Exploration and Verification of a Six-RNA Binding Proteins-based Prognosis Evaluating Model for Hepatocellular Carcinoma

Miao Chen

Nanjing university of Chinese Medicine <https://orcid.org/0000-0002-4718-376X>

Shujie Li

Nanjing university of Chinese medicine, Nanjing, Jiangsu province, China

Jiakang Zhang

Nanjing university of Chinese medicine, Nanjing, Jiangsu province, China

Jianbo Han

Department of hepatobiliary surgery, the affiliated Nanjing hospital of Nanjing university of Chinese medicine, Nanjing, Jiangsu province, China

Lili Wang

Clinical research center, the affiliated Nanjing hospital of Nanjing university of Chinese medicine, Nanjing, Jiangsu province, China

Dandan Yin

Clinical research center, the affiliated Nanjing hospital of Nanjing university of Chinese medicine, Nanjing, Jiangsu province, China

Changhua Yi

Clinical research center, the affiliated Nanjing hospital of Nanjing university of Chinese medicine, Nanjing, Jiangsu province, China

Yongxiang Yi (✉ 18120148097@189.cn)

Department of hepatobiliary surgery, the affiliated Nanjing hospital of Nanjing university of Chinese medicine, Nanjing, Jiangsu province, China <https://orcid.org/0000-0001-6133-1207>

Methodology

Keywords: RNA-binding protein, Hepatocellular carcinoma, Bioinformatics analysis, Prognostic model

Posted Date: December 3rd, 2020

DOI: <https://doi.org/10.21203/rs.3.rs-117829/v1>

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

[Read Full License](#)

Abstract

Background

RNA binding protein (RBP) plays a crucial role in tumorigenesis at post-transcriptional level in various cancer types. Nevertheless, the role of RBPs in liver hepatocellular carcinoma (LIHC) remains obscure. We attempted to uncover the association between RBPs and the prognosis of LIHC patients.

Methods

We analyzed the transcriptome and corresponding clinical data of LIHC patients from the cancer genome atlas (TCGA) (training cohort) and international cancer genome consortium (ICGC) (validating cohort) database with a series of bioinformatics methods. Differently expressed RNA-binding proteins (DERBPs) were screened and subjected to functional enrichment analysis and co-expression network establishment. Overall survival (OS) related DERBPs and our prognosis risk model were confirmed by univariate, LASSO and multivariate regression analysis in training cohort. Survival analysis, Receiver operating characteristic curve (ROC) and nomogram were conducted in both training and validating groups to confirm the performance of our model. Human protein atlas (HPA) database and Kaplan-Meier plotter were used to verify the expression and prognostic significance of the hub RBPs respectively.

Results

There were 330 RBPs were found significantly different in TCGA. Functional analysis indicated most of the DERBPs were majored in RNA processing, alternative splicing and metabolism, etc. 6 RBPs (UPF3B, MRPL54, ZC3H13, DHX58, PPARGC1A, EIF2AK4) were recognized as OS related and enrolled into our prognostic model. Survival analysis showed the risk signature was negatively correlated with the OS of LIHC patients in both training ($p = 5.808e-06$) and validating ($p = 3.38e-03$) groups. The area under curves (AUC) of the receiver operating characteristics (ROC) curve in training and validating cohorts was 0.756, 0.781 respectively which indicating the good performance of our model. The risk signature was an independent hazardous factor in multivariate COX regression analysis either in TCGA (HR = 1.626;95% CI 1.394 -1.897, $p < 0.001$) or ICGC (HR = 1.939;95% CI 1.324 -2.838, $p < 0.001$). Nomogram and calibration curve indicated our model had best performance in predicting 3-year survival rate.

Conclusions

We constructed a six-RBPs based risk signature model which had moderate efficiency in LIHC patients' prognosis forecasting which may assist practitioners to make better decision in the management of LIHC.

Background

Liver hepatocellular carcinoma (LIHC) is a disease that has plagued humans for long time. The incidence of HCC ranks sixth among that of malignant tumors and its cancer-related mortality places the top four.

In the meanwhile, its five-year survival rate merely meets 18%, second only to pancreatic cancer. Based on these tough truths, HCC poses a huge threat to the human life and social economy[1]. In 2020, the United States is expected to have 42,000 newly diagnosed liver cancer patients and 30,000 HCC related deaths [2]. Due to the huge population base in China, the situation is more serious than that in the United States. The year 2018 witnessed an increase of approximate 390 thousand newly diagnosed HCC patients and 270 thousand HCC related deaths in China [3]. At the same time, HCC is featured with high heterogeneity, apt to recur and metastasis after operation, the incidence rate has increased in recent years. The 5-year overall survival (OS) of LIHC at early phase after curative resection or radiofrequency ablation can reach 70% which emphasizes the importance of early detection [4]. However, owing to the lack of apparent early symptoms and the absence of efficient, sensitive and specific diagnostic methods, many of the LIHC patients are often in advanced phase when diagnosed, moreover, the 5-year OS of these patients is lower than 12.5% [5]. Thus, clarification of the underlying molecular mechanism of tumorigenesis in HCC can help practitioners to detect vulnerable population at early time and provide effective treatment to improve patients' outcome.

