Technology in Cancer Research & Treatment
Volume 19: 1-8

DOI: 10.1177/1533033820959353
journals.sagepub.com/home/tct
SAGE

Prediction of Poor Prognosis of HCC by Early Warning Model for Co-Expression of miRNA and mRNA Based on Bioinformatics Analysis

Zi-jian Su, MD[1], Chun-cheng Lin, MD[1], Jian-hui Pan, MD[1], Jian-hua Zhang, MD[1], Tao Han, MD, PhD[2], and Qunxiong Pan, MD[1]

Abstract

Objective: Hepatocellular Carcinoma (HCC) has the highest mortality rate worldwide with the intractability of its extremely complicated pathogenesis and unclear mechanism. The limited survival highlights the need for the further detection of prognosis for HCC. MicroRNAs (miRNAs) and messenger RNAs (mRNAs) have been identified as regulatory factors and target genes in human cancers, while some studies also found post-transcriptional modification plays a crucial role in the occurrence and development of HCC. The present study aimed to elucidate the prognostic significance of miRNA and mRNA models in HCC. **Methods:** Data were obtained from The Cancer Genome Atlas (TCGA), International Cancer Genome Consortium (ICGC), and Gene Expression Omnibus (GEO) databases. The miRNA and mRNA expressions were tested by the Wilcoxon and used funrich software to predict mRNA that might be related to miRNA. Then we determined the intersection with overlapped mRNA and miRNA Venn diagram, and screened out hub gene by using Degree algorithm in Cytoscape software. The COX models, with TCGA data as the training set and ICGC data as the test set, were constructed. All patients were divided into high-risk and low-risk groups. Data on overall survival of different groups were collected and analyzed by Kaplan-Meier method, and independent risk factors affecting prognosis were assessed by Cox analysis. **Results:** The miRNA and mRNA polygenic risk model showed a good true positive rate. Kaplan-Meier curve and Cox analysis suggested that the high-risk group was associated with poor prognosis, and the risk score could be used as an independent risk factor for HCC. **Conclusion:** Tumor risk models constructed in this study could effectively predict the prognosis of patients, which is expected to provide a reference for the prognostic stratification and treatment strategy development of HCC.

Keywords

hepatocellular carcinoma, prognostic signature, TCGA, miRNA, mRNA

Received: February 27, 2020; Revised: August 12, 2020; Accepted: August 21, 2020.

Introduction

MicroRNAs (miRNAs) are small RNAs with endogenous lengths of about 20 to 24 nucleotides, which have a variety of important regulatory functions in cells,[1,2] and regulate a third of human genes.[3] MicroRNAs play an important role in various biological regulatory mechanisms, including the time of development, host-pathogen interaction, cell differentiation, proliferation, apoptosis, and tumorigenesis.[4] Each miRNA can target to multiple genes, and several different miRNAs can also regulate the same gene, which forms an extensive regulatory network. This complex regulatory network can either regulate

[1] Hepatobiliary surgery, Quanzhou First Hospital Affiliated to Fujian Medical University, Quanzhou, Fujian, China
[2] Department of Oncology, Second Affiliated Hospital of Dalian Medical University, Dalian, China

Corresponding Authors:
Qunxiong Pan, Hepatobiliary surgery, Quanzhou First Hospital Affiliated to Fujian Medical University, Quanzhou, Fujian 362000, China.
Email: 309869785@qq.com

Tao Han, Department of Oncology, Second Affiliated Hospital of Dalian Medical University, Dalian 116023, China.
Email: than1984@sina.com

the expression of multiple genes through a single miRNA or fine-tune the expression of a gene through combinations of several miRNAs. However, a microRNAs serve as a class of post-transcriptional regulators and do not act directly on genes but on mRNA transcribed by genes, to down-regulate the expression of target genes. MicroRNAs that coupled with protein complexes can work through 1 of the 2 mechanisms: mRNA cleavage or translation inhibition. While, in the August 2010 issue of Nature, David Barteland and his colleagues applied ribosomal profiling techniques and found that microRNAs work mainly by degrading target mRNAs.[5] Nevertheless, miRNA and mRNA always maintain a stable negative regulatory relationship.

A large number of studies have proved that miRNAs in malignant tumor cells often behave abnormally. Abnormal expression of certain miRNAs is closely related to the diagnosis, treatment, and prognosis of different types of cancer in humans, with positive or negative regulatory effects such as maintenance of proliferative signaling, anti-apoptosis, induction of angiogenesis, and cancer cell invasion and metastasis.[6] Therefore, miRNAs have become important biomarkers for diagnosis and prognosis as well as therapeutic targets for cancer.[7]

The factors leading to HCC are numerous and complex.[8] A large number of data have confirmed that the relationship between miRNA and HCC is inseparable, and tumor development is always accompanied by imbalance and abnormal messenger RNA (mRNA) expression[9,10] So, it is extremely important to explore the role of miRNA in the development of HCC. The regulatory effect of miRNA and related mRNA in HCC is mainly achieved by targeting key genes in signaling pathways.

