doi:10.4149/neo_2018_181008N746

Prognostic risk model construction and molecular marker identification in glioblastoma multiforme based on mRNA/microRNA/long non-coding RNA analysis using random survival forest method

H. WANG*, D. Y. LIU, J. YANG

Department of Neurosurgery, Beijing Luhe Hospital, Capital Medical University, Tongzhou District, Beijing, 101149, China

**Correspondence: hwang5168@vip.163.com*

Received October 8, 2018 / Accepted December 19, 2018

We aim to identify novel molecular signatures for prognosis prediction in glioblastoma multiforme (GBM). The expression and microarray data of GBM were downloaded from The Cancer Genome Atlas (TCGA) and the Gene Expression Omnibus (GEO). Differentially expressed mRNAs, microRNAs (miRNAs) and long non-coding RNAs (lncRNAs) between GBM and normal samples were identified by differential expression analysis using Bayesian T-test. Functional enrichment analysis was performed to identify GBM associated functions and pathways. A subset of signature mRNAs was selected from differentially expressed mRNAs and used to build a risk model for GBM using random survival forest (RSF) method. The performance of the model in prognosis prediction was validated using an independent validation dataset. A competing endogenous RNA (ceRNA) network was then constructed and key prognostic markers were identified from the network by survival analysis. In total, 905 mRNAs, 24 miRNAs and 403 lncRNAs were identified to be differentially expressed between GBM and normal samples. Functional and pathway items such as p53 signaling and PI3K/Akt signaling were significantly enriched by differentially expressed mRNAs. The RSF risk model showed a high performance in prognosis prediction for both training and validation dataset. The ceRNA network provided a comprehensive view of the interplays between differentially expressed mRNAs, miRNAs and lncRNAs. Among the ceRNA network, p21 (RAC1) activated kinase 1 (*PAK1*) and synaptic vesicle glycoprotein 2B (*SV2B*) were identified as key prognosis associated markers. The RSF risk model and key prognostic markers may contribute to GBM diagnosis in future clinical practice.

Key words: glioblastoma multiforme, ceRNA, prognosis, biomarkers

Glioblastoma multiforme (GBM) is the most frequent and aggressive primary brain tumors in adults [1, 2]. GBM is characterized by increased proliferation, aggressive invasion, vigorous angiogenesis and remarkable heterogeneity [2]. Despite tremendous progress in medical care, the prognosis of GBM patients remains gloomy, with a median survival of only 14.6 months [1]. Currently, the management of GBM is limited by the insufficient accuracy of histopathologic diagnosis in clinical outcome prediction [3]. Molecular prediction tools with high accuracy are in urgent need in future clinical practice of GBM.

Expression alterations of genes involved in tumor suppressive and oncogenic pathways are common features of GBM [4]. For example, inhibitor of growth family member 4 (*ING4*), a tumor suppressor functioning by suppressing hypoxia inducible factor 1 (HIF) activation [5] and nuclear factor kappa B (NF-κB) pathway [6], is significantly downregulated in GBM [7]. Recently, a bioinformatic analysis of 123 GBM patients has established a 14-mRNA prognostic signature, which could be used to classify GBM patients into low and high risk groups [8]. Among these signature mRNAs, the tumor suppressor Insulin like growth factor binding protein like 1 (*IGFBPL1*) is downregulated whereas the oncogenic genes epidermal growth factor receptor (*EGFR*) and C-C motif chemokine ligand 2 (*CCL2*) are upregulated in high risk group [8].

Long non-coding RNAs (lncRNAs) and microRNAs (miRNAs) are non-coding RNAs playing essential roles in various biological processes, including cell division, proliferation, differentiation and apoptosis [9, 10]. They regulate gene expression at post-transcriptional level by targeting protein-coding RNAs [10]. Besides, miRNAs may also function via interacting with lncRNAs [10]. The coexpressed mRNAs and lncRNAs targeted by the same miRNAs are

considered as competing endogenous RNAs (ceRNAs) [10]. Recently, dysregulation of lncRNAs and miRNAs in various cancers has attracted increasing attention, as they appear to be key players in tumor development and progression [9, 11]. Expression alterations of miRNAs and lncRNAs in GBM are frequently observed and the feasibility to develop miRNA and lncRNA biomarkers has also been demonstrated [12–14]. For example, miR-128 and miR-342-3p are significantly downregulated in GBM and may serve as prognostic markers [12]. Expression alterations of *AC005013.5*, *UBE2R2-AS1*, *ENTPD1-AS1*, *RP11-89C21.2*, *AC073115.6* and *XLOC_004803* have also been observed in GBM and these lncRNAs have been used to develop a 6-lncRNA signature of GBM [13].

In order to develop novel molecular tools to predict the prognosis of GBM patients, we analyzed the expression and microarray data of GBM downloaded from The Cancer Genome Atlas (TCGA) and the Gene Expression Omnibus (GEO). Differentially expressed mRNAs, miRNAs and lncRNAs were screened by differential expression analysis. A subset of differentially expressed mRNAs was selected to construct a random survival forest (RSF) model, which was efficient in evaluating the risks of GBM patients. Moreover, a GBM associated ceRNA network were constructed to illuminate the interactions between differentially expressed miRNAs, mRNAs and lncRNAs. Among the network, p21 (RAC1) activated kinase 1 (*PAK1*) and synaptic vesicle glycoprotein 2B (*SV2B*) were identified to be key prognostic markers.

Materials and methods

Data source. RNAseqV2 exon data (level 3, Illumina HiSeq 2000 RNA Sequencing platform), mRNA microarray data (Affymetrix Human Exon 1.0 ST Array) and clinical data of GBM were downloaded from TCGA (https://tcga-data.nci.nih.gov/) in March 2018. A total of 158 samples (153 GBM and 5 normal samples) were included in the TCGA-seq dataset and 441 samples (431 GBM and 10 normal samples) were included in the TCGA-array dataset. miRNA microarray data (Illumina Human v2 MicroRNA expression beadchip) under the accession code of GSE25631 [14, 15] was downloaded from GEO (https://www.ncbi.nlm.nih.gov/geo/). A total of 87 samples (82 GBM and 5 normal samples) were included in the GSE25631 dataset.

Data preprocessing. RNAseqV2 exons in the TCGA-seq dataset were annotated by mapping their starting points and sequences to Genecode database [16] (https://www.gencodegenes.org/) and the exons were defined as lncRNAs or mRNAs according to the mapping results.

