Mining TCGA Database to Construct a RNA Binding Proteins-Related Prognostic Model for GBM

Wenjing GUO
hua zhong ke ji da xue tong ji yi xe yuan: Tongji Medical College
Rui Chen
Huazhong University of Science and Technology
Hui Deng
Huazhong University of Science and Technology
Mengxian Zhang (401785181@qq.com)
Huazhong University of Science and Technology

Research

Keywords: Glioblastoma, RNA binding protein, prognostic value, bioinformatics analysis,

Posted Date: April 5th, 2021

DOI: https://doi.org/10.21203/rs.3.rs-379855/v1

Abstract

Background: Glioblastoma(GBM) is a common primary malignant brain tumor with poor prognosis, and currently effective therapeutic strategies are still limited. RNA binding proteins(RBPs) dysregulation has been reported in various cancers and is closely related to tumor initiation and progression. However, little is known about the role of RBPs in GBM.

Methods: We downloaded RNA-seq transcriptome from TCGA database and differently expressed RBPs were screened between tumor and normal tissues. Then we performed functional enrichment analysis of these RBPs and based on univariate and multivariate cox regression analysis, hub RBPs were identified. Furthermore, we constructed a risk model based on hub RBPs and divided patients into high- and low-risk groups based on the median risk score. To validate the model, CGGA database were conducted as a training set and then both survival analysis and ROC curve were conducted. We also developed a nomogram based on five RBPs, which made more convenient to observe each patient's prognosis and validated the connection between patients survival and each hub RBP . Finally, we used GEPIA website to further explore the value of these hub RBPs.

Results: A total 309 differently expressed RBPs were identified, including 145 downregulated and 164 upregulated RBPs. and the result indicated that they were mainly enriched in mRNA processing, RNA splicing, RNA catabolic process, RNA transport, spliceosome, ribosome and mRNA surveillance pathway. Five hub RBPs were identified and we observed that patients with high risk score were related to poor overall survival and the AUC of ROC curve was 0.752 in TCGA. The result was subsequently proved by CGGA, showing the good prediction function of the model. Then GEPIA website suggested that MRPL41, MRPL36 and FBXO17 were closely associate with OS in GBM.

Conclusion: Our result may provide novel insights into pathogenesis of GBM and development of new therapeutic targets. However, further experiments in vitro and in vivo will be warranted.

1. Introduction

GBM is the most aggressive and common brain cancer in adults[1], with a dismal outcome, despite aggressive treatment including surgical resection and radiotherapy with concomitant chemotherapy. The high heterogeneity and complexity of GBM may contribute the little effect on the survival of the patients receiving the above treatments. Therefore, it's urgent to understand the molecular mechanism of GBM in order to find effective methods for diagnosis, treatment and ameliorating the quality of life and survival time of patients.

RNA-binding proteins are a class of protein widely involved in many post transcriptional regulation processes such as RNA splicing□transport□sequence editing□intracellular location and translation control by identifying the special RNA binding domain and RNA interaction[2]. Considering the importance of post-transcriptional regulation in life process, it is not a surprise that aberrantly altered expression of

RBPs is a common phenomenon during the development and progression of many human diseases[3-5]. Nevertheless, the general roles of RBPs in various tumors are still ambiguous.

In recent years, genome-wide analysis has identified many RBPs as key molecules in the development and progression of cancers. It is generally known that deregulation of RBPs in cancer cells is mainly induced by genomic alterations, epigenetic mechanisms and miRNA-mediated regulation[6, 7].What's more, there are many studies linking cancers to RBPs dysregulation. For example, NONO is proposed as a innovative diagnostic and therapeutic biomarker by affecting breast cancer cell proliferation[8]. NELFE has been confirmed to play an important role in pancreatic cancer metastasis and participates in the epithelial-to-mesenchymal transition by decreasing the stabilization of NDRG2 mRNA[9]. PTBP1 has a crucial role in the development and progression of HCC[10]. Moreover, studies has shown that some RBPs are closely related to the malignancy of glioma[11-13]. Nevertheless, the function of most RBPs in glioma remains ambiguous. A systematic functional study of RBPs will help us better understand their roles in glioma. Thus we downloaded GBM RNA-sequencing and corresponding clinical information from TCGA database. Then a series of bioinformatics methods were applied to analyze the differently expressed RBPs and 5 hub RBPs were finally selected, providing an implication for the diagnosis and prognosis of GBM patients.

2. Materials And Methods

2.1 Data processing

The RNA-seq transcriptome data of 169 GBM samples and 5 normal brain tissue samples as well as corresponding clinical information were downloaded from TCGA. The raw data was normalized by limma package and filtered out genes with an average count value less than 1. Then differentially expressed RBPs were identified based on a false discovery rate<0.05 and |log2fold change(FC)|>=1.

2.2 KEGG pathway and GO enrichment analysis

In order to further explore the biological functions of these differentially expressed RBPs, we conducted GO enrichment and KEGG pathway analysis. Both enrichment analyses were carried out using the WebGestalt(Web-based Gene Set Analysis Toolkit),p and FDR values were less than 0.05 as the significant threshold.

2.3 Protein-protein Interaction(PPI) Network construction and module selection

STRING database was used to identify protein-protein interaction information among the differentially expressed PBPs. Then the Cytoscape 3.6.1 software was used to further construct the PPI network. Simultaneously, we applied Molecular Complex Detection(MCODE) plug-in in cytoscape to identify three key modules and hub genes with both score and node counts more than 5. All p<=0.05 were considered as significant difference.

2.4 Construction of prognostic model

We performed the univariate Cox regression analyses on all vital RBPs in the top 3 modules of TCGA dataset using survival R package. Then a log-rank test was applied to screen the significant candidate genes. Subsequently, based on the above manipulation, we constructed a multivariate cox proportional hazards regression model and calculated a risk score of each patient. The risk score was: Risk score=β1*EXP1 +β2*EXP2+β3*EXP3+……+βn*EXPn, in which β represented the coefficient value and the EXP represented the gene expression level. GBM patients were divided into high-risk and low-risk groups based on median risk score, and the ROC curve was performed to predict the accuracy via SurvivalROC package. In addition, the prognostic model was validated by CGGA dataset, which includes 247 GBM patient samples with corresponding clinical information. In the end, the nomogram with calibration plots was conducted using rms R package to forecast the likelihood of OS. All $p<0.05$ was considered as a significant difference. Simultaneously we used GEPIA website to further explore the value of these genes.

