Research Paper

Weighted correlation gene network analysis reveals a new stemness index-related survival model for prognostic prediction in hepatocellular carcinoma

Qiujing Zhang[1,*], Jia Wang[1,2,*], Menghan Liu[3], Qingqing Zhu[1], Qiang Li[4], Chao Xie[1], Congcong Han[1], Yali Wang[1], Min Gao[5], Jie Liu[1]

[1]Department of Oncology, Shandong Cancer Hospital and Institute, Shandong First Medical University and Shandong Academy of Medical Sciences, Jinan 250117, Shandong, China
[2]Department of Oncology, Zibo Maternal and Child Health Hospital, Zibo 255000, Shandong, China
[3]Basic Medicine College, Shandong First Medical University, Taian 271016, Shandong, China
[4]Department of Oncology, Mengyin County Hospital, Linyi 276299, Shandong, China
[5]Department of Radiotherapy, Shandong Cancer Hospital and Institute, Shandong First Medical University and Shandong Academy of Medical Sciences, Jinan 250117, Shandong, China
*Equal contribution

Correspondence to: Jie Liu, Yali Wang, Min Gao; **email:** jieliu1012@163.com, yli_wang@126.com, mingao666@126.com

Keywords: hepatocellular carcinoma, mRNA expression-based stemness index, survival model, weighted correlation network analysis, The Cancer Genome Atlas
Received: February 13, 2020 **Accepted:** May 27, 2020 **Published:** July 8, 2020

ABSTRACT

In this study, we constructed a new survival model using mRNA expression-based stemness index (mRNAsi) for prognostic prediction in hepatocellular carcinoma (HCC). Weighted correlation network analysis (WGCNA) of HCC transcriptome data (374 HCC and 50 normal liver tissue samples) from the TCGA database revealed 7498 differentially expressed genes (DEGs) that clustered into seven gene modules. LASSO regression analysis of the top two gene modules identified *ANGPT2*, *EMCN*, *GLDN*, *USHBP1* and *ZNF532* as the top five mRNAsi-related genes. We constructed our survival model with these five genes and tested its performance using 243 HCC and 202 normal liver samples from the ICGC database. Kaplan-Meier survival curve and receive operating characteristic curve analyses showed that the survival model accurately predicted the prognosis and survival of high- and low-risk HCC patients with high sensitivity and specificity. The expression of these five genes was significantly higher in the HCC tissues from the TCGA, ICGC, and GEO datasets (GSE25097 and GSE14520) than in normal liver tissues. These findings demonstrate that a new survival model derived from five strongly correlating mRNAsi-related genes provides highly accurate prognoses for HCC patients.

INTRODUCTION

The incidence of new cases of liver cancer increased by 2% to 3% annually between 2007 and 2016, according to the cancer statistics reported in 2020 [1]. The mortality rate of liver cancer ranks second among all the cancers worldwide, and the five-year survival rate is only 18% [2]. Hepatocellular carcinoma (HCC) is the most common primary liver cancer that accounts for nearly 90% of all liver cancer patients [3]. The standard therapy for HCC is surgical resection [4]. However, most patients are not amenable for surgical resection therapy because of disease progression and extrahepatic metastasis [5]. Furthermore, the five-year recurrence rate after surgical resection is 70% for HCC, with tumor recurrence reported in nearly two-thirds of the patients within two years after surgery [6]. Moreover, the sensitivity or specificity of current diagnostic imaging and tumor biomarkers such as α-fetoprotein (AFP), Protein induced by vitamin K absence-II (PIVKA-II), and Des-gamma carboxyprothrombin (DCP) is extremely low and cannot detect early stages of HCC accurately [7]. Therefore, there is an urgent need to identify reliable prognostic models for early diagnosis and accurate prognosis of HCC.

Tumorigenesis involves malignant cells acquiring stem cell-like characteristics, including self-renewal and differentiation [8]. Malta et al. used a machine learning algorithm to quantify the stemness index of tumors based on their dedifferentiation characteristics; they also demonstrated that the stemness index correlates with the survival times of HCC patients [9]. The application of New Generation Sequencing (NGS) technology and open access to major databases has resulted in identification of several potential prognostic and early diagnostic biomarkers in HCC, including *Protocadherin 19* (*PCDH19*) gene hypermethylation [10], *Glypican-3* or *GPC3* [11] and *Cytochrome P450 Family 3 Subfamily A Member 4* or *CYP3A4* [12]. Moreover, the overexpression of *YTH N6-Methyladenosine RNA Binding Protein 1* or *YTHDF1* [13] and *DDB1 and CUL4 associated factor 13* or *DCAF13* [14] is associated with poor prognosis of HCC. However, the biological role of key genes that determine the stemness index in HCC has not been reported so far. Furthermore, recent studies have identified several potential prognostic biomarker genes based on differential expression in HCC [15, 16], but their mechanistic role remains to be investigated in greater detail. Weighted correlation network analysis (WGCNA) is a method that identifies gene modules (GMs) containing highly correlating genes with potentially similar biological functions [17]. It has been widely used in the identification of disease characteristics, cancer-related biomarkers and therapeutic target genes of several cancers, such as non-small cell lung cancer [18], rectum adenocarcinoma [19], uveal melanoma [20], bladder cancer [21], and clear cell renal cell carcinoma [22, 23]. Therefore, in this study, we used WGCNA to classify DEGs with closely related stemness index into GMs in HCC. Then, we identified five key genes linked to mRNA expression-based stemness index (mRNAsi) with similar biological characteristics using the least absolute shrinkage and selection operator (LASSO) regression analysis. Furthermore, we developed a survival model using these five genes and evaluated prognostic prediction accuracy of these mRNAsi-related genes in HCC patients. To our knowledge, this is the first time that WGCNA has been used to screen key mRNAsi-related genes and build a survival model to predict prognosis of HCC.

RESULTS

Identifying mRNAsi-related DEGs in HCC

Figure 1 shows the flowchart of data analysis in this study. We analyzed the mRNAsi status of genes expressed in HCC samples as previously reported by Malta et al [9] and found that the mRNAsi were significantly higher in the HCC tumor samples compared to the normal liver tissue samples (p=3.761e−09; Figure 2A). Then, we used the edgeR software to analyze the transcriptome of 374 HCC and 50 normal liver tissue samples from The Cancer Genome Atlas (TCGA) database and identified 7498 DEGs in HCC tumor tissues relative to normal liver tissues (Supplementary Table 1). The volcano plot in Figure 2B depicts the genes that are expressed significantly higher (red) or lower (green) in the HCC tumor tissues relative to normal liver tissues, including 7104 genes with high expression and 394 genes with low expression.

Identification of gene modules among DEGs in HCC using WGCNA

The 7498 DEGs combined with stemness index data were then analyzed using WGCNA with a soft threshold power (β) value of 8 (Figure 3A). Then, we constructed a cluster dendrogram that grouped co-expressing genes into seven gene modules (GMs) that are shown in different color codes, as analyzed using the hybrid dynamic cutting tree algorithm in Figure 3B. Then, we analyzed the module significance (MS) value by evaluating the correlation between each module and the mRNAsi or epigenetically regulated mRNAsi (EREG-mRNAsi). The modules showing a higher correlation value were ranked higher, thereby indicating the higher significance of the module. As shown in Figure 4A, the degree of correlation is indicated by the color depth and the color codes indicate positive (red) or negative (blue) correlation of the modules to the mRNAsi and the EREG-mRNAsi. Among the seven GMs, the purple module (n=116 DEGs) showed the highest correlation of 0.7 with the mRNAsi, followed by the cyan module (n=44 DEGs) with a correlation co-efficient 0.62. Hence, we chose the purple and cyan modules for further analyses. We constructed a scatter diagram to display the genes in these two modules based on the gene significance (GS) and the module membership (MM) of each gene (Figure 4B, 4C), The X axis in the

