A novel five-gene signature for predicting prognosis in liver cancer

Qiuyue Zhong
Zunyi Medical University https://orcid.org/0000-0003-2345-3531
Yu Shuai
Zunyi Medical University
Qiong Luo
Affiliated Zhuhai Hospital, Southern Medical University
Guangyong Feng
Second Affiliated Hospital of Zunyi Medical University
Mingna Wu
Second Affiliated Hospital of Zunyi Medical University
Erxi Fan
Zunyi Medical University
Qiying Chen
Zunyi Medical University
Guojun Yue
Second Affiliated Hospital of Zunyi Medical University
Guihai Zhang (zghzhuhai@163.com)
https://orcid.org/0000-0002-4661-8864

Primary research

Keywords: liver cancer, TCGA, a five-gene signature, prognosis, risk score

Posted Date: January 3rd, 2020

DOI: https://doi.org/10.21203/rs.2.19369/v2

Abstract

Purpose Liver cancer is one of the most common malignant tumors in China, ranked 5th among the malignant common tumors in the world, which is still difficult to diagnose early and treat effectively. Therefore, exploring some indicators for prognostic prediction is imperative in the treatment of liver cancer. Methods Liver cancer data was obtained from The Cancer Genome Atlas (TCGA). We obtained differentially expressed genes (DEGs) by R software from TCGA database. Risk scores were acquired to assess the weighted gene-expression levels by Cox regression analysis and predict the prognosis of patients with liver cancer. Using the KEGG and GO databases, pathway enrichment was performed by identifying the analysis of DEGs. The display of receiver-operating characteristic (ROC) curves and area under the curve (AUC) could show the validity and the prognostic value of this model in liver cancer. Results In total, 1897 DEGs of transcriptome genes in liver cancer and 1197 DEGs of clinical data were extracted from the TCGA database. We identified a novel five-gene signature associated with liver cancer, including CDCA8, NR0B1, GAGE2A, AC018641.1, and SPANXC. Among of them, CDCA8 and NR0B1 were negatively related to 5-year OS, displaying a worse prognosis ($P < 0.05$). In particular, we also found that GAGE2A is related to lymphatic metastasis from the clinical data analysis in liver cancer. Receiver-operating characteristic (ROC) curve assessed the accuracy and sensitivity of the gene signature. In the heat map, each of the five genes for patients was presented with the distribution of the risk score. Conclusions We figured out a novel five-gene signature for the prognosis of patients with liver cancer, which may be an effective predictor for patients' prognosis in the future.

Introduction

Liver cancer is one of the most common malignant tumors in China, which ranked 5th among the malignant common tumors in the world. Over 620,000 incipient liver cancer cases occur in the world, including above half of Chinese, 80% of which are hepatocellular carcinoma (HCC) [1]. Viral hepatitis, cirrhosis, chemicals such as aspergillus flavus, and water and environmental pollution are the main causes of liver cancer [2]. The Barcelona Clinic Liver Cancer (BCLC) staging system, as a commonly utilized staging system for HCC, classifing patients according to tumor stage, liver function, and cancer-related symptoms [3]. In the light of the BCLC staging system, specific treatment options are substantial for different stages in HCC [4]. There are multiple therapeutic methods for HCC, including surgical options (resection or liver transplantation (LT), ablative electrochemical therapies (e.g. radiofrequency ablation or ethanol injection), and stereotactic body radiotherapy (SBRT) [5]. Despite of the development of surveillance systems, the over 5-year survival is less than 20%. Although risk factors including late diagnosis at the first time, a series of declines in body function induced by ascites caused by liver cirrhosis, recurrence, and occur of lymph vascular, it limits the availability of clinical treatment and leads to the poor prognosis [6].

Widespread prognostic systems have been acknowledged, but none are widely suitable for predicting prognosis. Some researches do not take into account major prognostic predictors for the presence of

cancer. Identification of novel targets or predictors through the analysis based on molecular levels may figure out some new therapeutic methods for better outcome prediction.

Methods

Data extraction, quality assessment

Liver cancer patients (including transcriptome and clinical data) were downloaded from the TCGA database. In order to reorganize data, we used the Perl scripting tools. Patients could be divided into several classification with missing data including vague clinical stages and pathological grading were excluded. We compared 407 liver cancer cases with 58 normal controls and calculated the fold changes between the genes using the edge R package in the R language.

Data screened for differentially expressed genes (DEGs) and heat map

Subsequently, by means of the mRNA data in liver cancer,in comparison with matched normal tissues, we identified DEGs (fold change = 3, $P < 0.05$) using R package software (version 3.5.2). The heat map was constructed by the data reduction (407 liver cancer cases vs. 58 normal controls). Five genes containing the division of the risk score of patients was demonstrated by a heat map. The high/low gene expression level was represented by the dark/light color in the heat map. The gplots package was used to draw the heat maps.

Gene enrichment

As for Gene Ontology (GO) enrichment, we utilized the Database for Annotation, Visualization, and Integrated Discovery (DAVID; http://david. abcc.ncifcrf.gov/) to execute related pathway dissect. Three GO terms [biological process (BP), cellular component (CC) and molecular function (MF)] were used to notarize the enrichment of pointed genes. Besides, the analysis of Kyoto Encyclopedia of Genes and Genomes (KEGG) was presented for the role pathway of these genes by identifying DEGs ($P < 0.05$, counts = 1544). Taking advantage of the Cytoscape (version 3.7.0), more intuitive pathway was shown with up/down-regulated genes (version 3.7.0) (http://www.cytoscape.org/cy3.html) and R (version 3.5.2).

Gene functional annotation evaluation

Receiver-operating characteristic curve (ROC) was conducted by ‘pROC package’ to assess the accuracy and sensitivity about the gene signature. Univariate Cox analysis was used to select the related variables. Based on the multivariate Cox regression analysis, we calculated each patient’ s risk score, which would be the sign of weighted expression levels of these genes. Risk score = $\sum \beta_{genei} \times Exp_{genei}$, where β_{genei} indicates the coef for each of five genes and Exp_{genei} suggests the expression degree of liver cancer patients. The higher score is together with high risk mortality.

Statistical analysis

Edger R, characterized by log fold change (FC), log CPM (counts per million), P value, and FDR (false discovery rate), was suitable for us to distinguish the statistical significance in the process of data analysis (defined as a fold change = 3, $P < 0.05$). Derived from the univariate and multivariate Cox regression analysis, a five-gene signature would be considered as an indicative of prognostic predicting factor ($P < 0.05$).