RNA binding protein (RBP) is a cluster of molecules interacting with various kinds of RNA (mRNA, tRNA, rRNA, etc.) and mediating the function of them at post-transcriptional level [6]. The combination of these proteins and RNA is the prerequisite for the function of RBP, which accomplished through the specific domains [7] within the proteins including RNA recognition motif, hnRNP K homology, zinc fingers, S1 domains and double-stranded RNA-binding domain, etc. Existing articles have proved that RBPs take part in various biological processes, like splicing of the RNA[8], subcellular localization of mRNA [9], translation and m6A modification [10] of RNA, stabilizing of target RNA [11]. Considering the dysfunction of RBPs may disturb the balance of transcription or translation process and therefore induces tumorigenesis, the role of RBPs in malignant disease were investigated by researchers. RBP CPEB3 [12] inhibits the proliferation of colorectal cancer through the JAL/STAT pathway. Zmat3 [13] is a crucial factor in maintaining the stability of transcript which participates in the tumor suppressing process mediated by p53 in many types of cancers. RBPs NONO, QKI, and RBMX [14] may deteriorate the outcome of breast cancer patients by interacting with long non-coding RNA ST8SIA6-AS1. Furthermore, a recent study identified RBM43 [15] as an anti-oncogene in liver cancer as its knockdown can boost the proliferation of tumor cell both in vivo and vitro, however, the comprehensive understanding of RBPs in HCC is yet to be achieved.

On this basis, we attempted to explore the profile of RBPs in LIHC and the correlation between RBPs and patients' prognosis. Transcriptome and corresponding clinicopathological information of LIHC patients were extracted from online database TCGA for a series of bioinformatics analysis. RBPs with aberrant expression level in cancer tissue were screened and their biological function and related pathway were studied. Moreover, after COX regression analysis, six RBPs were highlighted as hub genes and used to reflect the risk signature of patients' outcome. The expression and clinical data from ICGC were used for external validation of our model. According to the outcome in our study, the RBP-based model had moderate performance in predicting prognosis of LIHC patients which may supplement our knowledge of managing HCC.

Results

Identification of DERBPs in LIHC

In order to identify the DERBPs between hepatocellular carcinoma and normal tissue, The transcriptome and relevant clinical data of LIHC patients were obtained from TCGA database, which consisted of 50 normal and 374 cancer samples. The primary expression files were merged and the initial gene ids were transformed into gene symbols by Perl scripts for further bioinformatics analysis. According to the 1542 reported RBPs (Additional file 1: Table S1), we extracted the RBP expression data. A total of 330 RBPs satisfied the significant standard of our study ($|\text{LogFC}| > 0.5$ & $P\text{-val} < 0.05$), including 122 downregulated and 208 upregulated RBPs (Additional file 2: Table S2). Furthermore, the DERBPs were visualized by volcano plot (Fig. 1a) and heatmap (Fig. 1b) through R packages.

Functional analysis of DERBPs

After the identification of the DERBPs, we intended to investigate the biological functions and potential mechanisms of them. The up and down-DERBPs were subjected to gene ontology and KEGG pathway enrichment analysis which conducted by R package 'clusterProfiler' and the results were shown by bar plots (Fig. 2) and bubble plots (Additional file 3: Fig S1). As shown in the graph, the up-DERBPs (Fig. 2a) were mainly enriched in the biological process (BP) such as non-coding RNA metabolic process, non-coding RNA processing, RNA splicing via transesterification reactions with bulged adenosine as nucleophile, mRNA splicing via spliceosome, ribonucleoprotein complex biogenesis, RNA 3'-end processing, tRNA metabolic process and RNA catabolic process while the down-DERBPs (Fig. 2b) were significantly enriched in the regulation of translation, regulation of cellular amide metabolic process, RNA phosphodiester bond hydrolysis, defense response to virus, nucleic acid phosphodiester bond hydrolysis. When it comes to the cellular component analysis, the majority of the up-DERBPs were involved in spliceosomal complex, cytoplasmic ribonucleoprotein granule, ribonucleoprotein granule, ribosome, small nuclear ribonucleoprotein complex and Sm-like protein family complex, albeit the down-DERBPs were significantly enriched in cytoplasmic ribonucleoprotein granule, P-body, CCR4-NOT complex, organellar ribosome and mitochondrial ribosome. In the section of molecular function analysis, the up-DERBPs apparently functioned in catalytic activity on RNA, mRNA 3'-UTR binding, catalytic activity on a tRNA and translation regulator activity as the down-DERBPs mainly performed in catalytic activity on RNA, single and double-stranded RNA binding, nuclease activity and ribonuclease activity. Furthermore, the KEGG analysis results (Fig. 2c, 2d) unveiled that these DERBPs participated in pathways related to RNA degradation, mRNA surveillance, RNA transport, Ribosome biogenesis in eukaryotes, Hepatitis C, Influenza A, RIG-I-like receptor signaling etc.

PPI networks and relevant subnetworks visualization

In the interest of uncover the internal relationship between the DERBPs in LIHC, all of the DERBPs were uploaded to the STRING database for knitting a PPI network. As a result, the PPI network was featured by

311 nodes and 2942 edges. Then the nodes and edges details were processed by Cytoscape software to plot a ring diagram (Fig. 3a). For the sake of screening the dense region of the PPI network, we adopted the ‘MCODE’ plugin to discover the key subnetworks. The top 3 important subnetworks were shown in the picture (Fig. 3b) and the details were in Additional file 4: Table S3. GO analysis indicated that the most important subnetwork was concentrated in ribonucleoprotein complex biogenesis, RNA splicing via transesterification reactions with bulged adenosine as nucleophile and mRNA splicing via spliceosome, furthermore, the KEGG results showed it was enriched in pathways associated with spliceosome, mRNA surveillance and Ribosome biogenesis in eukaryotes, etc. (Additional file 5: Table S4, Additional file 6: Table S5)

