Genetic expression and mutation profile analysis in different pathologic stages of hepatocellular carcinoma patients

Xingjie Gao
Tianjin Medical University
Chunyan Zhao
Tianjin Medical University
Xiaoteng Cui
Tianjin Medical University
Nan Zhang
Tianjin Medical University
Yuanyuan Ren
Tianjin Medical University
Chao Su
Tianjin Medical University
Shaoyuan Wu
Tianjin Medical University
Zhi Yao
Tianjin Medical University
Jie Yang (yangj@tmu.eud.cn)
Tianjin Medical University https://orcid.org/0000-0002-9669-1832

Primary research

Keywords: Expression, mutation, HCC, TCGA, pathologic stage

Posted Date: September 9th, 2020

DOI: https://doi.org/10.21203/rs.3.rs-71784/v1

Abstract

Background: The expression and mutation of multiple genes are involved in the complicated mechanism regarding the occurrence and development of hepatocellular carcinoma (HCC). The clinical pathological stage of HCC is closely linked to clinical prognosis of liver cancer. This study aims at analyzing the gene expression and mutation profile of different clinical pathological stages of HCC (stage I, II, III-IV), based on 367 HCC cases included in TCGA cohort.

Results: We identified a series of targeting genes with copy number variation (CNV), which is statistically associated with gene expression. For instance, compared withthe normal group, CCNE2 gene is highly expressed in the tumor group and specificstage I group, which are associated withthree CNV types of single deletion, single gain, and amplification mutations. Protein interaction network construction and followed "Molecular Complex Detection" analysis indicated that the high expression of some cell cycle-related genes in HCC, such as TTK, CDC20, ASPM, is positively correlated with CNV. Non-synonymous mutations mainly existed in some genes, such as TTN, TP53, CTNNB1, MUC16, andALB, however, we did not observe the association between thegene mutation frequency and the clinical pathological grade distribution. The rs121913396 and rs121913400 polymorphisms withintheCTNNB1 gene were associated with the high expression of CTNNB1 protein, but not linked to the clinical prognosis of HCC. We performed the random forest and decision tree approachesfor the modeling analysis and identified a group of genes related to different HCC pathological grades, such as the lowly expressed VIPR1, FAM99A, and GNA14 genes, or highly expressed CEP55, SEMA3F, and PRR11. Moreover, we conducted a principal component analysis (PCA) to obtain several genes associated with different pathological grades, including SLC27A5, ADAM17, SNRPA, SNRPD2, and ALDH2. Finally, we confirmed the highly expressed GAS2L3, SNRPA, SNRPD2 genes in the HCC tissues, for the first time, through a Chinese HLivH060PG02 cohort analysis.

Conclusions: The identification of the targeting genes, including GAS2L3, SNRPA, SNRPD2, provides insight into the molecular mechanisms associated with different prognosis of HCC.

Background

Hepatocellular carcinoma (HCC) is the primary histological subtype of liver cancer [1–3]. A series of factors, including the genetic, epigenetic changes, chronic hepatitis B/C virus infection, aflatoxin exposure, smoking, obesity, and diabetes, contribute to the progression, diagnosis, and prognosis of HCC [4–6]. The clinical pathological stage of HCC is closely linked to clinical prognosis of liver cancer [4, 5]. For the HCC cases with early pathological stage, the radical therapies (e.g., resection, radiofrequency ablation, transplantation, et al.) are valid and feasible [1]. It is thus meaningful to identify the potential genes, which is associated with the pathological stage I, II, III-IV of HCC.

The Cancer Genome Atlas (TCGA), a public database, provides the multiple-genomics data from more than thirteen types of cancer, including gene expression, copy number variation (CNV), simple nucleotide variation (SNV), single nucleotide polymorphism (SNP), DNA methylation, and clinical information, etc. [7]. TCGA cohorts enrolled a total of more than 360 HCC cases, and the related gene expression and mutation information are available. In this study, we first performed the statistical analysis, random forest, decision tree, and principal component analysis to identify the differential gene expression, CNV, SNV and SNP profiles, which are associated with the different pathologic stages of TCGA HCC cases. We also analyzed the expression levels of some targeting genes in a Chinese HLivH060PG02 HCC cohort.

Results

Different pathological grades of HCC in the TCGA cohort

We obtained the expression and clinical data of 367 hepatocellular carcinomas, 3 fibrolamellar carcinomas, 7 hepatobiliary mixed carcinomas, and 50 adjacent normal controls from the TCGA-LIHC project (Fig. 1A). We first investigated the association between the histological grades of HCC (Fig. 1B, G1, G2, G3, and G4) and clinical outcomes of liver cancer. As shown in Fig. 1C, there was no statistically significant difference in overall survival (OS) and disease-free survival (DFS) between different histological grades (*P* value for Log-rank analysis > 0.05). Figure 1D specifically shows the clinical pathological stages (stage I, II, III-IV) and TNM staging of HCC cases. As expected, stage III-IV or T4 patients had the worst prognosis, whereas stage I or T1 patients had a better prognosis (Fig. 1E, $P < 0.001$). Therefore, our study focused on HCC cases.

We analyzed the correlation between different pathological stage (stage I, II, III-IV) and clinical indicators. As shown in Figure S1A-F, the total bilirubin, albumin, fetoprotein, and platelet count indicators, but not creatine and protherombin time, showed a statistically significant association with the different HCC pathological grades. In addition, we did not observe a correlation between HCC pathological grades and other factors, including age, height, weight, race, cluster, and gender (Figure S1G-L). Our study aims at analyzing the gene expression and mutational profiles associated with different clinical pathological grades of HCC in the TCGA cohort.

Differential gene screening

First, we attempted to screen the genes that exhibit the increment or decrement trend in the normal, stage I, stage II, and stage III-IV groups. A range of differential genes between three comparison groups, including Tumor vs. Normal, stage II vs. stage III, stage III+IV vs. stage II, were identified, using the “EdgeR” package. Figure S2A presents the volcano plots for the above three sets. Then, we performed the intersection analysis of the up-regulated and down-regulated genes. As shown in Figure S2B-C, twelve up-regulated genes were screened, but no down-regulated genes were obtained. The twelve up-regulated

genes did not show the protein interaction relationship (Figure S2D) and mainly exist in stage III+IV, but not with a high proportion (Figure S2D). Figure S2E shows the full name information of these genes.

Next, based on the GEPIA online database, we analyzed the expression levels of these genes in normal and tumor, and different pathological grades of HCC. As shown in Figure S3A-B, except for the *CRTAC1* gene, other genes were highly expressed in the tumor group, compared with the normal control group. However, only the gene expressions of *DUCX2*, *IQCA1*, *PCSK1*, *HOXB9*, *KCNH2*, and *NPTX1* were statistically correlated with the stage I-IV distribution. The results of the overall survival and disease-free survival analysis further indicated that the high expression levels o*f CUZD1* and *IQCA1* were associated with poor prognosis of HCC (Figure S3C).

Copy number variation analysis

We performed the somatic copy number variation (CNV) profile and identified a total of 16,644 genes with CNV from the TCGA HCC dataset. And the Circos 2D track plot for the CNV distribution in the chromosomes was shown in Figure 2A. Then, we utilized the Kolmogorov-Smirnov test to analyze the correlation between CNV and gene expression and obtained a series of genes. After the GO and KEGG analysis, we found that most of these genes were involved in the cell division or cell cycle processes, such as organelle fission, nuclear division, and spindle location (Figure 2B-E). For instance, cell cycle-associated *CCNE2* gene in the Tumor, stage I, and stage II group exhibits the single deletion (sd) and single gain (sg) mutations, which are correlated with the gene expression of CCNE2 protein. However, the *GADD45G* expression level in HCC cases is higher than the negative controls, hinting the presence of other potential gene expression inhibition mechanisms (Figure 2G). Figure S4 shows some CNV-driven genes involved in the cell cycle pathway.