Results

1.Characteristics of the datasets

The transcriptome and clinical data were abstracted from the TCGA database, including 465 samples from patients with liver cancer. Among of them, 407 were tumor tissues and 58 were normal tissues. As for clinical data, there are totally 418 patients, consist of 146 women and 272 men, ranged from 16 to 90 years. 222 were White people, 164 were Asians, 20 were Black or African American, and 12 were from the other backgrounds. Amongst the 418 patients, 147 died with their lifespan post-diagnosis ranged from 0–3675 days. (Table 1)

Table 1
Demographic and clinical characteristics of the The Cancer Genome AtlasLiver cancer cohort (N = 418)

Characteristics	Number of sample, n(%)	
Age (years)	n	Percentage(%)
≦ 55	128	30.62%
> 55	248	59.33%
unknow	42	10.05%
Sex		
Male	272	65.07%
Female	146	34.93%
T stage		
T1	204	48.80%
T2	107	25.60%
T3	90	21.53%
T4	14	3.35%
TX	1	0.24%
unknow	2	0.48%
N stage		
N0	290	69.38%
N1	8	1.91%
NX	119	28.47%
M stage		
MO	303	72.49%
M1	8	1.91%
MX	107	25.60%
Stage		
I	194	46.41%
II	98	23.44%

Characteristics	Number of sample, n(%)	
III	90	21.53%
IV	12	2.87%
not available	24	5.74%
Histologic grade		
G1	55	13.16%
G2	180	43.06%
G3	124	29.67%
G4	13	3.11%
unknow	46	11.00%
Vital status		
Living	271	64.83%
Deceased	147	35.17%
Race		
White people	222	53.11%
American Indian or Alaska Native	2	0.48%
Asian	164	39.23%
Black or African American	20	4.78%
unknow	10	2.39%
Total	418	100.00%

2.Enrichment and visualization of signaling pathways

We identified 1897 DEGs by integrating the differential expression of 465 samples in the TCGA data. Of them, 77 were downregulated genes and 1820 were upregulated genes (Fig. 1). KEGG was presented for the role pathway of these genes by identifying DEGs. The 1897 DEGs were enriched with 6 signaling pathways significantly ($P < 0.05$, counts = 149) (Table 2).

Table 2. Enrichment analysis of differentially expressed genes (DEGs) in the KEGG pathways

ID	Description	GeneRatio	pvalue	p.adjust	qvalue	Count
hsa00280	Valine, leucine and isoleucine degradation	18/149	2.49E-08	1.67E-06	1.47E-06	18
hsa00071	Fatty acid degradation	15/149	1.69E-06	5.66E-05	4.98E-05	15
hsa00640	Propanoate metabolism	12/149	1.32E-05	0.0002957	0.00026012	12
hsa00650	Butanoate metabolism	10/149	6.49E-05	0.0010864	0.00095586	10
hsa00380	Tryptophan metabolism	12/149	0.000143329	0.0019206	0.00168977	12
hsa00410	beta-Alanine metabolism	10/149	0.000310915	0.0034719	0.0030546	10

The top six listed KEGG signaling pathways were shown as follows: hsa00280 (Valine, leucine and isoleucine degradation); hsa00071 (Fatty acid degradation); hsa00640 (Propanoate metabolism); hsa00650 (Butanoate metabolism); hsa00380 (Tryptophan metabolism) and hsa00410 (beta-Alanine metabolism) (Fig. 2). Taking advantage of the Cytoscape (version 3.7.0), more intuitive pathway was shown with up/down-regulated genes (Fig. 3). We then enriched DEGs according to the GO terms (Table 3). As shown in Fig. 4A and 4B, these specific genes were strikingly enriched in the GO terms, such as BP, CC and MF. Following, we also organized the graphs of the GO signaling pathways by using R package components (Cluster Profiler, Stringi, Pathview, and GOplot) (Fig. 5A, B).

Table 3. Enrichment analysis of differentially expressed genes (DEGs) in the GO terms

Term	Count	PValue	FDR
GO:0043565 ~ sequence-specific DNA binding	75	3.47E-15	5.27E-12
GO:0005576 ~ extracellular region	155	1.34E-12	1.88E-09
GO:0005615 ~ extracellular space	130	1.1E-10	1.54E-07
GO:0007586 ~ digestion	20	4.39E-10	7.92E-07
GO:0009952 ~ anterior/posterior pattern specification	22	9.28E-10	1.67E-06
GO:0005179 ~ hormone activity	23	2.5E-09	3.84E-06
GO:0048704 ~ embryonic skeletal system morphogenesis	15	6.92E-09	1.25E-05
GO:0007062 ~ sister chromatid cohesion	23	2.36E-08	4.26E-05
GO:0021520 ~ spinal cord motor neuron cell fate specification	7	0.000000163	0.000294
GO:0007067 ~ mitotic nuclear division	36	0.00000018	0.000325
GO:0007268 ~ chemical synaptic transmission	35	0.000000246	0.000443

A fivegene signature and risk score assessment

A five-gene signature model was drawn from the synthesis of univariate or multivariate Cox regression analysis. It revealed a five-gene signature as an prognostic predictor in liver cancer patients ($P < 0.05$). The five-gene signature model consists of CDCA8, NR0B1, GAGE2A, AC018641.1, and SPANXC. Of these, CDCA8 and NR0B1 were related to a worse prognosis ($P < 0.05$) (Fig. 6). In particular, we also found that GAGE2A is positively related to lymphatic metastasis through the data analysis in liver cancer. We calculated each patient' s risk score, which would be the sign of weighted expression levels of these five genes (Table 4). By means of median-risk score, patients were divided into high- and low-risk groups. A poorer prognosis could be represented by a higher risk score. These results may help clinicians develop adaptive treatment strategies for the liver cancer patients to predict prognosis.

Table 4. A five-gene signature identified by multivariate Cox regression analysis

Gene	coef	exp(coef)	se(coef)	z	p
CDCA8	0.2583	1.2947	0.076	3.4	0.00068
NR0B1	0.0551	1.0567	0.0316	1.74	0.08112
GAGE2A	0.1029	1.1084	0.04	2.57	0.0101
AC018641.1	0.1188	1.1261	0.0577	2.06	0.03968
SPANXC	0.3103	1.1261	0.1083	2.86	0.00417

Accuracy of TCGA datasets

The sensitivity and accuracy of data could be shown through ROC curve and AUC (= 0.714) ($P < 0.05$) (Fig. 7A). Risk score calculated was associated dramatically with 5-year OS (overall survival) by using univariate and multivariate Cox regression analysis ($P < 0.05$) (Fig. 7B). The heat map based on the risk scores in patients with liver cancer revealed deeply the obvious differences in expression level of these five genes between liver cancer and corresponding normal tissues (Fig. 8).

