Original Article
Identification and validation of a ten-gene set variation score as a diagnostic and prognostic stratification tool in hepatocellular carcinoma

Chanhua Zou1*, Chunling Yuan2*, Jiazhou Ye3, Ziyu Liu2, Xing Gao2, Xuemin Piao2, Rongyun Mai2, Yan Lin2, Donghua Zou4, Zhaoshan Fang4, Rong Liang2
Departments of 1General Medicine, 2Medical Oncology, 3Hepatobiliary Surgery, Guangxi Medical University Cancer Hospital, Nanning 530021, Guangxi, People’s Republic of China; 4The Fifth Affiliated Hospital of Guangxi Medical University, Nanning 530022, Guangxi, People’s Republic of China. *Equal contributors.

Abstract: We aimed to identify a hepatocellular carcinoma (HCC)-specific gene set during progression. Using the HCC data set from The Cancer Genome Atlas, we found that 10 genes were gradually upregulated with the progression of HCC and associated with survival, classified as HCC-unfavorable genes; 29 genes were gradually downregulated and associated with survival, classified as HCC-favorable genes. Gene set variation analysis (GSVA) was used to score individual samples against the two gene sets. Receiver operating characteristic (ROC) curve analysis showed that both the HCC-unfavorable GSVA score and HCC-favorable GSVA score were reliable biomarkers for diagnosing HCC. Moreover, tROC curve analysis and univariate/multivariate Cox proportional hazards analyses indicated that the HCC-unfavorable GSVA score was an independent prognostic biomarker. The results were validated in an external independent data set. Our results support a ten-gene set variation score as a diagnostic and predictive strategy tool in HCC.

Keywords: Hepatocellular carcinoma, prognostic stratification system, HCC, TCGA

Introduction
According to global cancer statistics from 2018, hepatocellular carcinoma (HCC) has become the sixth most common cancer and the fourth leading cause of cancer death in the world [1]. The main causes of HCC include chronic hepatitis B virus (HBV), hepatitis C virus (HCV) infection, aflatoxin contaminated food, heavy drinking, obesity, smoking, and type 2 diabetes [2, 3]. Approximately 80-90% of HCC patients have potential cirrhosis [4]. Although there are many treatments, such as hepatectomy, liver transplantation, radiofrequency ablation, embolization therapy, and molecule-targeted chemotherapy, therapeutic efficacy in advanced HCC is still limited [5]. Therefore, it is essential to explore molecular mechanisms in HCC and robust diagnostic and prognostic markers.

With the development of high-throughput sequencing technology, an increasing number of molecular diagnostic markers have been identified for HCC. Most studies on the prognosis of HCC focused on a single or several molecules [6-9], whereas less attention has been paid to characteristic gene sets related to HCC progression. To date, there is no widely accepted molecular prognostic biomarker for HCC.

In this study, we identified two HCC progression characteristic gene sets: an HCC-unfavorable gene set and an HCC-favorable gene set. Gene set variation analysis (GSVA) was used to score individual samples against the two gene sets. Both the HCC-unfavorable GSVA score and HCC-favorable GSVA score may serve as biomarkers for HCC prognosis.

Materials and methods
Materials acquiring
In The Cancer Genome Atlas (TCGA, https://www.cancer.gov/) [10], there are 171 HCC sam-
A ten-gene set variation score as a prognostic tool in hepatocellular carcinoma

5684
Am J Transl Res 2020;12(9):5683-5695

plies at stage I, 86 at stage II, 83 at III, and 5 at stage IV as well as 42 healthy liver tissue samples. In addition, GSE54236 based on the GPL6480 platform was downloaded from Gene Expression Omnibus (GEO, https://www.ncbi.nlm.nih.gov/geo/) [11], including 81 HCC samples and 80 healthy liver tissue samples, and was used to verify the prognostic value of the identified gene sets. The "normalize Between Arrays" function in the limma package [12] was applied to normalize gene expression profiles. If a gene responded to multiple probes, the average value of these probes was considered to be the expression value of the corresponding gene. The workflow of the present study is shown in Figure 1.

