Using Integrated Multi-Omics Data Analysis to Identify 5-gene Signature for Predicting Survival of Patients with Hepatocellular Carcinoma

Rulin Zhang (✉ rulin_zhang@163.com)

Shanghai General Hospital <https://orcid.org/0000-0003-3402-9090>

Junyi Wu

Shanghai General Hospital

Yingying Zhao

Shanghai General Hospital

Heng Quan

Shanghai General Hospital

Dongge Xia

Shanghai General Hospital

Ziguang Niu

Shanghai General Hospital

Xiang Xu

Shanghai General Hospital

Xiaolei Liu

Zhengzhou University First Affiliated Hospital

Jun Wu

Shanghai General Hospital

Research Article

Keywords: HCC, multi-omics, 5-gene signature, prognosis biomarkers, qPCR

Posted Date: July 26th, 2021

DOI: <https://doi.org/10.21203/rs.3.rs-710165/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Abstract

Due to the poor prognosis for hepatocellular carcinoma (HCC) presently, a systemic analysis supported by the multi-omics data is extremely necessary to search for gene markers for the clinical prognostic prediction of HCC. The data on RNA-seq, single nucleotide polymorphism (SNP), and copy number variation (CNV), etc. were downloaded from TCGA, leading to a final of 367 samples, which were divided into training set and testing set randomly. In the training set, both prognosis-related genes and those with SNP or CNV were screened, which were incorporated for feature selection using the random forest method. The testing and GEO verification sets (N = 265) were used to verify the constructed gene-related prognosis model. qPCR was used to detect the expression of 5 genes in clinical specimens. After including genomic variant and prognosis-related genes, we got 78 candidate genes and 5 feature genes (CISH, LHPP, MGMT, PDRG1, and LCAT) eventually through random forest feature selection. The 5-gene signature is an independent prognostic risk factor for HCC patients. In addition, the signature shows good predicting performance and clinical practicality in training set, testing set and external verification set. The results of qPCR based on clinical samples showed that the expression of PDRG1 was increased in colon cancer tissues and the expression of CISH, LHPP, MGMT and LCAT were decreased in colon cancer tissues. We identify the ability of 5-gene signature to serve as an innovative marker of survival prediction for patients with HCC.

Introduction

Hepatocellular carcinoma (HCC) as a highly heterogenous malignant tumor originates from the liver epithelial or mesenchymal tissues and accounts for 90% of primary liver cancer [1, 2]. With a considerable high prevalence, HCC is characterized by frequent metastasis and strong invasion capacity; besides, its low 5-year survival rate makes it the second major malignant tumor that threatens human life [3, 4]. Therefore, the discovery of the prognosis biomarker for HCC patients is desperately required to facilitate accurate clinical outcome prediction for clinicians and the provision of reference for individualized medicine.

At present, numerous studies have searched for survival predicting biomarkers and explored the long-term prognosis guidelines for HCC. These biomarkers can be classified as several groups, including single-molecule as the independent prognosis indicator [such as MEP1A, hepatitis B virus (HBV), or other currently investigated novel markers] [5, 6], and the gene markers developed by a number of prognosis genes based on high-throughput gene expression profile analysis [7, 8]. So far, several systemic biological methods have been utilized to recognize the HCC gene biomarkers related to prognosis and to establish the gene feature. For instance, Sun Y and colleagues confirmed a 5-lncRNA signature in the file of gene expression through selection operator Cox regression, weighted gene correlation network analysis, etc. [9]; Désert R et al. had utilized meta-analysis across gene expression to recognize an 8-gene signature [10]; Wang Y et al. had developed a risk score system based on 4-methylated mRNA signature to evaluate and predict the prognosis of HCC patients [11]. All the above studies had tested their gene signatures in the independent external data sets, however, without the receptive areas under the curve (AUC) at 3- to 5-year.

Therefore, it remains challenging to identify the biomarker or gene signature that successfully and accurately predicts the survival for HCC patients, and more cohorts are required for verification. All in all, it is extremely significant and urgent to identify the HCC prognosis-related gene signals through their biological functions from the bioinformatic analysis.

In this study, we proposed a systemic and multi-omics scheme aiming to identify a reliable and effective gene signature related to the prognosis of HCC. The gene expression profiles of HCC patients including the data on single nucleotide polymorphism (SNP) and copy number variation (CNV) were obtained from TCGA and GEO databases. Meanwhile, we established a 5-gene signature by integrating the genomic and transcriptomic data to screen the prognosis markers, whose efficacy in survival prediction was further verified internally in testing set and externally in verification set. The 5-gene signature was found by GO analysis to have important involvement in both HCC biological processes and pathways. Parallel results also came from the next GSEA analysis, indicating the ability of 5-gene signature to complete risk prediction and provision of solid evidence for well understanding the molecular mechanism of HCC.

Materials & Methods

Data collection and processing

UCSC cancer browser (<https://xenabrowser.net/datapages/>) was used to obtain the data on TCGA RNA-seq, clinical follow-up information, and CNV of the SNP 6.0 chip [12]. In addition, the GDC client and GEO database were used to acquire the mutation annotation file (MAF) and the profile data on both GSE40873 and GSE15654 expression and clinical follow-up information, respectively[13, 14]. Besides, we screened 367 tumor samples with follow-up information from the TCGA RNA-seq data, and then further divided them into two groups randomly, including 184 in training set and 183 in testing set. Moreover, we externally verified the established signature in two verification sets: GSE40873 (N=216) and GSE15654 (N=49).