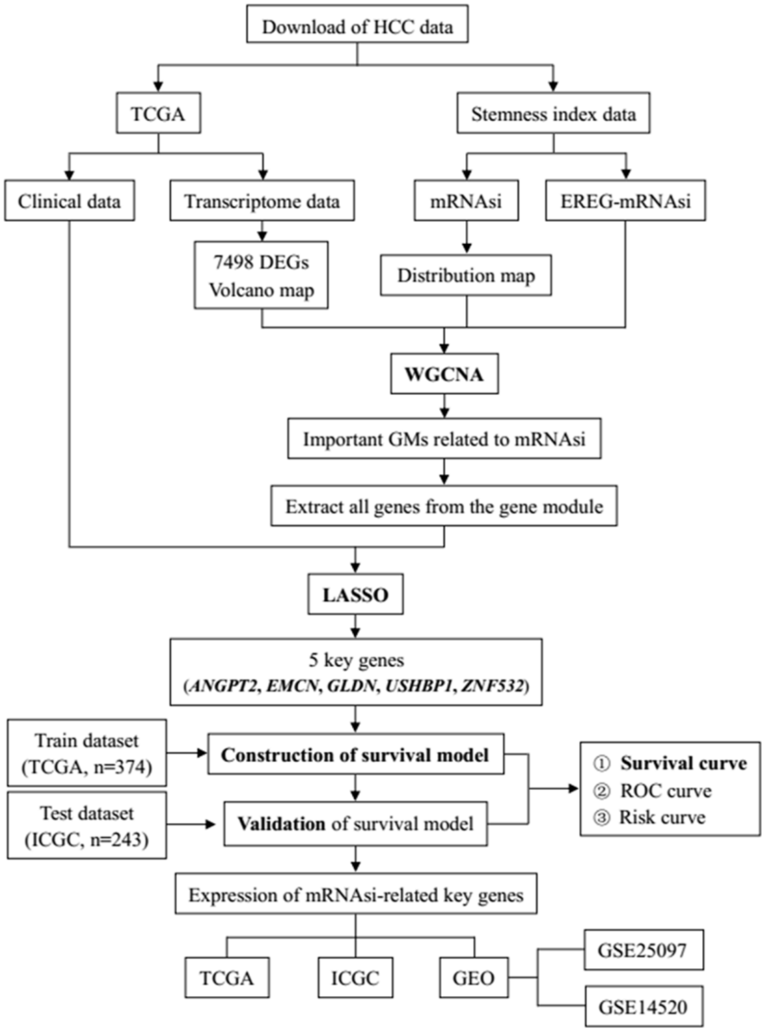


Figure 1. The flowchart of HCC data preparation, processing, analysis and validation.

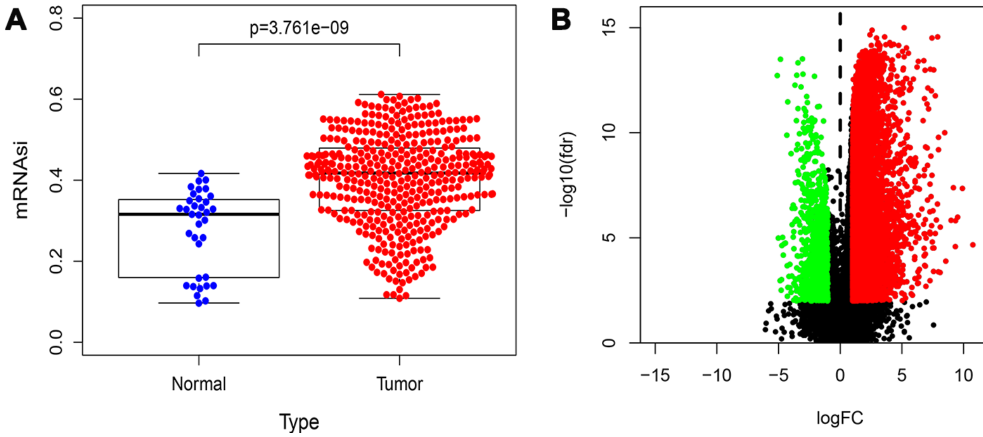


Figure 2. Distribution map of mRNAsi and DEGs in HCC. (**A**) Distribution map shows the mRNAsi of genes in HCC and control samples from the study published by Malta et al. The X axis is sample type (Normal or Tumor) and the Y axis is mRNAsi. (**B**) The volcano plot shows the expression profiles of 7498 DEGs in HCC samples compared to normal liver samples from the TGCA database. The low expressing genes (n=394) are shown in green and the high expressing genes (n=7104) are shown in red. The threshold criteria are FDR/fdr=0.01 and $\log_2$FC=1.

scatter diagram is MM in modules and the Y axis is GS for mRNAsi. The details of the genes in the scatter diagram are shown in Supplementary Table 2A, 2B.

The construction of survival model

We performed univariate Cox regression and LASSO regression analyses of the mRNAsi-related genes from the purple and cyan GMs and identified five key genes, *Angiopoietin 2* (*ANGPT2*), *Endomucin* (*EMCN*), *Gliomedin* (*GLDN*), *USH1 Protein Network Component Harmonin Binding Protein 1* (*USHBP1*) and *Zinc Finger Protein 532* (*ZNF532*), which were then used to construct the survival model (Table 1). The risk score for HCC patients based on this survival model was calculated according to the following formula: (0.154×*ANGPT2*) + (−0.138×*EMCN*) + (0.043×*GLDN*) + (−0.265×*USHBP1*) + (0.121×*ZNF532*). Each gene stands for the gene expression in the gene transcriptome data, and the number represents the model co-efficient of each gene.

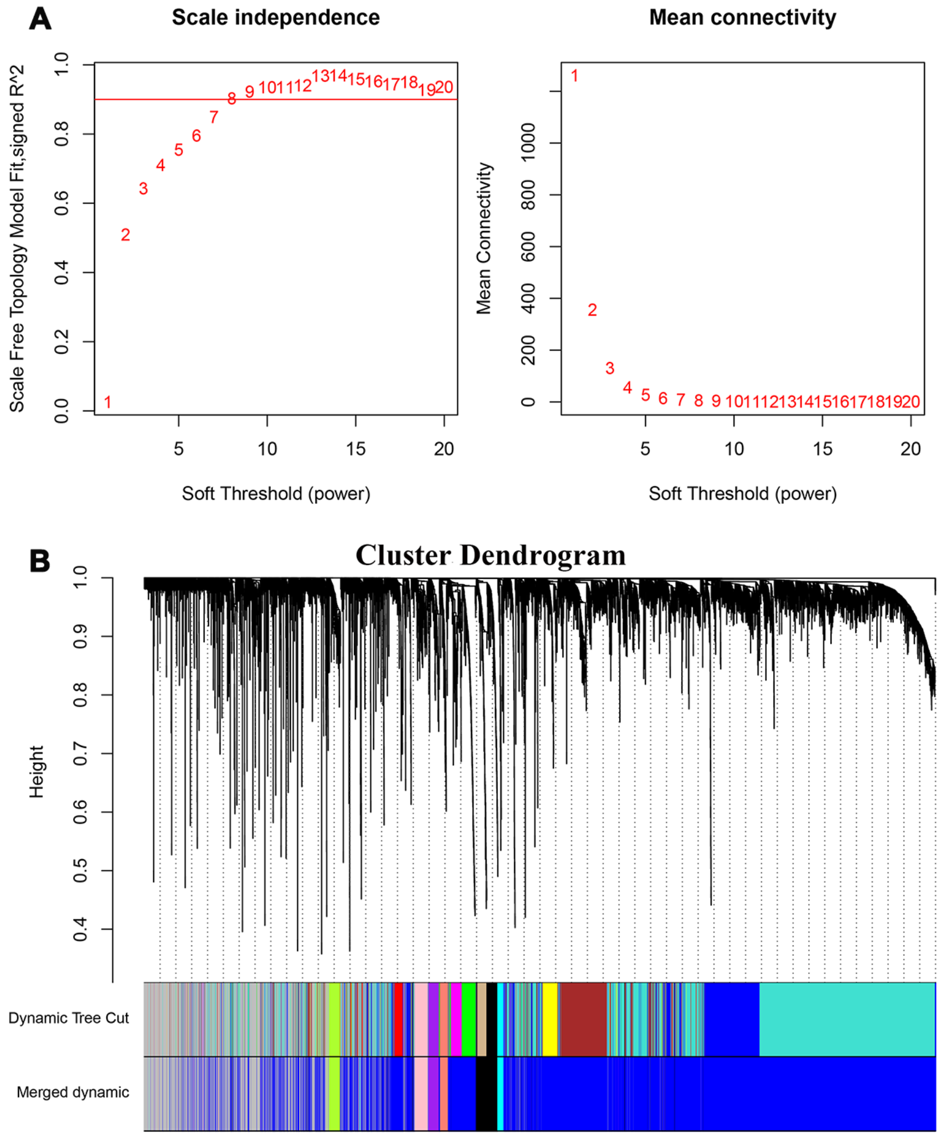


Figure 3. Weighted gene co-expression network analysis of HCC transcriptome. (**A**) The graph shows the scale-free fit index for various soft threshold powers to identify the optimal soft threshold power (β). In the graph on the left, the horizontal axis represents the soft threshold power or β values and the vertical axis represents the scale-free network index (R^2). The scale-free characteristics of the gene network are stronger when the R^2 value is higher. In the right graph, the horizontal axis represents the soft threshold power or β values, the vertical axis represents the means of all the gene adjacency functions in the corresponding gene module. (**B**) Identification of co-expressed gene modules in HCC. The different branches of the cluster dendrogram correspond to different gene modules that are represented by different colors. Each piece of the leaves corresponds to a single gene in the module.