Genes with low expression level were then removed from the TCGA-seq dataset. The preprocessing of lncRNA and mRNA sequencing data were preformed using the R package edgeR [17, 18] (Version 3.4, http://www.bioconductor.org/packages/release/bioc/html/edgeR.html). Specifically, the raw counts were normalized to logCPM (count per million) values for linear modeling and the mean-variance relationship was adjusted by the precision weights (voom algorithm). The preprocessing of miRNA microarray data included background correction by robust multi-array average (RMA) method [19, 20], quantile normalization and log transformation [21].

Screening of differentially expressed genes. Differential expression analysis for mRNA, lncRNA and miRNA data was performed using Bayesian T-test method of limma package [22] (version 3.10.3, http://www.bioconductor.org/packages/2.9/bioc/html/limma.html). A p-value was adjusted by applying multiple testing corrections of Benjamini-Hochberg [23]. The selection thresholds for differentially expressed mRNAs and lncRNAs in the TCGA-seq dataset were set as adj. p-value <0.05 and $|\log_2 FC(\text{fold change})| > 2$. The selection thresholds for differentially expressed mRNAs and lncRNAs in the TCGA-array dataset were set as adj. p-value <0.05 and $|\log_2 FC| > 1.5$. The selection thresholds for differentially expressed miRNAs in the TCGA-array dataset were set as adj. p-value <0.05 and $|\log_2 FC| > 1$. Bidirectional hierarchical clustering was performed based on differentially expressed mRNAs (TCGA-seq), lncRNAs and miRNAs using hclust algorithm of R.

Functional and pathway enrichment analysis. Gene Ontology (GO) [24] and Kyoto Encyclopedia of Genes and Genomes (KEGG) [25] enrichment analysis were performed for differentially expressed mRNAs using The Database for Annotation, Visualization and Integrated Discovery (DAVID) [26] (version 6.8, https://david.ncifcrf.gov/). A p-value was adjusted by Benjamini-Hochberg method and the selection criterion for significant functional and pathway items was set as adj. p-value <0.05.

Protein-protein interaction (PPI) network analysis. Mentha [27] (http://mentha.uniroma2.it/about.php), BioGRID [28] (version 3.4, https://wiki.thebiogrid.org/) and HPRD [29] (release 9, http://www.hprd.org/) are databases containing PPI data of *Homo sapiens*. PPIs appeared in all the three databased were combined and used as background PPIs. The differentially expressed mRNAs were then mapped to the background PPIs. A PPI network was constructed based on the resulting PPIs using Cytoscape [30] (http://www.cytoscape.org/). Topology analysis of the PPI network was performed using the plug-in CytoNCA [31] (version 2.1.6, http://apps.cytoscape.org/apps/cytonca) (parameter = 'without weight') and the degree of each node was thus acquired. Nodes with the highest degrees were considered to be central nodes or hub nodes [32]. Afterwards, function modules with biological significance were identified from the PPI network by the plug-in MCODE [33] of Cytoscape.

Construction of prognostic mRNA model. RSF is a survival analysis method based on random forest [34]. A prognostic mRNA model for GBM was constructed by applying RSF using the R package randomForestSRC (version 2.4.0, https://cran.r-project.org/web/packages/randomFor-

estSRC/index.html). Specifically, the TCGA-seq dataset was used as training dataset and N bootstrap samples were drawn from the dataset. A survival tree was built in each sample and the variable importance (VIMP) score [35] was acquired for each mRNA. The VIMP score of a mRNA was positively correlated with the prediction capacity of the variable. A VIMP score close or less than 0 indicated little predictive value. The top 20 ranked mRNAs were used to constructed a new prognostic RSF model. Risk scores of samples in the training dataset were calculated based on the cumulative hazard function (CHF) of the prognostic model [36]. A certain risk score was chosen as the cutoff to divide samples in the training dataset into high and low risk groups. The prognostic differences between the two groups were analyzed by log-rank test and Kaplan-Meier survival analysis.

A total of 70 samples were randomly extracted from the TCGA-array dataset and were used as validation dataset. The risk scores of samples in the validation dataset were calculated using the same method as the training dataset. Samples in the validation dataset were then divided into high and low risk groups, using the same cutoff as the training dataset. The prognostic differences between the two groups were further analyzed by log-rank test and Kaplan-Meier survival analysis.

Construction of ceRNA network. For TCGA-seq dataset, pearson correlation coefficient between each differentially expressed lncRNA and mRNA was calculated. The selection criteria for lncRNA-mRNA co-expression pairs were set as $|r| > 0.95$ and $p<0.05$.The targets of differentially expressed miRNAs were predicted using miRWalk2.0 [37] (http://zmf.umm.uni-heidelberg.de/apps/zmf/mirwalk2/), which could employ information from databases miRWalk, miRanda, miRMap, miRNAMap, RNA22 and Targetscan. A gene was considered to be a target of a miRNA when the gene was predicted by at least 4 of the 6 databases. The resulting miRNA-mRNA pairs were used as background. Differentially expressed miRNA-mRNA regulation pairs were acquired by mapping differentially expressed mRNAs to the background.

miRNA-lncRNA regulation pairs from starBase [38] (http://starbase.sysu.edu.cn/) and InCeDB [39] (http://gyanxet-beta.com/lncedb/) were integrated and the resulting miRNA-lncRNA pairs were used as background to identify differentially expressed miRNA-lncRNA pairs.

Based on the differentially expressed miRNA-mRNA pairs, miRNA-lncRNA pairs and mRNA-lncRNA pairs, lncRNA-mRNA pairs regulated by at least two common miRNAs were identified. These mRNA-lncRNA pairs and their common regulating miRNAs were used to construct a miRNA-lncRNA-mRNA network, which is also called ceRNA network. Hub nodes of the network were identified by topology analysis using CytoNCA. mRNAs in the network were subjected to pathway enrichment analysis using the R package clusterProfiler [40] (version 3.2.11, http://www.bioconductor.org/packages/release/bioc/html/clusterProfiler.html). The selection criterion was set as Benjamini-Hochberg adj. p-value <0.05.

Table 1. Statistics of differentially expressed miRNAs, mRNAs and lncRNAs.

	miRNA	mRNA-Seq	mRNA-array	lncRNA
Up-DEGs [a]	10	1033	322	125
Down-DEGs	14	1921	694	278
Total	24	2954	1016	403

[a] upregulated differentially expressed gene; [b] downregulated differentially expressed gene.