Materials and Methods

Data Mining and Collection

Data on clinical characteristic and miRNA and mRNA gene expression of HCC patients were downloaded from The Cancer Genome Atlas (TCGA) database. Gene chips (GSE112264) with miRNA expression data are download from Gene Expression Omnibus (GEO) (https://www.ncbi.nlm.nih.gov/geo/) database. Data on clinical characteristic and mRNA gene expression of HCC patients were downloaded from International Cancer Genome Consortium (ICGC) database (LICA-FR). The transcriptome data from different platforms are standardized by R v3.4.1 (https://www.r-project.org/) using the "SVA" package.

Bioinformatics Analysis

Data on microRNAs and mRNA gene expression in the TCGA database, miRNA gene expression in the GEO database, and mRNA gene expression in the ICGC database were divided into 2 groups according to cancer tissues and adjacent tissues, respectively. Wilcoxon test was used to analyze the expression of these genes in different tissues (FoldChange of >1 and P-value of <0.05). Funrich software was used for miRNA target gene prediction, and the Venn diagram was drawn to analyze the correlation of differential miRNA and differential mRNA in 3 databases. STRING (https://string-db.org/cgi/input.pl) was used in this study to analyze the protein-protein interactions of differentially expressed mRNA. Isolated nodes were removed, and the results of the interaction network were downloaded and then imported into Cytoscape (version 3.7.2) for subsequent analysis. After building the gene interaction network with the "cytoHubba" application, the network structure and weighted connection between nodes were calculated and analyzed based on degree algorithm. Genes in the network were ranked from high to low, and 20 hub genes were screened out. To elucidate the prognostic risk of genes, we constructed Cox models on the expression of miRNA and mRNA from TCGA sets and use the ICGC data set as the test set. The risk score was calculated using the following formula, where Coefi is the coefficient, and xi is the expression value of each selected molecule. According to the risk score, all patients were divided into high-risk or low-risk groups.

$$\text{Risk score} = \sum_{i=1}^{n} \text{Coefi} * \text{xi}$$

Further, the diagnostic capability of the model was evaluated by receiver-operating characteristic curve (ROC) through R v3.4.1 software of the time ROC and survival package. Univariate Cox analysis was used to select the correlation variables, and then multivariate Cox analysis was applied to analyze the independent clinical risk factors influencing the prognosis.

GSEA Functional Enrichment Analysis

Gene Set Enrichment Analysis (GSEA) is used to evaluate the distribution trend of genes in a predefined Gene Set and to determine its contribution to the phenotype through the order of phenotypic correlation. The whole gene expression data of patients in the various risk groups were uploaded to the GSEA v4.0.3 (www.gsea.com) and the Kyoto Encyclopedia of Genes and Genomes (KEGG) software for analysis, which was performed with 1,000 iterations. From the perspective of gene set enrichment, if we not limited to differential genes, theoretically, it is easier to include the effects of subtle but coordinated changes on the biological pathway.

Statistical Analysis

Statistical analysis of all RNA-seq transcriptome data was conducted using R v3.4.1. Differential mRNA and differential miRNA were analyzed using the Wilcoxon test according to the classification of cancer or adjacent tissues. Patients were divided into high-risk and low-risk groups based on the median. Pearson correlation coefficient method was used to test the correlation between gene expression and clinical data. Kaplan-Meier method was used to compare overall survival (OS) of patients between the high-risk group and low-risk group.