Verification of survival model

Next, we used the HCC tumor sample data from the TCGA database (training dataset) to verify if the prognostic prediction of this new survival model was accurate, specific and sensitive. We generated Kaplan-Meier survival curves, receiver operating characteristic (ROC) curve and the risk curve of high and low risk groups, which were classified based on the risk score formula of this new survival model (Figure 5A–5C). We observed that the difference in survival between high and low risk groups is statistically significant ($p<0.0001$; Figure 5A). Furthermore, ROC curve analysis showed that the survival model composed of *ANGPT2*, *EMCN*, *GLDN*, *USHBP1* and *ZNF532* showed good predictive value for survival when analyzed at 12 months (area under the curve (AUC)=0.713), 36 months (AUC=0.622), and at 60 months (AUC=0.751) for the HCC patients (Figure 5B). The risk curve indicated that the death toll of HCC patients increases with the increase of risk score (Figure 5C). Furthermore, we verified the new survival model in a test dataset of HCC patients from the International Cancer Genome Consortium (ICGC) database (Figure 5D-5F). The survival curve analysis showed statistically significant results in distinguishing high and low risk patient groups of the test dataset (Figure 5D). Moreover, ROC curve analyses showed good predictive value for survival with AUC values of 0.638, 0.625, and 0.593 at 12, 36 and 60 months, respectively (Figure 5E, 5F). The risk curve of the test dataset also showed the same trend with the training dataset from TCGA (Figure 5F). These results showed that the survival model constructed by mRNAsi-related genes, *ANGPT2*, *EMCN*, *GLDN*, *USHBP1* and *ZNF532* accurately predicted the survival of HCC patients.

The expression of the five survival model genes in HCC patient samples

Finally, we analyzed the expression of the five survival model genes using HCC patient data in the TCGA and

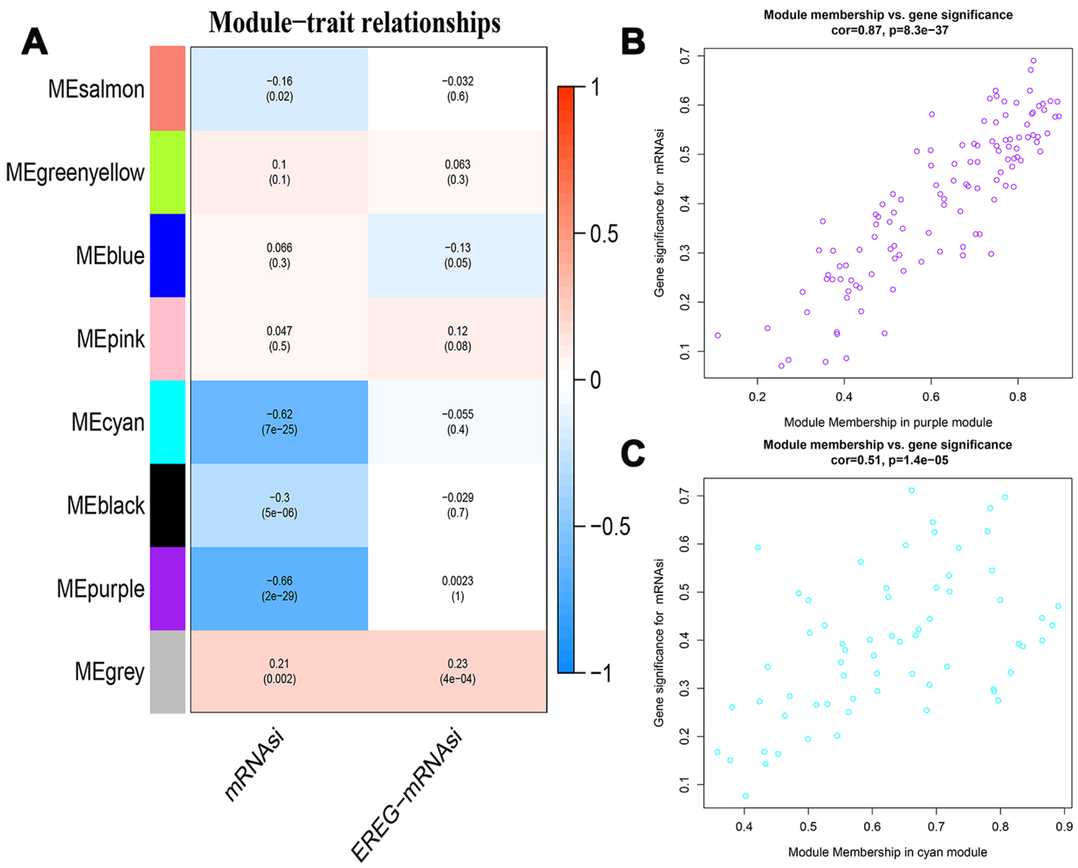


Figure 4. Identification of modules associated with stemness index in HCC. (**A**) The table shows the module-trait relationships of all gene modules, which are represented by different colors. Each cell in the table shows the correlation co-efficient and the p-value between the gene module in rows and the mRNAsi or EREG-mRNAsiin the columns. The degree of correlation is indicated by the color depth; red represents a positive correlation and blue represents a negative correlation. (**B**, **C**) The scatter plots of genes in the top 2 gene modules, purple (**B**, n=116) and cyan (**C**, n=44). The X axis is module membership in modules and the Y axis is gene significance for mRNAsi.

Table 1. The LASSO regression analysis results.

Gene	Co-efficient
ANGPT2	0.154
EMCN	-0.138
GLDN	0.043
USHBP1	-0.265
ZNF532	0.121

LASSO: the least absolute shrinkage and selection operator; ANGPT2: Angiopoietin 2; EMCN: Endomucin; GLDN: Gliomedin; USHBP1: USH1 Protein Network Component Harmonin Binding Protein 1; ZNF532: Zinc Finger Protein 532.

ICGC databases, which were used as the training and test datasets, respectively. The results showed that the expression of all the five genes was significantly higher in the HCC tissues from both the databases compared to the adjacent normal liver tissues (p<0.001; Figure 6A–6J). Moreover, we analyzed the expression of these genes in the GSE25097 and GSE14520 datasets from the Gene Expression Omnibus (GEO) database. The GSE25097 dataset of 557 samples included 268 HCC, 243 adjacent non-tumor liver tissues, 40 cirrhotic and 6