Protein-protein interaction network analysis

Based on the above identified genes, we constructed a protein-protein interaction (PPI) network using the “STRINGdb” package and “Cytoscape” software. We also performed the "Molecular Complex Detection" (MCODE) modular analysis to screen some hub genes within the PPI network. Figure 3A-B shows the two modules with the highest ratings. We further found that the expression levels of these hub genes within the two modules were statistically correlated with CNV. And the cell cycle-related genes, such as *TTK*, *CDC20* and *ASPM*, were highly expressed and exhibited a significant positive correlation with CNV (Figure 3C).

Genetic mutation analysis

We downloaded the HCC-related simple nucleotide variation data from the TCGA database, and selected the top 15 genes with the most frequent mutation frequency, such as *TTN*, *TP53*, *CTNNB1*, *MUC16*, and

ALB, to map the waterfall with clinical grading information. As shown in Figure S7A, the mutation types of these genes are mainly non-synonymous mutations; however, the gene mutation frequency in 286 HCC patients with mutations is not associated with the clinical pathology of HCC (stage I, II, III-IV). In addition, the high expression of CTNNB1 protein was related to the *CTNNB1* mutation in overall HCC, stage I, II, III-IV groups (Figure S7B). *TP53* gene mutation was associated with the reduced expression of TP53 in the overall HCC, stage I, II groups (Figure S7B). The *OBSCN* mutation also correlated with the low expression of *OBSCN* in overall HCC and specific stage I groups (Figure S7B).

We further conducted the waterfall map analysis on the above-mentioned *CENPF*, *ASPM*, *MELK*, *TTK*, *GADD45G*, *CDC20*, *CCNE2,* and other interesting genes, and did not observe the correlation between the low mutation frequency of these genes and pathological stages or gene expression, although mainly non-synonymous mutations as well (Figure S6). In addition, variations in the *CTNNB1*, *TP53*, *TTN*, and *OBSCN* genes were not found to be linked to the clinical prognosis in different pathological grades (Figure S7-8).

Next, we extracted the SNP data of HCC from the TCGA cohort and found that the rs121913396, rs121913400, rs121913407 SNP of *CTNNB1* and rs28934571 SNP of *TP53* gene were relatively high frequency (Figure S9A). There are more than ten types of SNP for *CTNNB1* gene (Figure S9B). Compared with the wild type group, the *CTNNB1* gene with rs121913396 and rs121913400 showed a higher expression level (Figure S9C). Nevertheless, we did not observe the positive correlation between the rs121913396, rs121913400, rs121913407 of *CTNNB1* gene, and HCC clinical prognosis (Figure S10A-C). Although we did not detect the relationship between *TP53* rs28934571 and gene expression (Figure S9C), the prognosis of AA and CA genotypes of *TP53* rs28934571 was poorer than that of wild type (Figure S10D).

Random forest and decision tree analysis

We combined the above clinical, mutation and expression information to perform the random forest modeling analysis. Multiple dimension scale plot in Figure S11A suggested the effective classification of negative normal and overall HCC group. AUC value of ROC equals to 0.956, indicating high classification accuracy (Figure S11B). We also showed the feature vectors extracted from the classification model in Figure S11C-D, and obtained the largely contributed genes, such as *ECM1*, *FCN2*, *ANGPTL6*, *OIT3*, *ADAMTS13* and *LRRC14*. Next, we performed the decision tree modeling analysis, based on the above genes. We first randomly selected 260 HCC cases for modeling, and then test other 125 cases, and finally found that the predicted rate of the genes was larger than 90% (Figure S11E). Meanwhile, we compared the expression of these genes in 50 HCC patients with adjacent non-tumor and found that *ECM1*, *FCN2*, *ANGPTL6*, *OIT3* genes in overall HCC tissue showed the higher expression level than the adjacent non-tumor tissue (Figure S11F, $P < 0.0001$).

Subsequently, we performed random forest modeling with different pathological stages, which is primarily based on TNM information. To prove the validity of this classification method, we performed random forest and decision tree modeling without removing TNM information and found that T1 and T2

information can effectively distinguish stage I, II, III-IV with the AUC value of 0.994 in ROC (Figure 4A-C). Then, we removed the TNM information for random forest modeling and found that the classification effect was reduced (Figure 4D-E, AUC=0.675). Figure 4F-G shows the genes that contribute significantly to the classification model. We then performed the decision tree analysis with 210 cases for training and 116 cases for testing. The result showed a prediction accuracy of 56.0% (Figure 4H). Compared with 50 normal adjacent controls, *VIPR1*, *FAM99A*, and *GNA14* genes were down-regulated in 50 HCC tissues (Figure 4I, $P<0.0001$), while CEP55, SEMA3F, and PRR11 genes were highly expressed ($P<0.0001$). The expression of these genes is closely related to different pathological stages (Figure 4J).

Principal component analysis

Finally, we used principal component analysis (PCA) to screen the target genes associated with different pathological grades of HCC. As shown in Figure 5A, the calculated variances of the principal component (PC) 1, 2, and 3 equaled to 9.4%, 8.1%, and 6.3%, respectively. Based on the PC1/2 (Figure 5B) and PC1/2/3 (Figure 5C), we can effectively classificate the negative normal and overall HCC groups, but not the stage I, II, III-IV groups. Figure 5D shows the top 10 genes that contributing mainly to PC1 and PC2. We analyzed the expression level of these genes between HCC tissue and adjacent normal tissue, or in different pathological stages. As shown in Figure 5E-F, compared with normal tissue, the *SLC27A5*, *ALDH2*, and *DCXR* genes were down-regulated ($P<0.0001$), while *LAMTOR4* ($P=0.003$), *SNRPA* ($P<0.0001$) *SNRPD2* ($P<0.0001$) genes were highly expressed, in overall HCC tissues. In addition, the expression of the *SLC27A5*, *ADAM17*, *SNRPA*, *SNRPD2*, and *ALDH2* genes was associated with different pathologic stages (Figure 5E-F).

HLivH060PG02 HCC cohort analysis

After the above analyses of TCGA cohort, we obtained a series of HCC pathological grade-associated genes. We further assessed the survival prognosis value and of these genes through GEO database (data not shown), and analyzed the research status of genes through the on-line PubMed database retrieval. Thus, seven interesting genes, including *GAS2L3*, *CUZD1*, *SNRPA*, *SNPRD2*, *SEMA3F*, *IQCA1*, *OIT3*, were screened out. We analyzed the expression difference of these genes between the HCC tissues and corresponding adjacent normal tissues, in the Chinese HLivH060PG02 HCC cohort. Unfortunately, due to the lower amplification efficiency of the *IQCA1* and *OIT3*, we finally analyzed the remaining five genes, namely *GAS2L3*, *CUZD1*, *SNRPA*, *SNPRD2*, and *SEMA3F*. Figure 6A illustrates the correlation between *GAS2L3*, *SNRPA*, *SNRPD2* and the prognosis of HCC, as example. As shown in Figure 6B, compared with adjacent normal tissues, we observed a highly expressed level of *GAS2L3* ($P=0.036$), *SNRPA* ($P<0.001$), and *SNRPD2* ($P=0.002$) genes in HCC tissues. Moreover, these three genes in pathologic stage III showed a higher expression trend then that in stage III, but statistical significance was only detected for the *GAS2L3* gene ($P=0.013$). Considering the small sample size, we do not rule out the correlation between *SNRPA*, *SNRPD2* gene*s* and pathological stage of HCC.