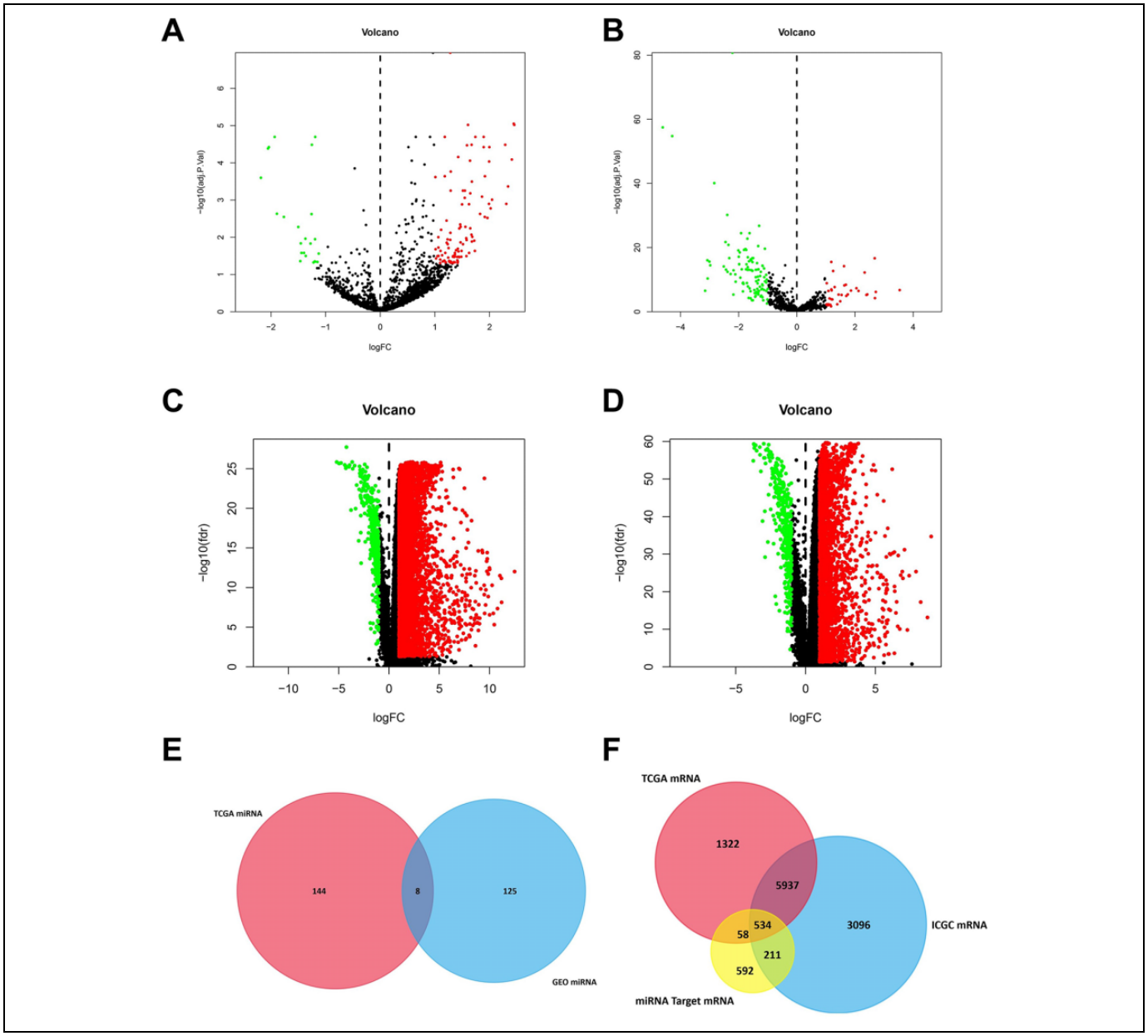

Figure 1. Expression of miRNA and mRNA in liver cancer and paracancerous tissues. (A)Volcano map of GEO miRNA differential expression. (B) Volcano map of TCGA miRNA differential expression. (C) Volcano map of TCGA mRNA differential expression. (D) Volcano map of ICGC mRNA differential expression. (E) Venn diagram of miRNA from TCGA and GEO database. (F) Venn diagram of mRNA from the miRNA targeting mRNA, TCGA, and ICGC database.

Results

miRNA and mRNA Expression in HCC

In order to comprehensively and accurately analyze the expression of miRNA and mRNA in HCC, we conducted researches on miRNA and mRNA, respectively, based on 2 databases. GEO and TCGA were used to verify the expression of miRNA in HCC. ICGC and TCGA were used to verify mRNA expression in HCC. We determined the difference in gene expression between HCC patients and normal people with the generated data presented in the volcano map. (Figure 1.A-D). The results showed that most of the differential miRNAs in the GEO database played a role in promoting tumor development. On the other aspect, for the differential miRNA in the TCGA database, the differential miRNA with inhibitory effect on tumor development accounted for a larger proportion. However, whether derived from ICGC or TCGA, most mRNAs have a down-regulating effect. For miRNAs, we obtained 152 differential miRNAs in TCGA and 125 differential miRNAs in GEO. In order to intuitively and effectively learn which differential genes are commonly owned in different databases, we use the Wayne diagram to obtain the overlapped genes. The results

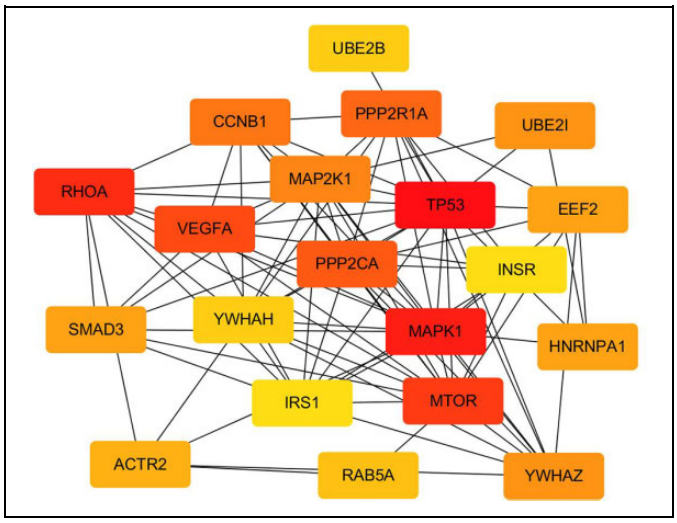

Figure 2. Network structure and weighted connectivity between the top 20 hub genes.

showed that 8 differential miRNAs were shared by GEO and TCGA (Figure 1 E). A total of 1395 mRNAs were predicted to be targeted by the 8 miRNAs using Funrich. On the mRNA side, the number of different mRNAs obtained from ICGC was 9778, and the number of different mRNAs obtained from TCGA was 7852. In the Venn diagram, we could clearly see the distribution of differential mRNAs and eventually obtained 534 crossed differential mRNAs (Figure 1 F).