3. Result

3.1 Identification of differently expressed RBPs in GBM patients

The transcriptome profiling of GBM and relevant clinical information were downloaded from TCGA database. A total of 169 GBM samples and 5 normal brain samples were analyzed. Then limma package was applied to normalize these data and finally got the differently expressed RBPs. Total 1373 RBPs were included in the analysis and only 309 RBPs met the standard of this study (FDR<0.05, |log2FC|>1), containing 145 downregulated and 164 upregulated RBPs. The expression distribution of these RBPs was shown in Figure1.

3.2 KEGG pathway and GO enrichment analysis of differently expressed RBPs

In order to further explore the potential function and molecular mechanisms of the identified RBPs, GO terms and KEGG pathway analysis were performed. Among upregulated differently expressed RBPs, for the biological processes(BP), RNA catabolic process, mRNA catabolic process, ribonucleoprotein complex biogenesis, ncRNA processing were the commonly enriched categories. In terms of the cellular component(CC) ontology, enriched categories were correlated with ribosomal subunit, ribosome, cytosolic part and cytosolic ribosome. With regards to the molecular function(MF), the differently expressed RBPs mainly showed enrichment in structural constituent of ribosome, catalytic activity, acting on RNA and nuclease activity. Similarly, downregulated differently expressed RBPs were enriched in mRNA processing, RNA splicing, mRNA splicing; cytoplasmic ribonucleoprotein granule, ribonucleoprotein granule; mRNA binding, catalytic activity, acting on RNA and mRNA 3'−UTR binding, respectively. Moreover, we found that upregulated differently expressed RBPs were mainly enriched in RNA transport, spliceosome, ribosome and mRNA surveillance pathway whereas downregulated RBPs were significantly enriched in RNA transport, mRNA surveillance pathway and Aminoacyl−tRNA biosynthesis. All above results were shown in figure 2.

3.3 PPI network construction and key modules screening

The PPI network of the differently expressed RBPs was constructed by using the STRING online database and cystoscope(version 3.6.1) which incorporated 309 nodes and 2619 edges based on the data from the STRING database(Figure 3A). MCODE plugin was used for module analysis of the PPI network and the most significant three modules were chosen for further analysis(Figure 3B).Function enrichment analysis revealed that the RBPs involved in module1 were related to nuclear-transcribed mRNA catabolic process, nonsense-mediated decay, cytosolic ribosome, structural constituent of ribosome and translation factor activity, RNA binding, while the RBPs in module2 were associated with RNA splicing, spliceosomal complex, catalytic step 2 spliceosome, snRNA binding and DNA-directed 5'-3' RNA polymerase activity, concurrently, piRNA metabolic process, cellular process involved in reproduction in multicellular organism, cytoplasmic ribonucleoprotein granule, helicase activity, catalytic activity, acting on RNA and ATP-dependent RNA helicase activity were enriched in module3 of the RBPs.

3.4 prognosis-model construction and analysis

A total of 309 critical differently expressed RBPs were identified from the PPI network. To investigate their prognostic value, we performed the univariate COX regression and identified 17 candidate RBPs (Figure 4). Subsequently, these 17 RBPs were analyzed by multivariate COX regression to explore their impact on the prognosis of patients. Finally, we obtained 5 hub RBPs which were found to be independent predictors in GBM(Figure 5)(Table 1). In addition, we constructed a prognostic risk model based on 5 RBPs as follows: risk score:(0.3049*MRPL41)+(-0.4194*MRPL36)+(0.2473*FBXO17)+(-0.4374*SRBD1)+(-0.8094*SARNP). Then we divided 160 GBM patients (deleting patients without survival time) into high-risk and low-risk subgroups on account of the median risk score. The GBM patients in high-risk group had statistically significantly worse overall survival than those in low-risk group with $p<0.05$ in TCGA cohort (Figure 6A). The area under the ROC curve(AUC) of this risk score model was 0.752(Figure 6B), indicating a better prognostic ability. What's more, the heatmap, survival status of patients and the risk score of each patients calculated by the model are displayed in Figure6C□D. Besides, CCGA database which is a web application for exploring brain tumors from Chinese cohorts, including 235 GBM samples and relevant clinical information was applied to validate the model. Similarly, the individual risk score of each patient was calculated and we found that patients in high-risk group had a poorer OS compared to those in low-risk group (Figure 7A□C□D). The AUC was 0.756, indicating a better specificity and sensitivity of the prognostic model (Figure 7B).

3.5 construction of a nomogram based on the 5 hub RBPs.

Based on the multivariate cox analysis, we established a nomogram employing the 5 hub RBPs. Use the point scale in the nomograph to assign points to variables. We draw a horizontal line to determine the points of each RBP, and calculated the total score of each patient by adding the points of all RBPs, normalizing it to a distribution of 0 to 100. Then we could get the survival rates for GBM patients at 1,2 and 3 years, which made the prognostic model more readable and convenient to evaluate patients (Figure 8).Besides, we evaluated the prognostic significance of different clinical features in GBM patients from CGGA by performing COX regression analysis. The result indicated that radiotherapy□chemotherapy□

IDHmut and risk score were correlated with OS, as well as the independent prognostic factors through multiple regression analysis.(Table 2)(Supplementary figure 1) Finally, the survival analysis was used to determine the relationship between hub RBPs and OS (Figure 9). It was consistent with our previous findings in which MRPL41、MRPL36 and FBXO17 were risk factors, SRBD1 and SARNP were protective factors yet, despite the fact that there was no statistical significance. Similarly, it was observed that the higher MRPL41、MRPL36 and FBXO17 expression, the worse the prognosis.

4. Discussion

GBM is a fatal primary brain tumor with life expectancy of only 12-15 month[14], despite the available therapeutic schedule. Therefore, it's imperative to find the genes and signaling pathways involved in the initiation and development of glioma. Recently, advances in bioinformatics and sequencing technology make it efficient and convenient to uncover the molecular mechanisms underlying glioma. It has been reported that RBPs show dysregulated expression in various tumors[15, 16]. However, the number of well characterized RBPs in gliomas is still relatively small. In this study, totally 309 differently expressed RBPs were identified between GBM and normal brain tissues, including 145 downregulated and 164 upregulated RBPs. We systematically investigated the relevant functional pathway and constructed a PPI network of these RBPs. Moreover, univariate COX regression analysis, multivariate COX regression analysis, ROC curve and survival analysis were used to further screen out hub RBPs and then we built a risk model based on the five hub RBPs. Our findings may provide implications for developing novel targets, with the expectation for improving patient survival time.