Discussion

Up to now, liver cancer is still hardly diagnose early and treat effectively. Hepatic resection, orthotopic liver transplantation, ablative therapies, chemoembolization and systemic therapies with cytotoxic drugs were still the main treatment strategies for liver cancer [7]. Following the past few years, increasing studies have shown that the molecular phenotype plays more and more imperative functions in the diagnosis and therapeutic reactions of liver cancer patients [8].

Several datasets about patients' genomes have been built, which makes researchers recognize the genomic changes between tumor and matched normal tissues. It becomes a crucial step to form a novel and effective identification for prognostic biomarkers on liver cancer outcomes. Meanwhile, accumulating studies have revealed that a variety of biomarkers are emerged as the indicators of prognosis by the analysis of dataset [9–11]. At this study, DEGs between tumor and normal tissues exerted according to a series of analytical processions based on the TCGA database. Especially, we obtained a five-gene signature by means of univariate and multivariate Cox regression analyses. Besides, the risk score was also calculated by this signature, which could be a possible marker to predict the prognosis of patients with liver cancer.

Cell division cycle associated 8 (CDCA8), is a member of the chromosomal passenger complex (CPC) crucial for transmission of the genome during cell division [12]. More studies illustrated that massive expression of CDCA8 was an urgent need to the formation and development of tumor [13]. Loss of CDCA8 could induce the proliferation of defective cell and early embryonic lethality [14]. That reminds us that CDCA8 could be a risk factor in liver cancer because of negative regulatory in the Kaplan–Meier curve at our studies. Nuclear Receptor Subfamily 0 Group B Member 1(NR0B1), as a Protein Coding gene, encodes a protein inferring a DNA-binding domain. The encoded protein acts as a dominant-negative regulator of transcription which is mediated by the retinoic acid receptor. The mutations of NR0B1 would trigger both X-linked congenital adrenal hypoplasia and hypogonadotropic hypogonadism [15]. Ho et al. found that NR0B1 expression is restricted to adrenal-glands, lung and pancreas [16]. In addition, NR0B1, associated with ERα, PR and AR expression [17], is also a positive prognostic factor in node-negative breast-cancer. It is correlated with smaller tumor-size, earlier disease-stage and increased survival [18]. With the negative survival correlation in this study, NR0B1 maybe a passitive prognostic factor in live cancer. Sperm Protein Associated With The Nucleus On The X Chromosome C (SPANX-C) is a member of the SPANX family, which is located in a gene cluster on chromosome X, whose overexpression has little influence about primary tumor growth in cancer cells, but may be sufficient to determine an invasive phenotypes, including the morphology of their nucleus and the diameter of vessel[19]. Studies of

Westcott et al. in vivo/vitro have shown that little SPANX-C expression reduced metastatic behavior of primary tumors and the number of potential metastatic cancer cells [20]. GAGE2A (G Antigen 2A), a protein coding gene, belongs to the GAGE family that code for peptide, containing an open reading frame coding for a protein of 138 amino acids, which are expressed in a significant proportion of melanomas(24%), sarcomas (25%), non-small cell lung cancer(19%), head and neck tumors(19%), and bladder tumors(12%)[21]. Mamsen et al. recognized that GAGE gene products could be an important part during early stages of embryonic development [22]. AC018641.1, the Aliases of ENSG00000226468 Gene, was poorly investigated, but along with the further investigations, it could be a prognostic possible marker for cancer detection and therapy.

Future research could investigate the biological characteristics of these five genes for their prognostic value in liver cancer. Besides, we would inform some limitations in this study as followed: the five-gene signature based on dataset analysis in liver cancer was initially useful, but it was not exhaustive. Deeply, it is necessary to have more comprehensive data sets obtained from other databases for supplementary authentication. Particularly, clinical practice would do to verify the five-gene signature mode, and experiments are needed to demonstrate our results. Consequently, this study provides us a guideline that roles of these five genes are related to the prognosis of patients in liver cancer. Although we have only limited knowledge of these five genes, these indicators would be a selecting assist on the cancer progression, monitor and treatment.

Conclusions

Totally, several examinations offer us further sustenance for the suggestions that the five-gene signature possess a meaningful relationship with the prognosis of the patients with liver cancer. We also hope that our studies will present long-term opinion as an effective predictor for the prognosis of patients with liver cancer in the future.

Declarations

Ethics approval and consent to participate

This study was approved by the Affiliated Hospital of Zunyi Medical University institutional review board (IRB). The data were obtained from TCGA (https://portal.gdc.cancer.gov/). Informed consent has been obtained from all individual participants included in the study was approved by the Institutional review Board.

Consent for publication

Not applicable.

Availability of data and materials

The datasets supporting the results of this article are publicly available at the TCGA (https://portal.gdc.cancer.gov/).

Competing interests

The authors declare that they have no competing interests.

Funding

Science and Technology Bureau funded project of Zunyi (No. ZSKHSZ201125)

Joint Fund Project of the Science and Technology Department in Guizhou Province (Qian Ke He LH [2014] 7585)

Joint Fund Project of the Science and Technology Department in Guizhou Province (No. Qian Ke He LKZ [2013] 10)

Authors' contributions

QZ conceptualized the main idea of the project. YS and QZ carried out the data processing, coding, and computational. GZ, QL, GY, GF, MW, QC, EF designed the study and supervised the project. QZ, YS, QL, GZ, EF, GF, MW, QC and GY wrote the paper. All authors read and approved the final manuscript.

Acknowledgements

We appreciate the generosity of TCGA network for sharing the huge amount of data.

References

1. El-Serag HB. Hepatocellular carcinoma. N Engl J Med. 2011;365(12):1118-1127.
2. WU MC. The current development in treatment and research for primary liver cancer of China. Journal of Chengdu Medical College. 2012;7(2):161-162.
3. Forner A, Gilabert M, Bruix J, et al. Treatment of intermediate-stage hepatocellular carcinoma. Nat Rev Clin Oncol. 2014;11(9):525-535.
4. J.M. Llovet, C. Bru, J. Bruix, et al. Prognosis of hepatocellular carcinoma: the BCLC staging classification. Semin. Liver Dis. 1999;19(3):329-338.
5. J Hartke J, Johnson M, Ghabril M, et al. The diagnosis and treatment of hepatocellular carcinoma. Semin Diagn Pathol. 2017;34(2):153-159.
6. Kudo M, Chung H, Osaki Y, et al. Prognostic staging system for hepatocellular carcinoma (CLIP score): its value and limitations, and a proposal for a new staging system, the Japan Integrated Staging Score (JIS score). J Gastroenterol. 2003;38(3):207-215.
7. Duseja.A. Staging of hepatocellular carcinoma. J.Clin.Exp.2014; 4:S74-79.