Discussion

Considering the complexity of etiology and pathogenesis of liver cancer, it is essential to continuously screen and analyze the target genes closely related to the pathogenesis of liver cancer. Based on the expression, mutation and clinical data of liver cancer cases in the TCGA cohort, we aim at identifying the potential liver cancer-associated targeting genes. It should be noted that the TCGA-LIHC project includes not only hepatocellular carcinomas (HCC) cases but also a small amount of fibrolamellar carcinomas and hepatobiliary mixed carcinomas cases. Taking account of the differences of distinct liver cancer types and the factor of sample size, we only selected the cases of HCC, the most common primary liver malignancy. The current reports regarding HCC-related target genes from the different clinical pathological stages (stage I, II, III, IV)or histological grades of (G1, G2, G3 and G4)HCC cases in TCGA cohort are very limited, even though several publications from other aspects or with different analysis strategies were retrieved [8–10]. We did not observe a correlation between histological grades of HCC and clinical outcomes through the survival curve analyses. As expected, different clinical pathological stages of HCC are closely related to clinical prognosis. Therefore, we attempted to screen the potential target genes associated with normal, stage I, II, III-IV pathological classification of HCC.

The alteration of the target gene expression level leads to the abnormality of the corresponding protein function, which is the critical mechanism underling the hepatocarcinogenesis [11–13]. Based on the sample size, we combined the data of stage III and IV, and focused on the differentially expressed genes associated with normal, stage I, stage II, and stage III-IV classifications. We tried toutilize the "EdgeR" package for the statistically significant differential genes in three comparisons (Tumor vs. Normal, stage II vs. stage III, stage III + IV vs. stage II), and further screen the intersection gene. Through this strategy, we did not identify the target genes with a decreasing trend in the groups of normal, stage I, stage II and stage III + IV), but several genes with increasing trend (e.g.*DUCX2*, *IQCA1*, *PCSK1*, etc.), which showed the low expression frequency and mainly gathered in stage III-IV. This analysis strategy is not effective. Subsequently, we used the Principal Component Analysis (PCA) approach [14, 15]to decrease the dimensionality of the datasets for the group of normal, stage I, stage II, and stage III-IV, and to screen for the genes that contributed largely to the main component. It was found that the normal and tumor group can be better distinguished by the principal components of 1, 2, and 3, but not the groups of stage I, stage II, and stage III-IV, may due to the low sample size and the complexity of different pathological staging mechanisms of HCC. Despite this, we also obtained the top 10 genes that contributed mostly to the principal component 1 and 2. Of them, the expression levels of the *SLC27A5*, *SNRPA*, *SNRPD2*, and *ALDH2* genes were significantly associated with different pathological stages of HCC. Up to now, only one study reported that DNA hypermethylation could reduce the expression of SLC27A5 in HCC, which contributing to HCC progression through NRF2/TXNRD1 pathway [16]. Two studies based on the mouse model indicated the potential role of *ALDH2* expression in the hepatocellular carcinogenesis [17, 18]. Our Chinese HLivH060PG02 HCC cohort analysis first provided the potential role of *GAS2L3* and two U1 snRNP component genes (*SNRPA* and *SNRPD2*)[19] in the HCC carcinogenesis, and there are still no relevant systematic reports. Therefore, it is meaningful to further explore the molecular mechanism of these genes in the progression and prognosis of HCC.

Genetic copy number variation (CNV) is caused by the genome rearrangement-induced the copy number amplification or deletion of a large genome fragment (> 1 kb) [20–22]. CNV-induced the change of gene expression level is an essential mechanism of tumorigenesis [23–25]. Even though one relevant study of liver cancer was reported [26],we utilized the distinct analysis strategy to screen out the interesting genes with CNV that are related to gene expression and clinical pathological stages of HCC in TCGA cohort. Our results identified a group of cell cycle or cell division-associated genes, such as *GADD45G* and *CCNE2*, exhibited the CNV-driven gene expression. In addition, we utilized MCODE modular analysis to screen out some key genes, such as *TTK* and *CDC20*, from the perspective of protein binding, which are also related to cell cycle and division behavior. It is worth noting that the expression levels of some down-regulated genes in HCC (e.g., *GADD45G*, *FPR2*, *PPBP*, etc.) are closely linked to the CNV in a dose dependent manner. Some other key inhibition mechanisms of gene expression, such as hypermethylation modification, may exist for these genes, apart from CNV.

Genetic mutations are considered as the key mechanisms of tumorigenesis[27, 28], and single nucleotide polymorphisms are closely related to the susceptibility of tumors in the population [29]. We performed a series of gene mutation analysis as well. We found that the major mutation types of these genes are non-synonymous mutations, and the gene mutation frequency is not associated with the clinical pathology of HCC. Even though there is a correlation between the overall variation and expression of *CTNNB1* and *TP53* genes in HCC and different pathological stages, no positive results were obtained for the mutations of specific site in the relatively low amount of cases. In addition, there are more than 10 SNPs for the*CTNNB1* gene in HCC, but with low frequency. We also did not observe the correlation between these SNPs and CTNNB1 high expression or clinical prognosis of HCC. More HCC cases may be required for the confirmation of this point.

We integrated the expression data of all target genes obtained from the above methods, somaticmutation data, and some clinical biochemical indicators, and then utilized the random forest, a robust classification and regression approach [30],for the classification analysis of normal, tumor and stage I, stage II, stage III-IV groups. Although the classification effect of stage I, II, III-IV is worse than the normal/tumor, we identified some key contributing genes (e.g., *VIPR1*, *FAM99A*, *GNA14CEP55*, *SEMA3F*, *PRR11*, etc.), which are closely related to different clinical stages. Interestingly, the interest genes screened by the random forest approach (e.g., *ECM1*, *FCN2*, *ANGPTL6*, *OIT3*, *ADAMTS13*, etc.) are mainly down-regulated in HCC, compared with the normal group.

Conclusion

In summary, based on the data of HCC cases in TCGA cohorts, we first conducted the statistical analysis, random forest, decision tree and principal component analysis to identify the differential gene expression, CNV, SNV and SNP profiles, which are associated with the different pathologic stage I, II, and III-IV. More molecular biology experiment so rclinical sample tests are required to further investigate whether the identified genes serve as the prognostic biomarker or therapeutic targets of HCC.

Materials And Methods

HCC pathological stage-related gene expression analysis

We utilized the “TCGAbiolinks” R package to download the liver cancer-associated mRNA, lncRNA expression data with the workflow type of “HTSeq-Counts” and clinical data from the TCGA-LIHC project within the TCGA database (http://tcga-data.nci.nih.gov/tcga/). We extracted the clinical information, including gender, age, race, ethnicity, height, weight, clinical pathologic stage, pathologic T/N/M stage, neoplasm histologic grade, survival status, follow-up time and various clinical, biochemical indicators. There groups of clinical pathologic stages, namely stage I, II and III-IV, were analyzed. Combined with other clinical indicators, we performed the Kruskal-Wallis test or chi-square test, using GraphPad Prism 5.1 software. We also conducted the log-rank test and Kaplan-Meier (KM) survival curve analysis using SPSS 20.0 statistical analysis software.