Functional enrichment analysis of these differently expressed RBPs showed that upregulated RBPs were significantly enriched in RNA catabolic process, ribonucleoprotein complex biogenesis, ribosome biogenesis, ncRNA processing, structural constituent of ribosome, ribosome and RNA transport. The downregulated RBPs were mainly enriched in mRNA processing, RNA splicing, mRNA binding, catalytic activity, RNA transport and mRNA surveillance pathway. In recent years, a large number of studies has been reported that abnormal expression of RBPs is a common phenomenon during the development and progression of cancers[17-19].Moreover, a great deal of RBPs involved in RNA splicing have been reported in various cancers. For instance, the RNA binding protein, QKI, serves as a critical regulator of splicing of NUMB by binding to two RNA elements[20]. In addition, altered expression of CUGBP1 can regulate IR-A:IR-B ratio via modulation of IR-A expression in breast cancer[21]. SRSF1 promotes cancer cell proliferation and progression by regulating the expression of LIG1 mRNA in non-small cell lung cancer[22]. In summary, RBPs malfunction could truly affect tumor cells growth, proliferation and invasion.

Subsequently, we performed univariate and multivariate cox regression to finally identify the five hub RBPs. From the forest map, we could observe that MRPL41, MRPL36 and FBXO17 served as protective factors in prognosis while SRBD1 and SARNP increased the risk in GBM. The conclusion was consistent with our survival analysis. It could be observed that patients with higher expression of MRPL41, MRPL36 and FBXO17 showed better OS. In contrast, in our study, higher level of SRBD1 and SARNP indicated

worse OS. Among the five hub RBPs, previous study has reported that MPRL41, encoding a mitochondrial ribosomal protein, induces apoptosis by increasing p53 stability in lymphoma[23].It has been reported that MRPL36 has a vital impact on the efficiency of mitochondrial translation[24]. Interestingly, the function of MRPL36 in tumors hasn't been reported yet and it could be a potential gene affecting tumor progression. A several studies have shown that FBXO17 plays a critical role in various cancers[25]. Tao Zhang et al found that SRBD1 promotes cell growth and inhibits cell apoptosis in non-small cell lung cancer, showing its potential value of diagnosis and treatment[26, 27]. Kang et al observed that SARNP was a contributor to the progression of breast cancer by suppression of E-cadherin expression and exerted a positive role on mRNA splicing and export[28].

In addition, we performed multivariate cox regression analysis and constructed a risk model based on the five hub RBPs to evaluate the prognosis in GBM patients. The ROC curve of prognostic model showed that the prediction model had a better performance for predicting OS with AUC=0.752 in TCGA cohort. To further prove the credibility of the model, CGGA database was acting as validation set. CGGA, a user-friendly web application for analysis to explore brain tumors from Chinese cohort, was utilized to testify the model. Similarly, it displayed a better accuracy with AUC=0.756. According to our result, patients with high risk score showed worse overall survival. Moreover, a nomogram was constructed to make 1,2,3 years OS more intuitive of each patient. Ultimately, we used GEPIA database to explore the five candidate RBPs in GBM. Survival analysis was performed to further assess the prognostic value, with the expectation of identifying new targets for therapies.

5. Conclusion

In this study, we conducted a comprehensive bioinformatics analysis of differently expressed RBPs and built a prognostic model based on five hub RBPs, which might serve as an independent prognostic factor for GBM patients. Furthermore, our study showed that MRPL41,MRPL36 and FBXO17, serving as risk factors, were tightly related to OS. These three genes may provide new therapeutic targets for GBM. However, further experiments were necessary for the identification of our conclusions.

Declarations

Ethics approval and consent to participate

Not applicable

Consent for publication

Not applicable

Availability of data and materials

The data used for analysis in this study are available from TCGA and CGGA database.

Competing interests

All authors declare that there are no conflicts of interest

Funding

This work was supported by the National Nature Science Foundation of China.(81772680)

Authors' contributions

Wenjing Guo designed the study, analysed the data and drafted the manuscript . Rui Chen, Hui Deng collected the data. Mengxian Zhang revised the manuscript. All authors have read and approved the manuscript.

Tables

Table 1	Multivariate cox regression analysis to identify hub RBPs				
gene	coef	HR	HR.95L	HR.95H	pvalue
MRPL41	0.304931	1.356532	1.040669	1.768265	0.024151
MRPL36	0.419407	1.52106	1.028673	2.249133	0.035586
FBXO17	0.247299	1.280562	1.018757	1.609647	0.034073
SRBD1	-0.43743	0.645693	0.462566	0.901318	0.010155
SARNP	-0.80944	0.445106	0.30032	0.659696	5.53E-05

Table 2 The prognostic value of different clinical factors						
	univariate analysis			multivariate analysis		
	HR	95% CI	p value	HR	95% CI	p value
age	1.01	0.99-1.02	0.087	1.01	0.99-1.02	0.179
gender	0.99	0.71-1.36	0.939	1	0.71-1.41	0.99
Radio	0.56	0.36-0.87	0.00989	0.68	0.41-1.14	0.142
Chemo	0.46	0.29-0.72	<0.001	0.53	0.32-0.88	0.015
IDHmut	0.49	0.32-0.78	0.002	0.58	0.36-0.92	0.019
MGMTmethy	0.89	0.64-1.22	0.471	0.93	0.67-1.31	0.686
Riskscore	1.93	1.55-2.42	<0.001	1.67	1.31-2.12	<0.001

Table 3 KEGG pathway and GO enrichment analysis of differently expressed RBPs			
	GO term	p value	FDR
Down-regulated RBPs	mRNA processing	6.40E-31	1.65E-17
	RNA splicing	1.96E-23	1.24E-20
	regulation of RNA splicing	1.65E-17	3.48E-15
	cytoplasmic ribonucleoprotein granule	3.95E-14	6.28E-12
	ribonucleoprotein granule	8.46E-14	6.73E-12
	mRNA binding	1.08E-17	2.04E-15
	catalytic activity, acting on RNA	2.64E-15	2.5E-13
	mRNA 3'-UTR binding	4.29E-15	2.71E-13
	RNA transport	2.37E-10	6.99E-09
	mRNA surveillance pathway	2.26E-09	3.33E-08
Up-regulated RBPs	RNA catabolic process	2.31E-46	2.97E-43
	mRNA catabolic process	1.3E-38	8.33E-36
	nuclear-transcribed mRNA catabolic process	1.35E-36	5.79E-34
	ribonucleoprotein complex biogenesis	2.86E-35	9.19E-33
	cytosolic ribosome	2.75E-34	4.03E-32
	ribosomal subunit	1.54E-32	1.12E-30
	structural constituent of ribosome	2.92E-19	2.38E-17
	RNA degradation	2.46E-07	2.78E-06
	mRNA surveillance pathway	3.36E-07	3.14E-06