Preliminary screening of prognostic related genes

It is necessary to use prognostic-related genes for follow-up analysis, and univariate Cox proportional hazards regression analyses could provide some Prognosis-related information. With the cutoff of $p < 0.01$, genes were seemed to be Prognosis-related. As previously reported by Guo's team [15], univariate analysis via Cox proportional hazard regression was performed to screen the potential genes which were markedly associated with patients' overall survival (OS) in training set, with $P < 0.01$ as the threshold to screen the prognosis related genes.

CNV and mutation analysis

For CNV, the loss and gain levels of copy-number changes have been identified using segmentation analysis and GISTIC algorithm. Therefore, in the present study, the significantly amplified or deleted genes were identified by GISTIC 2.0 software [16]. Fragments with either amplification or deletion length

>0.1 and P<0.05 were set as parameter thresholds. To identify the significantly mutant genes, downloading the MAF file of the mutation data from the GDC download tool (<https://portal.gdc.cancer.gov/>) of the TCGA database. the significantly mutant genes from the MAF of TCGA mutation data were identified by using Mutsig 2.0 software [17], with P<0.05 as the threshold.

Construction of the prognosis-related gene signature

We selected the significantly OS-related genes and those with CNV (amplification or deletion) and mutations, and the significance of the prognosis genes was sorted through a random survival forest algorithm [18]. R package random Survival Forest was applied for gene screen, which was similar to Meng's work [19]. The number of monte-carlo iterations was set as 100, with 5 forward steps, to identify genes with the significance of >0.65 as the feature genes.

The following risk scoring model was constructed by multivariate Cox regression analysis:

$$RiskScore = \sum_{k=1}^n Exp_k * e^{HR_k}$$

In which N refers to the predicted gene number, Exp_k refers to the gene expression value, and $HR_k = \sum_{i=1}^N \beta_{ik} * e^{\beta_{ik}}$ indicates the estimated gene regression coefficient in the established Cox regression model.

Analyses of functional enrichment

Package cluster profiler for genes in R language was utilized to conduct the enrichment analysis of Gene Ontology (GO) and Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway [20], to confirm the over-expressed GO terms in biological processes, KEGG pathways, and molecular function. A false discovery rate (FDR) lower than 0.05 was considered statistically significant for all analyses.

Gene Set Enrichment Analysis (GSEA)

JAVA program (<http://software.broadinstitute.org/gsea/downloads.jsp>) was used to conduct GSEA [21], together with the usage of MSigDB [22] as well as C2 Canonical pathways gene set collection. After 1000 permutations, gene sets with an FDR <0.05 were regarded as significantly enriched.

Real-time qPCR

Twenty five specimens of hepatocellular carcinoma were collected. The study was approved by the Ethics Committee of Shanghai General Hospital (the approval number is 020KY053). Normal tissues were matched with tumor biopsies of the same patient. Among the 25 patients, 10 patients were in stage 1, 10 patients were in stage 2, and 5 patients were in stage 3. These patients did not receive chemotherapy or radiotherapy before operation. We obtained the informed consent of each patient and obtained the approval of the human ethics review committee of Shanghai General Hospital. Real-time

qPCR was Power SYBR™ Green PCR Master Mix (No. A25742, Thermo Fisher Scientific, Waltham, MA, USA). Relative expression was calculated based on $2^{-\Delta\Delta C_t}$ method.

Statistical analysis

Setting the median risk score as cutoffs the Kaplan-Meier (KM) curve in each data set was plotted, to perform the comparison of survival risk between the high-risk and low-risk groups. The independent gene markers were identified by multivariate Cox regression analysis. A Kaplan-Meier (KM) curve is drawn when using the median risk score in each dataset as a cutoff to compare the survival risk between high-risk and low-risk groups. Multivariate Cox regression analysis is performed to test whether genetic markers are independent prognostic

Results

Identification of the patient OS-related gene set

Univariate regression analysis was applied to construct the relationship between gene expression and patient OS based on the TCGA training set samples. Table 1 shows the basic information of all samples in each group. A total of 1,265 genes with a univariate Cox regression log-rank p-value of <0.01 were identified as candidate prognostic genes. Table 2 presents the information of the top 20 genes with the closest correlation.

Identification of genomic variant genes with CNV and mutation

Genes with significant deletion or amplification were identified using GISTIC 2.0 based on the TCGA CNV data, and the fragment with the deletion or amplification length of over 0.1 ($p < 0.05$) was seen as the parameter threshold. The significantly amplified fragments in the HCC genome are shown in Figure 1A, and the significantly amplified genes in each fragment are recorded in Table 3. For instance, VEGFA on the 6p21.1 region was significantly amplified (q value = $1.53E-15$), MYC on the 8q24.21 region was evidently amplified (q value = $7.39E-37$), CCND1 on the 11q13.3 region was also markedly amplified (q value = $1.51E-28$), and a total of 200 genes were therefore amplified. The significantly deleted fragments in the HCC genome are displayed in Figure 1B, and the markedly deleted genes in each fragment are recorded in Table 4. For example, RB1 on the 13q14.2 region was markedly deleted (q value = $6.95E-22$), CDKN2A on the 9p21.3 region was evidently deleted (q value = $4.65E-11$), and CD3D on the 11q25 region was significantly deleted (q value = 0.010127), resulting in a total of 1,485 deleted genes.

