Identification of Key Genes in HCV-induced HCC at Early Stage on the Background of HCV-cirrhosis

Jiao Nie
Department of Gastroenterology, Weihai Municipal Hospital, Shandong University, Weihai, Shandong, 264200, P.R. China
https://orcid.org/0000-0002-6103-1922
Chao Du
Department of Gastroenterology, Linyi People's Hospital, Shandong University, Linyi, Shandong, 276000, P.R. China
Lin Lu
Department of Gastroenterology, Linyi People's Hospital, Linyi, Shandong, 276000, P.R. China
Xiaozhong Gao (gxznj2020@163.com)
Department of Gastroenterology, Weihai Municipal Hospital, Shandong University, Weihai, Shandong, 264200, P.R. China
https://orcid.org/0000-0002-0141-1923

Research

Keywords: hepatitis C virus (HCV), HCV-induced hepatocellular carcinoma, HCV-induced cirrhosis, biomarker, risk score

Posted Date: September 15th, 2020

DOI: https://doi.org/10.21203/rs.3.rs-70171/v1

Abstract

Background: Hepatocellular carcinoma (HCC) is one of the most common and deadly malignant tumors worldwide. Hepatitis C virus (HCV)-induced cirrhosis (HCV-cirrhosis) is one of the leading causes of HCC occurrence. Due to the lack of effective biomarkers, most patients with HCV-induced HCC (HCV-HCC) are at advanced stages when they are first diagnosed. Our study aims to employ transcriptomic profiling dataset to identify potential biomarkers that can predict HCV-HCC at early stage on the background of HCV-cirrhosis.

Methods: The dataset incorporating HCV-cirrhosis and HCV-HCC subjects at different stages from Gene Expression Omnibus was analyzed to identify gene signature-related to HCV-HCC on the background of HCV-cirrhosis. Multiple-genes based risk score-related prediction model was established to predict HCV-HCC samples at different stages, and receiver-operating characteristic (ROC) curve analysis was used to select the best cutoff value of risk scores for prediction and to evaluate the prediction model. Samples with risk scores higher than cutoff value were defined as high score group.

Results: Highly clustered 20 genes were identified and were all gradually increased as HCV-HCC progressed from HCV-cirrhosis to very advanced HCC. Compared with HCV-cirrhosis, the prevalence of HCV-HCC at any of the stages in high score group was 100%. The risk score-related prediction model based on the 20 genes-biomarker got the accuracy of over 95.2 % with area under ROC (AUC) over 0.94. To implement the biomarker easily in real life and make it economical, we tried to limit the gene numbers. Interestingly, CDKN3 and FAM83D were significantly decreased in HCV-cirrhosis tissues as compared to normal tissues and then increased as HCV-HCC progressed. The results were attractive that the prediction model based on this 2 genes-biomarker indicated that prevalence of HCV-HCC in high score group was still 100% and got the predictive accuracy of over 95.2 % with AUC over 0.96, revealing even a better performance.

Conclusion: Our study indicated a 2-genes-based biomarker that could identify HCV-HCC at earlier stages on the background of HCV-cirrhosis, which might be a promising biomarker for early diagnosis of HCV-HCC and potential novel treatment target.

Background

Hepatocellular carcinoma (HCC) is a primary malignant tumor of the liver that threatens human health. HCC mainly occurs in patients with chronic liver disease (often cirrhosis), largely resulting from chronic hepatitis B virus (HBV) and/or hepatitis C virus (HCV) infection [1, 2]. With the introduction of universal immunization to eliminate transmission of HBV infection, HBV-induced HCC has been largely prevented [3]. However, the incidence of HCV-induced HCC (HCV-HCC) has increased by fifteen- to twenty-fold over a 30-year period [4]. The mortality of HCV-HCC is still very high, as incidence of HCV-induced cirrhosis (HCV-cirrhosis) continues to increase [5]. Once HCV-cirrhosis is established, the risk for developing HCC is about 1–4% annually [6]. The 5-year survival rates of HCC patients diagnosed at early stage exceed 70% with effective surgical treatment, while patients present at advanced stage have a median survival time of less than 12 months [7]. Early detection of HCC has important implications for clinical outcomes. Therefore, patients with HCV-cirrhosis should perform surveillance to increase early detection of HCC and survival [8].

However, because of the lack of accurate and effective biomarkers in the early stage of HCV-HCC on the background of cirrhosis, most patients with HCC are at advanced stages when they are first diagnosed. To improve the early diagnosis rate of HCV-HCC, identification of biomarkers with highest specificity and sensitivity is strongly required and is an active field of research. However, there is no good potentially useful molecular biomarkers to precisely predict HCC [9]. It is very difficult to diagnose HCC with a single gene, while several studies have shown that multiple genes could help to predict the occurrence of HCC [10–12]. In the past, studies mainly focused on identifying biomarkers by comparing the gene expression profiles between tumor and non-tumorous tissues [11, 13–15]. But there is no research that discusses HCV-HCC occurrence on the background of HCV-cirrhosis. Therefore, it is promising to identify multi-genes-based biomarker to predict HCV-HCC at early stage on the background of HCV-cirrhosis. Now, there are rich public repositories of RNA-seq and microarray datasets, such as Gene Expression Omnibus (GEO) database [16] and The Cancer Genome Atlas (TCGA) [17]. Further analysis of those datasets

combined with computational approach such as prediction model can provide the opportunity to establish a novel biomarker for disease diagnosis with high accuracy [18, 19].

In our study, an open source dataset with HCV-HCC tissues at different stages and HCV-cirrhosis tissues from GEO was included. We found 20 HCV-HCC-related genes that were significantly up-regulated and these genes were progressively increased from HCV-cirrhosis to very advanced HCC. More importantly, the risk score-associated prediction model based on the 20 genes got predictive accuracy of over 95% and area under the receiver operating characteristic curve (AUC) over 0.94, indicating very good predictive power of the 20-genes-based biomarker to distinguish HCV-HCC at earlier stages and HCV-cirrhosis. To optimize the biomarker with limited gene number for better clinical application in the future, we investigated the role of CDKN3 and FAM83D that were interestingly decreased in HCV-cirrhosis tissues as compared with normal tissues in predicting HCV-HCC. We were surprised to find that this two-genes-based biomarker showed excellent diagnostic ability that the risk score-associated prediction model got the predictive accuracy of over 95.2% with AUC over 0.96. Therefore, this two-gene-based biomarker might be a promising biomarker for early diagnosis of HCV-HCC on the background of HCV-cirrhosis and could also be the potential novel treatment target in the future.

Materials And Methods

Retrieval of datasets on HCV-cirrhosis and HCV-HCC from public database

We searched transcriptome profiles of HCV-cirrhosis and HCV-HCC tissues from GEO database. The search terms we used included “HCV”, “cirrhosis” and “HCC.” Our selection criteria are as follows: (1) studies involved the use of adult liver tissues; (2) studies included HCV-cirrhosis tissues and HCV-HCC tissues at different stages; (3) microarray or RNA-seq datasets. According to the retrieval criteria, the open source dataset GSE6764 from the GEO database was included [20]. The $\log_2$ of the total genes in the dataset was calculated and the expression was normalized. Demographics for GSE6764 were summarized in Table 1. HCV-cirrhosis tissues and very advanced HCV-HCC tissues were defined as the discovery group. While, there were four validated groups: validated group-1 (advanced HCC tissues and cirrhosis tissues), validated group-2 (early HCC tissues and cirrhosis tissues), validated group-3 (very early HCC tissues and cirrhosis tissues) and validated group-4 (normal tissues and cirrhosis tissues).We illustrated our workflow for analytical procedure of the dataset in Fig. 1.