Figures

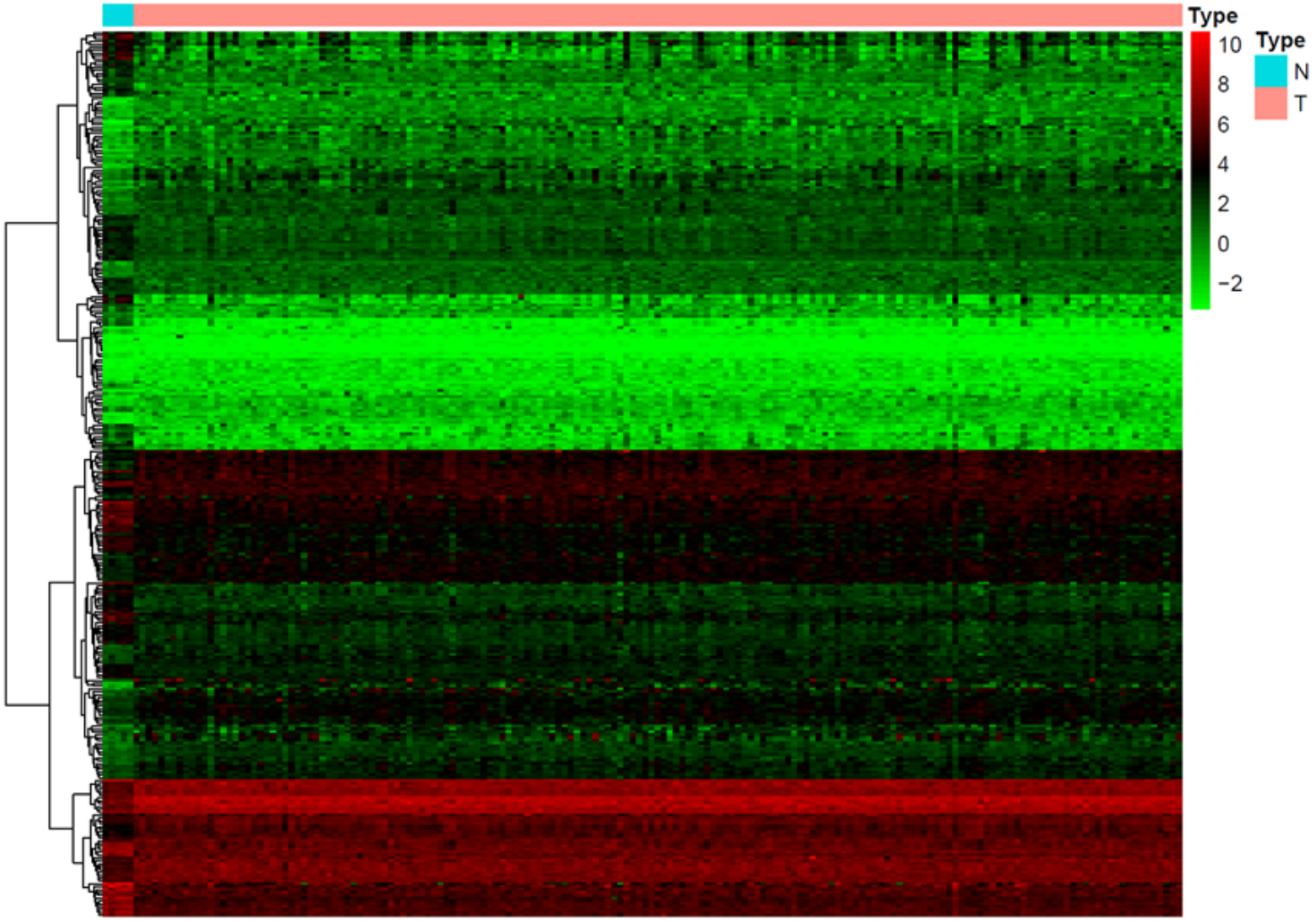

Figure 1

Heatmap of the differentially expressed RBPs in GBM

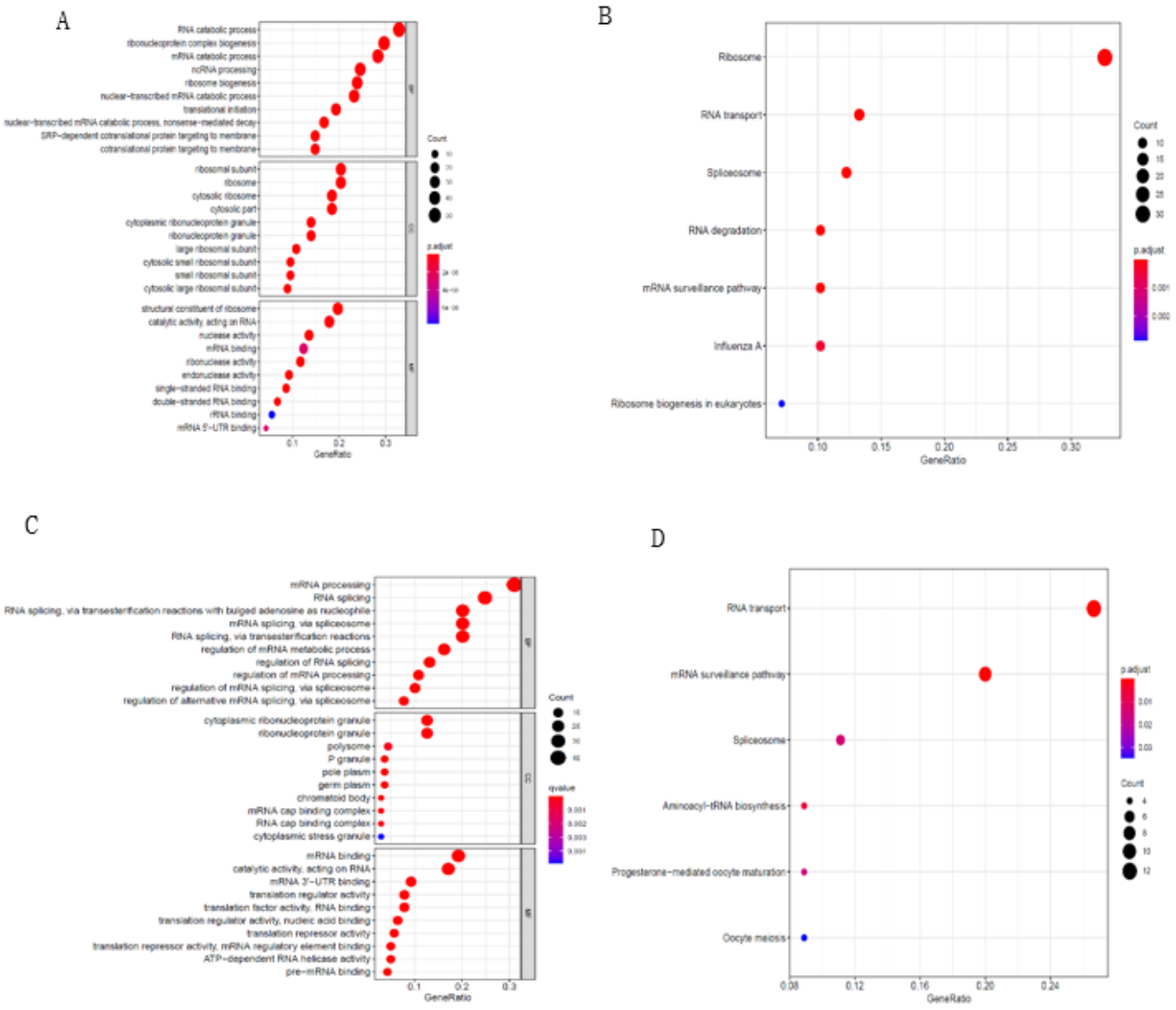

Figure 2

GO and KEGG pathway Functional analysis of differential expression RBPs (A,B) GO and KEGG pathway analysis of upregulated RBPs (C,D)GO and KEGG pathway analysis of downregulated RBPs