**Figure 1.** Flow chart of this study.

Differentially expressed gene (DEG) analysis

The RNA sequencing expression profile (displayed as read counts) of HCC in TCGA was downloaded. The voom function [13] in the limma package was employed to normalize the RNA sequencing data; the limma package [12, 14] was also used to identify DEGs in the 4 stages of HCC and healthy liver tissue samples. A DEG was considered at \(P < 0.01\) after FDR correction and \(|\log \text{FC}| > 1.5\) as the threshold. During HCC progression, if a DEG was gradually upregulated (\(\log \text{FC}_{\text{stage I vs control}} < \log \text{FC}_{\text{stage II vs control}} < \log \text{FC}_{\text{stage III vs control}} < \log \text{FC}_{\text{stage IV vs control}}\)) or gradually downregulated (\(\log \text{FC}_{\text{stage I vs control}} > \log \text{FC}_{\text{stage II vs control}} > \log \text{FC}_{\text{stage III vs control}} > \log \text{FC}_{\text{stage IV vs control}}\)), then it was considered to be a gene characteristic of HCC progression.

**Functional enrichment analysis**

Gene Ontology (GO) and Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway enrichment analyses of the gradually upregulated and downregulated genes were performed using the clusterProfiler package [15] in R. \(P < 0.05\) was considered significant.

**Survival analysis and HCC-unfavorable/favorable gene sets**

The median expression value for each gradually upregulated and gradually downregulated gene was used as the cutoff to dichotomize patients into high-expression and low-expression groups. We applied Kaplan-Meier survival analysis with the log-rank method to evaluate the association of a gene with prognosis. Survival analysis was performed using the survival package (https://CRAN.R-project.org/package=survival) in R. \(P < 0.01\) was considered to be significant. A gene that was gradually upregulated in HCC progression and associated with a poor prognosis was defined as an HCC-unfavorable gene; a gene that was gradually downregulated in HCC progression and associated with a good prognosis was defined as an HCC-favorable gene. Two HCC progression characteristic gene sets were established: an HCC-unfavorable gene set and an HCC-favorable gene set.
Calculation of HCC-unfavorable/favorable GSVA scores

The GSVA package implements a nonparametric unsupervised method, called Gene Set Variation Analysis (GSVA), for assessing gene set enrichment (GSE) in gene expression microarray data or RNA-seq data. The GSVA package [16] in R was used to calculate HCC-unfavorable GSVA and HCC-favorable GSVA scores for an individual sample.

Receiver operating characteristic (ROC) curve analysis, univariate/multivariate Cox proportional regression analysis and time-dependent ROC (tROC) curve analysis

The pROC package [17] was utilized to conduct ROC curve analysis to evaluate the ability of gene sets to diagnose HCC. Univariate/multivariate Cox proportional hazards analyses were used to compare the relative prognostic value of the two GSVA score systems with that of routine clinicopathological features. P < 0.05 was considered significant. tROC curve analysis was applied to evaluate the prognostic value regarding the 2-year survival rate of the independent prognostic factors.

Validation of the GSVA score system in an independent data set

The HCC-unfavorable/favorable GSVA score was calculated using the HCC cohort from TCGA, and ROC curve, tROC curve and survival analyses were performed in GSE54236.

Validation of aberrant expression of HCC-unfavorable genes at the protein level

The Human Protein Atlas (https://v15.protein-atlas.org/) [18] provides information on the tissue and cell distribution of all 24,000 human proteins. We scanned the Human Protein Atlas web tool to validate differential expression of HCC-unfavorable genes at the protein level.