Table 1

Demographics for GSE6764

Dataset: GSE6764	**Discovery group**		**Validated groups**							
	cirrhosis	Very advanced HCC	Advanced HCC vs cirrhosis		Early HCC vs cirrhosis		Very early HCC vs cirrhosis		Normal vs cirrhosis	
			Advanced HCC	cirrhosis	Early HCC	cirrhosis	Very early HCC	cirrhosis	normal	cirrhosis
sample number	13	10	7	13	10	13	8	13	10	13

Identification of gene signature-related to HCV-HCC on the background of HCV-cirrhosis

HCV-HCC-related genes were defined as highly clustered differentially expressed genes (DEGs) between HCV-cirrhosis and very advanced HCV-HCC tissues which might have pivotal implications in driving HCV-cirrhosis to HCV-HCC. Firstly, we used GEO2R

online software from GEO to determine the DEGs between cirrhosis and very advanced HCC tissues, and the cutoff value of adjust p-value was < 0.05 and |LogFC| was ≥ 4. Next, STRING database 11.0 (https://string-db.org/) combined with Cytoscape software 3.8.0 was used to show the interaction of the DEGs. The MCODE plug-in in the Cytoscape software was used to find the highly clustered DEGs which were the gene signature-related to HCV-HCC on the background of HCV-cirrhosis, and the cutoff criteria were 'maximum depth = 100', 'KCore = 2', 'node score cutoff = 0.2', and 'degree cutoff = 2' [21].

To visualize the expression pattern of the gene signature, heatmaps of the gene expression were generated using R 4.0.1 software to differentiate HCV-HCC at each stage from HCV-cirrhosis. Principal component analysis (PCA) was then used to investigate the classification performance of the gene signature visualized in 3D-plots by scatterplot3d package in R.

Establishing multiple-genes based prediction model

In order to verify that the gene signature could distinguish between HCV-cirrhosis and HCV-HCC, multiple-genes based risk score-related prediction model was established. Higher or lower than median value of risk score was used to differentiate high or low score group.

The formula for risk score is [22]:

$$\text{Risk Score}=\sum_{i=1}^{n} w_i\left(\frac{e_i-u_i}{s_i}\right)$$

(n: the count of genes; w_i: the weight value of the $_i$th gene (Table 2); e_i: the expression level of the $_i$th gene; u_i: mean value for the $_i$th gene among whole samples; s_i: standard deviation value for the $_i$th gene among whole samples.)

Table 2
HCC-related gene signatures

Gene symbol	weight
TOP2A	1
BIRC5	1
CDC20	1
CDK1	1
PTTG1	1
BUB1B	1
NEK2	1
TTK	1
MELK	1
CENPF	1
CDKN3	1
CCNB1	1
PRC1	1
KIF20A	1
PBK	1
ASPM	1
ANLN	1
UBE2T	1
NUF2	1
FAM83D	1

The receiver operating characteristic (ROC) curve was utilized to evaluate how well the gene signature-related risk scores could distinguish between HCV-HCC at different stages and HCV-cirrhosis by using pROC package in R software. We also used ROC curve analysis to determine the cutoff value of risk score. Then, we used higher or lower than the cutoff value of risk score to differentiate high or low score group.

To investigate the possible mechanism of the gene signature related to HCV-HCC on the background of HCV-cirrhosis, we used Database for Annotation, Visualization and Integrated Discovery (DAVID 6.8, https://david.ncifcrf.gov) to perform Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway analysis of the genes in the 20-gene signature. Pathways with p-value less than 0.05 were selected.

Results

Finding gene signature-related to HCV-HCC on the background of HCV-cirrhosis in the discovery group

According to the retrieval criteria, we included the open source dataset from GEO database identified under GSE6764 for HCV-cirrhosis and HCV-HCC tissues. We defined highly clustered DEGs between cirrhosis and very advanced HCC tissues as HCV-HCC-related genes, which might have important implications in driving HCV-cirrhosis to HCV-HCC. After GEO2R analysis, we found 94 genes in very advanced HCC as compared to cirrhosis in the discovery group (Fig. 2A). By using the STRING database, we found totally 252 interactions which was visualized by using Cytoscape software. A group of genes was showed to be highly connected, which might play a pivotal role in the network. Then, this highly clustered gene module was mined by using the MCODE plug-in, holding a high connectivity (cluster score = 19.895) with 20 nodes and 189 edges (Fig. 2B). What caught our attention was that the 20 genes were all up-regulated in very advanced HCC tissues as compared to cirrhosis tissues.

Then, heatmap was carried out to visualize the expression pattern of the 20 genes, and patients in the discovery group was obviously classified into two groups (Fig. 2C), most of which corresponded to the patients' diagnosis, indicating that the 20-gene signature might be able to differentiate very advanced HCC from cirrhosis. Then we performed PCA to verify the classification power of the gene signature. The PCA plot indicated that the 20-gene signature clearly distinguished very advanced HCC from cirrhosis (Fig. 2D). Therefore, this 20-gene signature might have a significant power of discriminating very advanced HCC from cirrhosis patients.

Verifying the gene-signature for discriminating HCV-HCC from HCV-cirrhosis using multiple-genes based prediction model

To verify the classification power of the gene signature in distinguishing between very advanced HCC and cirrhosis tissues, we established a multiple-genes based prediction model with a scoring system representing a linear combination of the 20-gene expression values with a weight value to allocate each sample with a score to measure the possibility of risk [22]. The risk scores from very advanced HCC were significantly higher than those in cirrhosis tissues, which was consistent with our expectations (Fig. 2E). The higher the score, the higher the probability of diagnosis of very advanced HCC. Next, we used higher or lower than median value of risk score to differentiate high or low score group, and the results showed that the prevalence of very advanced HCC in high score group was 100%, while that in low score group was 0%. We then utilized the ROC curve to evaluate this risk score model. The accuracy was 100% with specificity and sensitivity of 100%, and AUC for the ROC curve was 1 (Fig. 2F). Therefore, this risk score-related prediction model based on the 20 genes-biomarker could be used to predict very advanced HCV-HCC from HCV-cirrhosis.

Predicting HCV-HCC at earlier stage using the multiple-genes based prediction model in validated groups

The 20-gene signature-related risk score model was able to predict HCV-induced very advanced HCC from HCV-cirrhosis. However, we were more interested in the potential role of this 20-gene signature in prediction of advanced HCC, early HCC and even very early HCC in validated groups, which was significant in catching HCC as early as possible. We found that the 20 genes were gradually increased as HCC processed from cirrhosis to very advanced stage (Fig. 3). But there was no obvious difference between normal tissues and cirrhosis tissues. Therefore, the 20 genes were strongly associated with HCV-HCC progression and might be used to predict HCV-HCC at earlier stage on the background of HCV-cirrhosis.

Heatmaps was used to display the expression pattern of the 20 genes in validated groups. Except for validated group-4 that the expression profile was similar between normal tissues and cirrhosis tissues, the other validated groups showed obviously upregulated expression pattern of the 20 genes in advanced HCC, early HCC, as well as very early HCC compared to cirrhosis tissues (Fig. 4). We assumed that this 20-gene signature might also be able to discriminate advanced HCC, early HCC or even very early HCC from cirrhosis tissues. PCA plots showed that the 20-gene signature could clearly distinguish advanced HCC tissues, early HCC tissues, as well as very early HCC tissues from cirrhosis tissues, but could not distinguish between cirrhosis and normal tissues.