Significantly mutant genes were identified using Mutsig2 based on the TCGA mutation annotation data, and the threshold was set at $P < 0.05$, thereby obtaining a total of 344 genes with significant mutation frequency. Figure 2 presents the distribution of framework deletion or insertion, framework displacement, splice sites, nonsense mutation, missense mutation, synonymous mutation, and other non-synonymous mutation of the top 50 genes with the highest significance in the TCGA HCC patient samples. The number of samples with mutation in the 50 genes is indicated in the right histogram, while the upper histogram

represents the total number of non-synonymous and synonymous mutations in the 50 genes. Therefore, some of the 344 identified genes were found in previous studies to be closely correlated with cancer genesis and development, such as RB1, TP53, PTEN, CDKN2A, and CDKN1A [23-26].

Functional analysis of genomic variant genes with CNV and mutation

To explore the genomic variant genes (n = 2,029 in total), CNV identified amplified and deleted genes were integrated with significantly mutant genes for GO and KEGG enrichment analyses. Figure 3A shows that a total of 2,029 genes are markedly enriched in the cancer genesis and development-related pathways, such as Hepatocellular carcinoma, PI3K-AKT signaling pathway, FoxO signaling pathway, and Human T-cell leukemia virus 1 infection, etc. Figure 3B displays that the 2,029 genes are enriched in the biological processes of cancer genesis and development, including metabolic process, cellular process, cell differentiation, immune system process, etc.

Construction of the HCC prognosis-related gene signature

The gene set was first integrated with genomic CNV and mutation as well as the prognosis-related gene set, and the intersection set of the three gene sets was chosen as the candidate gene set, and thus 78 genes were obtained. The random forest was further applied for feature selection. Figure 4A displays the relationship between the number of classification trees and the error rate, and Table 4 exhibits 5 genes with the relative importance of >0.65. Figure 4B presents the 5 genes in the out-of-bag importance ranking. Then the multivariate Cox regression analysis method was used to construct the 5-gene signature, and the model is shown below:

$$\text{Risk} = -0.29823 \cdot \text{CISH} - 0.04233264 \cdot \text{LHPP} - 0.2659875 \cdot \text{MGMT} + 0.2069671 \cdot \text{PDRG1} - 0.2678402 \cdot \text{LCAT}$$

Through calculating the risk score of each training set sample, the samples were divided into groups with high or low risk given the median risk score (cutoff = -0.05475002). Figure 5 presents the classification effect of the 5-gene signature in the TCGA training set. As shown in Figure 5A, there is an evident difference between the high-risk group of 92 patients and the low-risk group of 92 patients ($P = 9.101663 \times 10^{-11}$). Figure 5B displays the ROC curve, with the 1-, 3- and 5-year AUC of 0.77, 0.82, and 0.86, respectively. According to Figure 5C, the survival time of death samples declined significantly with the increase in patient risk score, and more death samples were found in the high-risk group. Besides, the high expression of PDRG1 was identified to be the risk factor based on the variation in the expression of 5 different signature genes with the increase in the risk score. By contrast, the high expression of CISH, LHPP, MGMT, and LCAT was related to the low risk and thereby determined as protective factors.

Robustness verification of the 5-gene signature model

Through verification in the TCGA test set, the same model and cutoff were adopted as the TCGA training set to decide the 5-gene signature robustness. Figure 6 exhibits the classification effect of the TCGA test set. As shown in Figure 6A, there was a significant statistical difference between the high-risk group of 97 patients and the low-risk group of 86 patients ($P = 0.003221398$). Figure 6B shows the ROC curve, with

the 1-, 3- and 5-year AUC of 0.74, 0.69, and 0.59, respectively. Figure 6C shows similar results to those in the TCGA training set. To be specific, the survival time of death samples declined evidently with the increase in the risk score, and more death samples were found in the high-risk group. As expected, PDRG1 was identified as the risk factor; CISH, LHPP, MGMT, and LCAT were the protective factors.

To validate the classification performance of the 5-gene signature model in data from different data platforms, GEO platform data (GSE40873 and GSE15654) were treated as the external dataset to calculate the risk score of each sample using the model. Besides, samples from low-risk and high-risk groups were classified based on the training set cutoff. Figures 7A and 8A indicate that the low-risk group had a markedly better prognosis than that in the high-risk group. ROC analysis revealed similar 1-, 3-, 5-year AUC to those in the training set and test set, as presented in Figures 7B and 8B. Additionally, the relationship between the risk score and the expression of 5 genes was also in line with that between the test set and the training set, according to Figures 7C and 8C. To sum up, the 5-gene signature model displays favorable prognosis prediction capacity in both external and internal data.

