE12941	GPL5175	HCV	6	6	12	microarray
GSE69715	GPL570	HCV	37	66	103	microarray

HCV, Hepatitis C virus.

**Supplementary Table 2. Differential analysis of the overlapping 240 DEGs by the combined microarray dataset.**

**Supplementary Table 3. DEGs of the ICGC-LIRI-JP dataset. DEGs, differentially expressed genes.**

**Supplementary Table 4. Putative upstream regulatory transcription factors of the hub genes.**

<b>ID</b>	<b>Accession</b>	<b>Target</b>	<b>TargetID</b>	<b>Experiment</b>	<b>Literature</b>	<b>Tissue</b>
ATF1	466	TOP2A	7153	Unknown	12104051	Not Applicable
BRCA1	672	ASPM	259266	+	16123590	Not Applicable
BRCA1	672	CCNB1	891	Unknown	12647291	Not Applicable
CUX1	1523	RACGAP1	29127	Unknown	21886810	Not Applicable
E2F1	1869	AURKA	6790	+	20300951	Not Applicable
E2F1	1869	CCNB1	891	Unknown	22508987	Not Applicable
E2F1	1869	RACGAP1	29127	Unknown	21886810	Not Applicable
E2F1	1869	TOP2A	7153	+	11313881	Not Applicable
E2F3	1871	AURKA	6790	+	18776222	Not Applicable
E2F3	1871	CCNB1	891	Unknown	17098936	Not Applicable
E2F4	1874	CCNB1	891	Unknown	22508987	Not Applicable
FOXJ1	2302	ASPM	259266	Unknown	16809635	Not Applicable
FOXM1	2305	CCNB1	891	+	19276163	Not Applicable
IRF1	3659	CCNB1	891	-	22200613	Not Applicable
KLF4	9314	CCNB1	891	-	14627709	Not Applicable
KLF5	688	CCNB1	891	+	21951574	Not Applicable
MED1	5469	AURKA	6790	+	16574658	Not Applicable
MYC	4609	CCNB1	891	Unknown	11522645	Not Applicable
NFKB1	4790	CCNB1	891	+	11861406	Not Applicable
NFKB1	4790	CCNB1	891	Unknown	19610058	Not Applicable
OTX2	5015	AURKA	6790	+	21047732	Not Applicable
PTTG1	9232	CCNB1	891	+	22475756	Not Applicable
RELA	5970	CCNB1	891	+	11861406	Not Applicable
RELA	5970	CCNB1	891	Unknown	19610058	Not Applicable
TBP	6908	CCNB1	891	Unknown	17098936	Not Applicable
TFAP2A	7020	CCNB1	891	Unknown	7739559	Not Applicable
TP53	7157	CCNB1	891	-	11892838	Not Applicable
TP53	7157	PRC1	9055	-	12889596	Not Applicable
UHRF1	29128	TOP2A	7153	Unknown	10646863	Not Applicable
USF1	7391	CCNB1	891	Unknown	10548544	Not Applicable
YBX1	4904	CCNB1	891	-	20596676	Not Applicable
YBX1	4904	CDKN3	1033	+	20596676	Not Applicable
YBX1	4904	TOP2A	7153	Unknown	10597187	Not Applicable

**Supplementary Table 5. putative miRNA-mRNA pairs by experiments based on miRTarbase 8.0.**

**Supplementary Table 6. Anti-neoplastic drugs predicted by DGIdb.**