Next, we used risk score-related prediction model to investigate the role of the 20-gene signature in validated groups. The 20-gene-related risk scores from advanced HCC tissues were significantly higher than those in cirrhosis, while, more importantly, it was repeatedly observed in early HCC and very early HCC as compared to cirrhosis (Fig. 5A and D). Therefore, this 20-gene signature had the statistical power to predict HCC at earlier stage. We used higher or lower than median value of risk score to differentiate high or low score group, and the results showed that the prevalence of advanced HCC in high score group was 70%, while that in low score group was 0% (Fig. 5B). The prevalence of early HCC and very early HCC in high score group was 90.9% and 70% respectively, while that in low score group were 0% and 9.1% respectively. ROC curve evaluation showed that the risk score-related prediction model based on the 20 genes-biomarker got the accuracy of 100%, 100% and 95.2% in advanced HCC, early HCC and very early HCC respectively as compared with cirrhosis tissues in the validation groups, while AUC was 1, 1 and 0.94 respectively. ROC analysis showed that the cutoff value of risk score in the group with cirrhosis and all HCC tissues was − 9.01 (Fig. 5E). Then we used higher or lower than − 9.01 to differentiate high or low score group. The results indicated that the prevalence of advanced HCC, early HCC and very early HCC in high score group was all 100%, while that in low score group was 0%, 0%, and 7.1% respectively (Fig. 5C). Therefore, the risk score-related prediction model based on the 20-genes-biomarker combined with ROC analysis could better discriminate HCC at each stage from cirrhosis. The 20 genes might be associated with HCV-HCC onset and progression on the background of cirrhosis, and they had the statistical power to predict HCV-HCC at earlier stage.

Identifying the core genes among the 20-gene signature related to HCV-HCC on the background of HCV-cirrhosis

In order to explore interactions among the 20 genes, we used DAVID online software to analyze KEGG pathway related to the 20 genes and the top 6 hits were shown in Table 3. The results showed that cell cycle that played an important role in the carcinogenesis or progression of tumors was the most significant pathway, which was consistent with the significant role of cell cycle-related pathogenesis in HCC [23]. Therefore, these genes could not only predict the occurrence of early HCV-HCC, but also might be potential therapeutic targets. In our study, we found CDKN3 and FAM83D were the only two genes that were statistically significant between the normal tissues and the cirrhosis tissues. Interestingly, the expression of these two genes was decreased in HCV-cirrhosis tissues as compared to normal tissues, and was gradually increased as HCV-HCC developed. CDKN3 and FAM83D are the important regulators of cell cycle that however have received little attention, especially in HCV-HCC occurrence and progression on the background of HCV-cirrhosis [24–26]. We hypothesized that CDKN3 and FAM83D played a more significant role in HCV-HCC occurrence and progression and might be better biomarkers for early detection of HCV-HCC on the background of HCV-cirrhosis.

Table 3
KEGG Pathway Enrichment Analysis of the 20 genes

Category	Term	Pvalue	Genes
GOTERM_CC_DIRECT	GO:0030496 ~ midbody	2.30E-08	CDK1, PRC1, NEK2, BIRC5, ASPM, KIF20A
KEGG_PATHWAY	cfa04110: Cell cycle	3.78E-08	CCNB1, CDK1, TTK, BUB1B, CDC20, PTTG1
GOTERM_MF_DIRECT	GO:0005524 ~ ATP binding	4.53E-05	CDK1, NEK2, TTK, PBK, TOP2A, MELK, UBE2T, KIF20A
GOTERM_CC_DIRECT	GO:0005634 ~ nucleus	4.25E-04	CCNB1, NEK2, NUF2, BIRC5, PTTG1, CDKN3, TOP2A, ASPM, MELK, UBE2T
GOTERM_BP_DIRECT	GO:0000910 ~ cytokinesis	4.29E-04	PRC1, BIRC5, KIF20A
GOTERM_CC_DIRECT	GO:0005737 ~ cytoplasm	7.70E-04	CDK1, PRC1, NEK2, CDC20, BIRC5, PTTG1, CDKN3, ASPM, MELK, UBE2T

Investigating the role of risk score-related prediction model based on CDKN3 and FAM83D

In order to implement the biomarker easily in real life and make it economical, we tried to limit the gene numbers of the 20-genes-based biomarker. Then we established risk score-related prediction model based on CDKN3 and FAM83D according to the interesting result that the two genes were first downregulated in HCV-cirrhosis as compared to normal tissues and then increased as HCV-cirrhosis processed to HCV-HCC. CDKN3 and FAM83D-related risk scores from very advanced HCC tissues were significantly higher than those in cirrhosis, while, more importantly, it was repeatedly observed in advanced HCC, early HCC and very early HCC as compared to that in cirrhosis (Fig. 6A). The prediction model based on this 2 genes-biomarker got the predictive accuracy of over 95.2% with AUC over 0.96, revealing even a little bit better performance than that based on the 20-genes-based biomarker (Fig. 6B). Therefore, limiting the gene number to only two genes (CDKN3 and FAM83D) improved the performance of the multiple-gene based risk score-related prediction method. ROC analysis showed that the cutoff value of risk score in the group with cirrhosis and all HCC tissues was − 1.114. Then we used higher or lower than − 1.114 to differentiate high or low score group (Fig. 6C-6F). The results indicated that the prevalence of very advanced HCC, advanced HCC, early HCC and very early HCC in high score group was all 100%, while that in low score group was 0%, 0%, 0% and 7.1% respectively. Therefore, the 2-genes-based biomarker could well predict HCV-HCC at earlier stage on the background of HCV-cirrhosis, and might be the potential treatment target in the future.

Disscusion

HCC is one of the most common and deadly malignant tumors worldwide [27]. HCV infection is a pivotal cause of cirrhosis, with significantly increased incidence of HCC development [5]. In patients with HCV-HCC, the morbidity and mortality remain high with late diagnosis and poor outcome. Therefore, there is an urgent need to find potential markers for early diagnosis and treatment on the background of cirrhosis. Successful HCV-HCC animal models are more difficult to build. Meanwhile, liver biopsy is the most reliable method for the diagnosis of liver cancer, but it is an invasive test with potential complications. At present, a large amount of microarray data or RNA-seq data was published, but there is a lack of sufficient analysis to provide further guidance for clinical and scientific researches. Briefly, in our study, we investigated the data with HCV-cirrhosis tissues and HCV-HCC tissues from GSE6764 to compare the gene expression changes between HCC tissues at different stages and cirrhosis tissues. We found 20 hub genes which might be able to diagnose asymptomatic HCV-HCC patients at very early stage, more importantly, these genes could be used for screening HCV-HCC with relatively high accuracy. In order to implement the biomarker easily in real life and make it economical, we limited the gene numbers of the 20-genes-based biomarker to two genes (CDKN3 and FAM83D). The prediction model based on this 2 genes-biomarker showed even a little bit better performance than that based on the 20-genes-based biomarker. Therefore, CDKN3 and FAM83D played a significant role in HCV-HCC occurrence and progression, and this 2 genes-biomarker might be used to predict HCV-HCC at earlier stage in the future as well as new potential treatment target.

Here, firstly, the group with HCV-cirrhosis and very advanced HCV-HCC tissues were classified as the discovery group. A highly clustered module with 20 genes were identified in the discovery group. We established the risk score-related prediction model based on the 20-gene signature in the discovery group to investigate its discriminating role. The risk score in the very advanced HCC tissues was significantly higher than that in cirrhosis tissues. Interestingly, the prediction model got the accuracy of 100% with AUC for the ROC curve being 1. Therefore, the 20-gene signature could be used to predict very advanced HCC from cirrhosis, and this risk score-related prediction model based on multiple-genes might be used to investigate the role of the 20-gene signature in HCV-HCC at earlier stages.