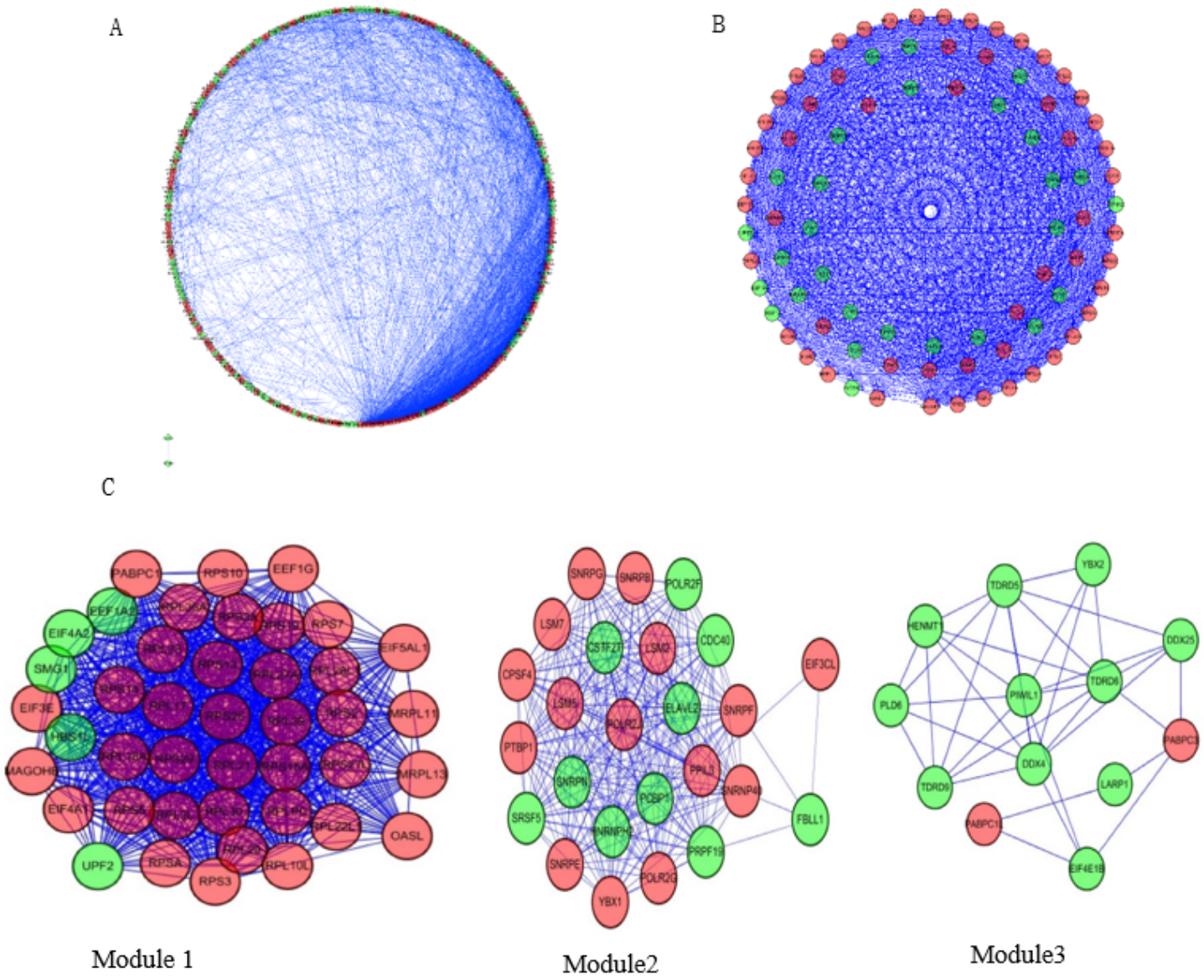

Figure 3

PPI network and module selection (A)protein-protein interaction of differential expression RBPs (B) the Top 3 modules from PPI network were selected via using the MCODE tool (C) Specific RBPs of the top3 modules from PPI network red and green represented up- and down- regulation respectively.