8. Sia D, Villanueva A, Friedman SL, et al. Liver Cancer Cell of Origin, Molecular Class, and Effects on Patient Prognosis. Gastroenterology.2017;152(4):745-761.
9. Lin SB, Zhou L, Liang ZY, et al. Expression of GRK2 and IGF1R in hepatocellular carcinoma: clinicopathological and prognostic significance.J Clin Pathol. 2017;70(9):754 -759.
10. Jiao Y, Fu Z, Li Y, et al. High EIF2B5 mRNA expression and its prognostic significance in liver cancer: a study based on the TCGA and GEO database. Cancer Manag Res. 2018;10:6003-6014.
11. Hu B, Chang X, Liu X. Predicting Functional Modules of Liver Cancer Based on Differential Network Analysis. Interdiscip Sci. 2019;11(4):636-644.
12. Hindriksen S, Meppelink A, Lens SM. Functionality of the chromosomal passenger complex in cancer. Biochem Soc Trans.2015;43:23-32.
13. Dai C, Miao CX, Xu XM, et al. Transcriptional activation of human CDCA8 gene regulated by transcription factor NF-Y in embryonic stem cells and cancer cells. J Biol Chem.2015;290:22423-34.
14. Li L, Li D, Tian F, et al. Hepatic Loss of Borealin Impairs Postnatal Liver Development, Regeneration, and Hepatocarcinogenesis. J Biol Chem. 2016;291:21137- 21147.
15. Bergadá I1, Andreone L, Bedecarrás P, et al. Seminiferous tubule function in delayed-onset X-linked adrenal hypoplasia congenita associated with incomplete hypogonadotrophic hypogonadism. Clin Endocrinol (Oxf). 2008;68(2):240-246.
16. Ho J, Zhang YH, Huang BL, et al. NR0B1A: an alternatively spliced form of NR0B1. Mol Genet Metab. 2004;83(4):330-336.
17. Chae BJ, Lee A, Bae JS, et al. Expression of nuclear receptor DAX-1 and androgen receptor in human breast cancer. J Surg Oncol. 2011;103(8):768 -772.
18. Zhang H, Slewa A, Janssen E, et al. The prognostic value of the orphan nuclear receptor DAX-1 (NROB1) in node-negative breast cancer. Anticancer Res. 2011;31(2):443- 449.
19. E.A. Maine, J.M. Westcott, A.M. Prechtl, et al. The cancer-testis antigens SPANX-A/C/D and CTAG2 promote breast cancer invasion. Oncotarget.2016;7:14708 -14726.
20. J.M. Westcott, A.M. Prechtl, E.A. Maine, et al. An epigenetically distinct breast cancer cell subpopulation promotes collective invasion. J. Clin. Invest. 2015;125:1927-1943.
21. Van den Eynde B, Peeters O, De Backer O, et al. A new family of genes coding for an antigen recognized by autologous cytolytic T lymphocytes on a human melanoma. J Exp Med. 1995;182(3):689-98.
22. Mamsen LS, Ernst EH, Borup R, et al. Temporal expression pattern of genes during the period of sex differentiation in human embryonic gonads. Sci Rep. 2017;7(1):15961.