Identification of prognostic mRNAs and lncRNAs from ceRNA network. Clinical characteristics, including overall survival (OS), OS status, disease-free survival (DFS) and DFS status, were used for the identification of prognostic mRNAs and lncRNAs from ceRNA network. Based on the mean expression value of each differentially expressed lncRNA or mRNA, tumor samples were divided into high and low expression group. The correlation between each lncRNA or mRNA and prognosis was evaluated by log-rank test and Kaplan-Meier survival analysis. The threshold of statistical significance was set as $p<0.05$.

Results

Differentially expressed mRNAs, lncRNAs and miRNAs. A total of 2631 lncRNAs and 18201 mRNAs were obtained from the TCGA-seq dataset after reannotation and data filtering. Among them, 403 lncRNAs and 2954 mRNAs showed significant expression differences between tumor and normal samples (Table 1). Further differential expression analysis also identified 1016 differentially expressed mRNAs from the TCGA-array dataset and 24 differentially expressed miRNAs from the GSE25631 dataset (Table 1). The intersection of differentially expressed mRNAs from TCGA-seq and TCGA-array dataset consisted of 905 (267 upregulated and 638 downregulated) mRNAs (Figure 1A). Further bidirectional hierarchical clustering analysis based on differentially expressed mRNAs, miRNAs and lncRNAs showed that tumor samples could be clearly distinguished from normal samples (Figure 1B).

Functional annotation of differentially expressed mRNAs. In order to reveal biological functions and pathways deregulated in GBM, functional annotation was performed to acquire GO and KEGG terms enriched by differentially expressed mRNAs (Table 2). The terms enriched by upregulated mRNAs included GO:0051301~cell division, GO:0007067~mitotic nuclear division, hsa04151:PI3K-Akt signaling pathway and hsa04115:p53 signaling pathway. The terms enriched by downregulated mRNAs included GO:0007268~chemical synaptic transmission, GO:0007269~neurotransmitter secretion and GO:0017157~regulation of exocytosis.

PPI network. In order to identify functionally important genes in GBM, a PPI network was built for differentially expressed mRNAs. In total, 1216 PPIs were acquired

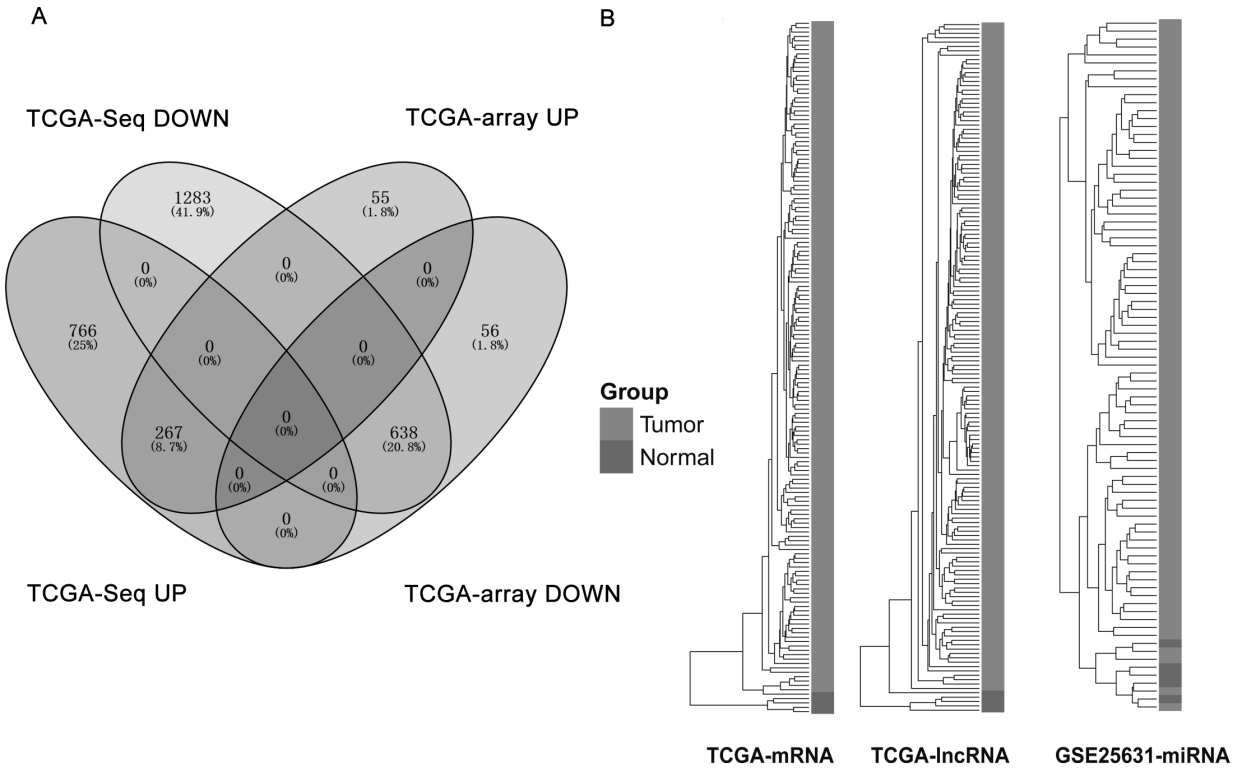

Figure 1. Differentially expressed mRNAs and bidirectional hierarchical clustering. A) Venn diagram of differentially expressed mRNAs in TCGA-seq dataset and TCGA-array dataset. B) Bidirectional hierarchical clustering of samples based on differentially expressed mRNAs (left), miRNAs (middle) and lncRNAs (right). Tumor and normal samples were represented as light grey and dark grey, respectively.

Table 2. Functional and pathway items enriched by differentially expressed mRNAs.