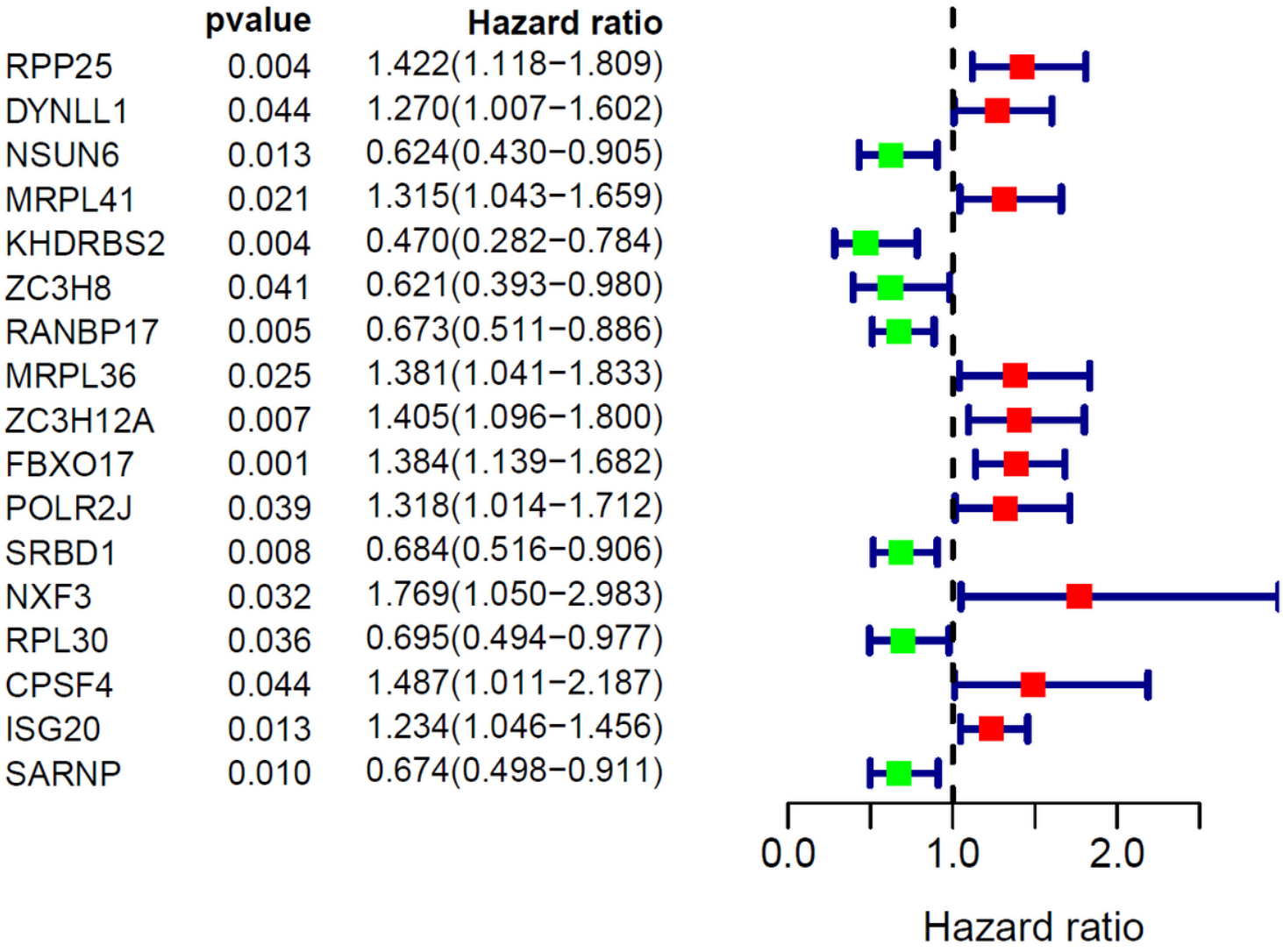

Figure 4

the forest plot of 17 hub RBPs identified by univariate cox proportional hazards regression

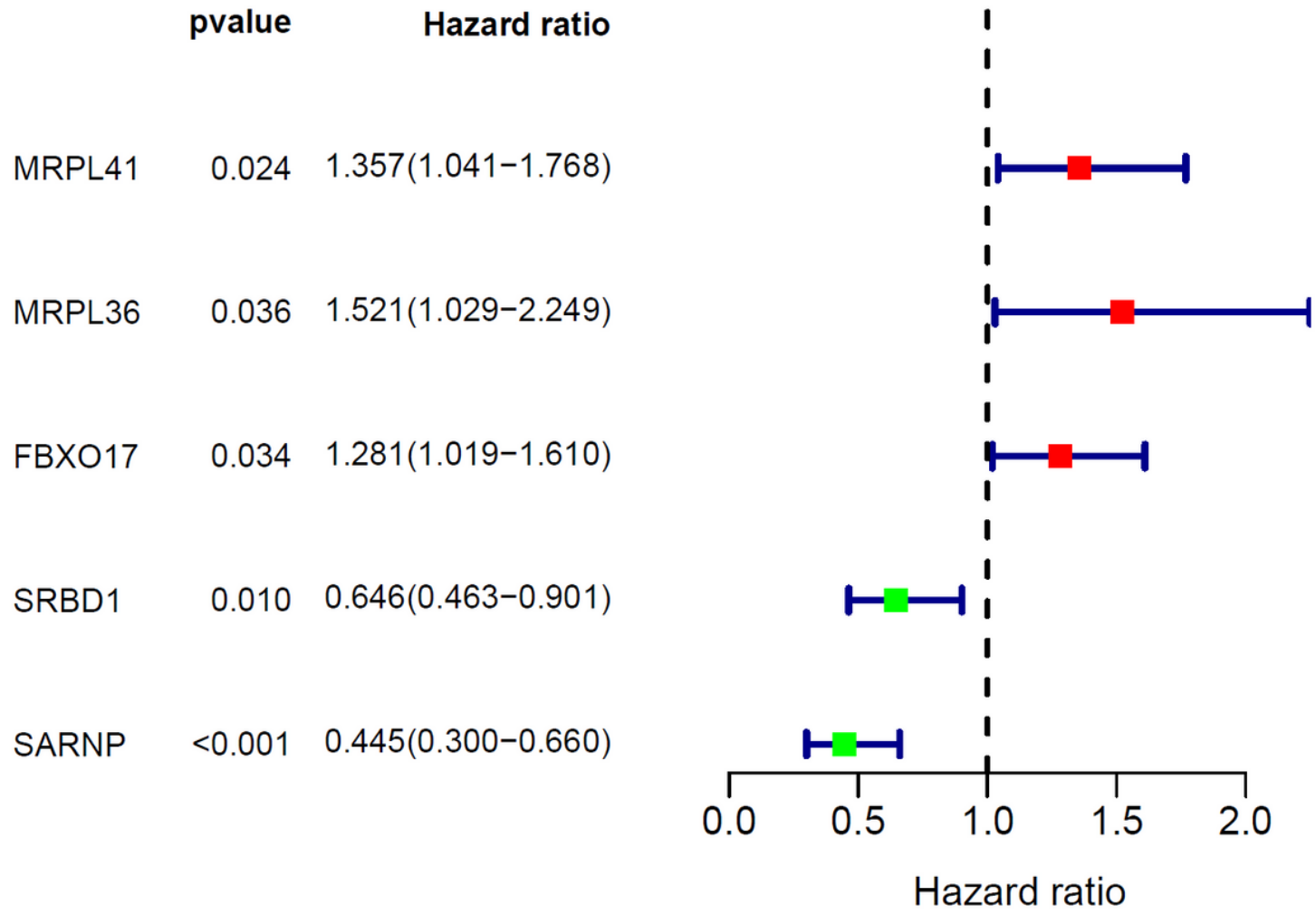

Figure 5

the forest plot of 5 prognostic RBPs screened out by multivariate cox proportional hazard regression