Prognostic RBPs filtration and Prognostic model construction

For the convenience of subsequent analysis, the expression data and patients’ survival data were combined. All of the node DERBPs (n = 311) were subjected to univariate cox regression analysis and 20 DERBPs (Fig. 4a) were obtained as prognosis-related altogether (p < 0.001). Then the 20 DERBPs endured further filtration by LASSO regression analysis at 1000 mixit through the ‘glmnet’ R package and at last we got 14 candidates (Fig. 4b, 4c, Additional file 7: Table S6). These 14 candidates were analyzed by multiple stepwise Cox regression. Finally, 6 hub genes appeared (Fig. 4d), including UPF3B (UPF3B regulator of nonsense mediated mRNA decay), MRPL54 (Mitochondrial ribosomal protein L54), ZC3H13 (Zinc finger CCCH-type containing 13), DHX58 (DExH-box helicase 58), PPARGC1A (PPARG coactivator 1 alpha) and EIF2AK4 (Eukaryotic translation initiation factor 2 alpha kinase 4), the details of them were shown in Table (Table 1). These six DERBPs were exploited to build the prognostic model, the risk score calculating formula was that : Risk score = 0.3750 * Exp-val of UPF3B + (-0.3908 * Exp-val of MRPL54) + (-0.2019 * Exp-val of ZC3H13) + (-0.2282 * Exp-val of DHX58) + (-0.1438 * Exp-val of PPARGC1A) + (-0.5214 * Exp-val of EIF2AK4).

Table 1
Overview of the six hub RBPs

ID	coef	HR	HR.95L	HR.95H	pvalue
UPF3B	0.375064	1.45508453	1.01053138	2.09520559	0.0437704
MRPL54	-0.390805	0.67651163	0.50197593	0.91173293	0.01026037
ZC3H13	-0.201938	0.8171456	0.63319873	1.05452981	0.12068099
DHX58	-0.228279	0.79590151	0.59731792	1.06050596	0.11904026
PPARGC1A	-0.143802	0.86605891	0.74338511	1.0089764	0.06499514
EIF2AK4	-0.521421	0.59367634	0.36499338	0.96563834	0.03565464

Based on the algorithm above, the training cohort containing 370 HCC patients was divided into high and low risk subgroups according to median of the risk value acquired. Kaplan-Meier survival analysis indicated that the subgroup with lower risk score had favorable overall survival rate compared with high risk group (Fig. 5a). Receiver operating characteristics (ROC) curve was conducted to verify the performance of our prognosis signature model and the area under the curve (AUC) in the training cohort was 0.756 (Fig. 5b). For the purpose of confirming whether our model still worked in other cohort, we also downloaded the expression and clinical data of HCC in ICGC database ([LIRI-JP] Liver Cancer - RIKEN, JP) as validation cohort which consisted of 232 cancer patients. The data was handled the same as the training sets, the results of survival curve (Fig. 6a) and ROC analysis (AUC = 0.781) (Fig. 6b) of the validation group further testified the power of our model in predicting prognosis. Therefore, we plotted the heatmap (Fig. 5c, 6c) of the six hub genes, the distribution of risk score and patients' status (Fig. 5d, 6d) in both training and validation group.

The risk signature is an independent prognostic factor

In our next step, we explored the relationship between survival prognosis and clinical traits including age, gender, histologic grading and TNM stage. Uni- and multivariate cox regression analysis were conducted in both training and validating cohort. In training group, the clinical-pathological parameters TNM stage and risk signature were significant hazardous factors through univariate cox regression (Fig. 7a), while age, TNM stage and risk score were all significant in multivariate cox regression (Fig. 7b). While in validating cohort, clinical traits including gender, TNM stage and risk score were all related to patients' overall survival either in uni- or multivariate cox regression (Fig. 7c, 7d).

Construction of the nomogram and online verification of the of hub RBPs

Here we also provided the nomogram for the calculation of survival prognosis signature based on the six RBPs. Every single value obtained by the prognosis model formula had its corresponding normalized value on the horizontal axis at the top, then the possibility of 1-, 3- and 5-year survival could be figured out (Fig. 8a). Furthermore, the calibration curves indicated the nomogram had the best performance in predicting 3-year overall survival (Fig. 8b). In the following step, we verified the respective prognostic significance of the six hub RBPs (UPF3B, MRPL54, ZC3H13, DHX58, PPARGC1A, EIF2AK4) by online tool Kaplan-Meier Plotter (Fig. 9). The results indicated that all of the six RBPs were OS related. So as to ascertain the expression of the six RBPs at tissue level, the online immunohistochemistry database HPA was utilized for validation work (Fig. 10), however, the IHC slices of PPARGC1A was still on the waiting list. As it turned out, UPF3B, MRPL54 and EIF2AK4 had higher expression in HCC tissue than normal tissue, ZC3H13 and DHX58 were relatively decreased in LIHC samples.

Discussion

Nowadays, hepatocellular carcinoma still threatens millions of people's life every year as the sixth prevalent malignant disease [16]. The current therapy on HCC mainly consists of surgery, ablation, interventional therapy and systematic treatment [17]. Unfortunately, these methods have their limitation because of the heterogeneity of liver cancer. Liver transplantation [18] is supposed to be most effective in countering HCC, however, the long waiting time and the contradiction between shortage of donors and huge population of patients restrict its application. Thus, early detection of the disease is still highlighted in HCC management, nevertheless effective biomarkers are lacked. The dysfunction of various RNA is pivotal in tumorigenesis, which exists in many cancer types [19–20]. RBPs are the main actor in regulating the performance of RNA at post transcription level [21] which accomplished by a series of modification including RNA alternative splicing, polyadenylation, localization and stabilization, etc. Existing literature have concluded that RBPs are crucial contributors in various phenotypes of malignant disease, like proliferation [22], apoptosis escape [23], angiogenesis [24], etc. However, the role of RBPs in HCC has not been systemically explored. So as to investigate the potential of RBPs as biomarkers in HCC, we analyzed the transcriptome and clinical data of LIHC patients from TCGA. According to our filtration standard, there were 330 DERBPs recognized between normal and cancerous samples. Then the biological processes and metabolic pathways involved these DERBPs were globally analyzed and the interaction between them was displayed by PPI network. After combining the sequencing and clinical data, survival-related RBPs were screened by univariate COX regression and KM analysis. These RBPs were next enrolled to LASSO and multivariate COX regression to get the core RBPs behind them. At last we build our six RBPs-based risk model, which had well performance in forecasting LIHC patients' prognosis.