Clinical independence of the 5-gene signature

To determine the independence of the 5-gene signature model in clinical application, univariate and multivariate Cox regression analyses of clinical information were carried out in the TCGA training and test sets, GSE40873 and GSE15654 data sets for an analysis of the related HR, 95% CI of HR and p-value. In addition, the clinical information recorded in the TCGA, GSE40873, and GSE15654 samples was systemically analyzed, which included age, sex, HBV, HCV, platelet count, tumor stage, and pathological M stage, N stage, T stage. Table 5 shows the grouping information of the 5-gene signature. The univariate Cox regression analysis of the TCGA training set suggested that age, tumor stage III/IV, pathologic M1/T3, and the high-risk group were markedly associated with survival. However, clinical independence was only observed in the high-risk group (HR = 9.51, 95%CI = 1.78-50.63, P = 0.00828), according to the corresponding multivariate Cox regression analysis. The univariate Cox regression analysis of the TCGA test set suggested that the high risk group and pathologic M1/T3/T4 were significantly correlated with survival, but clinical independence was only found in the gender male (HR = 4.71, 95%CI = 1.08-20.55, P = 0.0389) and the high-risk group (HR = 3.38, 95%CI = 1.08-10.57, P = 0.0358), according to the corresponding multivariate Cox regression analysis. In GSE40873, the high risk group was found to be significantly related with survival by the univariate Cox regression analysis, and clinical independence was also observed in the high risk group (HR = 3.64, 95%CI = 1.27-10.42, P = 0.0158), according to the corresponding multivariate Cox regression analysis. In GSE15654, the analysis suggested that the high risk group, varices presence, and platelet count $>100000/\text{mm}^3$ were markedly correlated with survival, and clinical independence was observed in the platelet count $>100000/\text{mm}^3$ (HR = 3.03, 95%CI = 1.69-5.39, P = 0.000178) and the high risk group (HR = 1.69, 95%CI = 1.02-2.81, P = 0.040943), according to the corresponding multivariate Cox regression analysis (Table 4). To sum up, the 5-gene signature model is a prognostic index and shows independent predictive performance.

Pathway difference in GSEA enrichment analysis between high and low risk groups

In the high and low risk groups of the TCGA training set, pathways with significant enrichment were analyzed using GSEA. Table 5 shows that a total of 39 pathways were obtained, including those closely related to tumor genesis, development, and metastasis, such as adherens junction, cell cycle, endocytosis, and pathways in cancer. The significant enrichment results of these pathways in high-risk samples are shown in Figure 9.

Expression Levels of 5 Genes

The bioinformatics analysis confirmed that the expression of 5 genes was verified in twenty-five normal tissues and HCC tissues. The results in Figure 10 showed that the mRNA expression of PDRG1 was decreased in HCC tissues and the mRNA expression of CISH, LHPP, MGMT and LCAT were increased in HCC tissues ($p < 0.05$). It was consistent with that analyzed using bioinformatic analysis.

Discussion

From the perspective of prognosis, HCC is a highly heterogeneous disease as HCC patients at similar stages have distinctly different survival times [27]. Increasingly, HCC is discovered and treated at the early stage, and conventional clinicopathological indicators, like TNM stage, vascular invasion, and cancer embolus, can hardly adapt to the current requirement of individual result prediction. Particularly, no universally applied therapeutic strategy has been verified to be effective for risk stratification [28]. For the individualized prevention and treatment of HCC patients, therefore, it is of great significance to screen the prognosis molecular markers that can sufficiently reflect the biological features of tumors [29–31]. In this paper, the expression profiles of HCC samples from GEO and TCGA were analyzed to identify the stable and reliable 5-gene signature related to OS, which was independent of clinical factors.

So far, some gene signatures have been applied clinically, such as the Oncotype DX for disease recurrence grading constructed based on the expression of 21 genes in breast cancer [32], and the Coloprint constructed based on the expression of 18 genes in colon cancer [33]. Thus, it is suggested that screening novel prognostic markers in cancer through gene expression profiles has become the most promising high-throughput molecular identification method. Up to now, several systemic biological methods have been utilized to identify HCC prognosis-related gene biomarkers and structure gene features. However, the situation of the low 3-5-year AUC can be observed occasionally after verification using the excessive gene number or external data set, which is not conducive to further promoting and verifying big data [34, 35]. By contrast, the 5-gene signature in this study possessed high AUC and fewer genes, which facilitated clinical conversion. In the 5-gene signature, PDRG1 was the risk factor, while CISH, LHPP, MGMT, and LCAT were the protective factors. Several studies have demonstrated that PDRG1 is a marker for multiple malignant tumors, which is closely correlated with the prognosis of multiple tumors, such as colorectal cancer, ovarian cancer, cervical cancer, breast cancer, gastric cancer, and lung cancer [36–38]. Besides, MGMT is tightly related to the prognosis of breast cancer, esophageal cancer, and glioma [39–41]; LCAT is verified to be markedly correlated with the prognosis of HCC, which has constituted a 4-gene signature with SPINK1, TXNRD1, and PZP [42]. Additionally, CISH is associated with

the prognosis for patients with gallbladder carcinoma, breast cancer, prostate cancer, and multiple myeloma [43–45].

Overall, PDRG1, CISH, MGMT, and LCAT have been gradually verified to be closely correlated with the tumor prognosis, though LHPP is not reported to be related to tumor at present. This study proves that the mRNA expression of PDRG1 were decreased in HCC tissues and the mRNA expression of CISH, LHPP, MGMT and LCAT were increased in HCC tissues. This study first put forward that the 5 genes could serve as novel prognosis markers for HCC; at the same time, the pathways enriched by the 5-gene signature were markedly related to the genesis, development pathway, and biological process of HCC, according to the results of GSEA analysis. These results suggested that the model could be potentially applied in clinical practice, providing a potential target for the diagnosis of clinical patients. Several studies have shown that GSEA enrichment analysis results is closely related to the progression and metastasis of COAD [46]. Examples include KEGG ERBB SIGNALING PATHWAY, KEGG COLORECTAL CANCER, KEGG p53 SIGNALING PATHWAY, and KEGG TGF BETA SIGNALING PATHWAY.