Next, we were more interested in the potential role of this 20-gene signature in prediction of HCC at earlier stages in validated groups, which was significant in catching HCC as early as possible. The results showed that the 20 genes were progressively increased from cirrhosis to advanced HCC stages. Heatmap and PCA plots showed that these genes could discriminate advanced HCC, early HCC or even very early HCC from cirrhosis tissues. The risk score-related prediction model based on the

20 genes-biomarker got the accuracy of over 95.2% with AUC over 0.94, indicating that these HCV-HCC-related genes could be used for early detection of HCV-HCC. Then we used higher or lower than the cutoff value of risk score to differentiate high or low score group, the prevalence of HCV-HCC at any of the stages in high score group was 100%. The risk score-related prediction model based on multiple-genes could well predict HCV-HCC at earlier stage on the background of HCV-cirrhosis, which might be also used in other studies to investigate the prediction role of multiple-genes biomarker.

However, it is better to limit the gene number of multiple-genes biomarker to implement the biomarker easily in real life and to make it economical. Among the 20 genes, interestingly, the expression of CDKN3 and FAM83D was decreased in cirrhosis tissues as compared to normal tissues, and gradually increased during HCC progression. Therefore, these two genes might have better predictive ability to identify HCV-HCC at early stage on the background of HCV-cirrhosis. Consistent with our finding, CDKN3 and FAM83D were important regulators of cell cycle, and studies had also shown that they were highly expressed in tumor tissues compared to normal tissues, and overexpression of CDKN3 and FAM83D was correlated with the poor outcome in HCC patients [24, 29–33]. However, studies focusing on the role of these two genes in HCV-HCC occurrence and progression on the background of HCV-cirrhosis are limited. Then we established the risk score-related prediction model based on CDKN3 and FAM83D, indicating that prevalence of HCV-HCC in high score group was still 100% and the prediction model got the predictive accuracy of over 95.2% with AUC over 0.96, showing even a little bit better performance than that based on the 20-genes-based biomarker. Therefore, CDKN3 and FAM83D played a significant role in HCV-HCC occurrence and progression, and this 2 genes-biomarker was able to predict HCV-HCC at earlier stage with high accuracy using the risk score-related prediction model.

In the past, studies focusing on analysis of HCC-related genes usually included patients with HCC caused by various factors such as chronic hepatitis B or C virus infection and alcoholic liver disease [18, 19, 34]. But there is limited research that only discusses HCV-related liver cancer on the background of HCV-cirrhosis, so we re-analyzed the data in GSE6764. We investigated the development of HCC at different stages on the background of cirrhosis which were all caused by hepatitis C. This was better for us to understand development of HCV-HCC from cirrhosis and HCC progression process. Compared with other studies, we included normal liver tissues instead of liver tissues adjacent to tumors,which helped us to detect hub genes between cirrhosis and normal tissues to investigate whether the genes also played a role in development of cirrhosis. No single biomarker has adequate sensitivity or specificity for HCC diagnosis. More and more studies are beginning to notice the use of genome-wide expression analysis to predict and diagnose diseases [35, 36]. Compared with the single-gene biomarker, our multiple-genes based risk score-related prediction model got high accuracy with high sensitivity and specificity, which could be also used in other studies.

Taken together, our study found 20 hub genes related to the progression of HCV-HCC on the background of HCV-cirrhosis, while more importantly, the risk score-related prediction model based on the 20 genes got high accuracy in discriminating HCV-HCC at earlier stage on the background of HCV-cirrhosis. It is more important that we went further to identify CDKN3 and FAM83D-based biomarker which could have better predictive accuracy to discriminate HCV-HCC at earlier stage from HCV-cirrhosis. Actually, there are a lot of researches focusing on developing novel biomarkers for HCC with few studies focusing on HCV-HCC, but further work is always needed to implement the utilization of novel biomarker to meet the real clinical demand. Now, identification of cost-efficient novel biomarkers with predictive technology such as risk score-related prediction model used in our study for the detection of diseases will be promising. In this study, only GSE6764 met the screening criteria, making it difficult to verify our result in other studies currently. Further work about the expression profile and predictive effect of this 2-genes signature in HCV-HCC on the background of HCV-cirrhosis is needed to be performed with a larger sample size.

Conclusion

In conclusion, this study found that the upregulated 20 hub genes were closely related to the entire spectrum of development of HCV-HCC on the background of HCV-cirrhosis. Multiple-genes based risk score-related prediction model was established to investigate the role of the 20-genes-based biomarker as well as CDKN3- and FAM83D-based biomarker in identifying HCV-HCC at earlier stage. Interestingly, CDKN3 and FAM83D that were both important regulators of cell cycle could better predict HCV-

HCC at earlier stage on the background of HCV-cirrhosis with better performance using the multiple-genes based risk score-related prediction model. Therefore, CDKN3- and FAM83D-based biomarker might be used to predict HCV-HCC at earlier stage on the background of HCV-cirrhosis in the future as well as new potential treatment target in HCV-HCC.

Abbreviations

HCV: hepatitis C virus; HCC: Hepatocellular carcinoma; ROC: receiver-operating characteristic; GEO: Gene Expression Omnibus; DEGs: differentially expressed genes; PCA: Principal component analysis (PCA)

Declarations

Ethics approval and consent to participate

Not applicable

Consent for publication

Not applicable

Availability of data and materials

The data of this study are from GEO database. The data that support the findings of this study are available from the corresponding author upon reasonable request.

Authors' contributions

Chao Du and Xiaozhong Gao conceived of and designed the study. Lin Lu performed literature search. Jiao Nie generated the figures and analyzed the data. Jiao Nie wrote the manuscript and Chao Du critically reviewed the manuscript. Xiaozhong Gao supervised the research. Jiao Nie and Chao Du contributed equally to this paper. All authors read and approved the final manuscript.

Acknowledgements

Not applicable

Funding

This work was supported by National Natural Science Foundation of China (NO. 8150040611).

Competing interests

The authors declare that they have no competing interests.

References

1. Elia G, Fallahi P. Hepatocellular carcinoma and CXCR3 chemokines: a narrative review. Clin Ter. 2017;168(1):e37-e41.
2. Arzumanyan A, Reis HM, Feitelson MA. Pathogenic mechanisms in HBV- and HCV-associated hepatocellular carcinoma. Nat Rev Cancer. 2013;13(2):123-35.
3. European Association for the Study of the Liver. Electronic address eee, European Association for the Study of the L. EASL Clinical Practice Guidelines: Management of hepatocellular carcinoma. J Hepatol. 2018;69(1):182-236.
4. El-Serag HB. Epidemiology of viral hepatitis and hepatocellular carcinoma. Gastroenterology. 2012;142(6):1264-73 e1.