As what was shown in bar and bubble plot, GO functional analysis pointed out that a large proportion of DERBPs took part in non-coding RNA metabolic processing, non-coding RNA processing, RNA 3'-end processing, tRNA metabolic process, RNA splicing, ribonucleoprotein complex biogenesis, RNA catabolic process, regulation of translation, regulation of cellular amide metabolic process, RNA phosphodiester bond hydrolysis and defense response to virus, etc. In line with other studies, alternative splicing of RNA, RNA processing and translational regulation were critical in the occurrence and progression of malignant diseases [21, 25]. The elevated expression of RBP KHSRP can facilitate the proliferation of small cell lung cancer by promoting the maturation of miR-26a, which suppressed the function of PTEN subsequently [26]. Aberrant change of splicing target of RBP SRSF1 induces the proliferation of HCC via generating the anti-apoptotic isoforms of target genes [27]. By interacting with the internal ribosome entry site of target mRNA laminin B1 at translation level, RBP La protein increases the expression of laminin B1 and therefore promotes the invasiveness of HCC [28]. Ribonucleoprotein is the functional state of RBP, its formation relies on the specific domains on RBP. The dysfunction of ribonucleoprotein is reported in various cancer types [6]. In addition, the outcome of KEGG pathway enrichment indicated most of DERBPs were involved in pathways associated with RNA spliceosome, mRNA surveillance, RNA degradation, Ribosome biogenesis in eukaryotes and RIG – I-like receptor signaling pathway, etc.

We also investigated the co-expressional relations of these DERBPs and the STRING returned us a PPI network with 311 nodes. The top three important modules with MCODE score over 10 were further

analyzed by GO and KEGG. According to the report, module 1 was mainly involved in ribonucleoprotein complex biogenesis, various kinds of RNA splicing and processing. Module 2 was associated with defense response to virus response, interferon production and the signaling of interferon, etc. In the meanwhile, module 3 majored in translational determination and elongation, process of mitochondrial translation, etc. In agreement with previous studies, some of these RBPs functioned in liver cancer. BOP1 was frequently expressed in HCC and promoted the invasiveness of cancer cells, which indicated poor prognosis in most cases [29]. High expression level of IFIT3 amplified the therapeutic effect of interferon- α in combating HCC, thus can be utilized to screen LIHC patients adapted to interferon treatment [30]. EIF5A2 was associated with poor survival and can advance the proliferation of HCC by accelerating the metabolic rate of cancer cell [31]. These researches corroborate the important role of RBPs in HCC.

In order to further investigate the relationship between these DERBPs and outcome of patients, we performed uni-, multi-COX regression and LASSO regression. There were six RBPs highlighted as core genes, including UPF3B, MRPL54, ZC3H13, DHX58, PPARGC1A and EIF2AK4. Previous study has demonstrated that ZC3H13 was associated with OS of LIHC patients [32] and PPARGC1A was related to people's susceptibility of HCC [33], however the roles of the others in HCC are yet to be confirmed. Hereafter, these six RBPs together with their coefficient returned by multi-step COX regression were enrolled into our risk model, which was trained in TCGA datasets. As the result of time-dependent ROC test, our model showed well performance in distinguishing patients of different prognosis in both training and validating cohorts. Furthermore, a nomogram based on these hub RBPs were plotted for the purpose of approximately calculating the likelihood of patients' survival. At last, the reliability of these six RBPs as prognostic genes were verified in Kaplan-Meier plotter and HPA database and the outcome agreed with our study.

Despite the encouraging results of our study, there were some unavoidable limitations. First, we got all the data in our study from public databases like TCGA and ICGC and we did not verify the prognostic model in other datasets or our own reliable prospective clinical trial. Second, some of the samples bonded with incomplete clinical information were abandoned during the analysis which may increase the bias error of our study.

Conclusions

Our study suggested RBPs as prognostic genes in HCC for the crucial function of them in various cancers, and provided a six-RBPs based risk model which had moderate efficiency in prognosis forecasting. Our work may enrich the knowledge of the management of LIHC patients and contribute to the understanding of the underlying mechanism of HCC initiation, nevertheless, uncovering the role of RBPs in liver cancer requires further efforts henceforth.

Methods

Data downloading and processing

The 50 normal and 374 LIHC patients' mRNA expression profiles and the parallel clinical documents were obtained from TCGA (<https://portal.gdc.cancer.gov>). All the individual sequencing file were merged into an expression matrix by Perl script and the probe ids were transformed into gene symbols by referring to the GTF file gained from GENECODE (<https://www.gencodegenes.org>, release 35). The list of 1542 RNA binding proteins with reference were picked up from reported article. Their expression information was winkled out for variance analysis. The whole TCGA cohort was used as training set. Then we extracted the sequencing data and corresponding donors' details (LIRI-JP) from International Cancer Genome Consortium (ICGC, <https://icgc.org>) which included 232 liver cancer patients for external validation. The handling of the data resembled TCGA.