In this study, bioinformatics has been applied to identify potential candidate genes related to tumor prognosis through big data analysis, but there are still some limitations that should also be noted. Firstly, some clinical follow-up information was lacking for samples; as a result, no consideration was given to whether there were biomarkers that could distinguish the prognosis of patients based on other health status factors of patients. Secondly, the results obtained from bioinformatic analysis alone were not sufficient, which should be verified through clinical investigation and experiments. Therefore, further studies with a larger sample size should be conducted for experimental verification.

Conclusions

To sum up, in this study, we have developed a 5-gene signature prognosis stratification system, which has good AUC in both the training set and verification set and is independent of the clinical features. Compared with clinical features, gene classifier can improve the survival risk prediction. Therefore, we recommend that such classifier should be used as the molecular diagnosis test to evaluate the prognosis risk of HCC patients.

Declarations

Funding

This work was supported in part by National Natural Science Foundation of China (No. 82072892), Natural Science Foundation of Shanghai (No. 21ZR1454900) and Key Discipline Project of Shanghai Jiading District (No. 2020-jdyxzdxx-13).

Conflict of interest

The authors have declared that no competing interest exists.

Contributions

Conceived and designed the study: Ru-lin Zhang, Xiao-lei Liu, Jun Wu. Collected the data and clinical samples: Ru-lin Zhang, Jun-yi Wu, Heng Quan, Dong-ge Xia, Zi-guang Niu and Xiang Xu. Data analyses: Ru-lin Zhang, Ying-ying Zhao and Jun Wu. Wrote the manuscript: Ru-lin Zhang, Jun-yi Wu, Xiao-lei Liu and Jun Wu. All authors read and approved the final manuscript.

Ethical approval

qPCR research samples were isolated from the human HCC tissues, as permitted by The Ethics Committee of Shanghai General Hospital. The Approval number is 020KY053.

Availability of data and materials

All data generated or analysed during this study are included in this published article.

Author's statement

The Author(s) declare that the paper is being submitted for consideration for publication in Molecular Biology Reports, that the content has not been published or submitted for publication elsewhere. This study has not been submitted in as a brief abstract in the proceedings of a scientific meeting or symposium.

References

1. Singal AG, Murphy CC: Hepatocellular Carcinoma: A Roadmap to Reduce Incidence and Future Burden. *J Natl Cancer Inst* 2019, 111:527-528.
2. Baecker A, Liu X, La Vecchia C, Zhang ZF: Worldwide incidence of hepatocellular carcinoma cases attributable to major risk factors. *Eur J Cancer Prev* 2018, 27:205-212.
3. Mancebo A, Varela M, Gonzalez-Dieguez ML, Navascues CA, Cadahia V, Mesa-Alvarez A, Rodrigo L, Rodriguez M: Incidence and risk factors associated with hepatocellular carcinoma surveillance failure. *J Gastroenterol Hepatol* 2018, 33:1524-1529.
4. White DL, Thrift AP, Kanwal F, Davila J, El-Serag HB: Incidence of Hepatocellular Carcinoma in All 50 United States, From 2000 Through 2012. *Gastroenterology* 2017, 152:812-820 e815.
5. OuYang HY, Xu J, Luo J, Zou RH, Chen K, Le Y, Zhang YF, Wei W, Guo RP, Shi M: MEP1A contributes to tumor progression and predicts poor clinical outcome in human hepatocellular carcinoma. *Hepatology* 2016, 63:1227-1239.
6. Zhang T, Liu Z, Zhao X, Mao Z, Bai L: A novel prognostic score model based on combining systemic and hepatic inflammation markers in the prognosis of HBV-associated hepatocellular carcinoma patients. *Artif Cells Nanomed Biotechnol* 2019, 47:2246-2255.