5. Axley P, Ahmed Z, Ravi S, Singal AK. Hepatitis C Virus and Hepatocellular Carcinoma: A Narrative Review. J Clin Transl Hepatol. 2018;6(1):79-84.
6. Omland LH, Krarup H, Jepsen P, Georgsen J, Harritshoj LH, Riisom K, et al. Mortality in patients with chronic and cleared hepatitis C viral infection: a nationwide cohort study. J Hepatol. 2010;53(1):36-42.
7. Dhir M, Lyden ER, Smith LM, Are C. Comparison of outcomes of transplantation and resection in patients with early hepatocellular carcinoma: a meta-analysis. HPB (Oxford). 2012;14(9):635-45.
8. Choi DT, Kum HC, Park S, Ohsfeldt RL, Shen Y, Parikh ND, et al. Hepatocellular Carcinoma Screening Is Associated With Increased Survival of Patients With Cirrhosis. Clin Gastroenterol Hepatol. 2019;17(5):976-87 e4.
9. West CA, Black AP, Mehta AS. Analysis of Hepatocellular Carcinoma Tissue for Biomarker Discovery. Hepatocellular Carcinoma. Molecular and Translational Medicine2019. p. 93-107.
10. Wurmbach E, Chen Y-b, Khitrov G, Zhang W, Roayaie S, Schwartz M, et al. Genome-wide molecular profiles of HCV-induced dysplasia and hepatocellular carcinoma. Hepatology. 2007;45(4):938-47.
11. Kaur H, Dhall A, Kumar R, Raghava GPS. Identification of Platform-Independent Diagnostic Biomarker Panel for Hepatocellular Carcinoma Using Large-Scale Transcriptomics Data. Front Genet. 2019;10:1306.
12. Nault JC, De Reynies A, Villanueva A, Calderaro J, Rebouissou S, Couchy G, et al. A hepatocellular carcinoma 5-gene score associated with survival of patients after liver resection. Gastroenterology. 2013;145(1):176-87.
13. Xia Q, Li Z, Zheng J, Zhang X, Di Y, Ding J, et al. Identification of novel biomarkers for hepatocellular carcinoma using transcriptome analysis. Journal of Cellular Physiology. 2019;234(4):4851-63.
14. Ji J, Chen H, Liu XP, Wang YH, Luo CL, Zhang WW, et al. A miRNA Combination as Promising Biomarker for Hepatocellular Carcinoma Diagnosis: A Study Based on Bioinformatics Analysis. J Cancer. 2018;9(19):3435-46.
15. Komatsu H, Iguchi T, Masuda T, Ueda M, Kidogami S, Ogawa Y, et al. HOXB7 Expression is a Novel Biomarker for Long-term Prognosis After Resection of Hepatocellular Carcinoma. Anticancer Res. 2016;36(6):2767-73.
16. Barrett T, Wilhite SE, Ledoux P, Evangelista C, Kim IF, Tomashevsky M, et al. NCBI GEO: archive for functional genomics data sets–update. Nucleic Acids Res. 2013;41(Database issue):D991-5.
17. Cancer Genome Atlas Research N, Weinstein JN, Collisson EA, Mills GB, Shaw KR, Ozenberger BA, et al. The Cancer Genome Atlas Pan-Cancer analysis project. Nat Genet. 2013;45(10):1113-20.
18. Zhou Z, Li Y, Hao H, Wang Y, Zhou Z, Wang Z, et al. Screening Hub Genes as Prognostic Biomarkers of Hepatocellular Carcinoma by Bioinformatics Analysis. Cell Transplant. 2019;28(1_suppl):76S-86S.
19. Wu M, Liu Z, Zhang A, Li N. Identification of key genes and pathways in hepatocellular carcinoma: A preliminary bioinformatics analysis. Medicine (Baltimore). 2019;98(5):e14287.
20. Wurmbach E, Chen YB, Khitrov G, Zhang W, Roayaie S, Schwartz M, et al. Genome-wide molecular profiles of HCV-induced dysplasia and hepatocellular carcinoma. Hepatology. 2007;45(4):938-47.
21. Bader GD, Hogue CW. An automated method for finding molecular complexes in large protein interaction networks. BMC Bioinformatics. 2003;4:2.
22. Feng A, Rice AD, Zhang Y, Kelly GT, Zhou T, Wang T. S1PR1-Associated Molecular Signature Predicts Survival in Patients with Sepsis. Shock. 2020;53(3):284-92.
23. Tripathi V, Shen Z, Chakraborty A, Giri S, Freier SM, Wu X, et al. Long noncoding RNA MALAT1 controls cell cycle progression by regulating the expression of oncogenic transcription factor B-MYB. PLoS Genet. 2013;9(3):e1003368.
24. Wang L, Sun L, Huang J, Jiang M. Cyclin-dependent kinase inhibitor 3 (CDKN3) novel cell cycle computational network between human non-malignancy associated hepatitis/cirrhosis and hepatocellular carcinoma (HCC) transformation. Cell Prolif. 2011;44(3):291-9.
25. Nalepa G, Barnholtz-Sloan J, Enzor R, Dey D, He Y, Gehlhausen JR, et al. The tumor suppressor CDKN3 controls mitosis. J Cell Biol. 2013;201(7):997-1012.

26. Fulcher LJ, He Z, Mei L, Macartney TJ, Wood NT, Prescott AR, et al. FAM83D directs protein kinase CK1alpha to the mitotic spindle for proper spindle positioning. EMBO Rep. 2019;20(9):e47495.
27. Jemal A, Bray F, Center MM, Ferlay J, Ward E, Forman D. Global cancer statistics. CA Cancer J Clin. 2011;61(2):69-90.
28. Cazzagon N, Trevisani F, Maddalo G, Giacomin A, Vanin V, Pozzan C, et al. Rise and fall of HCV-related hepatocellular carcinoma in Italy: a long-term survey from the ITA.LI.CA centres. Liver Int. 2013;33(9):1420-7.
29. Wang L, Sun L, Huang J, Jiang M. Cyclin-dependent kinase inhibitor 3 (CDKN3) novel cell cycle computational network between human non-malignancy associated hepatitis/cirrhosis and hepatocellular carcinoma (HCC) transformation. Cell Proliferation. 2011;44(3):291-9.
30. Babic M, Dimitropoulos C, Hammer Q, Stehle C, Heinrich F, Sarsenbayeva A, et al. NK cell receptor NKG2D enforces proinflammatory features and pathogenicity of Th1 and Th17 cells. J Exp Med. 2020;217(8).
31. Wu M, Liu Z, Li X, Zhang A, Lin D, Li N. Analysis of potential key genes in very early hepatocellular carcinoma. World J Surg Oncol. 2019;17(1):77.
32. Wang D, Han S, Peng R, Wang X, Yang XX, Yang RJ, et al. FAM83D activates the MEK/ERK signaling pathway and promotes cell proliferation in hepatocellular carcinoma. Biochem Biophys Res Commun. 2015;458(2):313-20.
33. Lin B, Chen T, Zhang Q, Lu X, Zheng Z, Ding J, et al. FAM83D associates with high tumor recurrence after liver transplantation involving expansion of CD44+ carcinoma stem cells. Oncotarget. 2016;7(47):77495-507.
34. Li L, Lei Q, Zhang S, Kong L, Qin B. Screening and identification of key biomarkers in hepatocellular carcinoma: Evidence from bioinformatic analysis. Oncol Rep. 2017;38(5):2607-18.
35. Yin L, He N, Chen C, Zhang N, Lin Y, Xia Q. Identification of novel blood-based HCC-specific diagnostic biomarkers for human hepatocellular carcinoma. Artif Cells Nanomed Biotechnol. 2019;47(1):1908-16.
36. Jiang CH, Yuan X, Li JF, Xie YF, Zhang AZ, Wang XL, et al. Bioinformatics-based screening of key genes for transformation of liver cirrhosis to hepatocellular carcinoma. J Transl Med. 2020;18(1):40.