Based on the R language software (https://www.r-project.org/), we combined the expression data and clinical data, and removed the non-HCC case data. An “EdgeR” package was then utilized for TMM data standardization and differential gene screening. The gene expression level was processed into logarithm base 2 (log 2). We obtained the Volcano maps through a “ggplot” package. We also utilized the online venn tool (http://bioinformatics.psb. ugent.be/ webtools/Venn/) for intersection analysis to obtain the intersection genes of different groups, and then used the Morpheus online software [https://software.broadinstitute.org/ Morpheus/] to draw a heat map of cluster analysis. ID conversion is implemented based on the gene ID conversion tool of DAVID (https://david.ncifcrf.gov/conversion.jsp). Protein-protein interaction network analysis of intersection genes was performed based on the STRING online analysis tool (https://string-db.org/). The survival curves and the expression status of specific genes in the groups of total HCC, control, stage I, stage II and stage III-IV were analyzed through the GEPIA, a web server for cancer and normal gene expression profiling and interactive analyses.

Copy number variation analysis

We first downloaded the copy number variation (CNV) data of the TCGA-LIHC project from the TCGA database page with the type of masked copy number segment, and then used the Perl script to add the according gene annotation based on the CNV chromosome location information. Segment_mean value between -0.2 and +0.2 will be considered as no variation and will be marked as “0”. CNV contains the double deletion (dd, “-2”), single deletion (sd, “-1”), single gain (sg, “+1”), and amplication (A,”+2 or +>2”)] of gene copy number. The chi-square test and the Bonferroni-adjusted *P* value correction method were utilized to obtain the CNV differential targeting genes between HCC and the normal control group, and then the "RCircos" package was used to obtain the Circos 2D track plot.

Also, we combined gene expression data with CNV differential targeting gene data, and performed Kolmogorov-Smirnov test for correlation analysis to identify the expression-correlated targeting genes with CNV. Then, the enrichGO () function was used for the Gene Ontology (GO) analysis, while

enrichKEGG () function was for the Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway enrichment analysis. We used the “STRINGdb” package to build a protein interaction network and visualized the results using Cytoscape software. Based on the "Molecular Complex Detection" (MCODE) modular analysis, we screened key hub genes in the PPI network using default settings.

Random forest and decision tree analysis

Upon the combination of the above clinical, mutation, and expression information, we used the “random Forest” package to perform the random forest modeling analysis. The specific gene profiles of normal, overall HCC and HCC with different clinical pathological grades were effectively classified by the principles of mean decrease accuracy and mean decrease Gini, and visualized by a “ggpubr” package. MDSplot () function was used to obtain the Multi Dimension Scale. Using the “pROC” package, the receiver operating characteristic (ROC) curve is plotted, and the area under the ROC curve (AUC) value is calculated. Based on the results of random forest analysis, the Decision tree modeling analysis is performed, using “rpart” and “rpart.plot” packages.

Genetic mutation analysis

We directly downloaded the simple nucleotide variation (SNV) data of the TCGA-LIHC project with the type of masked somatic mutation, and extracted the mutation data using the Perl script. According to the mutation rate, the top 15 genes were selected, and the “GenvisR” package was used to draw a waterfall map with clinical grading information. We also extracted the single nucleotide polymorphism (SNP) data, and performed the wilcox test and boxplot () function for the correlation analysis of gene mutation and expression in overall HCC and different pathological stages. The “survminer” package was further used to correlate the specific gene mutations in overall HCC and different pathological stages with the clinical prognosis, and drew the corresponding survival curves.

Principal component analysis

Based on the gene expression matrix, we used the prcomp () function for the principal component analysis (PCA) to screen the genes associated with different pathological stages of HCC. The "factoextra" and "ggplot2" packages were utilized to obtain the principal component (PC) gravity and gene contribution maps. A three-dimensional map (PC1, PC2, and PC3) is drawn using a “scatterplot3d” package; while a two-dimensional map (PC1, PC2) is obtained by a “ggord” package. In addition, for specific genes selected by a decision tree, random forest, and principal component analysis, we utilized the R language to obtain the expression data of overall HCC tissue and adjacent normal tissue, and perform the t test using GraphPad Prism 5.1 software. We also obtained the expression data of stage I, II, III and IV through the GEPIA.

HLivH060PG02 HCC cohort analysis

Based on the enrolled HCC patients in the Chinese population, namely HLivH060PG02 cohort (Shanghai Outdo Biotech Co., Ltd, Shanghai, China), we analyzed the expression levels of five targeting genes (*GAS2L3*, *CUZD1*, *SNRPA*, *SNPRD2*, *SEMA3F*), and their correlation with pathologic stages. The clinical characteristics of HCC cases were shown in Table S1, and the use of human biological materials (Number: YB M-05-02) was approved by the Use Ethics Committee of Shanghai Outdo Biotech Company. RNA samples were extracted from a total of 30 HCC tissues and 30 corresponding adjacent normal tissues, respectively. Bases on the synthesized cDNA, a quantitative real-time PCR (qPCR) assay was performed with a TB Green™ Premix Ex Taq™ II (Takara, RR820A), using an ABI 7500 Real-Time PCR System (Thermo Fisher Scientific).

The primer sequence information is listed: *GAS2L3*: 5'-CTGAGGACCCTCCTTGTAGTTG-3' (Forward), 5'-CCTTGAAGAGTATCCCAGCCTC-3' (Reverse); *CUZD1*: 5'-CCAGCCTTTCAACAGTGTGC-3' (Forward), 5'-GCCACGAGGTAGCATTTCCT-3' (Reverse); *SNRPA*: 5'-ACCCGCCCTAACCACACTAT-3' (Forward), 5'-GGAGAAGATGGCGTACAGGG-3' (Reverse); *SNPRD2*: 5'-CAAGTGCTCATCAACTGCCGCA-3' (Forward), 5'-GCGGTCTTTGTTGACTGGCTTG-3' (Reverse); *SEMA3F*: 5'-CAAGGATGTCAACGGCGAGT-3' (Forward), 5'-TGAGTCTGGGTCCATGGTGT-3' (Reverse); *beta-actin*: 5'-GAAGAGCTACGAGCTGCCTGA-3' (Forward), 5'-CAGACAGCACTGTGTTGGCG-3' (Reverse). Student's t-test was performed by GraphPad Prism 7.0.4 (San Diego, California USA). Differences with $P < 0.05$ were considered significant.

Declarations

Ethics approval and consent to participate

The studies involving human participants were reviewed and approved by the clinical research ethics Committee of Shanghai Outdo Biotech Company.

Consent for publication

Not applicable.

Availability of data and materials

The datasets generated and/or analysed during the current study are available in the The Cancer Genome Atlas (TCGA) database.

Competing interests

The authors declare that they have no competing interests.

Funding

This work was supported by grants from the Innovation Team Development Plan of the Ministry of Education (IRT13085 to JY), National Nature Science Foundation of China (31670759 to JY, 31571380 to XG, 81572882 to ZY, 31701182 to CS), Excellent Talent Project of Tianjin Medical University (to JY), Tianjin Enterprise Science and Technology Commissioner Project (18JCTPJC59400 to XG).

Author Contributions

Conceptualization, XJG and JY; Methodology, XJG, CYZ and SYW; Software, XJG and XYC; Validation, YYR and CS; Formal Analysis, XJG and CYZ; Data Curation, XJG, NZ and ZY; Writing-Original Draft Preparation, XJG and CYZ; Writing-Review & Editing, XJG and JY; Supervision, JY.

Acknowledgements

Not applicable.