DERBPs determination

The differently expressed RBPs (DERBPs) between liver cancer and normal tissue in TCGA were filtered through Wilcoxon sum rank test by R. The Limma package (<https://bioconductor.org/packages/release/bioc/html/limma.html>) was used to average the expression value of repeated RBPs and normalization. The DERBPs met the standard $|\log_2 \text{fold change (FC)}| > 0.5$ & False Discovery Rate (FDR) < 0.05 were considered significant in this study. The profile of the DERBPs was displayed in heatmap and volcano plot by R package 'pheatmap'.

GO and KEGG analysis

The Gene oncology (GO) and Kyoto encyclopedia of genes and genomes (KEGG) pathway enrichment analysis of identified DERBPs were further analyzed with R package 'clusterProfiler' and 'org.Hs.eg.db'. The GO analysis consisted of biological processes (BP), cellular component (CC) and molecular function (MF), only when $P\text{-val} < 0.05$ was determined significant [34]. Referring to p value, the top ten significant results were listed in bubble plots and bar plots drawn by R package 'ggplot2' and 'enrichplot'. The enrichment analysis of sub networks of the PPI was performed in the same way.

PPI network construction and visualization

Soon after we attained the name list of DERBPs, we uploaded them to STRING website [35](<https://string-db.org>) in order to investigate the interaction between them, moreover, the isolated nodes were rejected. Then the protein-protein interaction (PPI) network was established in circular layout by Cytoscape software (version 3.7.1) based on the tsv file exported from STRING. The plugin 'MCODE' [36] was installed for detection of crucial sub networks with the MCODE score > 5 and nodes more than 10. The top three sub modules were shown in nested loops.

Prognostic model construction and validation

The expression files of DERBPs screened from STRING were combined with corresponding clinical data. The samples with fragmentary clinical information were abandoned. Then the expression-clinical matrix was subjected to univariate cox regression analysis by 'survival' R package. When threshold of $P\text{-val}$ at 0.001, there were 20 RBPs selected which further underwent LASSO regression [37] ($\text{maxit} = 1000$, by 'glmnet' R package) to avoid overfitting of the curve. Afterwards, we performed multivariate cox

regression on the remaining RBPs to compose our prognostic risk model. The algorithm for calculating the risk score was: $\text{Coef } 1 * \text{Exp-val } 1 + \text{Coef } 2 * \text{Exp-val } 2 \dots + \text{Coef } n * \text{Exp-val } N$. The 'Coef' stood for the coefficient of each gene returned by multivariate cox regression and the 'Exp-val' represented the expression value of the gene in the expression matrix. Both the TCGA training and ICGC external validation cohort were divided into low and high-risk subgroups based on their respective risk-score midpoint calculated by the formula. The overall survival rate was compared between two subgroups by Log-Rank test through 'survival' and 'survminer' R packages. Furthermore, with the intention of confirming the accuracy of prognosis forecasting of our risk model, the Receiver Operating Characteristic curve (ROC) was conducted through 'survivalROC' R package in both training and validation datasets. Besides, the heatmap of the candidate genes, the allocation of the risk signature and the status of LIHC patients in both TCGA and ICGC units were drawn through 'pheatmap' R package.

Recognition of independent prognostic factors and nomogram plotting

So as to assess the independence of our risk model as hazardous factor, clinical parameters like age, gender, TNM stage and riskscore were analyzed by uni- or multivariate cox regression in both TCGA and ICGA samples and $p\text{-val} < 0.05$ was considered significant. In addition, the nomogram and calibration curves were plotted through 'rms' R package for prediction of patients' survival rate of different year.

Online confirmation of the six hub genes

For preliminary validation of the expression of the hub genes in our model, we referred to the immunohistochemical results from Human Protein Atlas [38] (<https://www.proteinatlas.org>). Also, the Kaplan-Meier plotter [39] (<https://kmplot.com>) was used for overall survival analysis.

Abbreviations

LIHC

Liver hepatocellular carcinoma

RBP

RNA binding protein

PPI

Protein-protein interaction

TCGA

The cancer genome atlas

ICGC

International cancer genome consortium

GO

Gene ontology

KEGG

Kyoto Encyclopedia of Genes and Genomes

AUC

Area under curve

ROC

Receiver operating characteristic curve

DERBP

Differently expressed RNA binding protein

OS

Overall survival

IHC

Immunohistochemistry

UPF3B

UPF3B regulator of nonsense mediated mRNA decay **MRPL54**:Mitochondrial ribosomal protein L54

ZC3H13

Zinc finger CCH-type containing 13

DHX58

DExH-box helicase 58

PPARGC1A

PPARG coactivator 1 alpha

EIF2AK4

Eukaryotic translation initiation factor 2 alpha kinase 4

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Availability of data and materials

The dataset supporting the conclusions of this article are available in the TCGA (<https://portal.gdc.cancer.gov/repository>) and ICGC repository (<https://icgc.org>).

Competing interests

The authors declare that they have no competing interests.

Funding

Jiangsu Science and Technology Project of Clinical Medicine Foundation, Science and Technology Department of Jiangsu Province (BL2014005).

Nanjing technology development plan for 2020 (ZX2020009).

Authors' contribution

The study was designed by Steven.Chen, LSJ and ZJK. CM and LSJ wrote the scripts and analyzed the data. Data analysis was performed by CM and YDD. The initial manuscript was written by CM, HJB and YCH. WLL and YYX were responsible for the revision of the manuscript. All of the authors approved the final article.