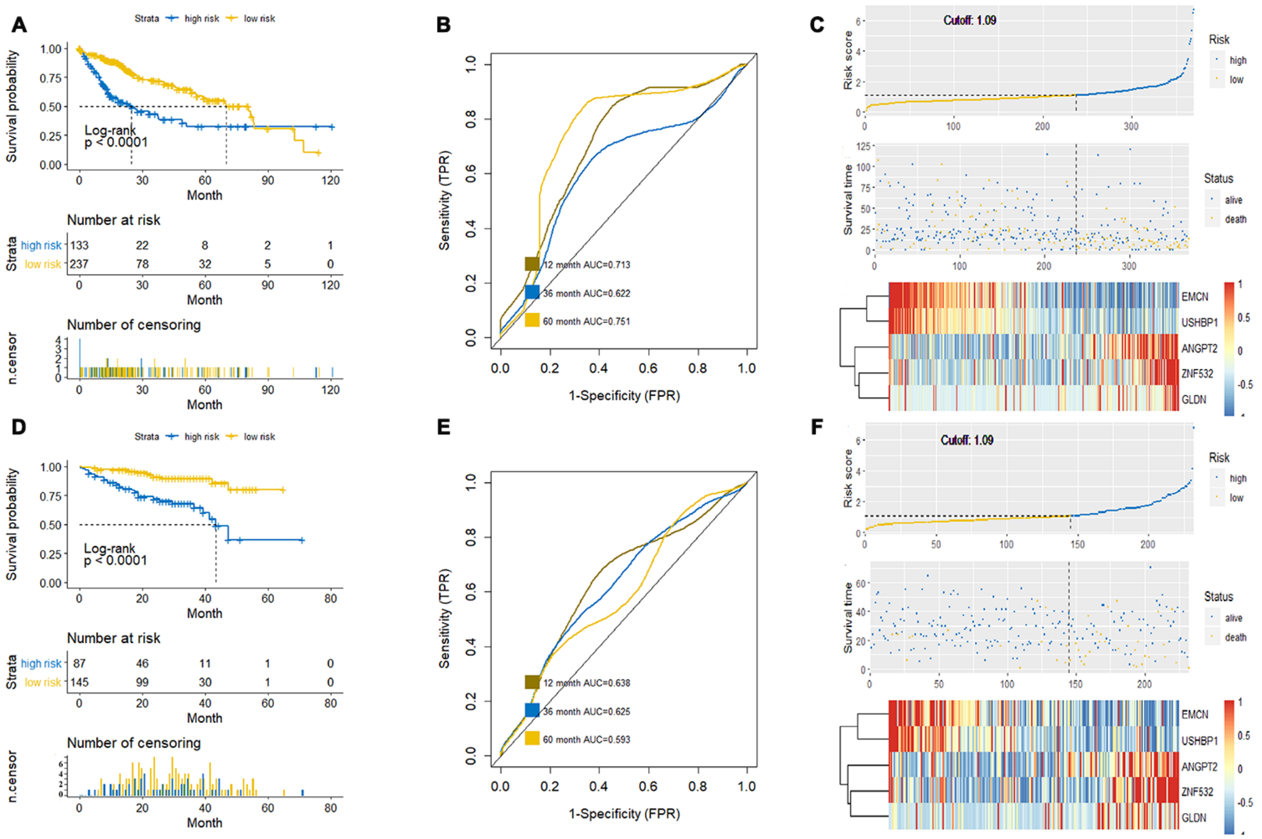


Figure 5. Verification of the prognostic prediction accuracy of the new survival model. (**A**, **C**) The Kaplan-Meier survival curve (**A**), ROC curve (**B**) and Risk curve (**C**) analyses of the high-risk and low-risk HCC patients of the training dataset from the TCGA database based on the new survival model is shown. (**D**–**F**) The Kaplan-Meier survival curve (**D**), ROC curve (**E**) and Risk curve (**F**) of the high-risk and low-risk HCC patients in the test dataset from the ICGC database based on the new survival model is shown. The horizontal axis of the Kaplan-Meier survival curve is survival time (month) and the vertical axis is patient survival, which is used to evaluate the prognostic prediction ability of the new model (P < 0.05 is considered to be statistically significant); the ROC curve evaluates the sensitivity and specificity of the model, in which the Abscissa is the specificity of the model and the ordinate is the sensitivity; moreover, the risk curve shows that the risk of death increases with the increase of the risk score of the new survival model.

healthy liver samples. The expression of *ANGPT2*, *GLDN*, *and ZNF532* was significantly higher in the 268 HCC tumor tissues of the GSE25097 dataset compared to the 243 non-tumor liver tissue samples (p < 0.001; Figure 6K–6M). However, the expression of *EMCN* and *USHBP1* genes was similar in both HCC and adjacent normal liver tissues samples (Supplementary Figure 1). The GSE14520 dataset (225 HCC tumor tissues and 220 liver non-tumor tissues) lacked the data for *GLDN* and *USHBP1* expression, but the expression of the other three genes *ANGPT2* (p=0.001), *EMCN* (p < 0.001), *ZNF532* (p < 0.001) were significantly higher in the HCC tissues compared to the adjacent normal liver tissue samples (Figure 6N–6P). As shown in Figure 7, we analyzed the patient sample data of other cancers (bladder cancer, breast cancer, cervical cancer, colorectal cancer, esophageal cancer, gastric cancer, head and neck cancer, kidney cancer, leukemia, liver cancer, lung cancer, lymphoma, etc.) in the Oncomine database and found that the expression of *ANGPT2*, *EMCN*, *GLDN*, *USHBP1*, *ZNF532* gene was significantly upregulated in most tumor tissue samples compared to the adjacent normal tissue samples.

DISCUSSION

HCC is a highly malignant cancer with high morbidity and mortality rates [22]. Currently, there is an urgent need to identify new molecular biomarkers that can improve early diagnosis as well as accurate prognosis prediction that can guide appropriate treatment to improve survival rates. Although several prognostic and diagnostic biomarkers have been reported for HCC, their reliability and efficacy remain to be verified for clinical applications. Moreover, the previous prognostic models ignore the correlation between genes. A recent study by Malta et al demonstrated the correlation between mRNAsi-related genes and the survival and

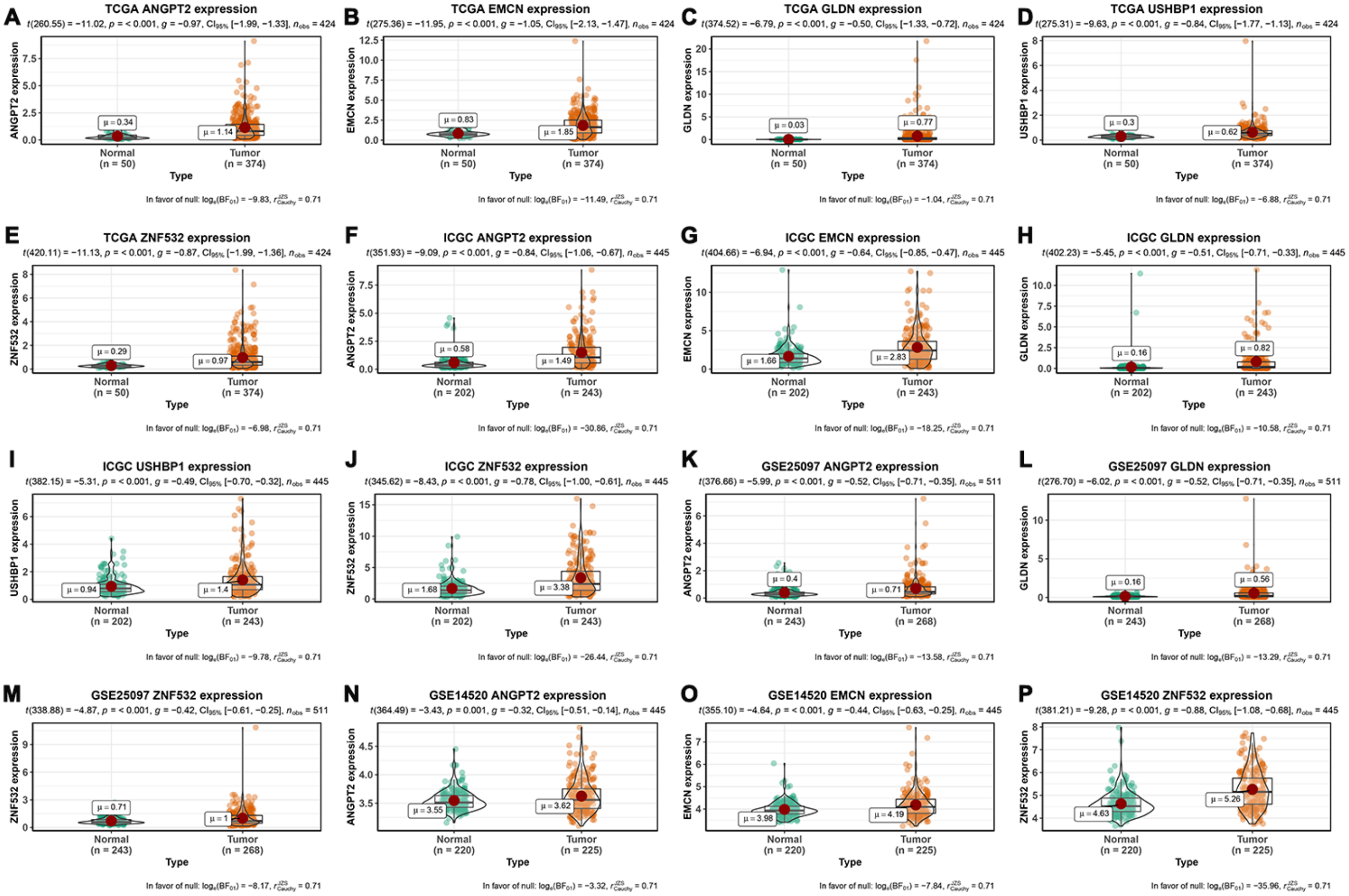
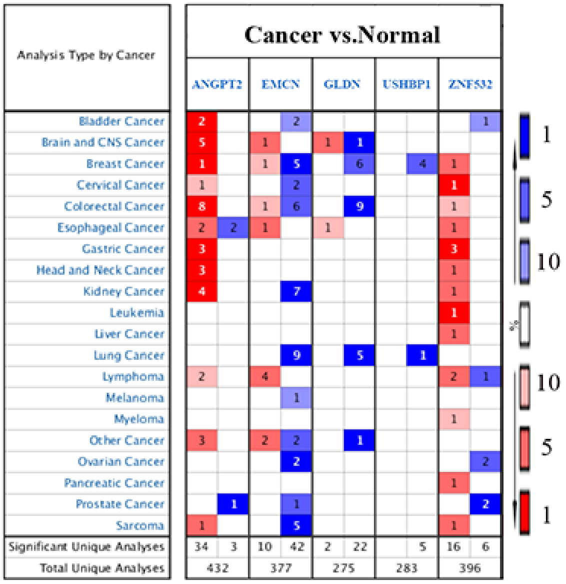