Figures

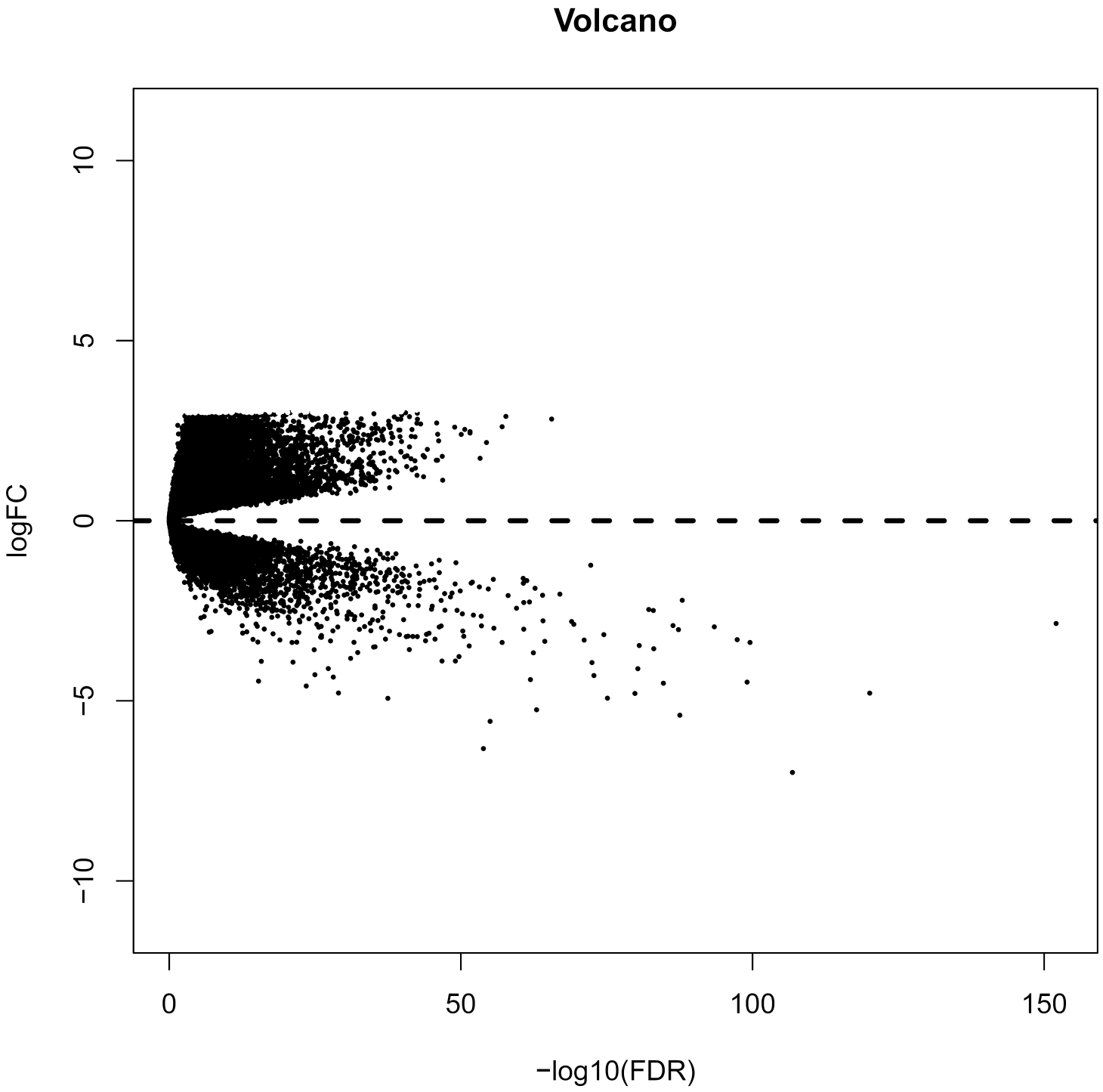

Figure 1

Volcano plot of differentially expressed genes (DEGs). We detected 1897 DEGs in liver cancer from normal samples; each green dot shows a downregulated gene and each red dot shows an upregulated gene (fold change>3, P< 0.05)

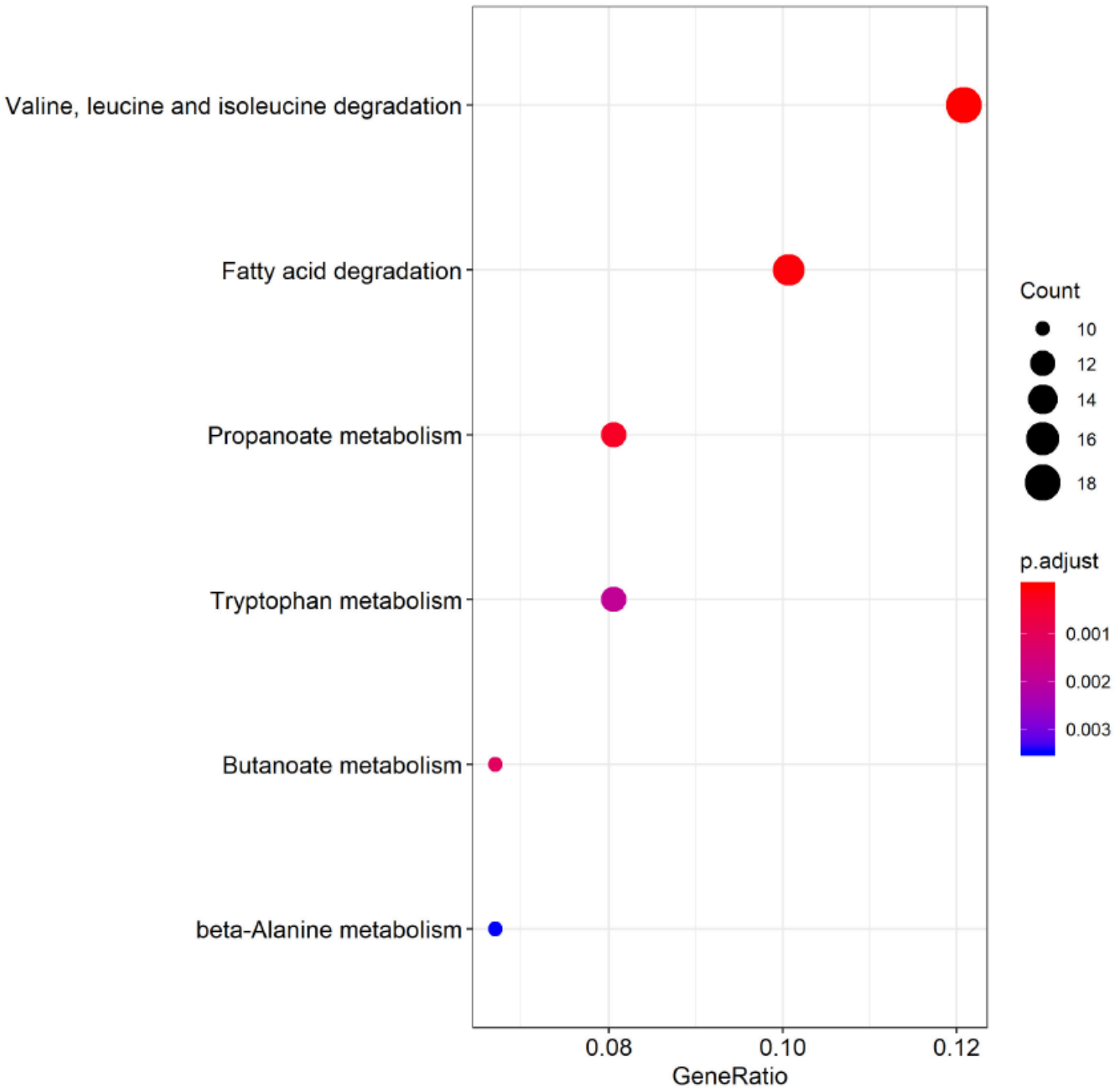

Figure 2

Enrichment analysis performed on DEGs using the KEGG database. Dot plot showing the top 6 KEGG pathway permutations according to the number of DEGs. The correlation is more significant as the red/blue ratio increases (P <0.05, counts= 149)

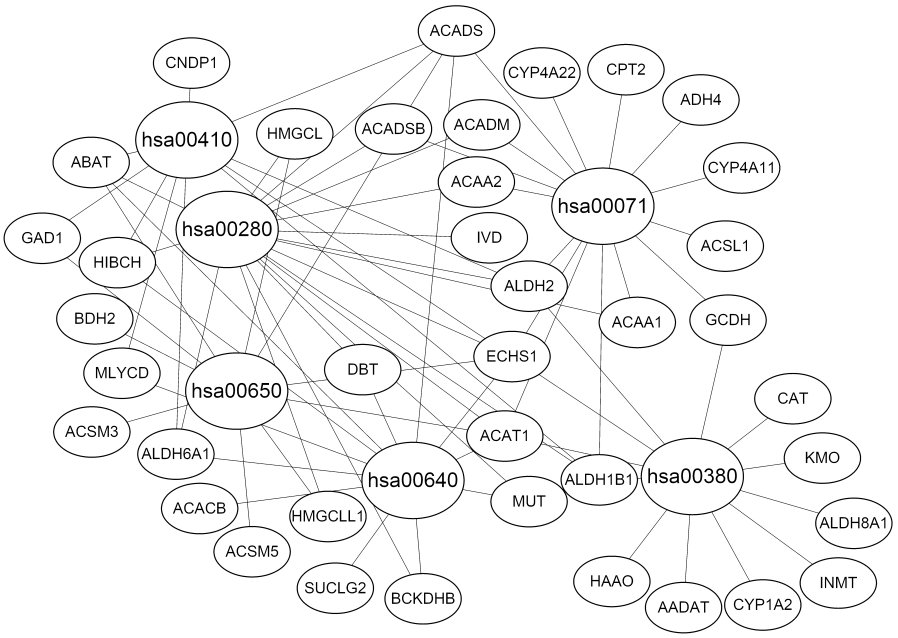

Figure 3

Enrichment analysis data (as visualized by cytoscape) showing 6 KEGG functional pathways. A part of genes were downregulated represented in green, others were upregulated represented in red