Acknowledgments

We appreciate the contribution of the data providers on TCGA and ICGC, and thank the paper reviewers for their precious comments.

Author information

¹Nanjing university of Chinese medicine, Nanjing, Jiangsu province, China.

²Department of hepatobiliary surgery, the affiliated Nanjing hospital of Nanjing university of Chinese medicine, Nanjing, Jiangsu province, China

References

1. El-Serag HB, Kanwal F. Epidemiology of hepatocellular carcinoma in the United States: Where are we? Where do we go? *Hepatology*. 2014;60(5).
2. Siegel RL, Miller KD, Jemal A. Cancer statistics. 2020. *CA: A Cancer Journal for Clinicians*. 2020;70(1):7–30.
3. Feng RM, Zong YN, Cao SM, Xu RH. Current cancer situation in China: good or bad news from the 2018 Global Cancer Statistics? *Cancer Commun (Lond)*. 2019 Apr 29;39(1):22.
4. Tsuchiya N, Sawada Y, Endo I, Saito K, Uemura Y, Nakatsura T. Biomarkers for the early diagnosis of hepatocellular carcinoma. *World Journal of Gastroenterology*. 2015;(37):10573–83.
5. Wang W, Wei C. Advances in the early diagnosis of hepatocellular carcinoma. *Genes Diseases*. 2020;7(3):308–19.
6. Gerstberger S, Hafner M, Tuschl T. A census of human RNA-binding proteins. *Nat Rev Genet*. 2014;15(12):829–45.
7. Lunde BM, Moore C, Varani G. RNA-binding proteins: modular design for efficient function. *Nat Rev Mol Cell Biol*. 2007 Jun;8(6):479–90.

8. Mei C, Song P-Y, Zhang W, Zhou H-H, Li X, Liu Z-Q. Aberrant RNA Splicing Events Driven by Mutations of RNA-Binding Proteins as Indicators for Skin Cutaneous Melanoma Prognosis. *Front Oncol.* 2020;10:568469.
9. Dermit M, Dodel M, Lee FCY, Azman MS, Schwenzer H, Jones JL, et al. Subcellular mRNA Localization Regulates Ribosome Biogenesis in Migrating Cells. *Dev Cell.* 2020 Nov 9;55(3):298–313.e10.
10. Yang C, Hu Y, Zhou B, Bao Y, Li Z, Gong C, et al. The role of m6A modification in physiology and disease. *Cell Death Dis.* 2020 Nov 8;11(11):960.
11. Pereira B, Billaud M, Almeida R. RNA-Binding Proteins in Cancer: Old Players and New Actors. *Trends Cancer.* 2017;3(7):506–28.
12. Fang Y, Zhong Q, Wang Y, Gu C, Liu S, Li A, Yan Q. CPEB3 functions as a tumor suppressor in colorectal cancer via JAK/STAT signaling. *Aging (Albany NY).* 2020 Nov 3;12(21):21404–21422.
13. Biegging-Rolett KT, Kaiser AM, Morgens DW, Boutelle AM, Seoane JA, Van Nostrand EL, et al. Zmat3 Is a Key Splicing Regulator in the p53 Tumor Suppression Program. *Mol Cell.* 2020 Nov 5;80(3):452–469.e9.
14. Chen Z, Huang J, Feng Y, Li Z, Jiang Y. Profiling of specific long non-coding RNA signatures identifies ST8SIA6-AS1 as a novel target for breast cancer. *J Gene Med.* 2020 Oct 10;e3286.
15. Feng H, Liu J, Qiu Y, Liu Y, Saiyin H, Liang X, et al. RNA-binding motif protein 43 (RBM43) suppresses hepatocellular carcinoma progression through modulation of cyclin B1 expression. *Oncogene.* 2020 Aug;39(33):5495–506.
16. Plaz Torres MC, Bodini G, Furnari M, Marabotto E, Zentilin P, Strazzabosco M, et al. Surveillance for Hepatocellular Carcinoma in Patients with Non-Alcoholic Fatty Liver Disease: Universal or Selective? *Cancers (Basel).* 2020 May 31;12(6).
17. Villanueva A. Hepatocellular Carcinoma N, *Engl JM.* 2019 Apr 11;380(15):1450–1462. doi: 10.1056/NEJMra1713263.
18. Mazzaferro V, Sposito C, Zhou J, Pinna AD, De Carlis L, Fan J, et al. Metroticket 2.0 Model for Analysis of Competing Risks of Death After Liver Transplantation for Hepatocellular Carcinoma. *Gastroenterology.* 2018;154(1):128–39.
19. Uddin MN, Wang X. The landscape of long non-coding RNAs in tumor stroma. *Life Sci.* 2020 Nov 6;118725.
20. Li Y, Yin Z, Fan J, Zhang S, Yang W. The roles of exosomal miRNAs and lncRNAs in lung diseases. *Signal Transduct Target Ther.* 2019 Nov;13(1):47. 4(.
21. Kang D, Lee Y, Lee J-S. RNA-Binding Proteins in Cancer: Functional and Therapeutic Perspectives. *Cancers (Basel).* 2020 Sep 21;12(9).
22. Wang ZL, Li B, Luo YX, Lin Q, Liu SR, Zhang XQ, Zhou H, Yang JH, Qu LH. Comprehensive Genomic Characterization of RNA-Binding Proteins across Human Cancers. *Cell Rep.* 2018 Jan 2;22(1):286–298. doi: 10.1016/j.celrep.2017.12.035.