Results

Various genes differentially expressed with HCC progression

PCA analysis of TCGA data showed that the expression patterns of global genes (Figure 2A) could not distinguish HCC from controls. Compared to control samples, there were 2114 DEGs in stage I HCCs (Figure 2B), 2714 DEGs in stage II HCCs (Figure 2C), 2871 DEGs in stage III HCCs (Figure 2D) and 3718 DEGs in stage IV HCCs (Figure 2E). A total of 1273 common DEGs were identified in stage I-IV HCCs (Figure 2F). Among them, 82 DEGs were gradually upregulated and 176 DEGs gradually downregulated with HCC progression. PCA analysis showed that the expression patterns of these genes could distinguish HCC from controls (Figure 2G).

Gradually upregulated/downregulated genes involved in multiple HCC-related pathways

Functional enrichment analysis was used to explore the biological functions and related pathways of gradually the upregulated and downregulated genes. The results of GO analysis revealed that the gradually upregulated genes were significantly enriched in negative regulation of megakaryocytes, olfactory bulb interneuron differentiation, endothelial growth factor stimulus and other biological processes (Figure 3A). In contrast, the gradually downregulated genes are mainly involved in xenobiotic metabolic processes, responses to xenobiotic stimuli, cellular responses to xenobiotic stimuli and other biological processes (Figure 3B). Furthermore, the gradually upregulated genes are significantly involved in multiple pathogens of HCC-related pathways, such as viral carcinogenesis and alcoholism (Figure 3C) and the gradually downregulated genes in the PPAR signaling pathway, retinol metabolism, steroid hormone biosynthesis, bile secretion, and ABC transporter pathways (Figure 3D).

HCC-unfavorable/favorable gene set

A total of 10 gradually upregulated genes (ACP4, ATP6 V0D2, BRSK1, CHGA, CLEC2L, CREG2, CYP19A1, PNCK, STEAP1B, TMC7) were associated with poor overall survival and classified as an HCC-unfavorable gene set (Figure 4A). Moreover, 4 genes (BRSK1, CLEC2L, PNCK and TMC7) were included in the Human Protein Atlas, and all of them were highly expressed in HCC compared to normal liver (Figure 4B), which is consistent with our findings. Twenty-nine gradually downregulated genes were associated with good prognosis and classified as the HCC-favorable gene set (Table 1). CAMK4, DMGDH, IYD, CCDC42,
A ten-gene set variation score as a prognostic tool in hepatocellular carcinoma

Figure 2. Differential expression gene (DEG) analysis and principal component analysis (PCA). A. PCA of TCGA HCC gene expression profiles. B. Volcano plot of differentially expressed genes between stage I HCC and normal liver tissue. C. Volcano plot of differentially expressed genes between stage II HCC and normal liver tissue. D. Volcano plot of differentially expressed genes between stage III HCC and normal liver tissue. E. Volcano plot of differentially expressed genes between stage IV HCC and normal liver tissue. Red represents upregulated genes, blue represents downregulated genes, and gray represents no significantly differentially expressed genes. F. Common differentially expressed genes in HCC stages I-IV. Intersected genes represent genes that are dysregulated in all four stages of liver cancer development. G. PCA of gradually upregulated and downregulated genes.
A ten-gene set variation score as a prognostic tool in hepatocellular carcinoma
ESR1, CPEB3, CYP3A43, VIPR1, AKR1D1 and ADRA1A were the ten genes with the most significant association with a good prognosis (Figure 4C).

**HCC-unfavorable and HCC-favorable GSVA scores are biomarkers of HCC, and the HCC-unfavorable GSVA score is an independent prognostic factor**

The GSVA package was applied to calculate the HCC-unfavorable GSVA score and HCC-favorable GSVA score for all samples. Obviously, the HCC-favorable GSVA score decreased but the HCC-unfavorable GSVA score increased with HCC progression (Figure 5A). ROC curve analysis indicated that both HCC-unfavorable and -favorable GSVA scores are biomarkers for HCC, with AUC = 0.962 and AUC = 0.992, respectively (Figure 5B), as verified in GSE54236, with AUC = 0.679 and AUC = 0.862, respectively (Figure 5C). All HCC samples in TCGA were separated into low- and high-score groups according to the median GSVA score. Both GSVA score systems were associated with prognosis in univariate Cox proportional regression analysis. In addition, multivariate Cox proportional regression analysis indicated that the HCC-unfavorable GSVA score was an independent prognostic factor of HCC compared to clinicopathological features (Table 2), with AUC = 0.704 for the 2-year tROC curve (Figure 5D). As expected, a high HCC-unfavorable GSVA score was associated with poorer overall survival (Figure 5E), which was validated in GSE54236 (Figure 5F).