Figures

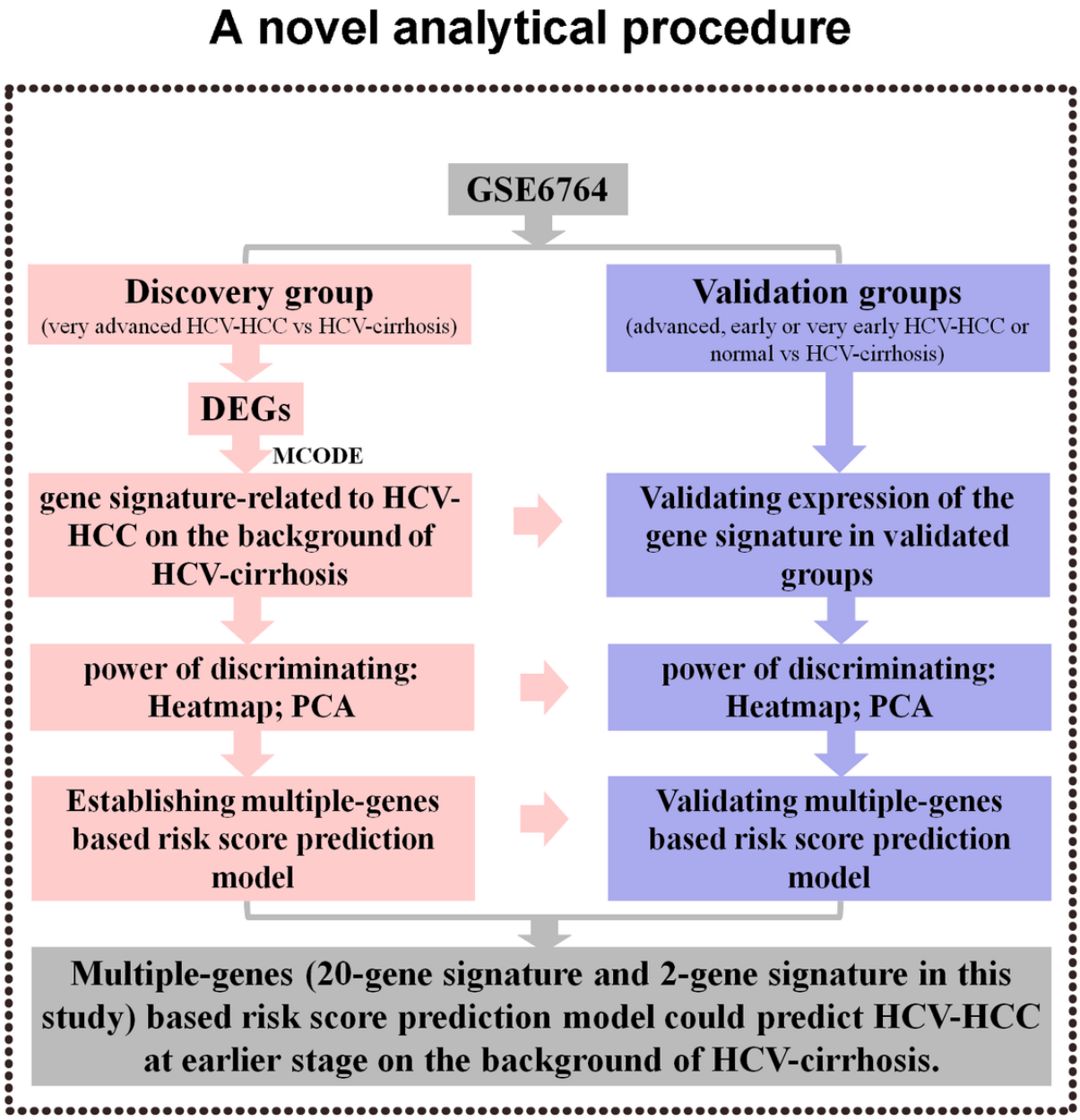

Figure 1

Workflow

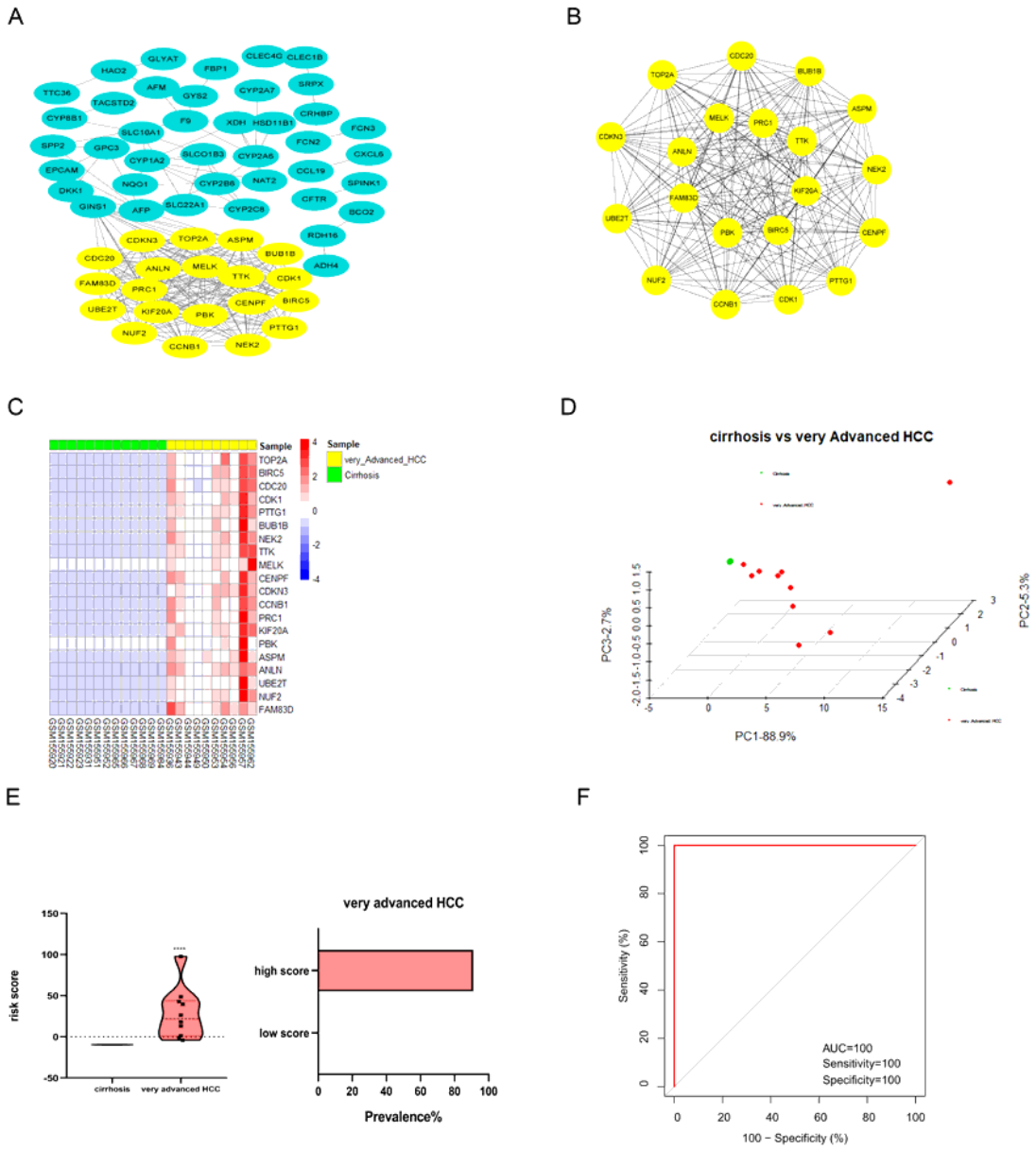

Figure 2

The 20-gene signature differentiated very advanced HCC from cirrhosis in discovery group. (A) HCC-related network on the background of cirrhosis was visualized using Cytoscape software. (B) HCV-HCC-related gene-signature. The 20 genes in this module were highly connected. (C)Heatmap showed the 20-gene signature expression in the discovery group. (D)PCA plot indicated that the 20-gene signature clearly distinguished very advanced HCC from cirrhosis. (E)Violin plot of the risk scores in discovery group. (F)ROC curves of the 20-gene signature in distinguishing very advanced HCC from cirrhosis.