7. Li N, Li L, Chen Y: The Identification of Core Gene Expression Signature in Hepatocellular Carcinoma. *Oxid Med Cell Longev* 2018, 2018:3478305.
8. Gao Q, Wang XY, Qiu SJ, Zhou J, Shi YH, Zhang BH, Fan J: Tumor stroma reaction-related gene signature predicts clinical outcome in human hepatocellular carcinoma. *Cancer Sci* 2011, 102:1522-1531.
9. Sun Y, Zhang F, Wang L, Song X, Jing J, Zhang F, Yu S, Liu H: A five lncRNA signature for prognosis prediction in hepatocellular carcinoma. *Mol Med Rep* 2019, 19:5237-5250.
10. Desert R, Mebarki S, Desille M, Sicard M, Lavergne E, Renaud S, Bergeat D, Sulpice L, Perret C, Turlin B, Clement B, Musso O: "Fibrous nests" in human hepatocellular carcinoma express a Wnt-induced gene signature associated with poor clinical outcome. *Int J Biochem Cell Biol* 2016, 81:195-207.
11. Wang Y, Ruan Z, Yu S, Tian T, Liang X, Jing L, Li W, Wang X, Xiang L, Claret FX, Nan K, Guo H: A four-methylated mRNA signature-based risk score system predicts survival in patients with hepatocellular carcinoma. *Aging (Albany NY)* 2019, 11:160-173.
12. Goldman M, Craft B, Swatloski T, Cline M, Morozova O, Diekhans M, Haussler D, Zhu J: The UCSC Cancer Genomics Browser: update 2015. *Nucleic Acids Res* 2015, 43:D812-817.
13. Kudo A, Mogushi K, Takayama T, Matsumura S, Ban D, Irie T, Ochiai T, Nakamura N, Tanaka H, Anzai N, Sakamoto M, Tanaka S, Arai S: Mitochondrial metabolism in the noncancerous liver determine the occurrence of hepatocellular carcinoma: a prospective study. *J Gastroenterol* 2014, 49:502-510.
14. Hoshida Y, Villanueva A, Sangiovanni A, Sole M, Hur C, Andersson KL, Chung RT, Gould J, Kojima K, Gupta S, Taylor B, Crenshaw A, Gabriel S, Minguez B, Iavarone M, Friedman SL, Colombo M, Llovet JM, Golub TR: Prognostic gene expression signature for patients with hepatitis C-related early-stage cirrhosis. *Gastroenterology* 2013, 144:1024-1030.
15. Guo JC, Wu Y, Chen Y, Pan F, Wu ZY, Zhang JS, Wu JY, Xu XE, Zhao JM, Li EM, Zhao Y, Xu LY: Protein-coding genes combined with long noncoding RNA as a novel transcriptome molecular staging model to predict the survival of patients with esophageal squamous cell carcinoma. *Cancer communications (London, England)* 2018, 38:4.
16. Mermel CH, Schumacher SE, Hill B, Meyerson ML, Beroukhi R, Getz G: GISTIC2.0 facilitates sensitive and confident localization of the targets of focal somatic copy-number alteration in human cancers. *Genome Biol* 2011, 12:R41.
17. Pickering CR, Zhou JH, Lee JJ, Drummond JA, Peng SA, Saade RE, Tsai KY, Curry JL, Tetzlaff MT, Lai SY, Yu J, Muzny DM, Doddapaneni H, Shinbrot E, Covington KR, Zhang J, Seth S, Caulin C, Clayman GL, El-Naggar AK, Gibbs RA, Weber RS, Myers JN, Wheeler DA, Frederick MJ: Mutational landscape of aggressive cutaneous squamous cell carcinoma. *Clin Cancer Res* 2014, 20:6582-6592.
18. Taylor JM: Random Survival Forests. *J Thorac Oncol* 2011, 6:1974-1975.
19. Meng J, Li P, Zhang Q, Yang Z, Fu S: A four-long non-coding RNA signature in predicting breast cancer survival. *J Exp Clin Cancer Res* 2014, 33:84.
20. Yu G, Wang LG, Han Y, He QY: clusterProfiler: an R package for comparing biological themes among gene clusters. *Omics : a journal of integrative biology* 2012, 16:284-287.

21. Subramanian A, Kuehn H, Gould J, Tamayo P, Mesirov JP: GSEA-P: a desktop application for Gene Set Enrichment Analysis. *Bioinformatics* 2007, 23:3251-3253.
22. Liberzon A, Subramanian A, Pinchback R, Thorvaldsdottir H, Tamayo P, Mesirov JP: Molecular signatures database (MSigDB) 3.0. *Bioinformatics* 2011, 27:1739-1740.
23. Edamoto Y, Hara A, Biernat W, Terracciano L, Cathomas G, Riehle HM, Matsuda M, Fujii H, Scoazec JY, Ohgaki H: Alterations of RB1, p53 and Wnt pathways in hepatocellular carcinomas associated with hepatitis C, hepatitis B and alcoholic liver cirrhosis. *Int J Cancer* 2003, 106:334-341.
24. Gouas DA, Shi H, Hautefeuille AH, Ortiz-Cuaran SL, Legros PC, Szymanska KJ, Galy O, Egevad LA, Abedi-Ardekani B, Wiman KG, Hantz O, Caron de Fromentel C, Chemin IA, Hainaut PL: Effects of the TP53 p.R249S mutant on proliferation and clonogenic properties in human hepatocellular carcinoma cell lines: interaction with hepatitis B virus X protein. *Carcinogenesis* 2010, 31:1475-1482.
25. Hou W, Liu J, Chen P, Wang H, Ye BC, Qiang F: Mutation analysis of key genes in RAS/RAF and PI3K/PTEN pathways in Chinese patients with hepatocellular carcinoma. *Oncol Lett* 2014, 8:1249-1254.
26. Biden K, Young J, Buttenshaw R, Searle J, Cooksley G, Xu DB, Leggett B: Frequency of mutation and deletion of the tumor suppressor gene CDKN2A (MTS1/p16) in hepatocellular carcinoma from an Australian population. *Hepatology* 1997, 25:593-597.
27. Zucman-Rossi J, Villanueva A, Nault JC, Llovet JM: Genetic Landscape and Biomarkers of Hepatocellular Carcinoma. *Gastroenterology* 2015, 149:1226-1239 e1224.
28. Guichard C, Amaddeo G, Imbeaud S, Ladeiro Y, Pelletier L, Maad IB, Calderaro J, Bioulac-Sage P, Letexier M, Degos F, Clement B, Balabaud C, Chevet E, Laurent A, Couchy G, Letouze E, Calvo F, Zucman-Rossi J: Integrated analysis of somatic mutations and focal copy-number changes identifies key genes and pathways in hepatocellular carcinoma. *Nat Genet* 2012, 44:694-698.
29. Woo HG, Park ES, Lee JS, Lee YH, Ishikawa T, Kim YJ, Thorgeirsson SS: Identification of potential driver genes in human liver carcinoma by genomewide screening. *Cancer Res* 2009, 69:4059-4066.
30. Kwon SM, Kim DS, Won NH, Park SJ, Chwae YJ, Kang HC, Lee SH, Baik EJ, Thorgeirsson SS, Woo HG: Genomic copy number alterations with transcriptional deregulation at 6p identify an aggressive HCC phenotype. *Carcinogenesis* 2013, 34:1543-1550.
31. Woo HG, Choi JH, Yoon S, Jee BA, Cho EJ, Lee JH, Yu SJ, Yoon JH, Yi NJ, Lee KW, Suh KS, Kim YJ: Integrative analysis of genomic and epigenomic regulation of the transcriptome in liver cancer. *Nat Commun* 2017, 8:839.
32. Olsson-Brown A, Piskilidis P, O'Hagan J, Thorp N, Robson P, Innes H, Wong H, Cicconi S, Jackson R, Kiernan T, Holcombe C, O'Reilly S, Palmieri C: The impact of the 21-gene recurrence score (Oncotype DX) on concordance of adjuvant therapy decision making as measured by the Liverpool Systemic Therapy Adjuvant Decision Tool. *Breast (Edinburgh, Scotland)* 2019, 44:94-100.
33. Tan IB, Tan P: Genetics: an 18-gene signature (ColoPrint(R)) for colon cancer prognosis. *Nat Rev Clin Oncol* 2011, 8:131-133.