References

1. Gibbs P, Tie J, Bester L: **Radioembolization for hepatocellular carcinoma: current role and future directions - the medical oncologist's perspective**. *Hepat Oncol* 2015, **2**(2):117-132.
2. Hartke J, Johnson M, Ghabril M: **The diagnosis and treatment of hepatocellular carcinoma**. *Semin Diagn Pathol* 2017, **34**(2):153-159.
3. El Jabbour T, Lagana SM, Lee H: **Update on hepatocellular carcinoma: Pathologists' review**. *World J Gastroenterol* 2019, **25**(14):1653-1665.
4. Raza A, Sood GK: **Hepatocellular carcinoma review: current treatment, and evidence-based medicine**. *World J Gastroenterol* 2014, **20**(15):4115-4127.
5. Gomaa AI, Waked I: **Recent advances in multidisciplinary management of hepatocellular carcinoma**. *World J Hepatol* 2015, **7**(4):673-687.
6. Wallace MC, Preen D, Jeffrey GP, Adams LA: **The evolving epidemiology of hepatocellular carcinoma: a global perspective**. *Expert Rev Gastroenterol Hepatol* 2015, **9**(6):765-779.
7. Wang Z, Jensen MA, Zenklusen JC: **A Practical Guide to The Cancer Genome Atlas (TCGA)**. *Methods Mol Biol* 2016, **1418**:111-141.
8. Xu B, Lv W, Li X, Zhang L, Lin J: **Prognostic genes of hepatocellular carcinoma based on gene coexpression network analysis**. *J Cell Biochem* 2019.
9. Wu P, Xiao Y, Guo T, Wang Y, Liao S, Chen L, Liu Z: **Identifying miRNA-mRNA Pairs and Novel miRNAs from Hepatocelluar Carcinoma miRNomes and TCGA Database**. *J Cancer* 2019, **10**(11):2552-2559.

10. Agarwal R, Narayan J, Bhattacharyya A, Saraswat M, Tomar AK: **Gene expression profiling, pathway analysis and subtype classification reveal molecular heterogeneity in hepatocellular carcinoma and suggest subtype specific therapeutic targets**. *Cancer Genet* 2017, **216-217**:37-51.
11. Cai J, Li B, Zhu Y, Fang X, Zhu M, Wang M, Liu S, Jiang X, Zheng J, Zhang X *et al*: **Prognostic Biomarker Identification Through Integrating the Gene Signatures of Hepatocellular Carcinoma Properties**. *EBioMedicine* 2017, **19**:18-30.
12. Shangguan H, Tan SY, Zhang JR: **Bioinformatics analysis of gene expression profiles in hepatocellular carcinoma**. *Eur Rev Med Pharmacol Sci* 2015, **19**(11):2054-2061.
13. Pinato DJ, Pirisi M, Maslen L, Sharma R: **Tissue biomarkers of prognostic significance in hepatocellular carcinoma**. *Adv Anat Pathol* 2014, **21**(4):270-284.
14. Marziali G, Buccarelli M, Giuliani A, Ilari R, Grande S, Palma A, D'Alessandris QG, Martini M, Biffoni M, Pallini R *et al*: **A three-microRNA signature identifies two subtypes of glioblastoma patients with different clinical outcomes**. *Mol Oncol* 2017, **11**(9):1115-1129.
15. Jolliffe IT, Cadima J: **Principal component analysis: a review and recent developments**. *Philos Trans A Math Phys Eng Sci* 2016, **374**(2065):20150202.
16. Gao Q, Zhang G, Zheng Y, Yang Y, Chen C, Xia J, Liang L, Lei C, Hu Y, Cai X *et al*: **SLC27A5 deficiency activates NRF2/TXNRD1 pathway by increased lipid peroxidation in HCC**. *Cell Death Differ* 2019.
17. Hou G, Chen L, Liu G, Li L, Yang Y, Yan HX, Zhang HL, Tang J, Yang YC, Lin X *et al*: **Aldehyde dehydrogenase-2 (ALDH2) opposes hepatocellular carcinoma progression by regulating AMP-activated protein kinase signaling in mice**. *Hepatology* 2017, **65**(5):1628-1644.
18. Seo W, Gao Y, He Y, Sun J, Xu H, Feng D, Hee Park S, Cho YE, Guillot A, Ren T *et al*: **ALDH2 deficiency promotes alcohol-associated liver cancer by activating oncogenic pathways via oxidized DNA enriched extracellular vesicles**. *J Hepatol* 2019.
19. Azam S, Hou S, Zhu B, Wang W, Hao T, Bu X, Khan M, Lei H: **Nuclear retention element recruits U1 snRNP components to restrain spliced lncRNAs in the nucleus**. *RNA Biol* 2019, **16**(8):1001-1009.
20. Zarrei M, MacDonald JR, Merico D, Scherer SW: **A copy number variation map of the human genome**. *Nat Rev Genet* 2015, **16**(3):172-183.
21. Nowakowska B: **Clinical interpretation of copy number variants in the human genome**. *J Appl Genet* 2017, **58**(4):449-457.
22. Liang L, Fang JY, Xu J: **Gastric cancer and gene copy number variation: emerging cancer drivers for targeted therapy**. *Oncogene* 2016, **35**(12):1475-1482.
23. Takai A, Dang HT, Wang XW: **Identification of drivers from cancer genome diversity in hepatocellular carcinoma**. *Int J Mol Sci* 2014, **15**(6):11142-11160.
24. Gu DL, Chen YH, Shih JH, Lin CH, Jou YS, Chen CF: **Target genes discovery through copy number alteration analysis in human hepatocellular carcinoma**. *World J Gastroenterol* 2013, **19**(47):8873-8879.

25. Chen Y, Chen C: **DNA copy number variation and loss of heterozygosity in relation to recurrence of and survival from head and neck squamous cell carcinoma: a review**. *Head Neck* 2008, **30**(10):1361-1383.

26. Lu X, Ye K, Zou K, Chen J: **Identification of copy number variation-driven genes for liver cancer via bioinformatics analysis**. *Oncol Rep* 2014, **32**(5):1845-1852.

27. Zucman-Rossi J, Villanueva A, Nault JC, Llovet JM: **Genetic Landscape and Biomarkers of Hepatocellular Carcinoma**. *Gastroenterology* 2015, **149**(5):1226-1239.e1224.

28. Schulze K, Nault JC, Villanueva A: **Genetic profiling of hepatocellular carcinoma using next-generation sequencing**. *J Hepatol* 2016, **65**(5):1031-1042.

29. Koberle B, Koch B, Fischer BM, Hartwig A: **Single nucleotide polymorphisms in DNA repair genes and putative cancer risk**. *Arch Toxicol* 2016, **90**(10):2369-2388.