Screening of Hub Genes

We used degree algorithm to calculate and analyze the network structure and the weighted connection between nodes to screened out the 20 hub genes. The interaction between them was shown in Figure 2.

A Multi-Gene Prediction Model Was Constructed for the Prognostic Analysis of HCC

We also evaluated the role of miRNA and mRNA in the prognostic risk of HCC. Firstly, Cox risk models were constructed using differential miRNAs from the TCGA database, and 16 risk miRNAs were screened. The coefficients for these risk miRNAs are shown in Table 1. Similarly, we constructed a Cox risk model using the TCGA database of hub mRNAs as a training set and screened out 6 risk mRNAs. The coefficients of these risk mRNAs are shown in Table 2. The ICGC data were used as a test set to verify the stability and specificity of the constructed model. As shown in Figure 3A-C, the ROC of the predictive model was executed. The ROC curve validated the predictive accuracy of the risk model, and the area under the curve (AUC) was between 0.65 and 0.8, which indicates a high diagnostic ability. The survival curve showed a significant difference in OS between the 2 groups (Figure 3D-F, $P < 0.01$). Survival was significantly lower in the high-risk group than that in the low-risk group.

Table 1. TCGA miRNA Polygenic Risk Model.

Gene	Coefficient	HR	HR.95L	HR.95H	P
hsa-mir-139	-0.882	0.414	0.144	1.188	0.101
hsa-mir-621	-0.218	0.804	0.606	1.068	0.132
hsa-mir-3607	-1.353	0.259	0.118	0.564	0.001
hsa-mir-326	0.838	2.311	1.450	3.683	0.000
hsa-mir-195	-1.450	0.235	0.053	1.043	0.057
hsa-mir-10b	1.060	2.886	0.936	8.896	0.065
hsa-mir-5589	-0.231	0.793	0.641	0.982	0.033
hsa-mir-497	1.195	3.304	0.835	13.077	0.089
hsa-mir-1269a	0.179	1.196	0.933	1.534	0.159
hsa-mir-592	0.225	1.253	1.020	1.540	0.032
hsa-mir-150	-1.597	0.202	0.082	0.499	0.001
hsa-mir-105-2	-0.464	0.629	0.438	0.904	0.012
hsa-mir-105-1	0.511	1.667	1.136	2.447	0.009
hsa-mir-9-1	7.274	1.442	0.891	2.333	0.054
hsa-mir-9-3	-6.451	0.002	0.000	2.408	0.085
hsa-mir-512-1	0.150	1.162	0.951	1.419	0.143

Table 2. TCGA mRNA Polygenic Risk Model.

Gene	Coefficient	HR	HR.95L	HR.95H	P
CCNB1	0.043	1.043	1.024	1.063	>0.001
MAP2K1	-0.050	0.952	0.905	1.001	0.053
YWHAZ	0.006	1.006	0.998	1.014	0.121
ACTR2	0.017	1.017	1.002	1.031	0.022
VEGFA	0.020	1.020	1.001	1.039	0.040
TP53	-0.045	0.956	0.921	0.992	0.017

Independent Risk Factors Were Used for Prognostic Analysis and Clinical Correlation Analysis of HCC

Seven univariate Cox analysis of age, gender, grade, stage, risk score, vascular invasion and alpha-fetoprotein (AFP) was done to analyze TCGA database data. Relevant variables were selected and multivariate Cox analysis was performed. The results of all COX regression analyses suggested that the risk score could serve as an independent risk factor for HCC prognosis ($p < 0.05$, $HR > 1$) (Figure 4A-F). The higher the risk score, the higher the patient's risk of death. Next, we performed a correlation analysis of clinicopathological characteristics and risk score. Analysis of the TCGA database suggested that patients in the high-risk group were more likely to show poorly differentiated tumor and high AFP levels ($p < 0.05$), suggesting a poor clinical prognosis (Figure 4G, H). However, the ICGC database did not show a correlation between clinical characteristics and risk score, which may be related to the fact that ICGC included less information about the clinical characteristics of patients (Figure 4I). The results from different databases were generally consistent, suggesting that the risk model we constructed is a valid indicator of clinical prognosis.

GSEA

KEGG enrichment analysis and GO enrichment analysis was performed on the gene set, and the targets were the differential