Figure 6. Expression of mRNAsi-related key genes in HCC and normal liver tissues. (**A**–**E**) The expression of *ANGPT2* (A), *EMCN* (B), *GLDN* (**C**), *USHBP1* (**D**) and *ZNF532* (**E**) genes in 374 HCC and 50 non-cancer tissues from the TCGA database. (**F**–**J**) The expression of *ANGPT2* (**F**), *EMCN* (**G**), *GLDN* (**H**), *USHBP1* (**I**) and *ZNF532* (**J**) genes in 243 HCC and 202 normal liver tissues from the ICGC database. (**K**–**M**) The expression of *ANGPT2* (**K**), *GLDN* (**L**) and *ZNF532* (**M**) genes in 268 HCC and 243 normal liver tissue samples from the GSE25097 dataset. (**N**–**P**) The expression of *ANGPT2* (**N**), *EMCN* (**O**) and *ZNF532* (**P**) genes in 225 HCC and 220 normal liver tissue samples from the GSE14520 dataset. The X axis is sample type (Normal or Tumor) and the Y axis is gene expression.

prognosis of cancer patients in all TCGA tumors [9]. However, mRNAsi-related molecular markers have not been reported for HCC. Therefore, we performed WGCNA analysis of the microarray data of HCC patients and identified gene modules (GMs) with mRNAsi-related genes. Besides, LASSO regression analysis of the genes in the top 2 GMs identified five key genes, which were then used to construct the new survival model of HCC. Our study suggests that these 5 genes are potential prognostic and therapeutic targets for HCC. However, future investigations are necessary to demonstrate the clinical significance of these genes.

WGCNA is an algorithm that clusters genes with similar patterns of expression into GMs [17]. This allows establishing the correlation between GMs and the characteristics of patient samples in different stages of progression. Thus, WGCNA has been used extensively to study the prognostic potential of several genes that correlate with patient prognosis and survival [24].

In this study, we first identified 7498 HCC-related DEGs and used WGCNA to classify them into seven gene modules based on their correlation with the mRNAsi. Furthermore, genes in the purple module and cyan module showed the highest correlation with the mRNAsi. We then identified 5 key mRNAsi-related genes from these two models using LASSO regression analysis and then constructed a survival model with these five genes to predict the prognosis and survival of HCC patients. Then, we successfully verified that the survival model accurately predicts the prognosis of HCC patients by using patient's data from the TCGA and ICGC databases as the training and test groups, respectively. We also found that the expression of these 5 genes, namely, *ANGPT2*, *EMCN*, *GLDN*, *USHBP1* and *ZNF532*, was significantly upregulated in HCC tumor tissues compared to the adjacent normal liver tissues in the TCGA and ICGC datasets. We also verified the survival model using GSE25097 and GSE14520 datasets. The expression of *EMCN* and *USHBP1* was not statistically significant in the HCC patients compared to the controls from the GSE25097 dataset, but the expression of *ANGPT2*, *GLDN* and *ZNF532* was significantly higher than the controls. The reason for this discrepancy is not known and needs to be evaluated in future studies. On the other hand, the GSE14520 dataset lacked expression data for the *GLDN* and *USHBP1* genes. Nevertheless, the expression of *ANGPT2*, *EMCN* and *ZNF532* genes was significantly higher in the HCC tumor samples compared to the normal liver tissue samples. Furthermore, analysis of the expression profiles of these five genes in the Oncomine database demonstrated differential expression in several cancers. However, these 5 genes were not differentially expressed in the liver cancer samples of the Oncomine dataset. One plausible reason for this anomaly is that the liver cancer samples in the large Oncomine database may belong to different pathological types of liver cancer and therefore represents a heterogeneous dataset. Another plausible reason is that the threshold setting we used may not be appropriate for screening samples in the Oncomine database. Overall, our data suggests that the survival model constructed using the *ANGPT2*, *EMCN*, *GLDN*, *USHBP1* and *ZNF532* genes shows good predictive value and demonstrates potential for clinical use to evaluate the prognosis of patients with HCC.



Analysis Type by Cancer	Cancer vs.Normal									
	ANGPT2		EMCN		GLDN		USHBP1		ZNF532	
Bladder Cancer	2			2						1
Brain and CNS Cancer	5		1		1	1				
Breast Cancer	1		1	5		6		4	1	
Cervical Cancer	1			2					1	
Colorectal Cancer	8		1	6		9			1	
Esophageal Cancer	2	2	1		1				1	
Gastric Cancer	3								3	
Head and Neck Cancer	3								1	
Kidney Cancer	4			7					1	
Leukemia									1	
Liver Cancer									1	
Lung Cancer				9		5		1		
Lymphoma	2		4						2	1
Melanoma				1						
Myeloma									1	
Other Cancer	3		2	2		1				
Ovarian Cancer				2						2
Pancreatic Cancer									1	
Prostate Cancer		1		1						2
Sarcoma	1			5					1	
Significant Unique Analyses	34	3	10	42	2	22		5	16	6
Total Unique Analyses	432		377		275		283		396	

Figure 7. The expression of five survival model genes in various cancers in the Oncomine database. The expression of *ANGPT2*, *EMCN*, *GLDN*, *USHBP1*, and *ZNF532* genes in the tumor and control samples of different cancers (bladder cancer, breast cancer, cervical cancer, colorectal cancer, esophageal cancer, gastric cancer, head and neck cancer, kidney cancer, leukemia, liver cancer, lung cancer, lymphoma and other cancers) in the Oncomine database are shown.

An integrated analysis of genomic and expression profiling found that the high expression of *nucleophosmin* (*NPM1*) in HCC was associated with the prognosis of patients [25]. It is plausible that gene copy number variations may also influence the prognosis and survival of HCC patients. However, gene copy number variations of these 5 survival model genes need to be evaluated in the HCC patients.

As far as we know, except for *ANGPT2*, the remaining four genes have not been previously identified as biomarkers for HCC patients. *ANGPT2* encodes for the angiopoietin-2 protein, which competitively inhibits angiopoietin-1 by specifically binding to the angiopoietin receptor, and thereby modulates the growth and progression of several cancers [26–28]. A prospective study shows that angiogenesis-related genes, including *ANGPT2*, are independent factors that correlate with the tumor progression and prognosis of liver cancer patients [29]. Chen et al showed that serum *ANGPT2* levels represent a potential serum prognostic biomarker in liver cancer patients [30]. *ANGPT2* is an essential factor for the formation of vessels that encapsulate tumor clusters (VETC), which is a unique vascular pattern that is associated with HCC progression [31].

EMCN encodes a type I O-glycosylated sialic acid-rich glycoprotein called endomucin I, which is specifically expressed on the endothelial cells of veins and capillaries [32]. Endomucin I is a novel therapeutic target for angiogenesis-related diseases because it inhibits vascular endothelial growth factor (VEGF)-induced migration, growth and morphogenesis of endothelial cells by modulating vascular endothelial growth factor receptor 2 (VEGFR2) endocytosis and activity [33, 34]. Moreover, a study by Holmfeldt et al. identified *EMCN* as one of the 17 genes that regulates repopulation of murine hematopoietic stem cells [35].

GLDN is located on chromosome 15 and its protein product promotes the adhesion of heterogeneous cells by selectively binding to the extracellular protein complexes [36]. *GLDN* is a potential prognostic biomarker that predicts the overall survival (OS) of patients with colorectal cancer [37] and melanoma patients that may benefit from immunotherapy [38].

USHBP1 gene, also known as *MCC2* gene, is expressed in the heart, liver, small intestine, lung and other tissues [39]. A Genome-Wide Association Study (GWAS) study by Hass et al showed that *USHBP1* was involved in schizophrenia by regulating synaptic tissue development [40]. *ZNF532* encodes a protein that prominently interacts with the BRD4-NUT interacting fusion oncoprotein in the chromatin of NUT midline carcinoma cells and drives oncogenesis by propagating the oncogenic chromatin complex [41, 42].

WGCNA has recently been used to identify new gene targets that regulate gene progression for HCC prognosis and therapy [43, 44, 24]. Although mRNAsi has been shown to be related to prognosis and survival of HCC patients [9], the mRNAsi-related prognostic markers have not been studied. We used WGCNA algorithm to screen HCC-related mRNAsi genes for the first time and successfully constructed and verified a new survival model that can predict the prognosis of HCC patients. This prognostic model needs to be further confirmed using prospective multicenter randomized controlled trials. Moreover, the mechanism details of the five genes that have been used to develop this survival model needs to be further explored in HCC.