30. Liu Y, Zhao H: **Variable importance-weighted Random Forests**. *Quant Biol* 2017, **5**(4):338-351.

Figures

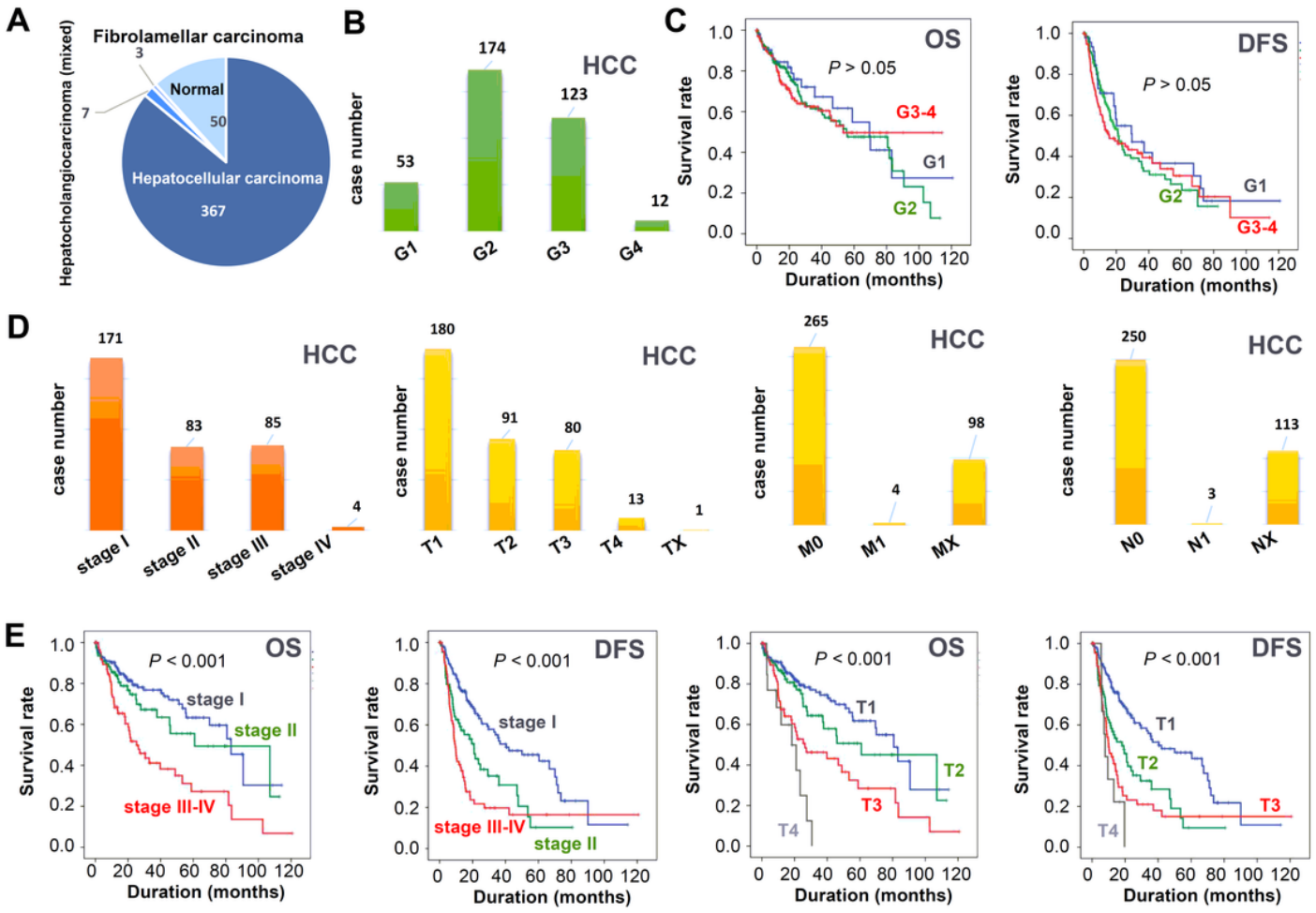

Figure 1

Figure 1

Survival curve analysis for the different pathologic stages of HCC patients in the TCGA database. (A) The included liver cancer cases and adjacent normal controls in TCGA cohorts. (B) The neoplasm histologic grades (G1, G2, G3 and G4) of HCC cases. (C) The log-rank test and KM survival curve analysis according to the histological grades of HCC were performed. (D) The clinical pathologic stages (stage I, stage II and stage III-IV) and T/N/M stage of HCC cases. (E) The survival curve analysis according to the stage I, II, III-IV and T1-T4 were performed.

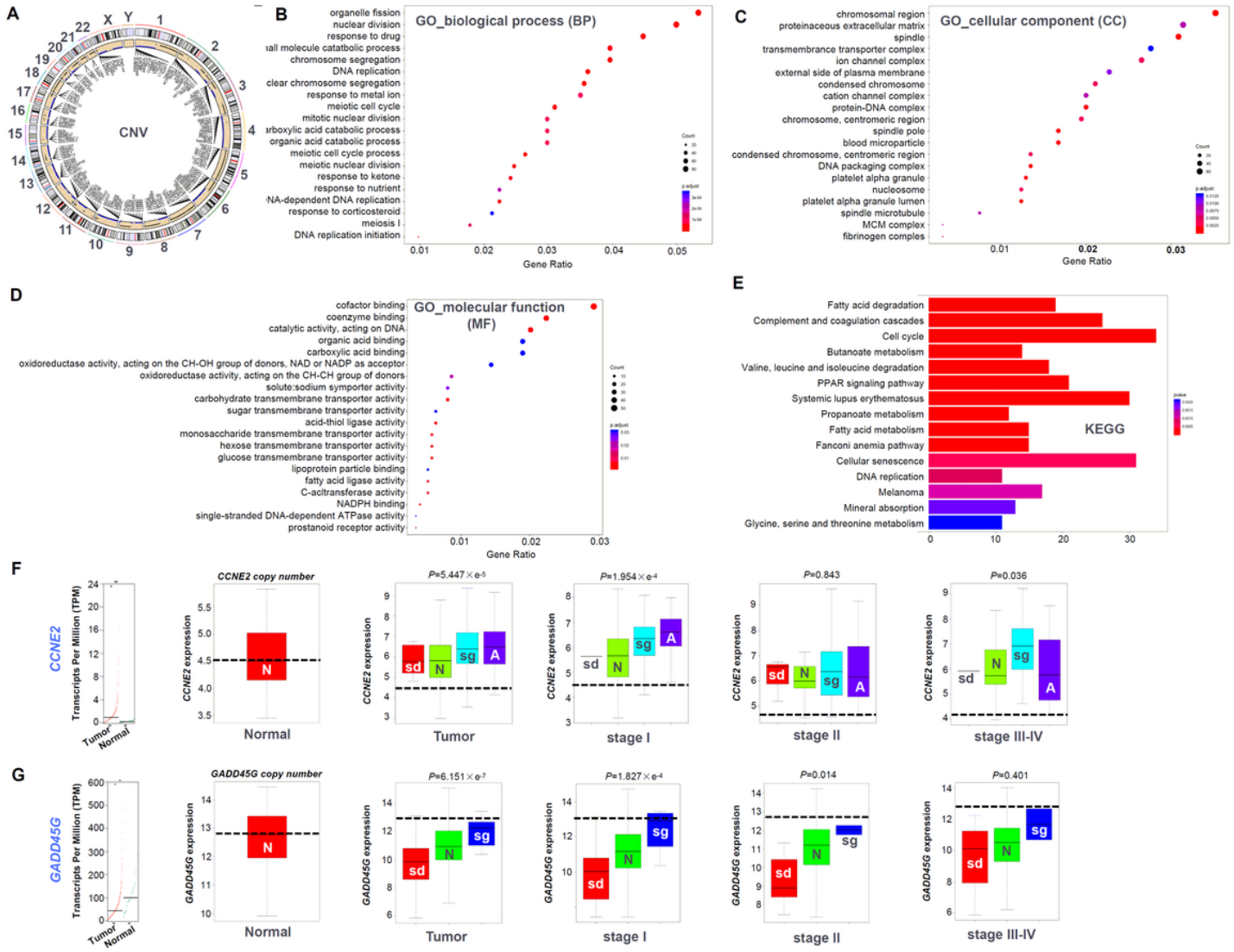

Figure 2

Figure 2

Genetic copy number variant analysis for the different pathologic stages of HCC patients. (A) Circos 2D track plot of CNV profile was showed. (B-D) GO and KEGG analysis data of the genes with CNV, which was correlated with the gene expression. (E-F) We analyzed the expression levels of CCNE2, GADD45G

genes in normal and tumor by GEPIA, and the correlation between gene expression and CNV in normal and different pathological grades of HCC.