**Discussion**

HCC is one of the most lethal malignant tumors in the world [19]. Most HCCs are diagnosed at stages III and IV, resulting in poor prognoses. Furthermore, the pathological mechanism of HCC remains elusive, and there are no reliable biomarkers for use in the clinic to predict the survival of patients with HCC. Many previous studies have mainly focused on a single gene or molecule, without considering simultaneous changes in multiple genes [20-22]. In the present study, we identified 82 gradually upregulated genes and 176 gradually downregulated genes with HCC progression, which revealed that the development of HCC results from synergistic effects of multiple genes. Functional enrichment analysis indicated that the gradually upregulated genes are significantly involved in multiple pathogen-related pathways, such as viral carcinogenesis [23] and alcoholism [24]. This may indicate that expression of pathogen-related genes in HCC reflects the progression of HCC, and it thus would be crucial to eliminate pathogens in the management of HCC.

Survival analysis showed that only a few upregulated/downregulated genes are associated with prognosis. To the best of our knowledge, this is the first report of an HCC-unfavorable gene set including 10 gradually upregulated genes and an HCC-favorable gene set including 29 gradually downregulated genes. Not surprisingly, we found that some of these genes have been associated with cancer. In the HCC-unfavorable gene set, STEAP1B, TMC7, CYP19A1 and PNCK have been associated with prostate cancer, pancreatic carcinoma, and breast cancer, respectively [25-28]. In our study, we found that these genes may also be associated with HCC. ACP4, BRSK1, CHGA, ATP6 V0D2, CLEC2L, and CREG2 may also be associated with HCC, which has been rarely reported. Many genes in the HCC-favorable gene set have been identified to be associated with HCC, such as VIPR1, CPEB3, HTR2A-AS1, ACSM3, ADRA1A, AKR1D1, BHMT, CD226, CD5 L, CYP3A4, CYP3A43, DMGDH, ESR1, GLYATL1 and RDH16. For example, low expression of VIPR1 has an adverse prognostic impact on HCC [29], loss of ACSM3 expression correlates with advanced HCC stages and poor survival [30], downregulation of BHMT in HCC is associated with poor prognosis [31], low expression of CD226 promotes proliferative, migrating, and invasive activities of HCC cells [32], and downregulation of CYP3A4 is an independent predictor of early recurrence of HCC [33]. The results of previous studies are consistent with our findings.
A ten-gene set variation score as a prognostic tool in hepatocellular carcinoma

A

Survival Analysis of ACP4
Strata: g1high, g1low

Survival Analysis of ATP6V0D2
Strata: g1high, g1low

Survival Analysis of BRSK1
Strata: g1high, g1low

Survival Analysis of CHGA
Strata: g1high, g1low

Survival Analysis of CLEC2L
Strata: g1high, g1low

Survival Analysis of CREG2
Strata: g1high, g1low

Survival Analysis of CYP19A1
Strata: g1high, g1low

Survival Analysis of PNCK
Strata: g1high, g1low

Survival Analysis of STEAP1B
Strata: g1high, g1low

Survival Analysis of TMC7
Strata: g1high, g1low

B

BRSK1: Antibody HPA061719
CLEC2L: Antibody HPA045050
A ten-gene set variation score as a prognostic tool in hepatocellular carcinoma