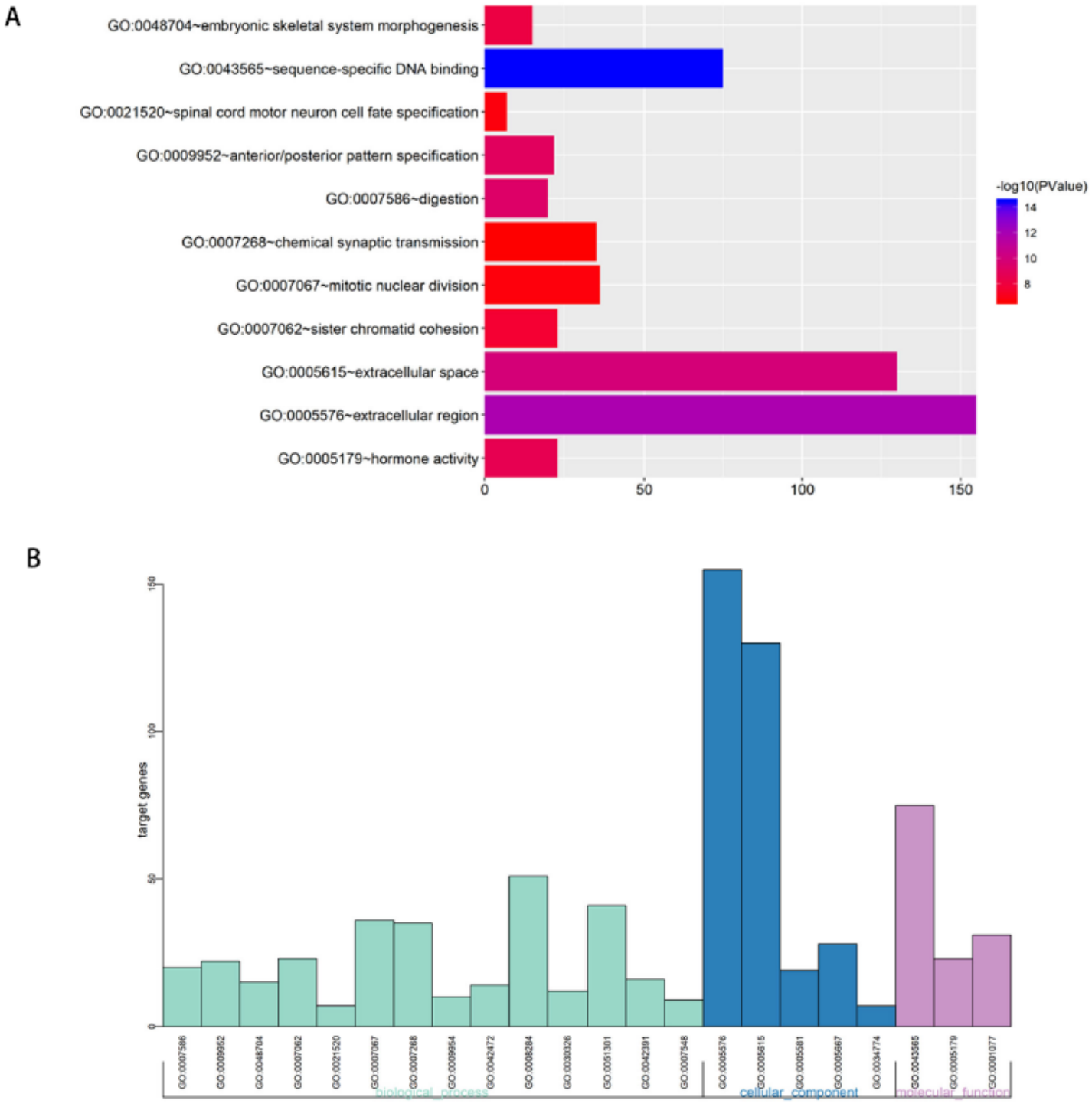

Figure 4

(A and B) mRNAs most highly enriched in the GO terms. GO, Gene Ontology. DAVID, Database for Annotation, Visualization and Integrated Discovery.