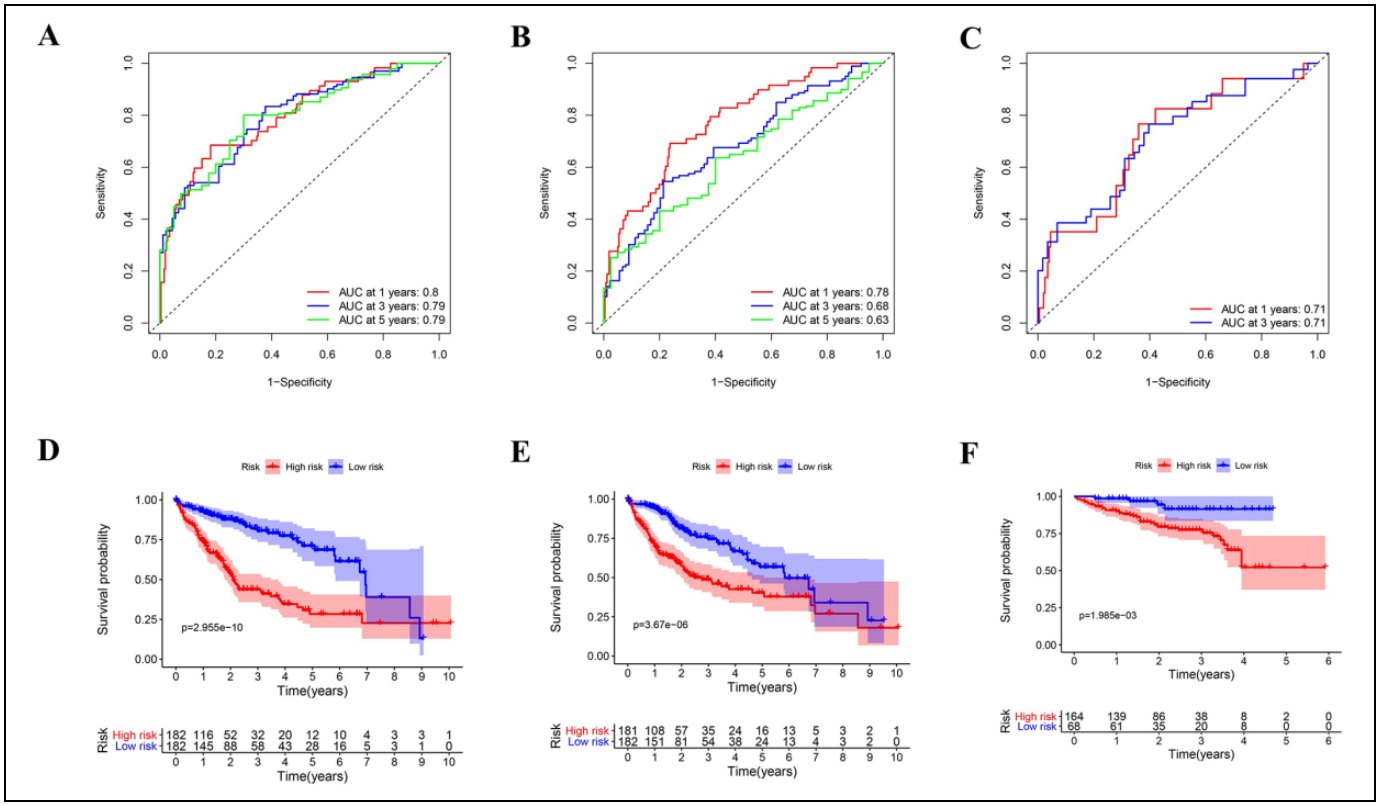

Figure 3. ROC and survival curves for different risk models. (A) ROC curve of TCGA miRNA. (B) ROC curve of TCGA mRNA. (C) ROC curve of ICGC mRNA. (D) Survival curves of TCGA miRNA. (E) Survival curves of TCGA mRNA. (F) Survival curves of ICGC mRNA.

miRNA and mRNA. GO enrichment analysis was performed on miRNAs derived from GEO and TCGA. The distribution results obtained are shown in the pie charts. Although the miRNA data of different databases were not the same, they are enriched on many of the same pathways, such as regulation of nucleobase, nucleoside, nucleotide, and nucleic acid metabolism, signal transduction, transport, cell communication pathway in the biological process(BP), nucleus, cytoplasm, golgi apparatus, lysosome pathway in the cellular component(CC), transcription factor activity, receptor signaling complex scaffold activity, GTPase activity, translation regulator activity pathway in the molecular function(MF) (Figure 5 A-H). It is suggested miRNAs may regulate the occurrence and development of HCC by influencing these pathways. Then, we also carried out GO enrichment analysis and KEGG enrichment analysis on mRNA, and the bubble diagram clearly showed that during the occurrence and development of HCC, mRNA was more involved in endocytosis regulation of protein serine/threonine kinase activity and the endosome membrane (Figure 5 I-J).

Discussion

In 2007, the international consortium for the collaboration of cancer genomes was established and began the new era on studies of cancer genomes. With the integration of data analysis followed by experimental research, more and more studies were performed on the basis of various databases and received reliable results for the further experiments. TCGA database is a multi-omics database, containing genome, transcriptome, proteome, epigenomic and related clinical data of various cancers, and also provides the data for tumorigenesis, development, metastasis, etc.[11] ICGC is a database that aims to map somatic gene mutations in 50 cancer samples from a total of 25,000 patients.[12] Gene Expression Omnibus is a gene expression database created and maintained by the National Center for Biotechnology Information. It was created in 2000 and contained high-throughput gene expression data submitted by institutions worldwide. We extracted gene expression information and clinical information of patients from these 3 databases and conducted a large number of comprehensive analyses, which reduced the contingency brought by the independent analysis of a single database, and verified the results from multiple databases to have more accurate results output.