34. Zhang J, Baddoo M, Han C, Strong MJ, Cvitanovic J, Moroz K, Dash S, Flemington EK, Wu T: Gene network analysis reveals a novel 22-gene signature of carbon metabolism in hepatocellular carcinoma. *Oncotarget* 2016, 7:49232-49245.
35. Qiao GJ, Chen L, Wu JC, Li ZR: Identification of an eight-gene signature for survival prediction for patients with hepatocellular carcinoma based on integrated bioinformatics analysis. *PeerJ* 2019, 7:e6548.
36. Pajares M: PDRG1 at the interface between intermediary metabolism and oncogenesis. *World journal of biological chemistry* 2017, 8:175-186.
37. Tao Z, Chen S, Mao G, Xia H, Huang H, Ma H: The PDRG1 is an oncogene in lung cancer cells, promoting radioresistance via the ATM-P53 signaling pathway. *Biomedicine & pharmacotherapy = Biomedecine & pharmacotherapie* 2016, 83:1471-1477.
38. Jiang L, Luo X, Shi J, Sun H, Sun Q, Sheikh MS, Huang Y: PDRG1, a novel tumor marker for multiple malignancies that is selectively regulated by genotoxic stress. *Cancer biology & therapy* 2011, 11:567-573.
39. An N, Shi Y, Ye P, Pan Z, Long X: Association Between MGMT Promoter Methylation and Breast Cancer: a Meta-Analysis. *Cellular physiology and biochemistry : international journal of experimental cellular physiology, biochemistry, and pharmacology* 2017, 42:2430-2440.
40. Zhao JJ, Li HY, Wang D, Yao H, Sun DW: Abnormal MGMT promoter methylation may contribute to the risk of esophageal cancer: a meta-analysis of cohort studies. *Tumour biology : the journal of the International Society for Oncodevelopmental Biology and Medicine* 2014, 35:10085-10093.
41. Zhang Y, Zhu J: Ten genes associated with MGMT promoter methylation predict the prognosis of patients with glioma. *Oncology reports* 2019, 41:908-916.
42. Zheng Y, Liu Y, Zhao S, Zheng Z, Shen C, An L, Yuan Y: Large-scale analysis reveals a novel risk score to predict overall survival in hepatocellular carcinoma. *Cancer management and research* 2018, 10:6079-6096.
43. Martínez-Baños D, Sánchez-Hernández B, Jiménez G, Barrera-Lumbreras G, Barrales-Benítez O: Global methylation and promoter-specific methylation of the P16, SOCS-1, E-cadherin, P73 and SHP-1 genes and their expression in patients with multiple myeloma during active disease and remission. *Experimental and therapeutic medicine* 2017, 13:2442-2450.
44. Ghafouri-Fard S, Oskooei VK, Azari I, Taheri M: Suppressor of cytokine signaling (SOCS) genes are downregulated in breast cancer. *World journal of surgical oncology* 2018, 16:226.
45. Stone L: Putting a SOCS in prostate cancer. *Nature reviews Urology* 2019, 16:147.
46. Yiping Lu , Si Wu ,Changwan Cui ,Miao Yu ,Shuang Wang Yuanyi Yue ,Miao Liu ,Zhengrong Sun Gene Expression Along with Genomic Copy Number Variation and Mutational Analysis Were Used to Develop a 9-Gene Signature for Estimating Prognosis of COAD.*OncoTargets and Therapy* 2020,13: 10393–10408.

Tables

Figures

Figure 1

Genes with significant deletion or amplification were identified using GISTIC 2.0 based on the TCGA CNV data, and the fragment with the deletion or amplification length of over 0.1 ($p < 0.05$) was seen as the parameter threshold. The significantly amplified fragments in the HCC genome are shown in Figure 1A, and the significantly amplified genes in each fragment are recorded in Table 3. For instance, VEGFA on the 6p21.1 region was significantly amplified (q value = 1.53×10^{-15}), MYC on the 8q24.21 region was evidently amplified (q value = 7.39×10^{-37}), CCND1 on the 11q13.3 region was also markedly amplified (q value = 1.51×10^{-28}), and a total of 200 genes were therefore amplified. The significantly deleted fragments in the HCC genome are displayed in Figure 1B, and the markedly deleted genes in each fragment are recorded in Table 4. For example, RB1 on the 13q14.2 region was markedly deleted (q value = 6.95×10^{-22}), CDKN2A on the 9p21.3 region was evidently deleted (q value = 4.65×10^{-11}), and CD3D on the 11q25 region was significantly deleted (q value = 0.010127), resulting in a total of 1,485 deleted genes.