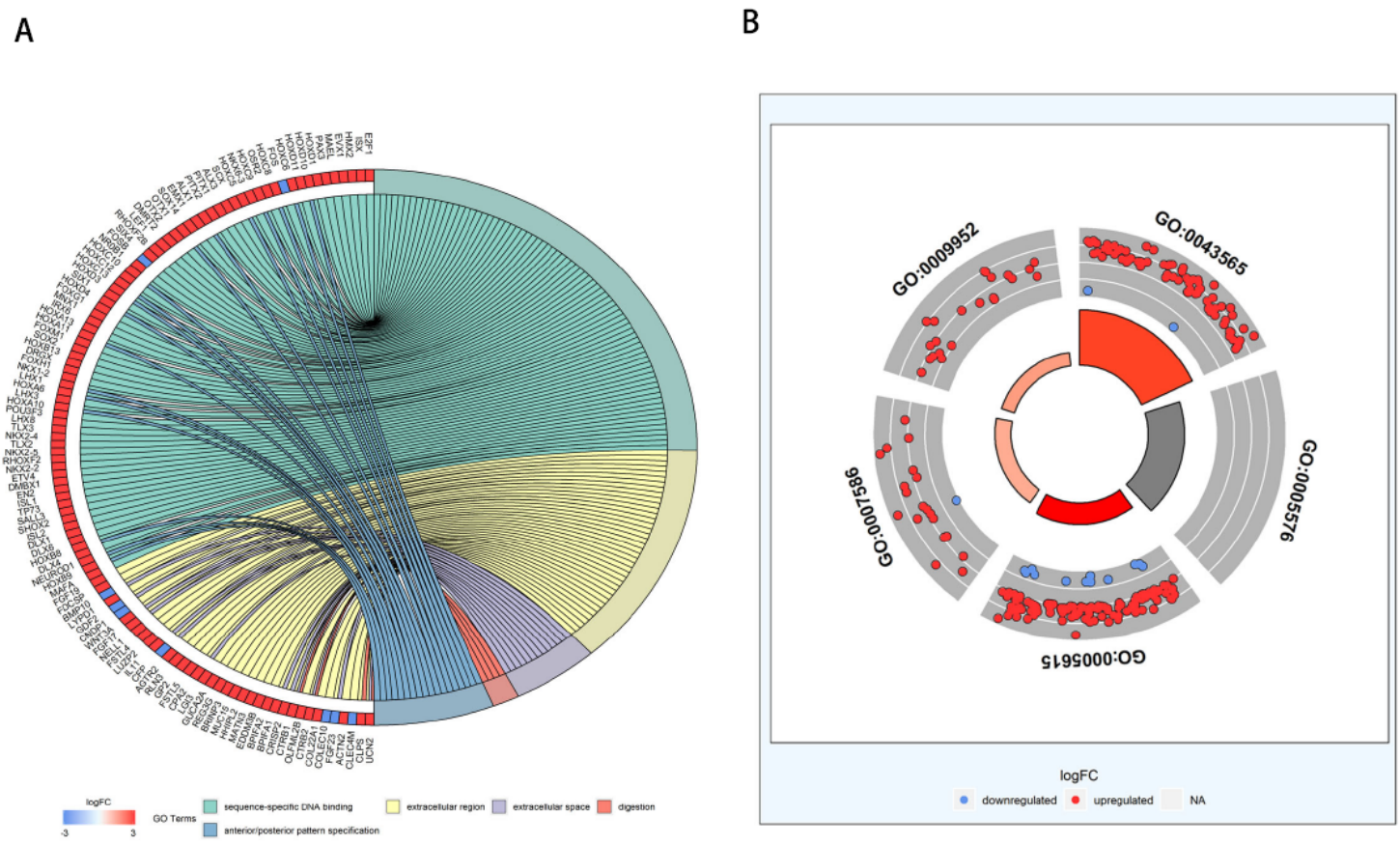

Figure 5

(A) Several genes are enrich in signal pathways by GO terms; (B GO enrichment analysis data also show five classical pathways. Most of genes were upregulated represented in red; others were downregulated represented in blue

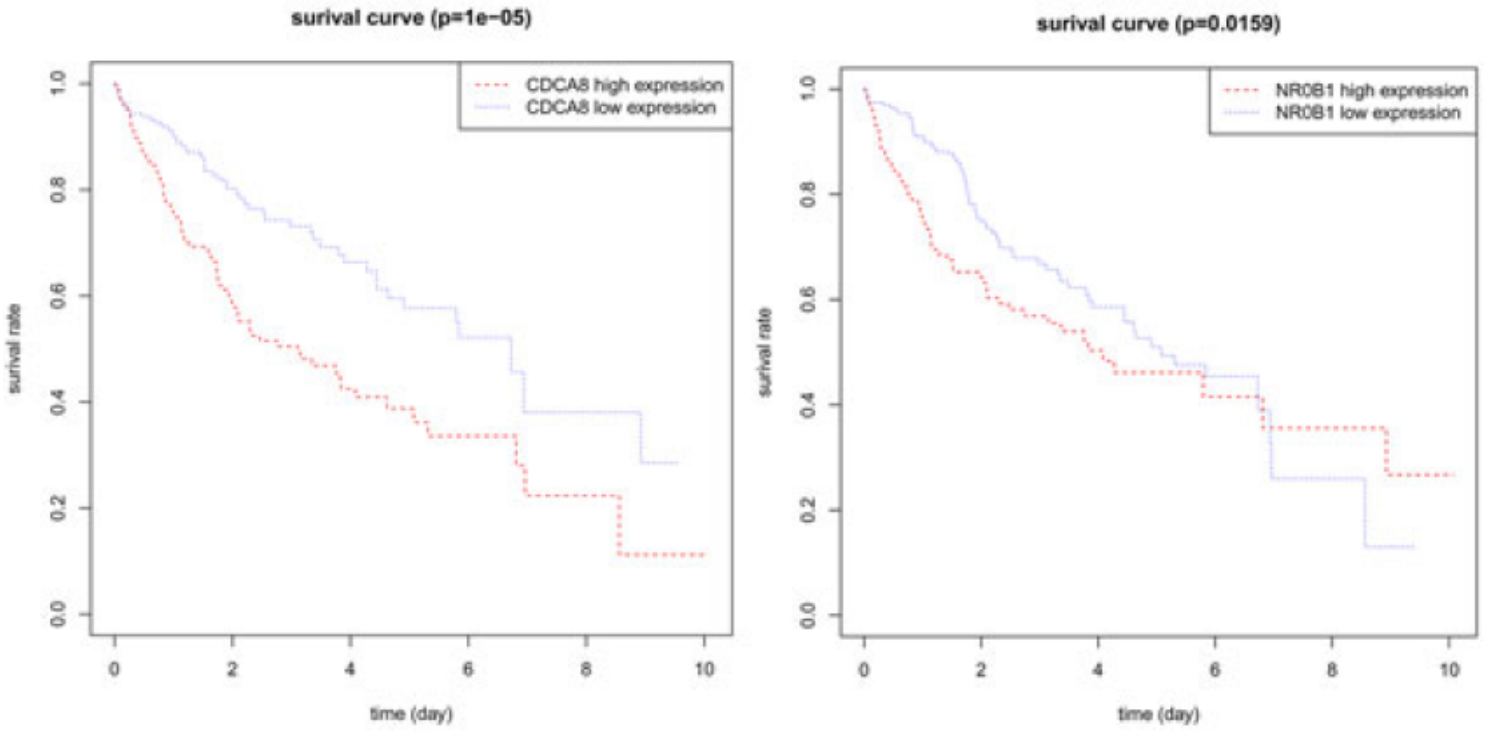

Figure 6

Survival cures of the CDCA8 and NR0B1 are showing on it, which were significantly correlated with gene expression ($P < 0.05$)