Figure 4. Survival analysis and immunohistochemistry. A. Survival curves of 10 HCC-unfavorable gene sets. High expression of these genes is associated with a poor prognosis in HCC. B. High expression of genes by immunohistochemistry. Normal liver tissue samples are on the left, and HCC samples are on the right. C. Survival curves of the 10 genes most significantly correlated with prognosis in the HCC-favorable gene set. High expression of these genes is associated with a better prognosis in HCC.
A ten-gene set variation score as a prognostic tool in hepatocellular carcinoma

Table 1. The HCC-unfavorable gene set and HCC-favorable gene set

<table>
<thead>
<tr>
<th>Gene set</th>
<th>Gene symbol</th>
</tr>
</thead>
<tbody>
<tr>
<td>HCC-unfavorable gene set</td>
<td>ACP4, ATP6 V0D2, BRSK1, CHGA, CLEC2 L, CREG2, CYP19A1, PNCK, STEAP1B, TMC7</td>
</tr>
<tr>
<td>HCC-favorable gene set</td>
<td>ACSM3, ADRA1A, AKR1D1, BHMT, CAMK4, CCDC42, CD226, CD5 L, CLEC12A, CPEB3, CYP3A4, CYP3A43, DMGDH, ESR1, ETFDH, GHR, GLYATL1, GRAMD1C, HTR2A-AS1, IYD, LGI1, LINCO0885, NDST3, NR1I2, NUGGC, RANBP3 L, RDH16, SRD5A2, VIPR1</td>
</tr>
</tbody>
</table>

Figure 5. Evaluating the diagnostic and prognostic abilities of HCC-unfavorable and HCC-favorable GSVA scores. A. The HCC-unfavorable GSVA score gradually increased and the HCC-favorable GSVA score gradually decreased with HCC progression. B. ROC curves of HCC-unfavorable and HCC-favorable GSVA scores. C. ROC curves of HCC-unfavorable and HCC-favorable GSVA scores in GSE54236. D. tROC curves of the HCC-unfavorable GSVA score in TCGA. E. Survival analysis of the HCC-unfavorable GSVA score in the HCC cohort from TCGA. A high HCC-unfavorable GSVA score is associated with a poor prognosis in HCC. F. Survival analysis of the HCC-unfavorable GSVA score in the HCC cohort from TCGA and GSE54236. A high HCC-unfavorable GSVA score is associated with a poor prognosis in HCC.

Several gene signatures have been reported to predict prognosis in HCC [34-37]. In these studies, a coefficient for a gene in the training set was based on Cox regression analysis or another method, and the coefficient was various. However, due to the limitations of the sample...
size in previous studies as well as tumor heterogeneity, we may never obtain the real coefficient of a gene. Therefore, GSVA was used in our study to score individual samples against gene sets (HCC unfavorable and favorable). ROC curve analysis suggested that both the HCC-unfavorable GSVA score and HCC-favorable GSVA score have strong diagnostic capacity in HCC, and tROC curve analysis showed that the HCC-unfavorable GSVA score can be a prognostic biomarker. Univariate and multivariate Cox regression analyses further suggested that the HCC-unfavorable GSVA score is an independent factor for overall survival in HCC.

Tumor markers play a crucial role in the diagnosis of HCC, especially for early asymptomatic microfocal tumors. When imaging cannot be obtained, the abnormality of tumor markers plays an important role as a reference value. Various diagnostic markers, such as alpha-fetoprotein (AFP), α-L-fucosidase (AFU), desaturated-γ-carboxy-prothrombin (DCP) and phosphatidylinositol proteoglycan-3 (GPC-3), have been developed for HCC [38, 39]. AFP is currently the most widely used HCC tumor marker worldwide, but its sensitivity and specificity are not very satisfactory, especially in the early stage [40]. As AFU will increase to a certain extent in diabetes, pancreatitis, and hypothyroidism, it is not very specific for the early diagnosis of HCC, and it needs to be combined with other tumor markers to effectively detect HCC [41]. DCP is an abnormal prothrombin lacking coagulation activity detected in patients with liver cancer: it is related to the size and grade of the tumor and can be used to determine patient prognosis [42, 43]. GPC-3 is a type of heparan sulfate glycoprotein on the surface of the cell membrane that is not expressed in the normal human liver but is found at significantly higher levels in HCC patients than in those with benign liver disease [44]. Therefore, GPC-3 is helpful for the early diagnosis of HCC and differential diagnosis of benign and malignant liver tumors [45, 46]. In this study, the gene set score we developed was based on multiple genes. It can be used not only for the diagnosis of patients with early HCC but also for predicting the survival status of all HCC patients, and it is easy to obtain from peripheral blood samples.