	Term	Count	Adjusted p-value
UP regulated mRNAs			
GO-BP [a]	GO:0051301~cell division	36	4.73E-16
GO-BP	GO:0007067~mitotic nuclear division	29	9.08E-14
GO-BP	GO:0030198~extracellular matrix organization	25	1.03E-12
GO-BP	GO:0007062~sister chromatid cohesion	15	1.59E-07
GO-BP	GO:0050900~leukocyte migration	14	1.10E-05
KEGG [b] pathway	hsa04512:ECM-receptor interaction	14	2.06E-07
KEGG pathway	hsa04110:Cell cycle	16	1.07E-07
KEGG pathway	hsa04151:PI3K-Akt signaling pathway	21	3.85E-05
KEGG pathway	hsa04510:Focal adhesion	16	5.04E-05
KEGG pathway	hsa04115:p53 signaling pathway	10	4.23E-05
DOWN regulated mRNAs			
GO-BP	GO:0007268~chemical synaptic transmission	76	2.92E-47
GO-BP	GO:0007269~neurotransmitter secretion	21	1.15E-13
GO-BP	GO:0007399~nervous system development	43	2.45E-12
GO-BP	GO:0017157~regulation of exocytosis	15	1.30E-11
GO-BP	GO:0034220~ion transmembrane transport	34	1.90E-10
KEGG pathway	hsa04727:GABAergic synapse	31	3.34E-19
KEGG pathway	hsa04723:Retrograde endocannabinoid signaling	32	3.95E-18
KEGG pathway	hsa05033:Nicotine addiction	22	6.33E-18
KEGG pathway	hsa04724:Glutamatergic synapse	33	9.65E-18
KEGG pathway	hsa05032:Morphine addiction	28	1.23E-15

[a] Gene Ontology biological process; [b] Kyoto Encyclopedia of Genes and Genomes.

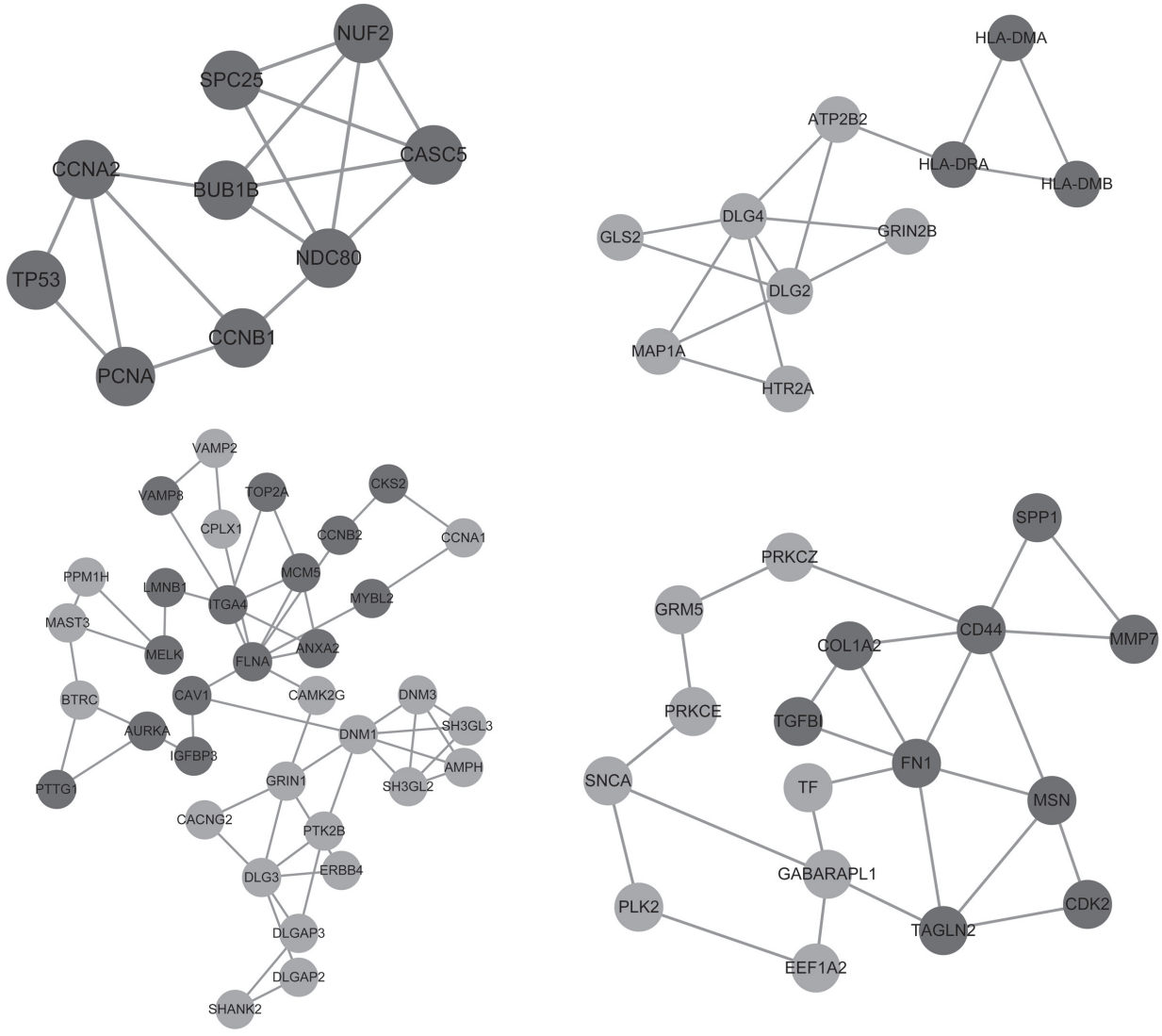

Figure 2. The top 4 modules identified from the protein-protein interaction (PPI) network of differentially expressed mRNAs. Upregulated and downregulated mRNAs were shown as dark grey and light grey nodes, respectively. Interactions between nodes were shown as gray lines.

for 557 (191 upregulated and 366 downregulated) mRNAs. Topological analysis of the PPI network showed that tumor protein p53 (TP53), cyclin-dependent kinase 2 (CDK2) and Fibronectin 1 (FN1) were nodes with the highest connectivity degree and were considered to be the hub nodes (Table 3). In addition, 9 modules could be identified from the PPI network. According to their scores, the top 4 ranked modules were considered to be functionally important for GBM (Figure 2).

Prognostic RSF model of GBM. In order to discriminate low from high risk GBM samples, a prognostic RSF model was further constructed using the training dataset (n=151). According to the VIMP scores of differentially expressed mRNAs, the top 20 mRNAs were used as signature mRNAs to construct the prognostic RSF model. Based on the CHF

Table 3. The top 12 nodes in the protein-protein interaction (PPI) network.

Rank	Gene	Degree
1	TP53	52
2	CDK2	45
3	FN1	36
4	CALM3	35
5	FLNA	29
6	PLK1	28
7	DLG4	26
8	YWHAH	24
9	GABARAPL1	23
10	VIM	23
11	AURKA	23
12	STX1A	23

of the RSF model, the risk scores of samples in the training dataset were calculated. The risk score 51 was set as the cutoff to divide samples into high risk (risk score ≥51, n=89) and low risk (risk score <51, n=62) groups, as the mortality ratio was high and reached a platform at where risk score = 51 (Figure 3A). The expression levels of the 20 mRNAs were displayed as heatmap in Figure 3B. Survival analysis showed that the prognosis of low risk group was significantly better than that of high risk group (log rank p<0.0001) (Figure 3C).