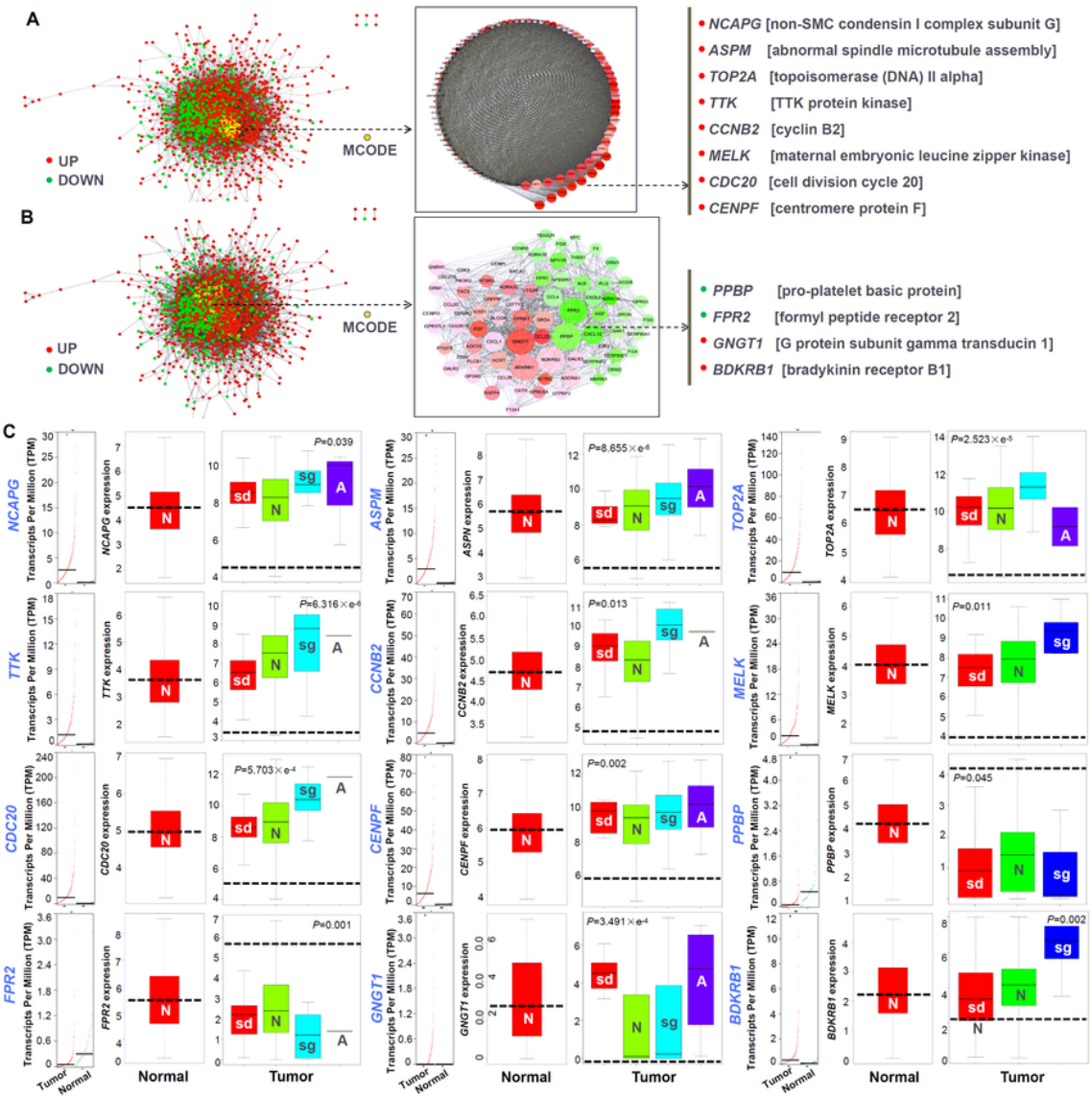

Figure 3

Figure 3

The protein-protein interaction network. (A-B) “STRINGdb” package and Cytoscape software, and "Molecular Complex Detection" (MCODE) were utilized for the construction of protein-protein interaction (PPI) network and the identification of hub genes. (C) The expression levels in normal and tumor, and the

correlation between gene expression and CNV of some hub genes in normal and different pathological grades of HCC were analyzed.

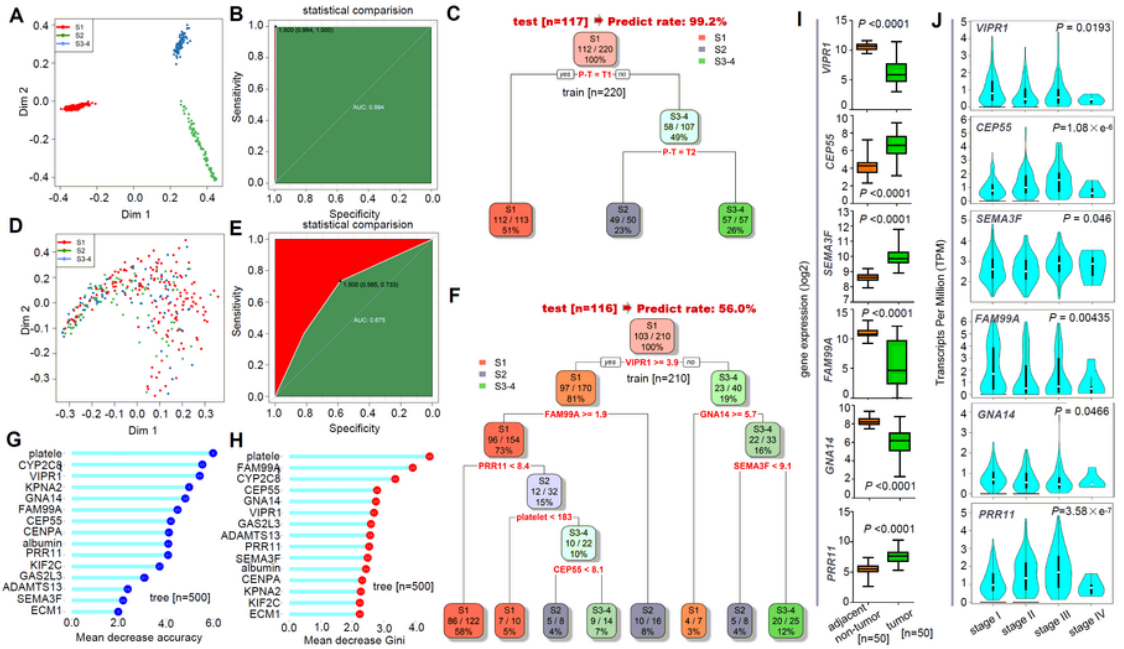

Figure 4

Figure 4

The decision tree and random forest analysis for the different pathologic stages of HCC patients in the TCGA cohort. (A) We combined the clinical, mutation and expression information to perform the random forest modeling analysis. Multiple dimension scale plot was provided. (B) ROC curve is plotted, and the AUC value is calculated. (C) Decision tree modeling analysis is performed. (D-F) We removed the TNM

information to complete the random forest modeling, again. (G-H) Based on the principles of mean decrease accuracy and mean decrease Gini, we identified the largely contributed genes. (I-J) We compared the expression of these genes in 50 HCC tissues with adjacent non-tumor tissues and obtained the expression data of stage I, II, III, and IV through the GEPIA.