In conclusion, our study used WGCNA and LASSO regression analyses to identify five mRNAsi-related genes, namely, *ANGPT2*, *EMCN*, *GLDN*, *USHBP1* and *ZNF532*. We then constructed a survival model with these five genes and successfully verified their accuracy, sensitivity and specificity to predict the prognosis of HCC patients in TGCA, ICGC and GEO databases. We postulate that these five survival model genes are potential therapeutic targets of HCC.

MATERIALS AND METHODS

HCC data download and processing

We downloaded the transcriptome and clinical data of 374 HCC and 50 paracancerous patient samples from the TCGA [45] database (https://portal.gdc.cancer.gov) using "TCGA-LIHC" (TCGA-Liver hepatocellular carcinoma) as the project id, "liver and intrahepatic bile ducts" as the primary site, and "HTSeq-FPKM" as the workflow type on December 18, 2019. The sample identifiers of the TCGA data are shown in Supplementary Table 3. The stemness index data for HCC, including their mRNAsi and EREG-mRNAsi was downloaded from the study published by Malta et al [9] and is listed in Supplementary Table 4. After downloading the mRNAsi data, we analyzed the distribution of the mRNAsi in the normal and HCC samples. Then, we used the edgeR software package version: 3.26.5 [46] to clean and filter the downloaded transcriptome data of HCC. Finally, the DEGs between the normal and HCC samples was obtained using the following threshold parameters: false discovery rate (FDR) = 0.01 and $\log_2$ fold change in gene expression (FC) = 1.

Gene module construction using WGCNA

WGCNA [17] was used to perform co-expression scale-free network analysis and identify gene modules containing strongly correlating genes. We imported the DEGs into the WGCNA software R package version: 1.68 [47] and determined that the soft power value was 0.8 based on the scale-free topology fit model index (R^2), which was achieved along with a mean connectivity value below 100. Then, the difference between a pair of genes was calculated using the topological overlap method to construct the cluster dendrogram. We then re-analyzed the module eigengenes (MEs) according to the standard of the hybrid dynamic cutting tree and merged two or more modules that were close to each other into a new module.

We used the gene significance (GS) index to determine the strength of the correlation between every single gene and the mRNAsi or EREG-mRNAsi. We also used the module membership (MM) value to measure the importance of genes in the corresponding modules. The method to obtain GS is use the modeEigengenes function in WGCNA software package to calculate the characteristic genes of the module firstly, then take the correlation value between the expression of DEGs and the module eigengenes (MEs) as the GS. In addition, MM is calculated by taking the correlation between the expression of DEGs and the mRNAsi or EREG-mRNAsi of the corresponding samples downloaded so that GS and MM be accurately assigned to each gene in the module. The module significance (MS) of each module was determined by calculating the GS between sample traits (mRNAsi or EREG-mRNAsi) and gene expression. Subsequently, P < 0.05 was used as the statistically significant standard to screen important GMs. Finally, scatter diagram was constructed based on the correlation between GS and MM in the top 2 GMs to identify the key genes.

Survival model construction

We performed univariate Cox hazard analysis [48] with P < 0.05 as a threshold parameter for all the genes in the top 2 GMs. Then, the lambda value with the minimum average error obtained from the cross-validation method was fitted into the LASSO regression analysis [49] to obtain key genes related to mRNAsi. These key genes were then used to construct the survival model of HCC. We determined the risk scores based on the expression of key genes in the 374 HCC tumor samples downloaded from TCGA database (training dataset), and grouped all the samples into high- and low-risk groups based on the scores. Then, we used the clinical information of these HCC patients in the high- and low-risk groups to generate the Kaplan-Meier survival curve and the ROC curves to determine the survival parameters as well as the AUC value, respectively, in order to determine the prognostic performance of the survival model.

Verification of survival model

To independently verify the reliability of the survival model, we downloaded the transcriptome data and clinical information of 202 normal paracancerous samples and 243 HCC samples on November 27, 2019 from the LIRI-JP (https://dcc.icgc.org/releases/current/Projects/LIRI-JP) project in the ICGC database version: release_28 (https://icgc.org/). The sample identifiers of ICGC data are shown in Supplementary Table 5. The 243 HCC samples were selected as the test dataset and were analyzed similar to the training dataset as described above.

Expression of the five survival model genes in different datasets

We used the cowplot (version: 1.0.0) and Ggstatsplot (version: 0.1.3) software packages to determine the expression of key mRNAsi-related genes that are included in the survival model in two randomly selected HCC patient datasets, GSE25097 and GSE14520 in the GEO (http://www.ncbi.nlm.nih.gov/geo/) database [50]. There were 268 and 225 HCC samples, 243 and 220 normal liver tissue samples in GSE25097 dataset and GSE14520 dataset, respectively. Furthermore, we retrieved the expression of these five mRNAsi-related genes in several cancer types from the Oncomine (http://www.oncomine.org) database [51] on December 24, 2019. We used "Cancer vs. Normal Analysis" as the analysis type and "p-value = 1E-4, FC = 2, gene rank = top 10%, and data type = all" as the threshold parameters.

Abbreviations

HCC: hepatocellular carcinoma; DEGs: differentially expressed genes; AFP: α-fetoprotein; PIVKA-II: Protein induced by vitamin K absence-II; DCP: Des-gamma carboxyprothrombin; NGS: new generation sequencing; *PCDH19*: *Protocadherin 19*; *GPC3*:*Glypican-3*; *CYP3A4*: *Cytochrome P450 Family 3 Subfamily A Member 4*; *YTHDF1*: *YTH N6-Methyladenosine RNA Binding Protein 1*; *DCAF13*: *DDB1 and CUL4 associated factor 13*; WGCNA: Weighted correlation network analysis; mRNAsi: mRNA expression-based stemness index; LASSO: the least absolute shrinkage and selection operator; TCGA: The Cancer Genome Atlas; GMs: gene modules; MS: module significance; EREG-mRNAsi: epigenetically regulated mRNAsi; GS: gene significance; MM: module membership; *ANGPT2*: *Angiopoietin 2*; *EMCN*: *Endomucin*; *GLDN*: *Gliomedin*; *USHBP1*: *USH1 Protein Network Component Harmonin Binding Protein 1*; *ZNF532*: *Zinc Finger Protein 532*; ROC: receiver operating characteristic; AUC: area under the curve; ICGC: International Cancer Genome Consortium; GEO: the Gene Expression Omnibus; *NPM1*: *nucleophosmin*; VETC: vessels that encapsulated tumor clusters; VEGF: vascular endothelial growth factor; VEGFR2: vascular endothelial growth factor receptor 2; OS: overall survival; GWAS: Genome-Wide Association Study; LIHC: liver hepatocellular carcinoma; FDR: false discovery rate; FC: fold change; MEs: module eigengenes.

AUTHOR CONTRIBUTIONS

JW and JL designed the research; QJZ and MHL performed the data analysis and collation; QJZ and

QQZ drafted the manuscript; QL, CX, CCH generated the figures and tables; JW, YLW and MG analyzed of the results; YLW, MG and JL reviewed and revised the manuscript. All the authors approved the final manuscript.

CONFLICTS OF INTEREST

The authors declare that there are no potential conflicts of interest.

FUNDING

This study was funded by a grant from the Key Research and Development Program of Shandong Province, China (Grant No. GG201710060039).

REFERENCES

1. Siegel RL, Miller KD, Jemal A. Cancer statistics, 2020. CA Cancer J Clin. 2020; 70:7–30. https://doi.org/10.3322/caac.21590 PMID:31912902

2. Bray F, Ferlay J, Soerjomataram I, Siegel RL, Torre LA, Jemal A. Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA Cancer J Clin. 2018; 68:394–424. https://doi.org/10.3322/caac.21492 PMID:30207593

3. Llovet JM, Zucman-Rossi J, Pikarsky E, Sangro B, Schwartz M, Sherman M, Gores G. Hepatocellular carcinoma. Nat Rev Dis Primers. 2016; 2:16018. https://doi.org/10.1038/nrdp.2016.18 PMID:27158749

4. Maluccio M, Covey A. Recent progress in understanding, diagnosing, and treating hepatocellular carcinoma. CA Cancer J Clin. 2012; 62:394–99. https://doi.org/10.3322/caac.21161 PMID:23070690

5. European Association for the Study of the Liver. Electronic address: easloffice@easloffice.eu; European Association for the Study of the Liver. EASL Clinical Practice Guidelines: Management of hepatocellular carcinoma. J Hepatol. 2018; 69:182–236. https://doi.org/10.1016/j.jhep.2018.03.019 PMID:29628281

6. Kulik L, El-Serag HB. Epidemiology and management of hepatocellular carcinoma. Gastroenterology. 2019; 156:477–91.e1. https://doi.org/10.1053/j.gastro.2018.08.065 PMID:30367835