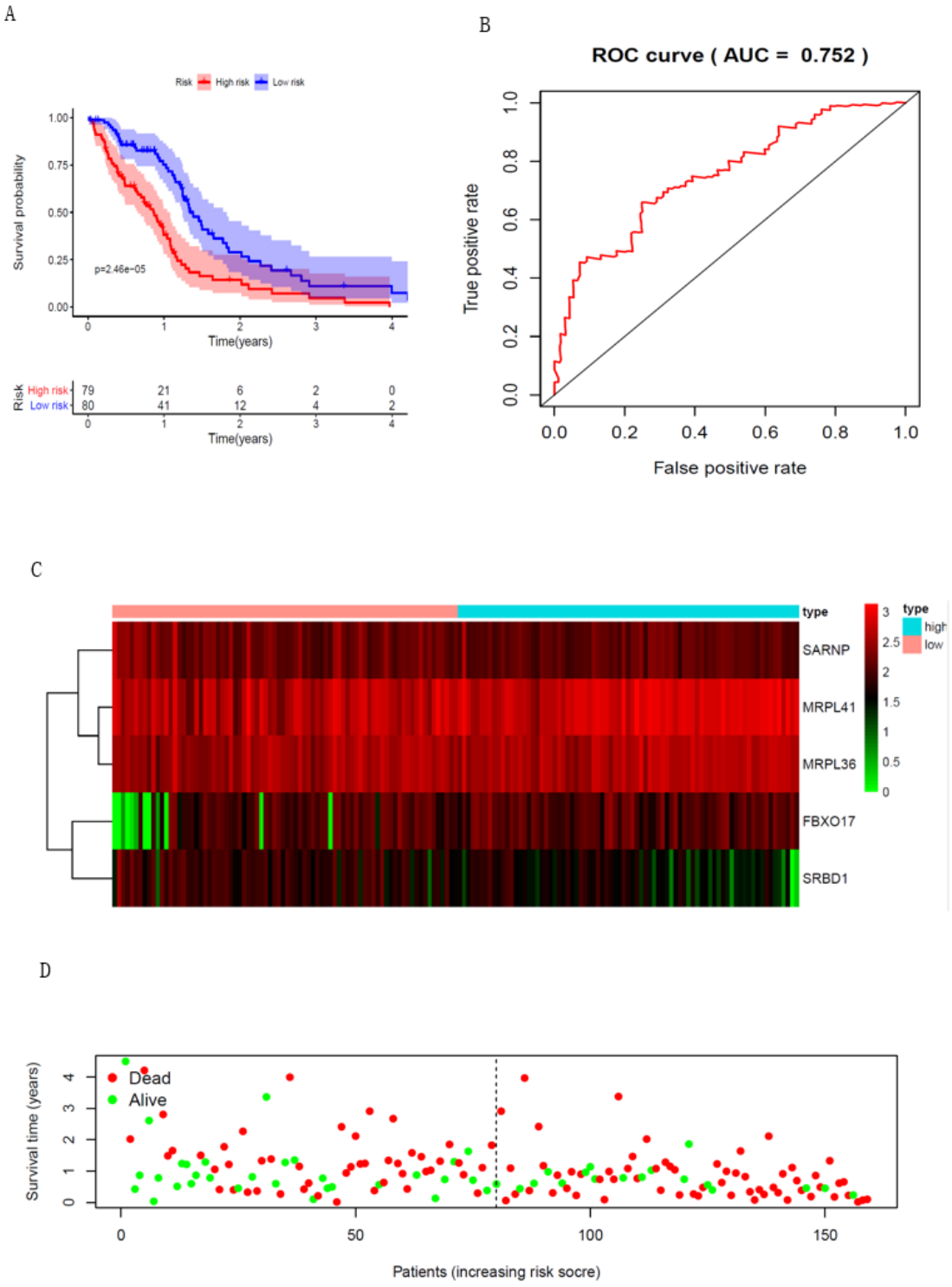

Figure 6

risk score model analysis based on the 5 hub genes in TCGA (A)survival curve between high-risk and low-risk patients (B)The ROC curve analysis for forecasting OS based on risk score (AUC=0.752) (C□D)the heatmap and risk status of 5 RBPs of the risk model between high-risk and low-risk patients