The performance of the prognostic RSF model was validated using validation dataset (n=70). The validation samples were divided into high (n=21) and low (n=49) risk groups using the same cutoff (risk score=51) as the training dataset (Figure 3A). The expression levels of the 20 mRNAs were displayed as heatmap in Figure 3E. Consistent with the training dataset, the prognosis of low risk group was also significantly higher than that of high risk group (log rank p<0.0001) in the validation dataset (Figure 3C).

GBM associated ceRNA network. A total of 1987 lncRNA-mRNA co-expression pairs (|r|>0.95 and p<0.05) were identified between 56 lncRNAs and 235 mRNAs, according to the Pearson correlation coefficients between differentially

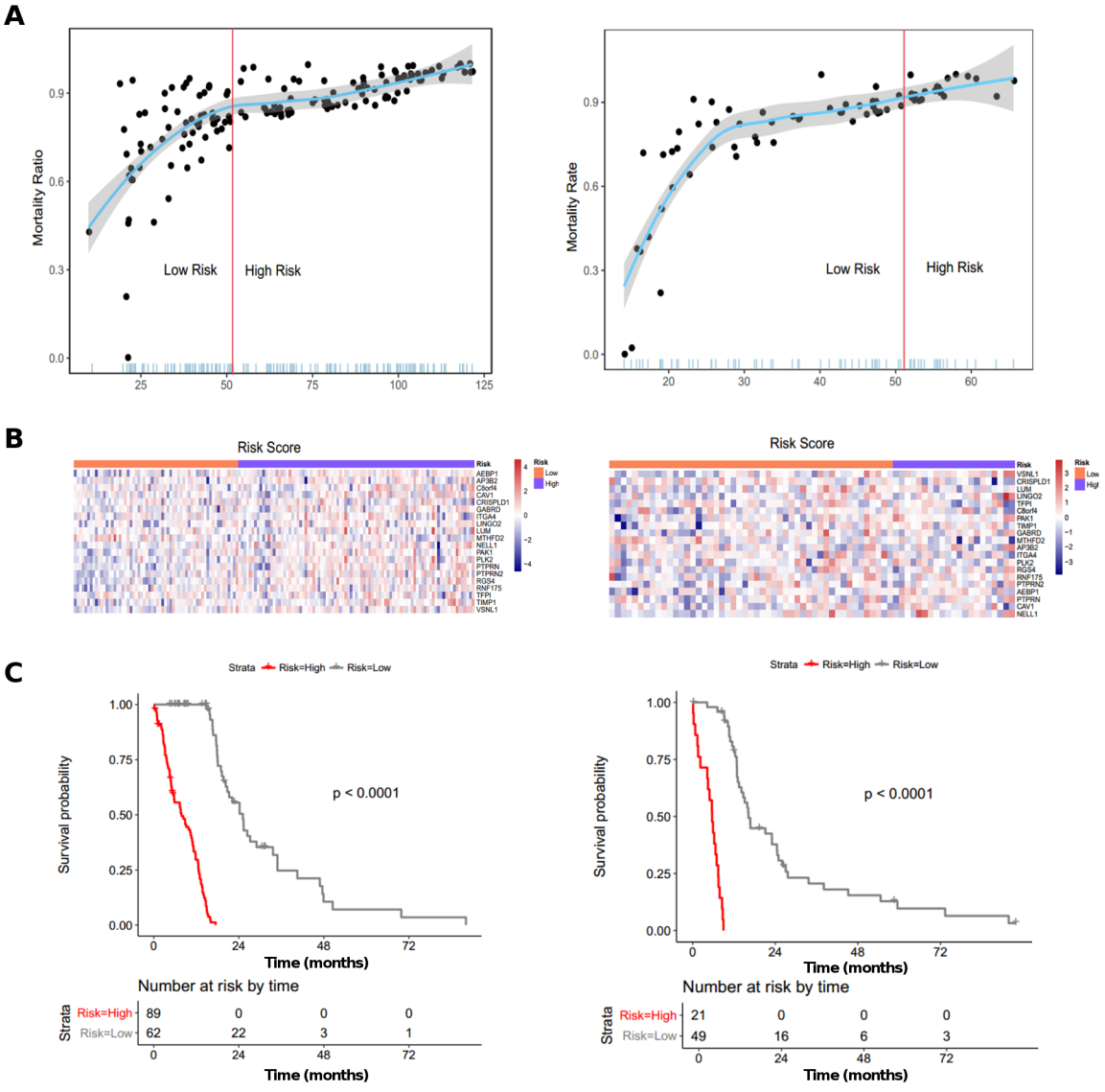

Figure 3. Survival analysis based on the random survival forest (RSF) risk model. A) The distribution of mortality ratios and risk scores of training samples (left) and validation samples (right). The red lines indicated where risk scores equaled 51. B) The heatmap showing the expression levels of the signature mRNAs in the training (left) and validation (right) dataset. Low and high-risk samples were indicated as orange and violet bars, respectively. The expression levels were indicated as colors from blue (low expression) to red (high expression). C) Kaplan-Meier curves of low-risk and high-risk groups in the training (left) and validation (right) dataset. The survival curves for low and high-risk groups were showed as grey and red lines, respectively.

expressed lncRNAs and mRNAs. The regulation relationships between differentially expressed miRNAs and lncRNAs were predicted on the basis of 6 databases, including miRWalk, miRanda, miRMap, miRNAMap, RNA22 and Targetscan. A total of 541 miRNA-lncRNA regulation pairs were identified between 32 miRNAs and 203 lncRNAs. Additionally, the regulation relationships between differentially expressed miRNAs and mRNAs were predicted based on the databases starBase and InCeDB. As a result, 5740 potential miRNA-mRNA regulation pairs were identified between 33 miRNAs and 851 mRNAs.

The miRNA-lncRNA regulation pairs, miRNA-mRNA regulation pairs and mRNA-lncRNA coexpression pairs identified above were further integrated. In total, 115 mRNA-lncRNA coexpression pairs were identified to be targeted by at least two common miRNAs. These mRNA-lncRNA pairs and the common miRNAs were used to construct a ceRNA network (Figure 4A), which consisted of 88 nodes (14 miRNAs, 65 mRNAs and 8 lncRNAs) and 241 edges. In the network, lncRNAs and mRNAs regulated by common miRNAs were considered as ceRNAs to each other.