Although we provide new insight into the HCC prognostic stratification system, several limitations were notable in the present study. First, the molecular mechanism requires experimental verification. Second, it is not clear whether the two gene sets are causal or merely markers for HCC and its prognosis. In addition, whether the gene set can distinguish other diseases, such as hepatitis or cirrhosis, remains to be further studied.

In conclusion, we identified an HCC-unfavorable gene set and an HCC-favorable gene set. The HCC-unfavorable and -favorable GSVA scores may serve as new biomarkers of HCC, and the HCC-unfavorable GSVA score is an independent biomarker for predicting prognosis.

Acknowledgements

This work was supported by the National Natural Science Foundation of China (NO. 8166-0498 and 81803007), China Postdoctoral Science Foundation (NO. 2019M663412), Guangxi Key Research and Development Plan (NO.

Table 2. Univariate and multivariate analyses of the two GSVA scores

<table>
<thead>
<tr>
<th>Factor</th>
<th>Univariate Cox analysis</th>
<th>Multivariate Cox analysis</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>β</td>
<td>P Value</td>
</tr>
<tr>
<td>Sex (male/female)</td>
<td>-0.232</td>
<td>0.228</td>
</tr>
<tr>
<td>Age (&gt;65 years/≤65 years)</td>
<td>0.239</td>
<td>0.202</td>
</tr>
<tr>
<td>T stage (T3-4/T1-2)</td>
<td>0.919</td>
<td>0.000</td>
</tr>
<tr>
<td>Lymph node stage (N2-3/N0-1)</td>
<td>0.684</td>
<td>0.341</td>
</tr>
<tr>
<td>Metastasis (M1/M0)</td>
<td>1.382</td>
<td>0.019</td>
</tr>
<tr>
<td>Pathological stage (III-IV/I-II)</td>
<td>0.911</td>
<td>0.000</td>
</tr>
<tr>
<td>Grade (G3-4/G1-2)</td>
<td>0.131</td>
<td>0.498</td>
</tr>
<tr>
<td>HCC-unfavorable GSVA score (high/low)</td>
<td>0.820</td>
<td>0.000</td>
</tr>
</tbody>
</table>
A ten-gene set variation score as a prognostic tool in hepatocellular carcinoma

Guangxi Scholarship Fund of Guangxi Education Department, Guangxi Medical University Training Program for Distinguished Young Scholars, Youth Talent Fund Project of Guangxi Natural Science Foundation (NO. 2018GXNSFBA281030, 2018-GXNSFBA281091), Guangxi Medical and Health Appropriate Technology Development and Application Project (No. S2017101, S2018062) and the Project of Nanning Scientific Research and Technology Development Plan (20153012).

Disclosure of conflict of interest

None.

Address correspondence to: Zhaoshan Fang, The Fifth Affiliated Hospital of Guangxi Medical University, Nanning 530022, Guangxi, People’s Republic of China. E-mail: fangzhsh_dr@126.com; Rong Liang, Department of Medical Oncology, Guangxi Medical University Cancer Hospital, NO. 71, Hedi Road, Nanning, Guangxi, People’s Republic of China. E-mail: ronglianggx@foxmail.com

References


A ten-gene set variation score as a prognostic tool in hepatocellular carcinoma


A ten-gene set variation score as a prognostic tool in hepatocellular carcinoma