As it is known that the occurrence of complex diseases is controlled by multiple loci of multiple genes, the effect of single or a few loci are weak, and hard to predict diseases accurately. In this situation, the polygenic risk score (PRS) takes its advantages, presenting as a common strategy to realize disease prediction in a better way.[13] The multi-gene risk score achieved great success in the study of genetic factors of complex diseases for the first time and has showed good application prospects in the risk prediction of complex diseases, such as coronary artery disease[14] and Alzheimer.[15] In recent years,

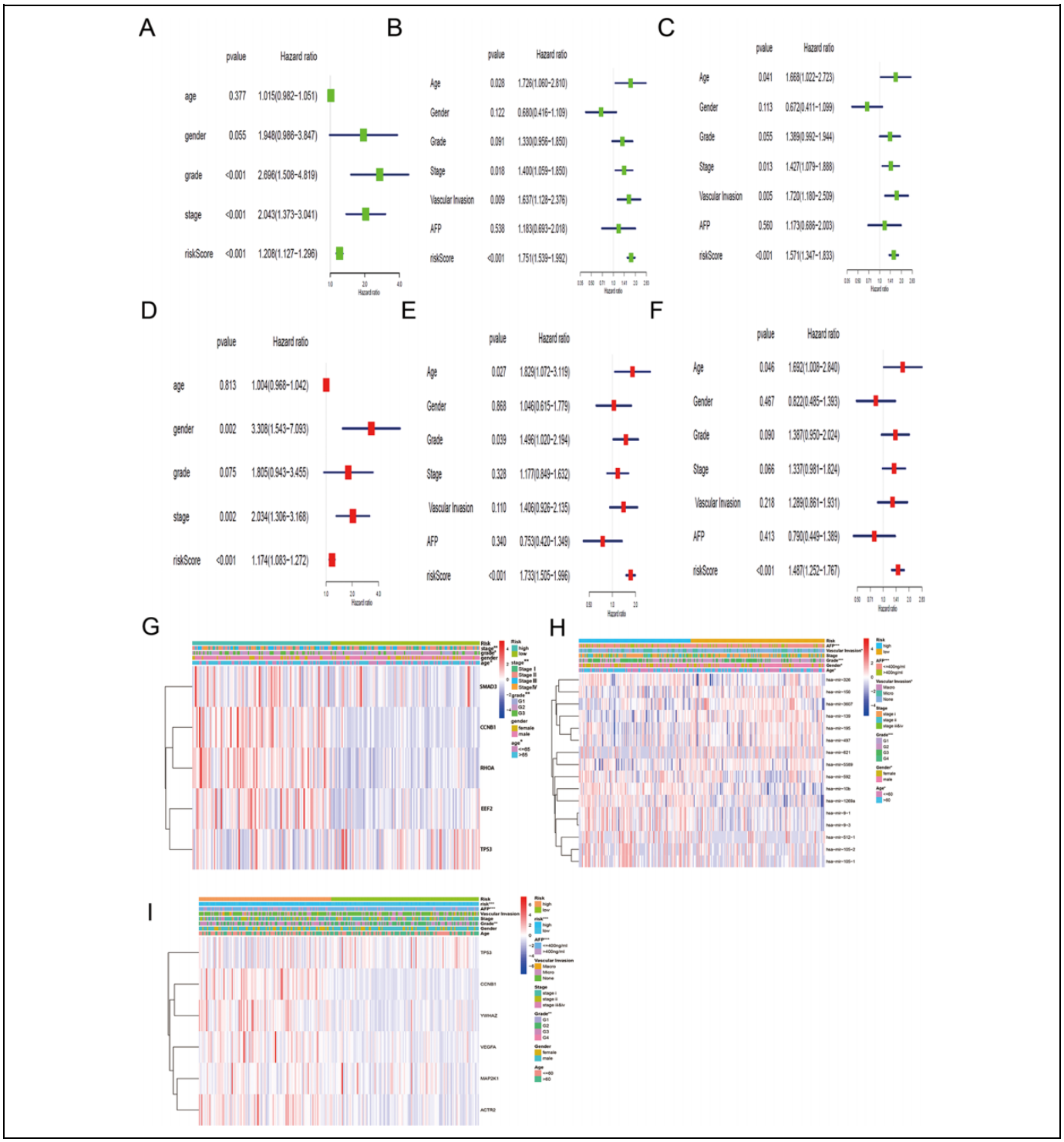

Figure 4. Correlation analysis of risk model and clinicopathological characteristics. (A-C) Univariate Cox regression analyses of TCGA miRNA, TCGA mRNA, and ICGC mRNA. (D-F) Multivariate Cox regression analyses of TCGA miRNA, TCGA mRNA, and ICGC mRNA. (G-I) Clinical correlation heatmaps of TCGA miRNA, TCGA mRNA, and ICGC mRNA. *P < 0.05, **P < 0.01, ***P < 0.001.

genomics research technology has entered a new stage with rapid development, and genome-wide association research, as an important method of molecular epidemiology research, has been gradually applied to the mechanism research of cancer.[16] Researchers from the Massachusetts general hospital and the Broad institute used PRS to identify people who have high risk of breast cancer.[17] Yiwey Shieh[18] revealed that breast cancer risk assessment could provide useful information to facilitate the decision on screening and prevention methods. Dai JC research group used a large prospective cohort to evaluate the