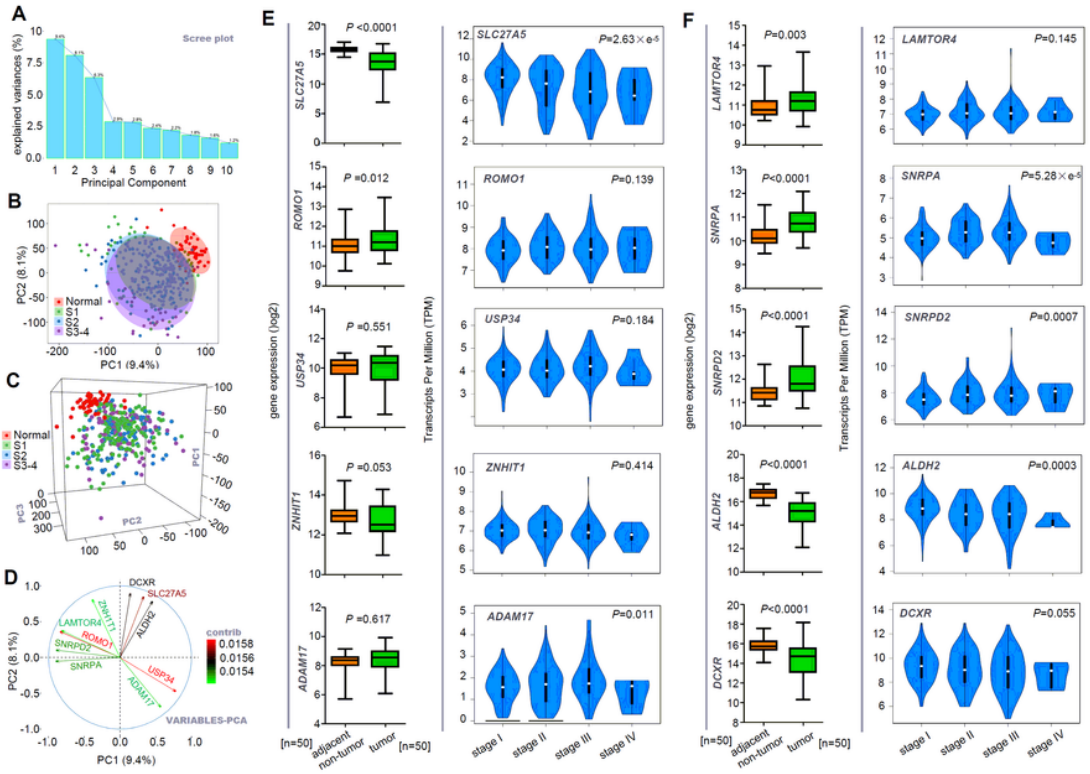

Figure 5

Figure 5

The PCA analysis for the different pathologic stages of TCGA HCC cases. We performed the principal component analysis (PCA) to screen the genes associated with different pathological stages of HCC. The

principal component (PC) gravity (A), a two-dimensional map (PC1/PC2) (B), a three-dimensional map (PC1/PC2/PC3) (C) and gene contribution maps (D) were provided. (E-F) We analyzed the expression level of overall HCC tissue, adjacent normal tissue, and stage I, II, III, and IV group.

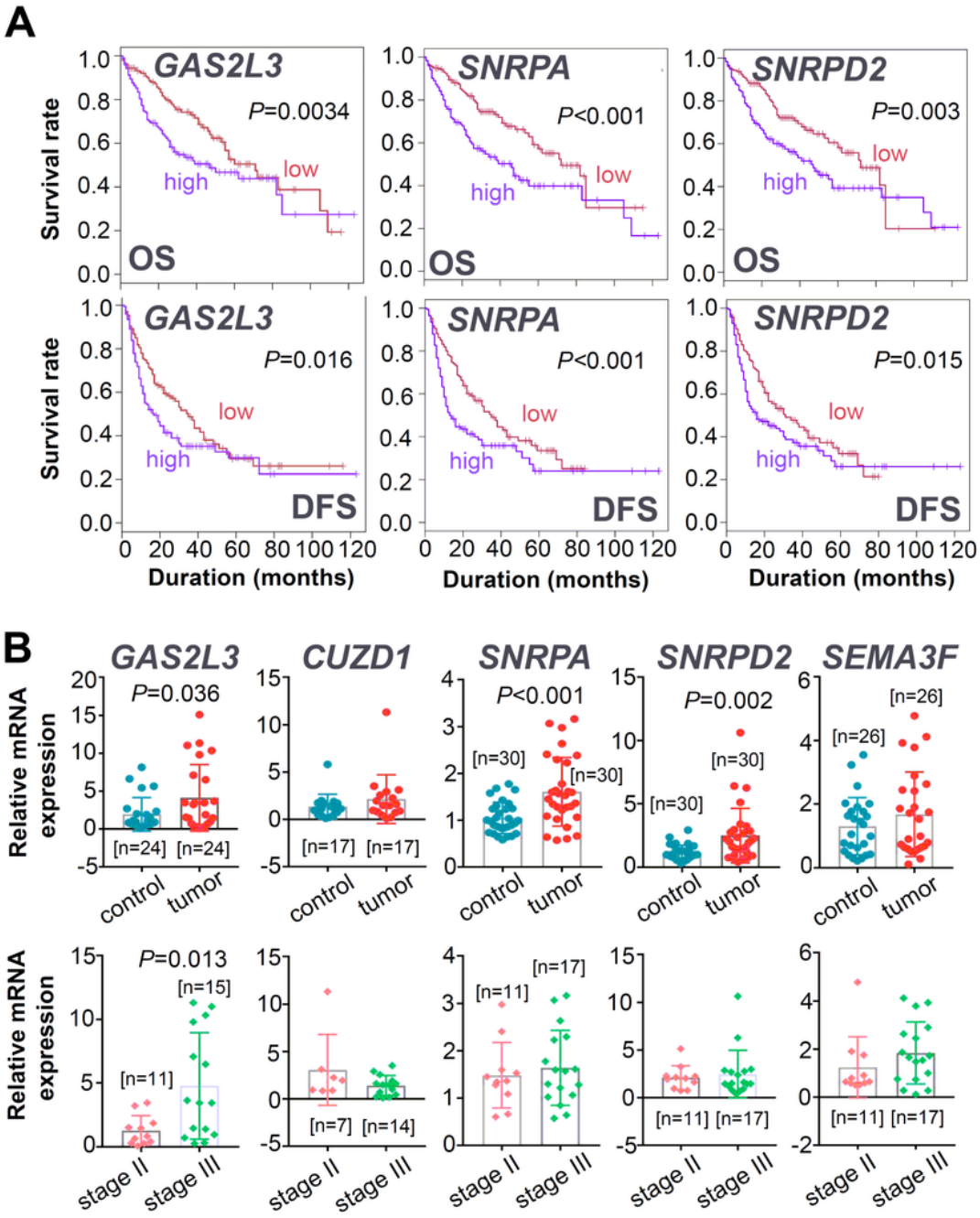

Figure 6

Figure 6

The expression level of five targeting genes in the HCC cases of Chinese population. (A), For the GAS2L3, SNRPA, and SNPRD2, we performed the Kaplan-Meier estimates of the overall survival (OS) or disease-

free survival (DFS) through GEPIA, based on the data in TCGA cohorts; (B) We further performed a qPCR assay to detect the expression levels of five targeting genes (GAS2L3, CUZD1, SNRPA, SNPRD2, SEMA3F) in the HLivH060PG02 HCC cohort of Chinese population, and their correlation with pathologic stages of HCC. Student's t-test was performed and significant differences were indicated.

Supplementary Files

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

- SUPPLEMENTARYMATERIAL.docx