7. Forner A, Reig M, Bruix J. Hepatocellular carcinoma. Lancet. 2018; 391:1301–14. https://doi.org/10.1016/S0140-6736(18)30010-2 PMID:29307467

8. Visvader JE, Lindeman GJ. Cancer stem cells: current status and evolving complexities. Cell Stem Cell. 2012; 10:717–28. https://doi.org/10.1016/j.stem.2012.05.007 PMID:22704512

9. Malta TM, Sokolov A, Gentles AJ, Burzykowski T, Poisson L, Weinstein JN, Kamińska B, Huelsken J, Omberg L, Gevaert O, Colaprico A, Czerwińska P, Mazurek S, et al, and Cancer Genome Atlas Research Network. Machine learning identifies stemness features associated with oncogenic dedifferentiation. Cell. 2018; 173:338–54.e15. https://doi.org/10.1016/j.cell.2018.03.034 PMID:29625051

10. Zhang T, Guan G, Chen T, Jin J, Zhang L, Yao M, Qi X, Zou J, Chen J, Lu F, Chen X. Methylation of PCDH19 predicts poor prognosis of hepatocellular carcinoma. Asia Pac J Clin Oncol. 2018; 14:e352–58. https://doi.org/10.1111/ajco.12982 PMID:29749051

11. Zhou F, Shang W, Yu X, Tian J. Glypican-3: a promising biomarker for hepatocellular carcinoma diagnosis and treatment. Med Res Rev. 2018; 38:741–67. https://doi.org/10.1002/med.21455 PMID:28621802

12. Ashida R, Okamura Y, Ohshima K, Kakuda Y, Uesaka K, Sugiura T, Ito T, Yamamoto Y, Sugino T, Urakami K, Kusuhara M, Yamaguchi K. CYP3A4 gene is a novel biomarker for predicting a poor prognosis in hepatocellular carcinoma. Cancer Genomics Proteomics. 2017; 14:445–53. https://doi.org/10.21873/cgp.20054 PMID:29109094

13. Zhao X, Chen Y, Mao Q, Jiang X, Jiang W, Chen J, Xu W, Zhong L, Sun X. Overexpression of YTHDF1 is associated with poor prognosis in patients with hepatocellular carcinoma. Cancer Biomark. 2018; 21:859–68. https://doi.org/10.3233/CBM-170791 PMID:29439311

14. Cao J, Hou P, Chen J, Wang P, Wang W, Liu W, Liu C, He X. The overexpression and prognostic role of DCAF13 in hepatocellular carcinoma. Tumour Biol. 2017; 39:1010428317705753. https://doi.org/10.1177/1010428317705753 PMID:28631558

15. Deng Z, Wang J, Xu B, Jin Z, Wu G, Zeng J, Peng M, Guo Y, Wen Z. Mining TCGA database for tumor microenvironment-related genes of prognostic value in hepatocellular carcinoma. Biomed Res Int. 2019; 2019:2408348.

https://doi.org/10.1155/2019/2408348 PMID:31828095

16. Bai Y, Long J, Liu Z, Lin J, Huang H, Wang D, Yang X, Miao F, Mao Y, Sang X, Zhao H. Comprehensive analysis of a ceRNA network reveals potential prognostic cytoplasmic lncRNAs involved in HCC progression. J Cell Physiol. 2019; 234:18837–48. https://doi.org/10.1002/jcp.28522 PMID:30916406

17. Langfelder P, Horvath S. WGCNA: an R package for weighted correlation network analysis. BMC Bioinformatics. 2008; 9:559. https://doi.org/10.1186/1471-2105-9-559 PMID:19114008

18. Niemira M, Collin F, Szalkowska A, Bielska A, Chwialkowska K, Reszec J, Niklinski J, Kwasniewski M, Kretowski A. Molecular signature of subtypes of non-small-cell lung cancer by large-scale transcriptional profiling: identification of key modules and genes by weighted gene co-expression network analysis (WGCNA). Cancers (Basel). 2019; 12:37. https://doi.org/10.3390/cancers12010037 PMID:31877723

19. Wu W, Yang Z, Long F, Luo L, Deng Q, Wu J, Ouyang S, Tang D. COL1A1 and MZB1 as the hub genes influenced the proliferation, invasion, migration and apoptosis of rectum adenocarcinoma cells by weighted correlation network analysis. Bioorg Chem. 2020; 95:103457. https://doi.org/10.1016/j.bioorg.2019.103457 PMID:31901757

20. Wan Q, Tang J, Han Y, Wang D. Co-expression modules construction by WGCNA and identify potential prognostic markers of uveal melanoma. Exp Eye Res. 2018; 166:13–20. https://doi.org/10.1016/j.exer.2017.10.007 PMID:29031853

21. Pan S, Zhan Y, Chen X, Wu B, Liu B. Identification of biomarkers for controlling cancer stem cell characteristics in bladder cancer by network analysis of transcriptome data stemness indices. Front Oncol. 2019; 9:613. https://doi.org/10.3389/fonc.2019.00613 PMID:31334127

22. Luo Y, Shen D, Chen L, Wang G, Liu X, Qian K, Xiao Y, Wang X, Ju L. Identification of 9 key genes and small molecule drugs in clear cell renal cell carcinoma. Aging (Albany NY). 2019; 11:6029–52. https://doi.org/10.18632/aging.102161 PMID:31422942

23. Xiao H, Chen P, Zeng G, Xu D, Wang X, Zhang X. Three novel hub genes and their clinical significance in clear cell renal cell carcinoma. J Cancer. 2019; 10:6779–91. https://doi.org/10.7150/jca.35223 PMID:31839812

24. Li B, Pu K, Wu X. Identifying novel biomarkers in hepatocellular carcinoma by weighted gene co-expression network analysis. J Cell Biochem. 2019. [Epub ahead of print]. https://doi.org/10.1002/jcb.28420 PMID:30746803

25. Zhou C, Zhang W, Chen W, Yin Y, Atyah M, Liu S, Guo L, Shi Y, Ye Q, Dong Q, Ren N. Integrated analysis of copy number variations and gene expression profiling in hepatocellular carcinoma. Sci Rep. 2017; 7:10570. https://doi.org/10.1038/s41598-017-11029-y PMID:28874807

26. Martinelli S, Kanduri M, Maffei R, Fiorcari S, Bulgarelli J, Marasca R, Rosenquist R. ANGPT2 promoter methylation is strongly associated with gene expression and prognosis in chronic lymphocytic leukemia. Epigenetics. 2013; 8:720–29. https://doi.org/10.4161/epi.24947 PMID:23803577

27. Chen Z, Zhu S, Hong J, Soutto M, Peng D, Belkhiri A, Xu Z, El-Rifai W. Gastric tumour-derived ANGPT2 regulation by DARPP-32 promotes angiogenesis. Gut. 2016; 65:925–34. https://doi.org/10.1136/gutjnl-2014-308416 PMID:25779598

28. Lin CY, Cho CF, Bai ST, Liu JP, Kuo TT, Wang LJ, Lin YS, Lin CC, Lai LC, Lu TP, Hsieh CY, Chu CN, Cheng DC, Sher YP. ADAM9 promotes lung cancer progression through vascular remodeling by VEGFA, ANGPT2, and PLAT. Sci Rep. 2017; 7:15108. https://doi.org/10.1038/s41598-017-15159-1 PMID:29118335

29. Villa E, Critelli R, Lei B, Marzocchi G, Cammà C, Giannelli G, Pontisso P, Cabibbo G, Enea M, Colopi S, Caporali C, Pollicino T, Milosa F, et al. Neoangiogenesis-related genes are hallmarks of fast-growing hepatocellular carcinomas and worst survival. Results from a prospective study. Gut. 2016; 65:861–69. https://doi.org/10.1136/gutjnl-2014-308483 PMID:25666192

30. Chen Y, Wu Y, Zhang X, Zeng H, Liu Y, Wu Q, Chen Y, Zhu G, Pan Q, Jin L, Guo L, Sun F. Angiopoietin-2 (Ang-2) is a useful serum tumor marker for liver cancer in the chinese population. Clin Chim Acta. 2018; 478:18–27. https://doi.org/10.1016/j.cca.2017.12.017 PMID:29253494

31. Fang JH, Zhou HC, Zhang C, Shang LR, Zhang L, Xu J, Zheng L, Yuan Y, Guo RP, Jia WH, Yun JP, Chen MS, Zhang Y, Zhuang SM. A novel vascular pattern

promotes metastasis of hepatocellular carcinoma in an epithelial-mesenchymal transition-independent manner. Hepatology. 2015; 62:452–65.
https://doi.org/10.1002/hep.27760 PMID:25711742

32. Morgan SM, Samulowitz U, Darley L, Simmons DL, Vestweber D. Biochemical characterization and molecular cloning of a novel endothelial-specific sialomucin. Blood. 1999; 93:165–75.
PMID:9864158