23. Wong RSY. Apoptosis in cancer: from pathogenesis to treatment. *J Exp Clin Cancer Res.* 2011 Sep;26:30:87.
24. Boriushkin E, Zhang H, Becker M, Peachey J, Shatat MA, Adams RH, Hamik A. Kruppel-like factor 4 regulates developmental angiogenesis through disruption of the RBP-J-NICD-MAML complex in intron 3 of Dll4. *Angiogenesis.* 2019 May;22(2):295–309. doi: 10.1007/s10456-018-9657-y.
25. van Kouwenhove M, Kedde M, Agami R. MicroRNA regulation by RNA-binding proteins and its implications for cancer. *Nat Rev Cancer.* 2011 Aug;5(9):644–56. 11(.
26. Tong L, Luo Y, Wei T, Guo L, Wang H, Zhu W, et al. KH-type splicing regulatory protein (KHSRP) contributes to tumorigenesis by promoting miR-26a maturation in small cell lung cancer. *Mol Cell Biochem.* 2016 Nov;422(1–2):61–74.
27. Malakar P, Shilo A, Mogilevsky A, Stein I, Pikarsky E, Nevo Y, Benyamini H, Elgavish S, Zong X, Prasanth KV, Karni R. Long Noncoding RNA MALAT1 Promotes Hepatocellular Carcinoma Development by SRSF1 Upregulation and mTOR Activation. *Cancer Res.* 2017 Mar 1;77(5):1155–1167. doi: 10.1158/0008-5472.CAN-16-1508.
28. Petz M, Them N, Huber H, Beug H, Mikulits W. La enhances IRES-mediated translation of laminin B1 during malignant epithelial to mesenchymal transition. *Nucleic Acids Res.* 2012 Jan;40(1):290–302.
29. Chung K-Y, Cheng IK-C, Ching AK-K, Chu J-H, Lai PB-S, Wong N. Block of proliferation 1 (BOP1) plays an oncogenic role in hepatocellular carcinoma by promoting epithelial-to-mesenchymal transition. *Hepatology.* 2011 Jul;54(1):307–18.
30. Yang Y, Zhou Y, Hou J, Bai C, Li Z, Fan J, et al. Hepatic IFIT3 predicts interferon- α therapeutic response in patients of hepatocellular carcinoma. *Hepatology.* 2017;66(1):152–66.
31. Cao T-T, Lin S-H, Fu L, Tang Z, Che C-M, Zhang L-Y, et al. Eukaryotic translation initiation factor 5A2 promotes metabolic reprogramming in hepatocellular carcinoma cells. *Carcinogenesis.* 2017;38(1):94–104.
32. Huang H, Bai Y, Lu X, Xu Y, Zhao H, Sang X. N6-methyladenosine associated prognostic model in hepatocellular carcinoma. *Ann Transl Med.* 2020 May;8(10):633.
33. Zhang S, Jiang J, Chen Z, Wang Y, Tang W, Chen Y, et al. Relationship of PPARG, PPARGC1A, and PPARGC1B polymorphisms with susceptibility to hepatocellular carcinoma in an eastern Chinese Han population. *Onco Targets Ther.* 2018;11:4651–60.
34. Yu G, Wang L-G, Han Y, He Q-Y. clusterProfiler: an R package for comparing biological themes among gene clusters. *OMICS.* 2012 May;16(5):284–7.
35. Szklarczyk D, Gable AL, Lyon D, Junge A, Wyder S, Huerta-Cepas J, Simonovic M, Doncheva NT, Morris JH, Bork P, Jensen LJ, Mering CV. STRING v11: protein-protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. *Nucleic Acids Res.* 2019 Jan 8;47(D1):D607-D613. doi: 10.1093/nar/gky1131.
36. Bader GD, Hogue CWV. An automated method for finding molecular complexes in large protein interaction networks. *BMC Bioinformatics.* 2003 Jan 13;4:2.

37. Engebretsen S, Bohlin J. Statistical predictions with glmnet. Clin Epigenetics. 2019 Aug 23;11(1):123. doi: 10.1186/s13148-019-0730-1.
38. Thul PJ, Åkesson L, Wiking M, et al. A subcellular map of the human proteome. Science. 2017;356(6340):eaal3321. doi:10.1126/science.aal3321.
39. Jia B, Zhao X, Wang Y, Wang J, Wang Y, Yang Y. Prognostic roles of MAGE family members in breast cancer based on KM-Plotter Data. Oncol Lett. 2019 Oct;18(4):3501–16.

Figures

Figure 1

Discrimination of the DERBPs in LIHC. a Volcano plot of the DERBPs ($|\log FC| > 0.5$ & $P < 0.05$) between LIHC and normal samples in TCGA, red and green dots represented the up- and down-regulated RBPs, respectively b Heatmap of DERBPs between normal (N) and tumour (T).

Figure 1

Discrimination of the DERBPs in LIHC. a Volcano plot of the DERBPs ($|\log_{10}(\text{FC})| > 0.5$ & $P < 0.05$) between LIHC and normal samples in TCGA, red and green dots represented the up- and down-regulated RBPs, respectively b Heatmap of DERBPs between normal (N) and tumour (T).

Figure 2

GO and KEGG pathway enrichment analysis of the DERBPs in TCGA cohort shown in bar plots. a GO analysis of up-DERBPs. b GO analysis of down-DERBPs. c KEGG analysis of up-DERBPs. d KEGG analysis of down-DERBPs.

Figure 2

GO and KEGG pathway enrichment analysis of the DERBPs in TCGA cohort shown in bar plots. a GO analysis of up-DERBPs. b GO analysis of down-DERBPs. c KEGG analysis of up-DERBPs. d KEGG analysis of down-DERBPs.