According to topological analysis, ST8 alpha-N-acetyl-neuraminide alpha-2,8-sialyltransferase 3 (*ST8SIA3*)-RFPL1 Antisense RNA 1 (*RFPL1S*) was the mRNA-lncRNA pair with the most common regulating miRNAs (n=4). *RP11-863P13.4* (n=9) was the lncRNA with most regulating miRNAs. DLG associated protein 2 (*DLGAP2*) (n=7) and *SV2B* (n=7) were the mRNAs with most regulating miRNAs. hsa-miR-485-5p (n=47), hsa-miR-339-5p (n=45) and hsa-miR-770-5p (n=38) were the miRNAs with most target mRNAs. According to pathway enrichment analysis, mRNAs in the ceRNA network mainly enriched in KEGG pathways such as nicotine addiction, morphine addiction, neuroactive ligand-receptor interaction (Figure 4B).

Prognosis related mRNAs and lncRNAs. In order to identify prognosis related mRNAs and lncRNAs from the ceRNA network, log rank test and Kaplan-Meier survival analysis were performed. As a result, 10 OS related mRNAs, 1 DFS related mRNA and 1 DFS related lncRNA were identified (Table 4). The OS related mRNAs were cyclin and CBS domain divalent metal cation transport mediator 1 (*CNNM1*), cellular repressor of E1A stimulated genes 2

Table 4. Prognostic markers identified from the competing endogenous RNA (ceRNA) network.

	Names	Log-rank p-value	High.50% (Months) [a]	Low.50% (Months) [b]
OS [c]				
mRNA	CNNM1	0.016263	12.55	15.93
	CREG2	0.033294	12.48	14.91
	PAK1	0.022436	12.55	15.77
	PSD	0.021277	12.48	15.93
	RXFP1	0.025564	11.83	14.91
	SLC12A5	0.040895	11.83	14.91
	SLC4A10	0.029783	11.83	14.91
	SV2B	0.00808	10.94	15.93
	SYT1	0.015276	11.83	15.77
	SYT13	0.034539	12.48	15.77
DFS [d]				
mRNA	PAK1	0.031982	6.41	7.62
lncRNA	TRHDE-AS1	0.014188	10.22	5.98

[a] Median survival time of high risk group; [b] Median survival time of low risk group; [c] Overall survival; [d] Disease-free survival.

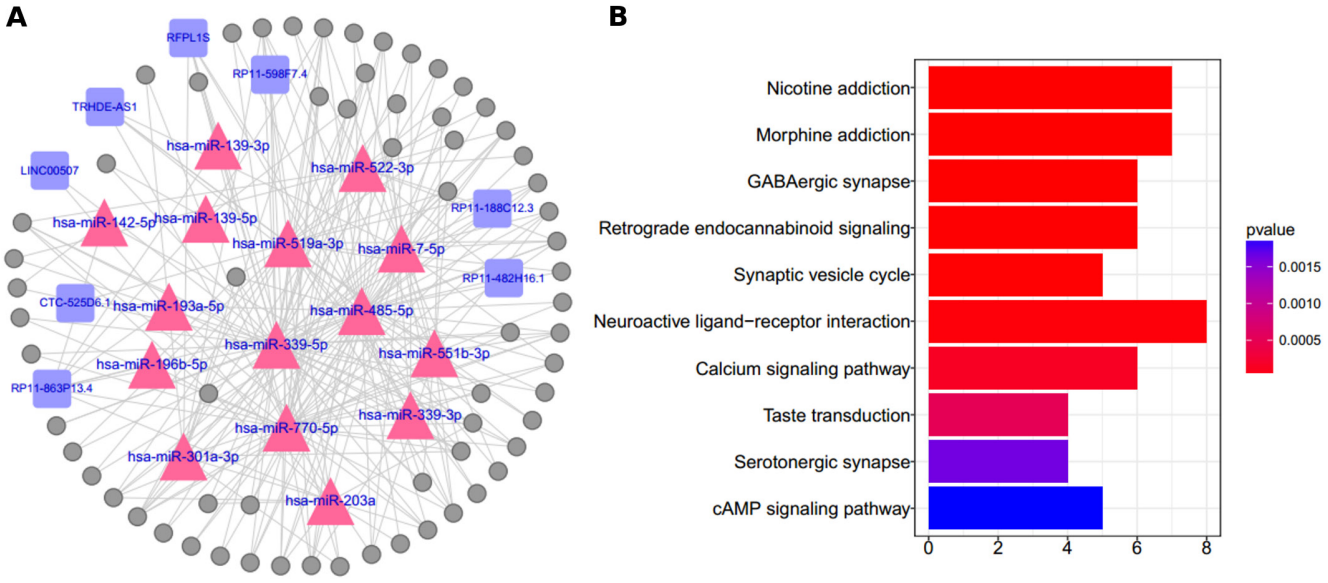

Figure 4. The GBM associated competing endogenous RNA (ceRNA) network. A) The ceRNA network. miRNAs, lncRNAs and mRNAs were shown as triangles, squares and circles, respectively. Regulation and coexpression relationships were shown as gray lines. B) Pathway items enriched by mRNAs in the ceRNA network. The vertical axis indicated the items and the horizontal axis indicated the number of mRNAs enriched.

(*CREG2*), *PAK1*) pleckstrin and Sec7 domain containing (*PSD*), relaxin family peptide receptor 1 (*RXFP1*), solute carrier family 12 member 5 (*SLC12A5*), solute carrier family 4 member 10 (*SLC4A10*), *SV2B*, synaptotagmin 1 (*SYT1*) and synaptotagmin 13 (*SYT13*). The DFS related mRNA and lncRNA were *PAK1* and TRHDE antisense RNA 1 (*TRHDE-AS1*), respectively.

Discussion

In the present study, we comprehensively analyzed the expression profiles of GBM and identified differentially expressed mRNAs, miRNAs and lncRNAs between tumor and normal samples. A subset of differentially expressed mRNAs was identified and used to build a GBM risk model, which was efficient and reliable in discriminating high risk from low risk GBM samples. In addition, a GBM associated ceRNA network was also constructed to delineate the interplays between differentially expressed mRNAs, miRNAs and lncRNAs.