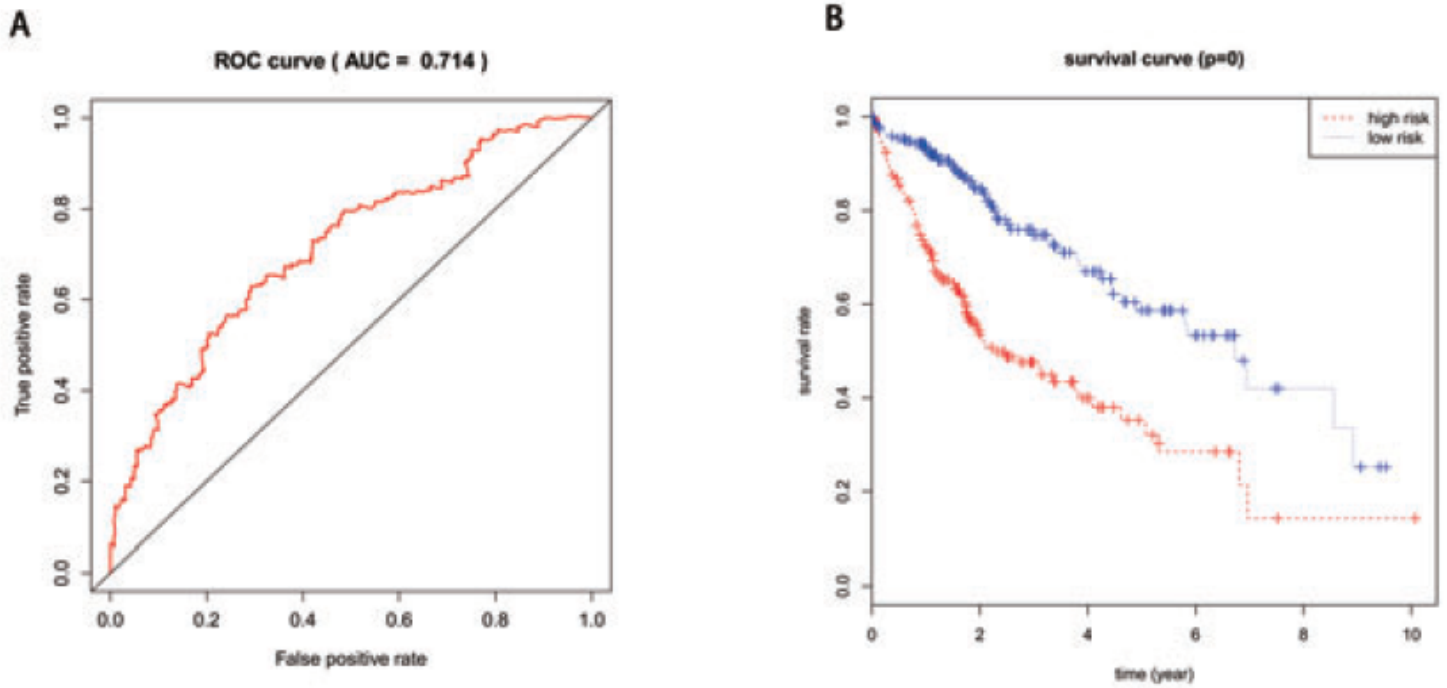

Figure 7

Validation of the five-gene signature. (A) ROC validation of the five-gene signature. The area under the curve (AUC) was 0.714 (P< 0.05), thus demonstrating that the five-gene signature had high sensitivity and specificity for the classification of liver cancer from normal controls; (B) Risk score was significantly correlated with overall survival (OS). Risk score was an independent prognostic factor for OS in multivariate Cox regression analysis (P=0). The median score divided patients into high- and low-risk groups.

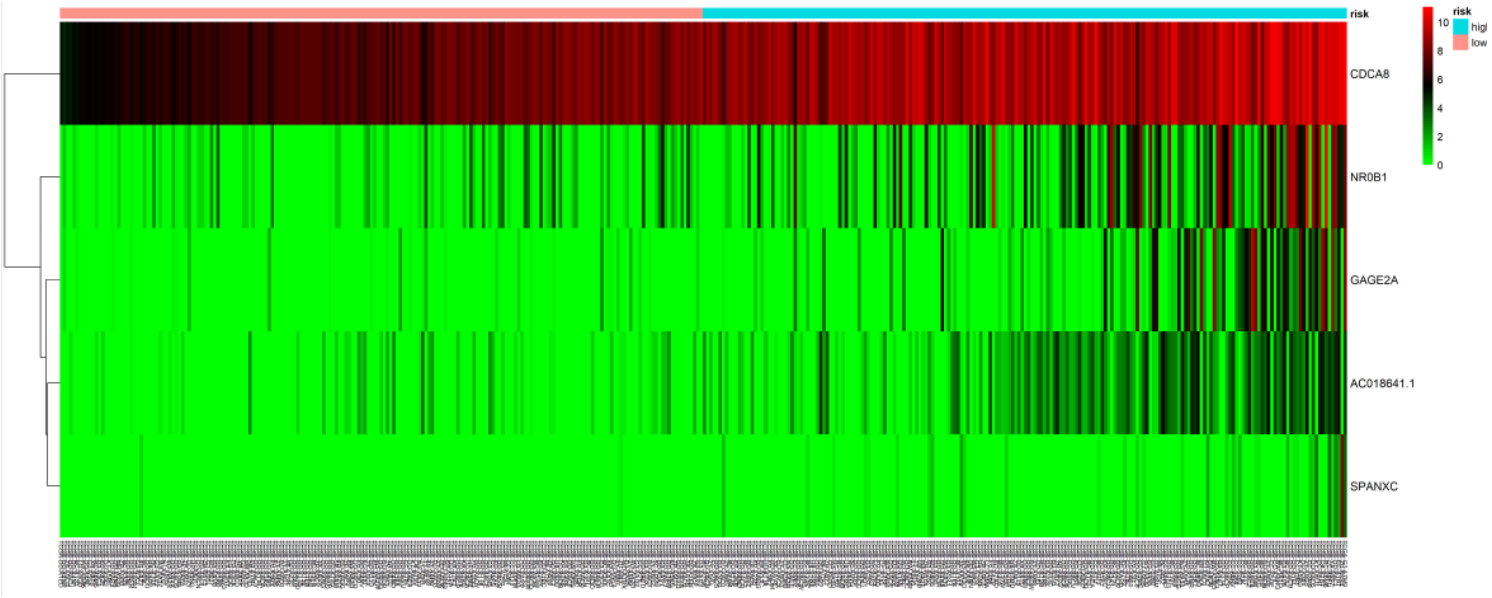

Figure 8

Heatmap of the five-gene signature in TCGA data sets. Each column represents a sample and each row represents one of the five genes. The expression levels of the five genes are shown in different colors, from orange to blue with increasing risk