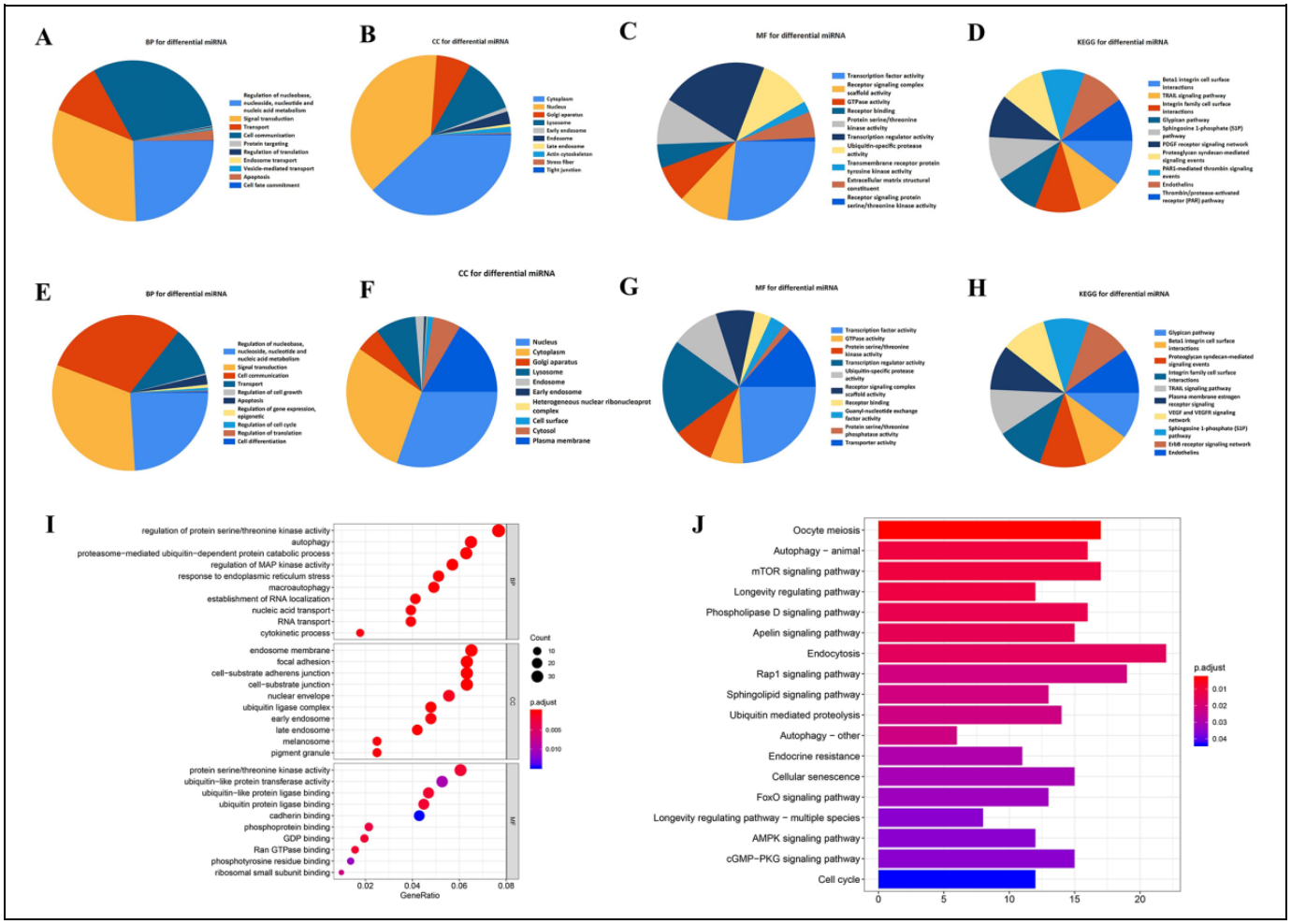

Figure 5. GO and KEGG analysis of differential miRNA and mRNA. (A-D) BP, CC, MF of GO enrichment analysis, and KEGG analysis for differential miRNA from GEO. (E-H) BP, CC, MF of GO enrichment analysis, and KEGG analysis for differential miRNA from TCGA. (I, J) GO enrichment analysis and KEGG analysis for differential mRNA.