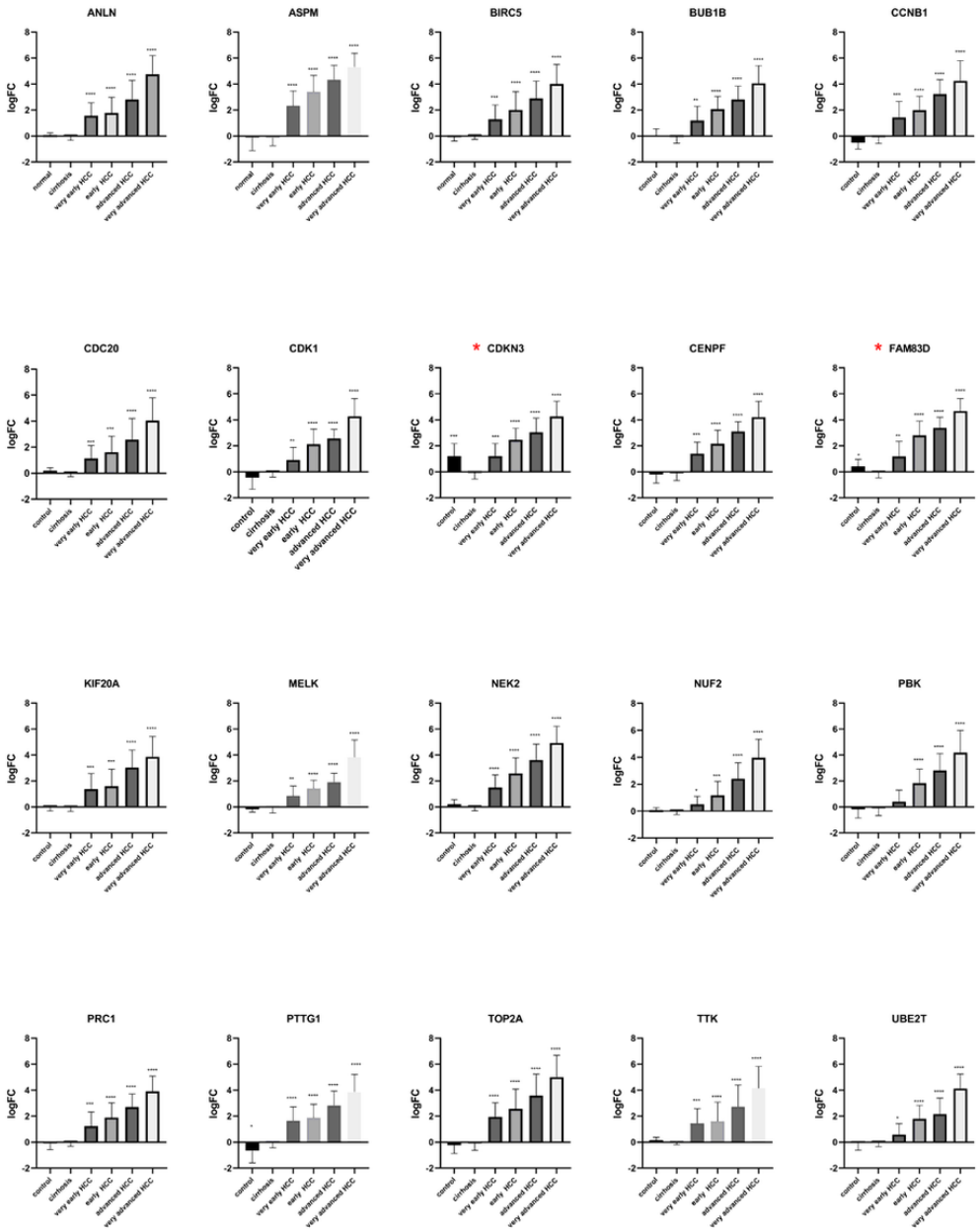

Figure 3

20 hub genes were gradually increased as HCV-HCC processed from HCV-cirrhosis to very advanced stage

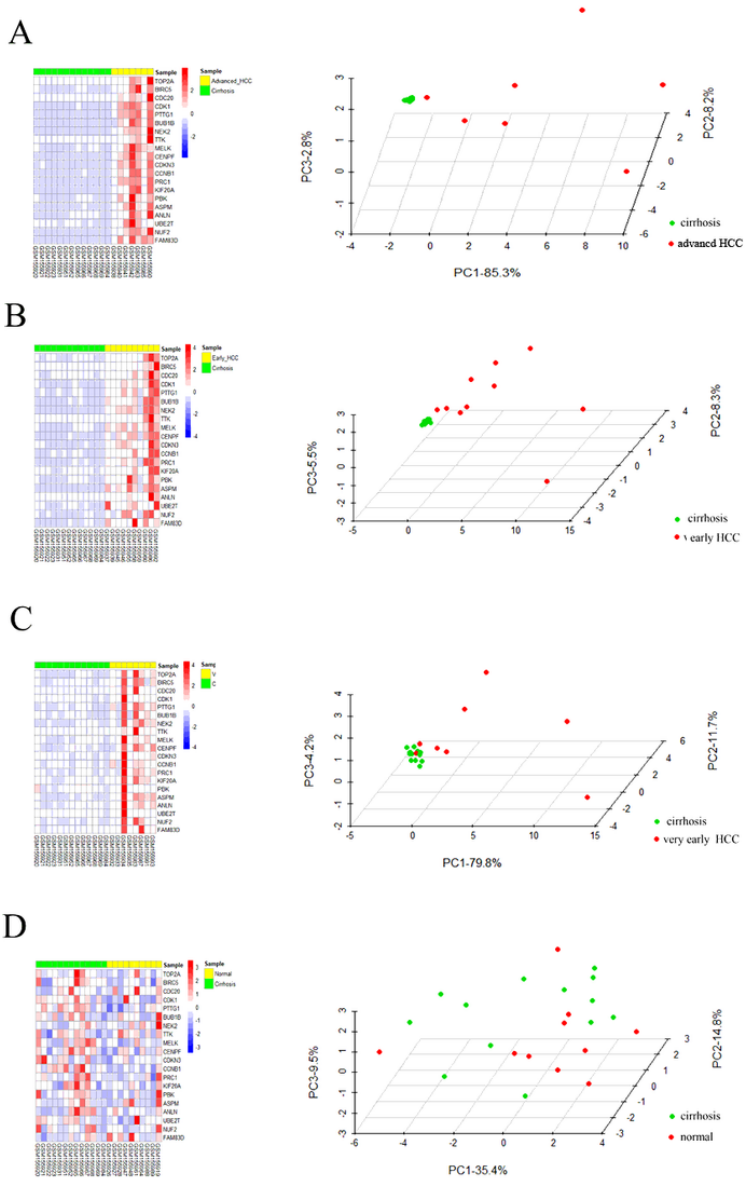

Figure 4

Heatmap and PCA plot of the 20-gene signature for the validation groups. (A) Heatmap and PCA plot of the 20-gene signature in advanced HCC vs cirrhosis. (B) Heatmap and PCA plot of the 20-gene signature in early HCC vs cirrhosis. (C)Heatmap and PCA plot of the 20-gene signature in very early HCC vs cirrhosis. (D) Heatmap and PCA plot of the 20-gene signature in normal vs cirrhosis.