RSF has been shown as a promising approach to identify disease associated molecular markers [41, 42]. In our study, 20 mRNAs were identified and used to build an RSF model, which showed a high performance in risk evaluation. When evaluated using the risk model, GBM samples could be classified to be either low or high-risk samples. According to survival analysis, high risk samples showed a significantly worse prognosis than low risk samples ($p<0.0001$) in both training and validation dataset. Therefore, our GBM risk model may serve as a novel diagnostic tool for GBM patients and may provide useful information for further treatment.

Deregulation of signaling pathways related to cancer progression is a common feature of cancers [43]. In accordance with this, our functional annotation analysis showed that p53 signaling pathway and PI3K-Akt signaling pathway were significantly enriched by differentially expressed mRNAs, indicating that these two pathways were dysregulated in GBM. P53 is a well characterized tumor suppressor, the function of which is essential for cell cycle arrest and apoptosis [44]. Impaired function of p53 is a frequent event that has been confirmed to significantly correlate with cancer cell invasion and metastasis in GBM [44]. Recently, restoring the function of p53 has been proposed as a reasonable approach for GBM treatment [45, 46]. Similar to attenuated p53 function in GBM, activation of PI3K/Akt signaling pathway is also frequently involved in GBM development and progression [47]. Inhibitors of PI3K and Akt have been considered as promising therapeutics of GBM [47–49]. Taken together, dysregulation of p53 signaling pathway and PI3K/Akt signaling pathway are major contributors to GBM progression.

According to survival analysis of the ceRNA network, *CNNM1*, *CREG2*, *PAK1*, *PSD*, *RXFP1*, *SLC12A5*, *SLC4A10*, *SV2B*, *SYT1*, *SYT13* were identified to be OS related mRNAs. Among these mRNAs, *PAK1* was noticeable and may be a key prognostic marker of GBM, as it was also a DFS related mRNA. According to our survival analysis, overexpression of *PAK1* correlated with better prognosis, suggesting that *PAK1* may serve as an oncogene in GBM. Consistent with this, PAK1 is a serine/threonine kinase downstream of PI3K/Akt signaling [50], further supporting a link between *PAK1* and GBM. Though the specific roles of *PAK1* in GBM remain elusive, *PAK1* plays important role in cell motility, invasion and metastasis in many other cancers [51–56]. For example, *PAK1* is strongly amplified in breast cancer and induces breast cancer cell transformation by activating MAPK and MET signaling [51, 52, 57]. *PAK1* also shows significant expression elevation in malignant colon carcinoma and promotes colon cancer progression by phosphorylating and activating β-catenin [53, 54]. Besides, upregulation of *PAK1* is also associated with non-small cell lung cancer cell invasiveness, whereas downregulation of *PAK1* inhibits non-small cell lung cancer progression [55, 56]. Therefore, we speculated that *PAK1* also plays essential role in GBM progression.

In addition to *PAK1*, *SV2B* may also serve as a key prognostic mRNA of GBM. *SV2B* was an OS related mRNA, as well as a hub node of the ceRNA network. SV2B is a membrane glycoprotein and is functionally important for neurotransmission processes [58, 59]. A recent bioinformatic analysis has shown that *SV2B* is associated with GBM [60]. Moreover, *SV2A*, a homologous gene of *SV2B*, is correlated with clinical response to levetiracetam treatment in glioma [61], further supporting a role of *SV2B* in GBM progression.

One of the main advantages of our study was the construction of a GBM risk model using RSF. The risk model was efficient and reliable for risk evaluation of GBM patients. In addition, a GBM associated ceRNA network was also constructed and provided a comprehensive view of the interplays between differentially expressed miRNAs, mRNAs and lncRNAs and may contribute to our understanding of the molecular mechanisms underlying GBM. Based on the ceRNA network, GBM associated prognostic markers such as *PAK1* and *SV2B* were further identified. However, there were also limitations in our study. Insufficient samples were included in our study and more samples should be included in future studies. Experimental studies are also needed to validate the involvement of prognostic markers in GBM and to provide a detailed understanding of the molecular mechanisms related to these markers.

In summary, we constructed a GBM risk model through RSF analysis. The risk model showed a high performance in discriminating GBM patients with different risk levels. In addition, we also identified *PAK1* and *SV2B* as key prognostic markers of GBM through ceRNA network analysis. Higher expression of *PAK1* and *SV2B* correlated with worse prognosis. Both the risk model and the prognostic markers may provide valuable information and contribute to outcome prediction of GBM in future clinical practice.

Acknowledgements: This study was supported by the science and technology planning project of Beijing Tongzhou district (KJ2016CX037-13).

References

[1] ALIFIERIS C, TRAFALIS DT. Glioblastoma multiforme: Pathogenesis and treatment. Pharmacol Ther 2015; 152: 63–82. https://doi.org/10.1016/j.pharmthera.2015.05.005

[2] HUSE JT, HOLLAND EC. Targeting brain cancer: advances in the molecular pathology of malignant glioma and medulloblastoma. Nature Rev Cancer 2010; 10: 319–331. https://doi.org/10.1038/nrc2818

[3] ADAMSON C, KANU OO, MEHTA AI, DI C, LIN N et al. Glioblastoma multiforme: a review of where we have been and where we are going. Expert Opin Investig Drugs 2009; 18: 1061–1083. https://doi.org/10.1517/13543780903052764

[4] HOLLAND EC. Gliomagenesis: genetic alterations and mouse models. Nat Rev Genet 2001; 2: 120–129. https://doi.org/10.1038/35052535

[5] OZER A, WU LC, BRUICK RK. The candidate tumor suppressor ING4 represses activation of the hypoxia inducible factor (HIF). Proc Natl Acad Sci U S A 2005; 102: 7481–7486. https://doi.org/10.1073/pnas.0502716102

[6] BYRON SA, MIN E, THAL TS, HOSTETTER G, WATANABE AT et al. Negative regulation of NF-kappaB by the ING4 tumor suppressor in breast cancer. PLoS One 2012; 7: e46823. https://doi.org/10.1371/journal.pone.0046823

[7] GARKAVTSEV I, KOZIN SV, CHERNOVA O, XU L, WINKLER F et al. The candidate tumour suppressor protein ING4 regulates brain tumour growth and angiogenesis. Nature 2004; 428: 328–332. https://doi.org/10.1038/nature02329