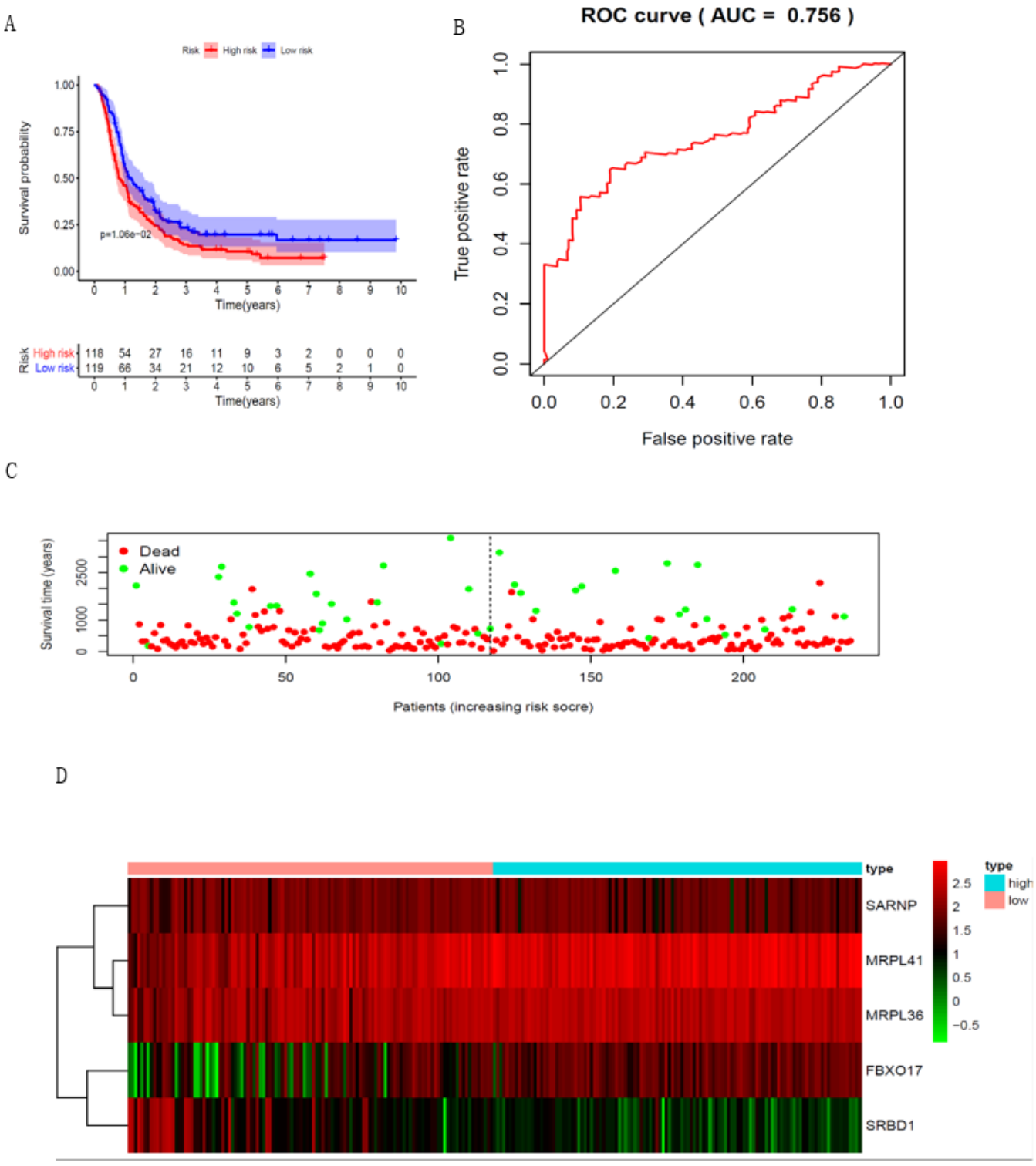

Figure 7

risk score model analysis based on the 5 hub genes in CCGA (A)survival curve between high-risk and low-risk patients (B)The ROC curve analysis for forecasting OS based on risk score (AUC=0.756) (C□D)the heatmap and risk status of 5 RBPs of the risk model between high-risk and low-risk patients

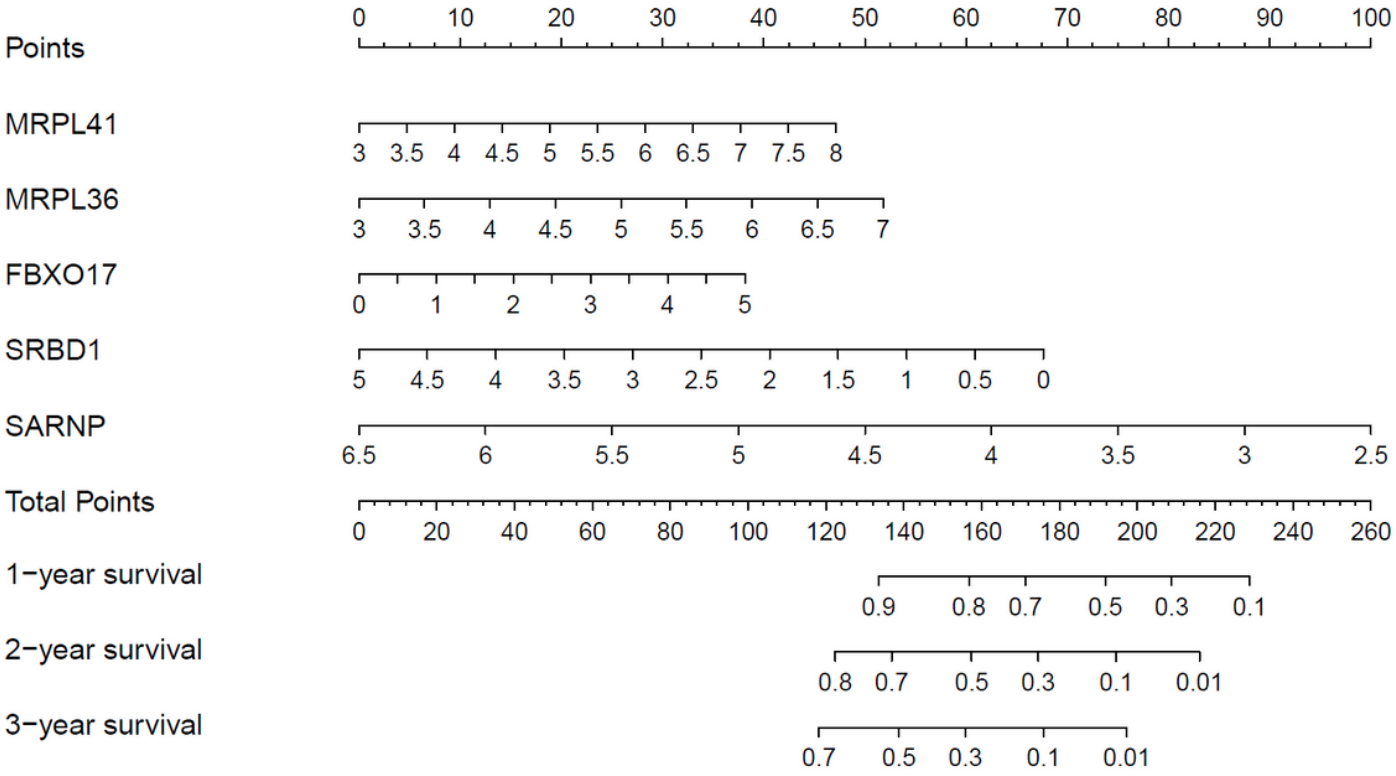

Figure 8

a nomogram for showing 1 year, 2 year and 3 year OS of GBM patients in TCGA

Supplementary Files

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

- Supp.png