33. Park-Windhol C, Ng YS, Yang J, Primo V, Saint-Geniez M, D'Amore PA. Endomucin inhibits VEGF-induced endothelial cell migration, growth, and morphogenesis by modulating VEGFR2 signaling. Sci Rep. 2017; 7:17138.
https://doi.org/10.1038/s41598-017-16852-x
PMID:29215001

34. LeBlanc ME, Saez-Torres KL, Cano I, Hu Z, Saint-Geniez M, Ng YS, D'Amore PA. Glycocalyx regulation of vascular endothelial growth factor receptor 2 activity. FASEB J. 2019; 33:9362–73.
https://doi.org/10.1096/fj.201900011R
PMID:31141406

35. Holmfeldt P, Ganuza M, Marathe H, He B, Hall T, Kang G, Moen J, Pardieck J, Saulsberry AC, Cico A, Gaut L, McGoldrick D, Finkelstein D, et al. Functional screen identifies regulators of murine hematopoietic stem cell repopulation. J Exp Med. 2016; 213:433–49.
https://doi.org/10.1084/jem.20150806
PMID:26880577

36. Shrivastava A, Rhodes RG, Pochiraju S, Nakane D, McBride MJ. Flavobacterium johnsoniae RemA is a mobile cell surface lectin involved in gliding. J Bacteriol. 2012; 194:3678–88.
https://doi.org/10.1128/JB.00588-12
PMID:22582276

37. Chen L, Lu D, Sun K, Xu Y, Hu P, Li X, Xu F. Identification of biomarkers associated with diagnosis and prognosis of colorectal cancer patients based on integrated bioinformatics analysis. Gene. 2019; 692:119–25.
https://doi.org/10.1016/j.gene.2019.01.001
PMID:30654001

38. Chen H, Yang M, Wang Q, Song F, Li X, Chen K. The new identified biomarkers determine sensitivity to immune check-point blockade therapies in melanoma. Oncoimmunology. 2019; 8:1608132.
https://doi.org/10.1080/2162402X.2019.1608132
PMID:31413919

39. Ishikawa S, Kobayashi I, Hamada J, Tada M, Hirai A, Furuuchi K, Takahashi Y, Ba Y, Moriuchi T. Interaction of MCC2, a novel homologue of MCC tumor suppressor, with PDZ-domain protein AIE-75. Gene. 2001; 267:101–10.
https://doi.org/10.1016/s0378-1119(01)00378-x
PMID:11311560

40. Hass J, Walton E, Kirsten H, Liu J, Priebe L, Wolf C, Karbalai N, Gollub R, White T, Roessner V, Müller KU, Paus T, Smolka MN, et al, and IMAGEN Consortium. A genome-wide association study suggests novel loci associated with a schizophrenia-related brain-based phenotype. PLoS One. 2013; 8:e64872.
https://doi.org/10.1371/journal.pone.0064872
PMID:23805179

41. Alekseyenko AA, Walsh EM, Zee BM, Pakozdi T, Hsi P, Lemieux ME, Dal Cin P, Ince TA, Kharchenko PV, Kuroda MI, French CA. Ectopic protein interactions within BRD4-chromatin complexes drive oncogenic megadomain formation in NUT midline carcinoma. Proc Natl Acad Sci USA. 2017; 114:E4184–92.
https://doi.org/10.1073/pnas.1702086114
PMID:28484033

42. French CA. NUT carcinoma: clinicopathologic features, pathogenesis, and treatment. Pathol Int. 2018; 68:583–95.
https://doi.org/10.1111/pin.12727
PMID:30362654

43. Yin L, Cai Z, Zhu B, Xu C. Identification of key pathways and genes in the dynamic progression of HCC based on WGCNA. Genes (Basel). 2018; 9:92.
https://doi.org/10.3390/genes9020092
PMID:29443924

44. Xu W, Rao Q, An Y, Li M, Zhang Z. Identification of biomarkers for barcelona clinic liver cancer staging and overall survival of patients with hepatocellular carcinoma. PLoS One. 2018; 13:e0202763.
https://doi.org/10.1371/journal.pone.0202763
PMID:30138346

45. Tomczak K, Czerwińska P, Wiznerowicz M. The cancer genome atlas (TCGA): an immeasurable source of knowledge. Contemp Oncol (Pozn). 2015; 19:A68–77.
https://doi.org/10.5114/wo.2014.47136
PMID:25691825

46. Robinson MD, McCarthy DJ, Smyth GK. edgeR: a bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics. 2010; 26:139–40.
https://doi.org/10.1093/bioinformatics/btp616
PMID:19910308

47. Dessau RB, Pipper CB. ["R"—project for statistical computing]. Ugeskr Laeger. 2008; 170:328–30.
PMID:18252159

48. Emura T, Matsui S, Chen HY. compound.Cox: univariate feature selection and compound covariate for predicting survival. Comput Methods Programs Biomed. 2019; 168:21–37.

https://doi.org/10.1016/j.cmpb.2018.10.020
PMID:30527130

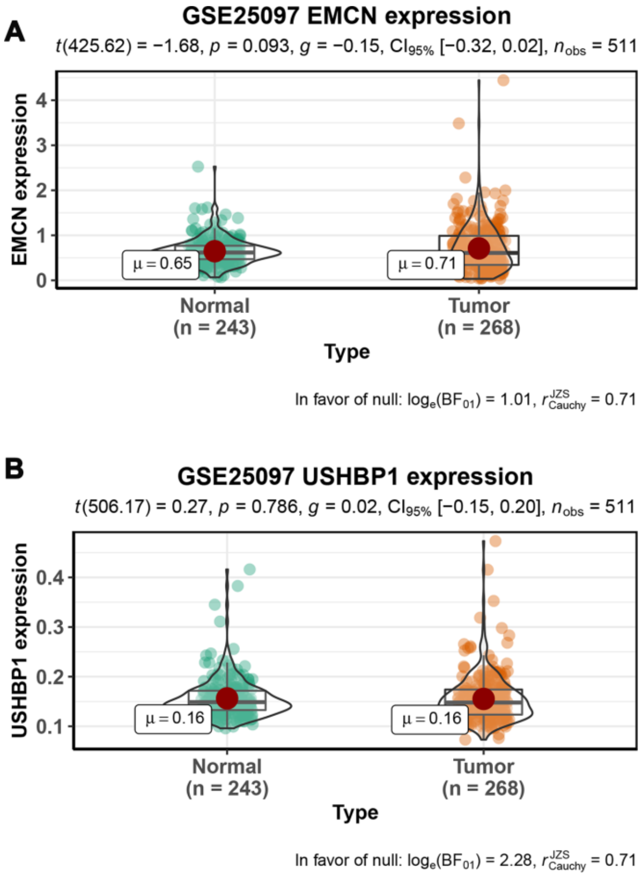
49. Tibshirani R. The lasso method for variable selection in the cox model. Stat Med. 1997; 16:385–95.
https://doi.org/10.1002/(sici)1097-0258(19970228)16:4<385::aid-sim380>3.0.co;2-3
PMID:9044528

50. Barrett T, Wilhite SE, Ledoux P, Evangelista C, Kim IF, Tomashevsky M, Marshall KA, Phillippy KH, Sherman PM, Holko M, Yefanov A, Lee H, Zhang N, et al. NCBI GEO: archive for functional genomics data sets—update. Nucleic Acids Res. 2013; 41:D991–95.
https://doi.org/10.1093/nar/gks1193
PMID:23193258

51. Rhodes DR, Kalyana-Sundaram S, Mahavisno V, Varambally R, Yu J, Briggs BB, Barrette TR, Anstet MJ, Kincead-Beal C, Kulkarni P, Varambally S, Ghosh D, Chinnaiyan AM. Oncomine 3.0: genes, pathways, and networks in a collection of 18,000 cancer gene expression profiles. Neoplasia. 2007; 9:166–80.
https://doi.org/10.1593/neo.07112
PMID:17356713

SUPPLEMENTARY MATERIALS

Supplementary Figure



Supplementary Figure 1. The expression of EMCN (**A**) and USHBP1 (**B**) in HCC and control samples in the GSE25097 dataset.

Supplementary Tables

Please browse Full Text version to see the data of Supplementary Tables 1 to 5.

Supplementary Table 1. List of 7498 DEGs in HCC tumor tissues.

Supplementary Table 2. The gene significance and module membership scores of all genes in the purple (A) and cyan (B) gene modules.

Supplementary Table 3. The sample identifiers of TCGA datasets.

Supplementary Table 4. The mRNAsi and EREG-mRNAsi of HCC samples.

Supplementary Table 5. The sample identifiers of ICGC datasets.