Figure 3

Protein-protein interaction network construction and crucial subnetworks visualization. a PPI network of the DERBPs, red and green nodes represented the up- and down- RBPs respectively, edges denoted the interaction. b The top 3 crucial subnetworks with MCODE score > 5.

Figure 3

Protein-protein interaction network construction and crucial subnetworks visualization. a PPI network of the DERBPs, red and green nodes represented the up- and down- RBPs respectively, edges denoted the interaction. b The top 3 crucial subnetworks with MCODE score > 5.

Figure 4

OS related hub genes filtration. a Recognition of 20 survival related RBPs by univariate Cox regression ($P < 0.001$). b Adjusting the selection of factors at 1000 maxit. c Coefficient profile of the OS related DERBPs after LASSO regression. d Six RBPs were recognized by multivariate COX regression and subjected to our risk model.

Figure 4

OS related hub genes filtration. a Recognition of 20 survival related RBPs by univariate Cox regression ($P < 0.001$). b Adjusting the selection of factors at 1000 maxit. c Coefficient profile of the OS related DERBPs after LASSO regression. d Six RBPs were recognized by multivariate COX regression and subjected to our risk model.

Figure 5

Verification of the risk model in training cohort (TCGA). a Kaplan-Meier survival analysis of the LIHC patients with low (blue) and high (red) risk in TCGA. b Accuracy of the risk model in training set analyzed by ROC curve (AUC=0.756). c Expressing profile of the six hub genes in validating group according the risk level. The horizontal bar at the top divided the group. (pink denoted low risk, blue denoted high risk). d Distribution of the risk score (upper plot), patients and corresponding living status (lower plot, green, red represented alive and dead, respectively) in TCGA cohort.

Figure 5

Verification of the risk model in training cohort (TCGA). a Kaplan-Meier survival analysis of the LIHC patients with low (blue) and high (red) risk in TCGA. b Accuracy of the risk model in training set analyzed by ROC curve (AUC=0.756). c Expressing profile of the six hub genes in validating group according the risk level. The horizontal bar at the top divided the group. (pink denoted low risk, blue denoted high risk). d Distribution of the risk score (upper plot), patients and corresponding living status (lower plot. green, red represented alive and dead, respectively) in TCGA cohort.

Figure 6

Verification of the risk model in validating cohort (ICGC). a Kaplan-Meier survival analysis of the LIHC patients with low (blue) and high (red) risk in ICGC. b Accuracy of the risk model in training set analyzed by ROC curve (AUC=0.781). c Expressing profile of the six hub genes in validating group according the risk level. The horizontal bar at the top divided the group. (pink denoted low risk, blue denoted high risk). d Distribution of the risk score (upper plot), patients and corresponding living status (lower plot. green, red represented alive and dead, respectively) in ICGC cohort.

Figure 6

Verification of the risk model in validating cohort (ICGC). a Kaplan-Meier survival analysis of the LIHC patients with low (blue) and high (red) risk in ICGC. b Accuracy of the risk model in training set analyzed by ROC curve (AUC=0.781). c Expressing profile of the six hub genes in validating group according the risk level. The horizontal bar at the top divided the group. (pink denoted low risk, blue denoted high risk). d Distribution of the risk score (upper plot), patients and corresponding living status (lower plot. green, red represented alive and dead, respectively) in ICGC cohort.

Figure 7

Relationships between the clinical parameters and patients' prognosis in training and validating cohorts. a univariate COX regression analysis in training set. b multivariate COX regression analysis in training set. c univariate COX regression analysis in validating set. d multivariate COX regression analysis in validating set.

Figure 7

Relationships between the clinical parameters and patients' prognosis in training and validating cohorts. a univariate COX regression analysis in training set. b multivariate COX regression analysis in training set. c univariate COX regression analysis in validating set. d multivariate COX regression analysis in validating set.

Figure 8

Construction of a nomogram based on TCGA cohort. Gray line represents the ideal nomogram and red line represents the actual efficiency of our nomogram. a Nomogram for forecasting OS of 1-, 3-, 5-years based on the six hub RBPs b Corresponding calibration curves for 1-, 3-, 5-years OS prediction.

Figure 8

Construction of a nomogram based on TCGA cohort. Gray line represents the ideal nomogram and red line represents the actual efficiency of our nomogram. a Nomogram for forecasting OS of 1-, 3-, 5-years based on the six hub RBPs b Corresponding calibration curves for 1-, 3-, 5-years OS prediction.

Figure 9

Survival analysis of the six hub RBPs in LIHC by Kaplan-Meier plotter. a UPF3B. b DHX58. c MRPL54. d ZC3H13. e PPARGC1A. f EIF2AK4.

Figure 9

Survival analysis of the six hub RBPs in LIHC by Kaplan-Meier plotter. a UPF3B. b DHX58. c MRPL54. d ZC3H13. e PPARGC1A. f EIF2AK4.

Figure 10

Confirmation of the expression of the six hub RBPs in LIHC by HPA database.

Figure 10

Confirmation of the expression of the six hub RBPs in LIHC by HPA database.

Supplementary Files

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

- [Additionalfile2.xls](#)
- [Additionalfile2.xls](#)
- [Additionalfile5.xls](#)
- [Additionalfile5.xls](#)
- [Additionalfile1.xls](#)
- [Additionalfile1.xls](#)
- [Additionalfile3.png](#)
- [Additionalfile3.png](#)
- [Additionalfile6.xls](#)
- [Additionalfile6.xls](#)
- [Additionalfile7.xls](#)
- [Additionalfile7.xls](#)
- [Additonalfile4.xls](#)
- [Additonalfile4.xls](#)