Figure 2

Significantly mutant genes were identified using Mutsig2 based on the TCGA mutation annotation data, and the threshold was set at $P < 0.05$, thereby obtaining a total of 344 genes with significant mutation frequency. Figure 2 presents the distribution of framework deletion or insertion, framework displacement, splice sites, nonsense mutation, missense mutation, synonymous mutation, and other non-synonymous mutation of the top 50 genes with the highest significance in the TCGA HCC patient samples. The number of samples with mutation in the 50 genes is indicated in the right histogram, while the upper histogram represents the total number of non-synonymous and synonymous mutations in the 50 genes. Therefore, some of the 344 identified genes were found in previous studies to be closely correlated with cancer genesis and development, such as RB1, TP53, PTEN, CDKN2A, and CDKN1A [23-26].

Figure 3

To explore the genomic variant genes (n = 2,029 in total), CNV identified amplified and deleted genes were integrated with significantly mutant genes for GO and KEGG enrichment analyses. Figure 3A shows that a total of 2,029 genes are markedly enriched in the cancer genesis and development-related pathways, such as Hepatocellular carcinoma, PI3K-AKT signaling pathway, FoxO signaling pathway, and Human T-cell leukemia virus 1 infection, etc. Figure 3B displays that the 2,029 genes are enriched in the biological processes of cancer genesis and development, including metabolic process, cellular process, cell differentiation, immune system process, etc.

Figure 4

The gene set was first integrated with genomic CNV and mutation as well as the prognosis-related gene set, and the intersection set of the three gene sets was chosen as the candidate gene set, and thus 78 genes were obtained. The random forest was further applied for feature selection. Figure 4A displays the relationship between the number of classification trees and the error rate, and Table 4 exhibits 5 genes with the relative importance of >0.65 . Figure 4B presents the 5 genes in the out-of-bag importance ranking.

Figure 5

Figure 5 presents the classification effect of the 5-gene signature in the TCGA training set. As shown in Figure 5A, there is an evident difference between the high-risk group of 92 patients and the low-risk group of 92 patients ($P = 9.101663e-11$). Figure 5B displays the ROC curve, with the 1-, 3- and 5-year AUC of 0.77, 0.82, and 0.86, respectively. According to Figure 5C, the survival time of death samples declined significantly with the increase in patient risk score, and more death samples were found in the high-risk group. Besides, the high expression of PDRG1 was identified to be the risk factor based on the variation in the expression of 5 different signature genes with the increase in the risk score. By contrast, the high expression of CISH, LHPP, MGMT, and LCAT was related to the low risk and thereby determined as protective factors.

Figure 6

gh verification in the TCGA test set, the same model and cutoff were adopted as the TCGA training set to decide the 5-gene signature robustness. Figure 6 exhibits the classification effect of the TCGA test set. As shown in Figure 6A, there was a significant statistical difference between the high-risk group of 97 patients and the low-risk group of 86 patients ($P = 0.003221398$). Figure 6B shows the ROC curve, with the 1-, 3- and 5-year AUC of 0.74, 0.69, and 0.59, respectively. Figure 6C shows similar results to those in the TCGA training set. To be specific, the survival time of death samples declined evidently with the increase in the risk score, and more death samples were found in the high-risk group. As expected, PDRG1 was identified as the risk factor; CISH, LHPP, MGMT, and LCAT were the protective factors.

Figure 7

Figures 7A and 8A indicate that the low-risk group had a markedly better prognosis than that in the high-risk group. ROC analysis revealed similar 1-, 3-, 5-year AUC to those in the training set and test set, as presented in Figures 7B and 8B. Additionally, the relationship between the risk score and the expression of 5 genes was also in line with that between the test set and the training set, according to Figures 7C and 8C. To sum up, the 5-gene signature model displays favorable prognosis prediction capacity in both external and internal data.

Figure 8

Figure 8A indicate that the low-risk group had a markedly better prognosis than that in the high-risk group. ROC analysis revealed similar 1-, 3-, 5-year AUC to those in the training set and test set, as presented in Figure 8B. Additionally, the relationship between the risk score and the expression of 5 genes was also in line with that between the test set and the training set, according to Figure 8C. To sum up, the 5-gene signature model displays favorable prognosis prediction capacity in both external and internal data.

Figure 9

In the high and low risk groups of the TCGA training set, pathways with significant enrichment were analyzed using GSEA. Table 5 shows that a total of 39 pathways were obtained, including those closely related to tumor genesis, development, and metastasis, such as adherens junction, cell cycle, endocytosis, and pathways in cancer. The significant enrichment results of these pathways in high-risk samples are shown in Figure 9.

Figure 10

The bioinformatics analysis confirmed that the expression of 5 genes was verified in twenty-five normal tissues and HCC tissues. The results in Figure 10 showed that the mRNA expression of PDRG1 was decreased in HCC tissues and the mRNA expression of CISH, LHPP, MGMT and LCAT were increased in HCC tissues ($p < 0.05$). It was consistent with that analyzed using bioinformatic analysis.

Supplementary Files

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

- [supplement.doc](#)
- [Table1.xlsx](#)
- [Table2.xlsx](#)
- [Table3.xlsx](#)
- [Table4.xlsx](#)
- [Table5.xlsx](#)
- [Table6.xlsx](#)
- [Table7.xlsx](#)