effect of PRS in lung cancer risk prediction for the first time in the world.[19] In the field of cancer research, there are few studies on the correlation between PRS and HCC. This study screened the clinical risk factors related to the prognosis of HCC through the construction of a multi-gene risk score, aiming to prove the role of PRS in the screening and prevention strategy decision-making that deserves further consideration. Multiple databases are validated against each other to ensure the accuracy of the models in this study. TCGA and ICGC were used to analyze mRNA expression in HCC, and TCGA and GEO were used to analyze the expression of miRNA in HCC. Then we determined the intersection by Venn diagram, obtained the overlapped mRNA and miRNA in the database, and screened out hub gene by using Degree algorithm in Cytoscape software. Several prognostic models were constructed based on the clinical information from the databases, and risk group and risk scores were calculated. The role of risk values that affecting prognosis was assessed. ICGC database was used for verification at the same time. Further results showed that when we creatively combined miRNA and mRNA expression to build a polygenic risk model, this model has a good true positive rate. Kaplan-Meier curve and Cox analysis suggested that the high-risk group was associated closely with poor prognosis, and the risk score could be used as an independent risk factor for HCC. In addition, gene set enrichment analysis showed miRNA was significantly enriched in regulation of nucleobase, nucleoside, nucleotide, and nucleic acid metabolism, signal transduction, nucleus, cytoplasm, and transcription factor activity. Messenger RNA enrichment was most significant in endocytosis, regulation of protein serine/threonine kinase activity and endosome membrane. In conclusion, the study results suggest that miRNAs and mRNAs are involved in the regulation of multiple biological system functions during tumor development. MicroRNA and mRNA models constructed in this study have certain reference value in the prognostic analysis of HCC.

Conclusion

Our study suggested that the early warning model constructed in this study for co-expression of miRNAs and mRNAs for prejudgment was accurate and effective to a certain degree in the analysis of multiple biological system functions during tumor development. And this is the first report to set up a model which co-expression of miRNA and mRNA were used to forecast the poor prognosis of HCC based on bioinformatics analysis.

Acknowledgments

We thank our anonymous reviewers for their valuable comments on this manuscript, which have led to much many improvements to the article.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Ethics statement

This study uses public databases and does not involve ethics.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This work was supported by grants from Quanzhou science and technology plan project (2019N013 S, 049), Quanzhou science and technology high level talent plan project (2019C032 R), Fujian Natural Science Foundation Project (2018J01198) and Fujian Medical Innovation Project (2017-CX-47).

ORCID iD


Tao Han https://orcid.org/0000-0003-3814-8492

References

1. Beyer C, Zampetaki A, Lin NY, et al. Signature of circulating microRNAs in osteoarthritis. *Ann Rheum Dis*. 2015;74(3):e18.
2. Roderburg C, Trautwein C. Cell-specific functions of miRNA in the liver. *J Hepatol*. 2017;66(3):655-656.
3. Krek A, Grün D, Poy MN, et al. Combinatorial microRNA target predictions. *Nat Genet*. 2005;37(5):495-500.
4. Cai Y, Yu X, Hu S, Yu J. A brief review on the mechanisms of miRNA regulation. *Genom Proteom Bioinf*. 2009;7(4):147-154.
5. Guo H, Ingolia NT, Weissman JS, Bartel DP. Mammalian microRNAs predominantly act to decrease target mRNA levels. *Nature*. 2010;466(7308):835-840.
6. Hanahan D, Weinberg RA. Hallmarks of cancer: the next generation. *Cell*. 2011;144(5):646-674.
7. Reddy KB. MicroRNA (miRNA) in cancer. *Cancer Cell Int*. 2015;15(1):38.
8. Gravitz L. Liver cancer. *Nature*. 2014;516(7529):S1.
9. Leibovitch M, Topisirovic I. Dysregulation of mRNA translation and energy metabolism in cancer. *Adv Biol Regul*. 2018;67:30-39.
10. Jiao Y, Fu Z, Li Y, Meng L, Liu Y. 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. Tomczak K, Czerwińska P, Wiznerowicz M. The Cancer Genome Atlas (TCGA): an immeasurable source of knowledge. *Contemp Oncol (Pozn)*. 2015;19(1A):A68-77.
12. Hu X, Yang H, He J, Lu Y. The cancer genomics and global cancer genome collaboration. *Sci Bull (Beijing)*. 2014;60(1):65-70.
13. Hang D, Shen HB. [Application of polygenic risk scores in risk prediction and precision prevention of complex diseases: opportunities and challenges]. *Zhonghua Liu Xing Bing Xue Za Zhi*. 2019;40(9):1027-1030.
14. Alain B.Letter by Braillon Regarding Article, "Polygenic risk score identifies subgroup with higher burden of atherosclerosis and greater relative benefit from statin therapy in the primary prevention setting." *Circulation*. 2017;136(22):2204-2205.
15. Valentina E, Myers Amanda J, Matt H, John H. Polygenic risk score analysis of pathologically confirmed Alzheimer disease. *Ann Neurol*. 2017;82(2):311-314.
16. Dai J, Shen W, Wen W, et al. Estimation of heritability for nine common cancers using data from genome-wide association studies in Chinese population. *Int J Cancer*. 2017;140(2):329-336.
17. Khera Amit V, Mark C, Aragam Krishna G, et al. Genome-wide polygenic scores for common diseases identify individuals with risk equivalent to monogenic mutations. *Nat Genet*. 2018;50(9):1219-1224.
18. Shieh Y, Hu D, Ma L, et al. Breast cancer risk prediction using a clinical risk model and polygenic risk score. *Breast Cancer Res Treat*. 2016;159(3):513-525.
19. Dai J, Lv J, Zhu M, et al. Identification of risk loci and a polygenic risk score for lung cancer: a large-scale prospective cohort study in Chinese populations. *Lancet Respir Med*. 2019;7(10):881-891.