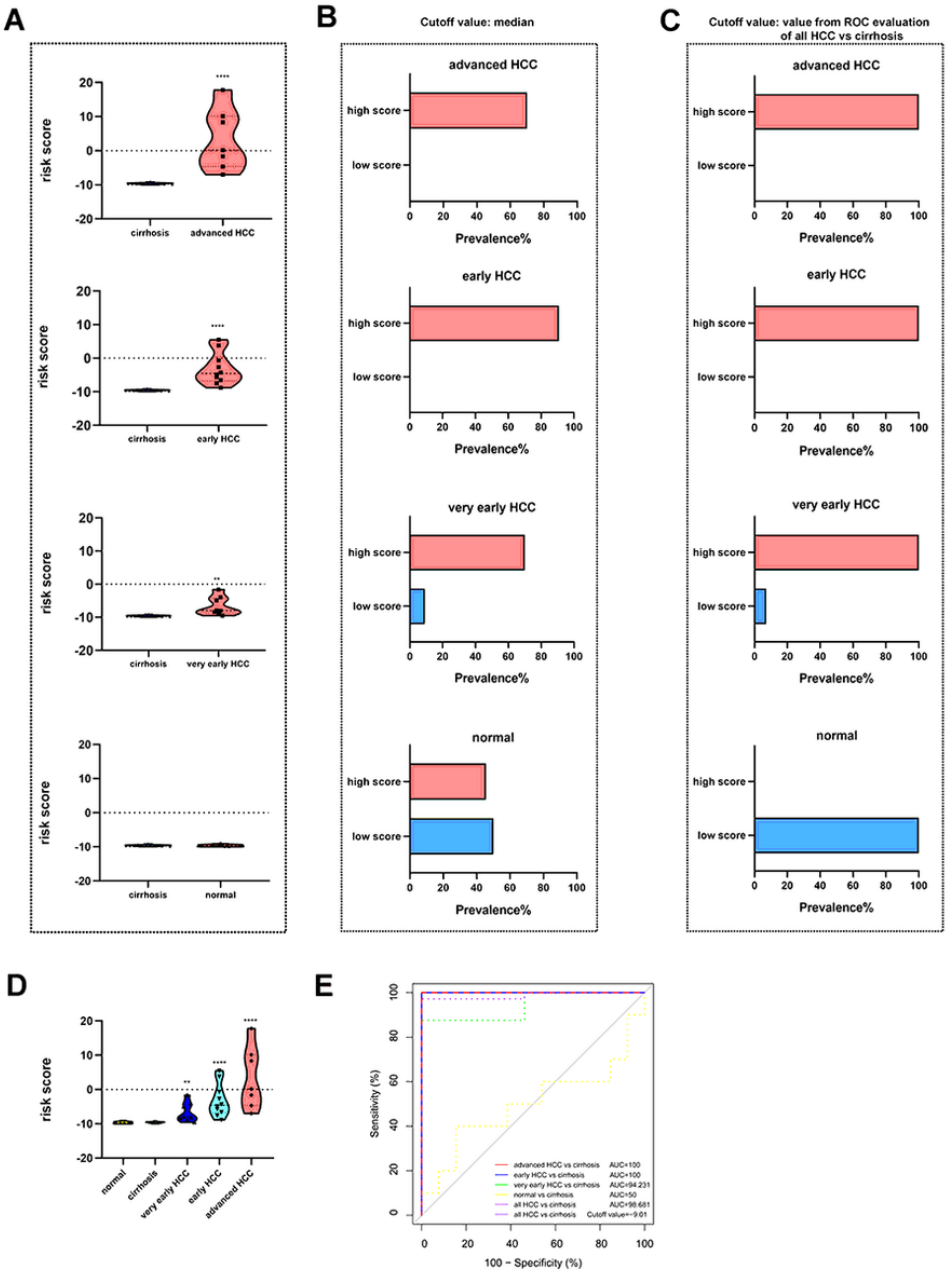

Figure 5

The 20-gene signature-based risk score differentiated HCV-HCC from cirrhosis in validation groups. (A) The 20-gene-related risk scores in validation groups were significantly higher than those in cirrhosis. (B) Prevalence of HCV-HCC. Higher or lower than median value of risk score was used to differentiate high or low score group. (C) Prevalence of HCV-HCC. Higher or lower than the cutoff value of risk score was used to differentiate high or low score group. (D)Summary of the 20-gene-related risk scores in validation groups. (E)ROC curves of the20-gene signature in distinguishing HCV-HCC from cirrhosis.

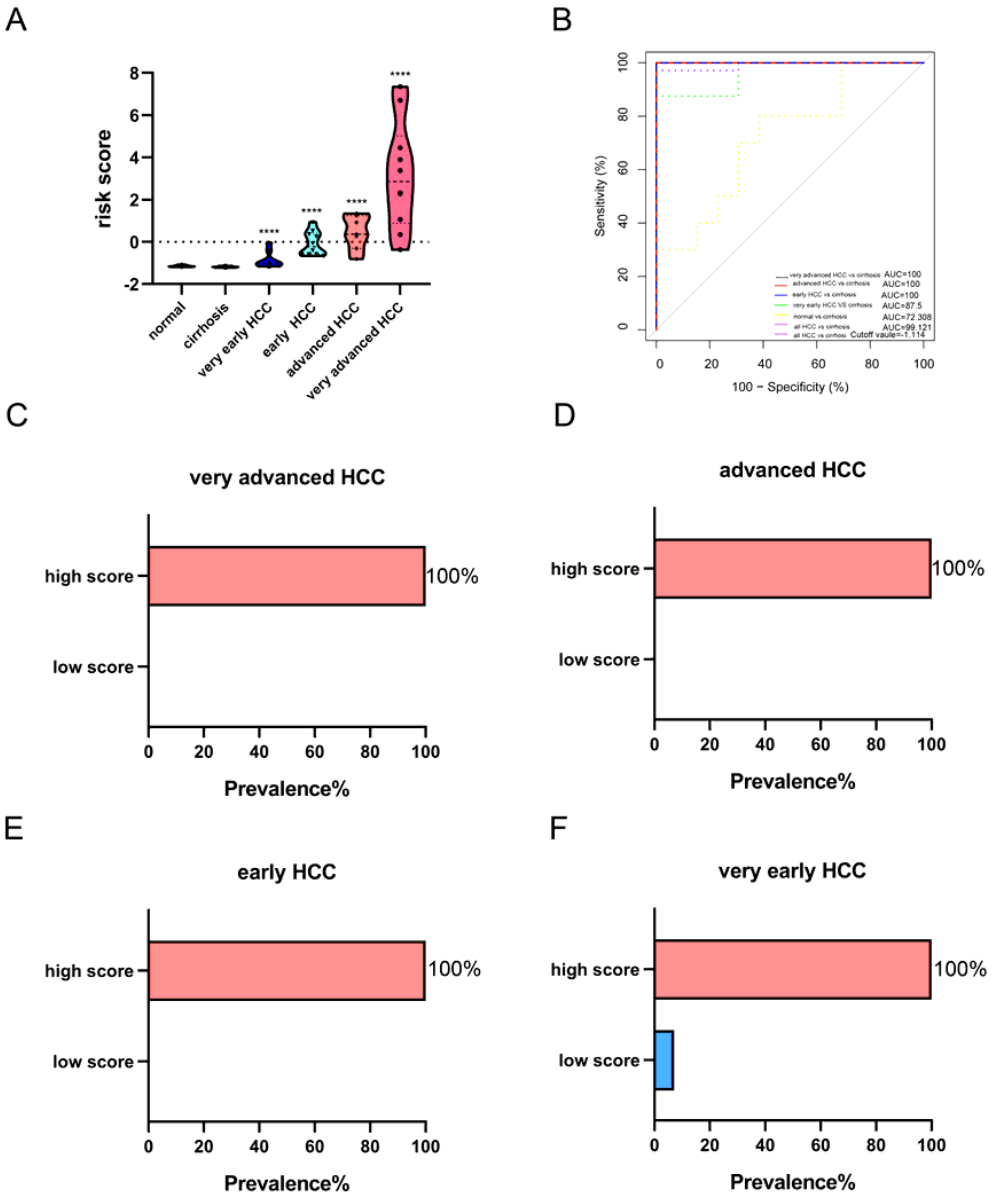

Figure 6

CDKN3 and FAM83D-based risk score differentiated HCV-HCC from cirrhosis. (A)The 2-gene-related risk scores in HCV-HCC were significantly higher than those in cirrhosis. (B) ROC curves of the 2-gene signature in distinguishing HCV-HCC from cirrhosis. (C-F) Prevalence of HCV-HCC.