[8] ARIMAPPAMAGAN A, SOMASUNDARAM K, THENNARASU K, PEDDAGANGANNAGARI S, SRINIVASAN H et al. A fourteen gene GBM prognostic signature identifies association of immune response pathway and mesenchymal subtype with high risk group. PLoS One 2013; 8: e62042. https://doi.org/10.1371/journal.pone.0062042

[9] ESTELLER M. Non-coding RNAs in human disease. Nat Rev Genet 2011; 12: 861–874. https://doi.org/10.1038/nrg3074

[10] LING H, FABBRI M, CALIN GA. MicroRNAs and other non-coding RNAs as targets for anticancer drug development. Nat Rev Drug Discov 2013; 12: 847–865. https://doi.org/10.1038/nrd4140

[11] LIZ J, ESTELLER M. lncRNAs and microRNAs with a role in cancer development. Biochim Biophys Acta 2016; 1859: 169–176. https://doi.org/10.1016/j.bbagrm.2015.06.015

[12] WANG Q, LI P, LI A, JIANG W, WANG H et al. Plasma specific miRNAs as predictive biomarkers for diagnosis and prognosis of glioma. J Exp Clin Cancer Res 2012; 31: 97. https://doi.org/10.1186/1756-9966-31-97

[13] ZHOU M, ZHANG Z, ZHAO H, BAO S, CHENG L et al. An Immune-Related Six-lncRNA Signature to Improve Prognosis Prediction of Glioblastoma Multiforme. Mol Neurobiol 2018; 55: 3684–3697. https://doi.org/10.1007/s12035-017-0572-9

[14] ZHANG W, ZHANG J, HOADLEY K, KUSHWAHA D, RAMAKRISHNAN V et al. miR-181d: a predictive glioblastoma biomarker that downregulates MGMT expression. Neuro Oncol 2012; 14: 712–719. https://doi.org/10.1093/neuonc/nos089

[15] CHEN L, ZHANG W, YAN W, HAN L, ZHANG K et al. The putative tumor suppressor miR-524-5p directly targets Jagged-1 and Hes-1 in glioma. Carcinogenesis 2012; 33: 2276–2282. https://doi.org/10.1093/carcin/bgs261

[16] HARROW J, FRANKISH A, GONZALEZ JM, TAPANARI E, DIEKHANS M et al. GENCODE: the reference human genome annotation for The ENCODE Project. Genome Res 2012; 22: 1760–1774. https://doi.org/10.1101/gr.135350.111

[17] ROBINSON MD, MCCARTHY DJ, SMYTH GK. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics 2010; 26: 139–140. https://doi.org/10.1093/bioinformatics/btp616

[18] MCCARTHY DJ, CHEN Y, SMYTH GK. Differential expression analysis of multifactor RNA-Seq experiments with respect to biological variation. Nucleic Acids Res 2012; 40: 4288–4297. https://doi.org/10.1093/nar/gks042

[19] BOLSTAD BM, IRIZARRY RA, ASTRAND M, SPEED TP. A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. Bioinformatics 2003; 19: 185–193.

[20] IRIZARRY RA, HOBBS B, COLLIN F, BEAZER-BARCLAY YD, ANTONELLIS KJ et al. Exploration, normalization, and summaries of high density oligonucleotide array probe level data. Biostatistics 2003; 4: 249–264. https://doi.org/10.1093/biostatistics/4.2.249

[21] SHI W, OSHLACK A, SMYTH GK. Optimizing the noise versus bias trade-off for Illumina whole genome expression BeadChips. Nucleic Acids Res 2010; 38: e204. https://doi.org/10.1093/nar/gkq871

[22] SMYTH GK. limma: Linear Models for Microarray Data, p 397–420. In: R. Gentleman, VJ. Carey, W. Huber, RA. Irizarry, S. Dudoit (Eds.), Bioinformatics and Computational Biology Solutions Using R and Bioconductor. Springer New York, NY, 2005, pp 465. ISBN 9870387251462.

[23] BENJAMINI Y HY, HOCHBERG Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. J R Statist Soc B 1995; 57: 289–300.

[24] ASHBURNER M, BALL CA, BLAKE JA, BOTSTEIN D, BUTLER H et al. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat Genet 2000; 25: 25–29. https://doi.org/10.1038/75556

[25] KANEHISA M, GOTO S. KEGG: kyoto encyclopedia of genes and genomes. Nucleic Acids Res 2000; 28: 27–30.

[26] HUANG DW, SHERMAN BT, TAN Q, COLLINS JR, ALVORD WG et al. The DAVID Gene Functional Classification Tool: a novel biological module-centric algorithm to functionally analyze large gene lists. Genome Biol 2007; 8: R183. https://doi.org/10.1186/gb-2007-8-9-r183

[27] CALDERONE A, CASTAGNOLI L, CESARENI G. mentha: a resource for browsing integrated protein-interaction networks. Nat Methods 2013; 10: 690–691. https://doi.org/10.1038/nmeth.2561

[28] CHATR-ARYAMONTRI A, BREITKREUTZ BJ, OUGHTRED R, BOUCHER L, HEINICKE S et al. The BioGRID interaction database: 2015 update. Nucleic Acids Res 2015; 43: D470–478. https://doi.org/10.1093/nar/gku1204
[29] KESHAVA PRASAD TS, GOEL R, KANDASAMY K, KEERTHIKUMAR S, KUMAR S et al. Human Protein Reference Database-2009 update. Nucleic Acids Res 2009; 37: D767–772. https://doi.org/10.1093/nar/gkn892
[30] SHANNON P, MARKIEL A, OZIER O, BALIGA NS, WANG JT et al. Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome Res 2003; 13: 2498–2504. https://doi.org/10.1101/gr.1239303
[31] TANG Y, LI M, WANG J, PAN Y, WU FX. CytoNCA: a cytoscape plugin for centrality analysis and evaluation of biological networks. Biosystems 2015; 127: 67–72. https://doi.org/10.1016/j.biosystems.2014.11.005
[32] HE X, ZHANG J. Why do hubs tend to be essential in protein networks? PLoS Genet 2006; 2: e88. https://doi.org/10.1371/journal.pgen.0020088
[33] BADER GD, HOUGE CW. An automated method for finding molecular complexes in large protein interaction networks. BMC Bioinformatics 2003; 4: 2. https://doi.org/10.1186/1471-2105-4-2
[34] ISHWARAN H, KOGALUR UB, BLACKSTONE EH, LAUER MS. Random survival forests. Ann Appl Stat 2008; 2: 841–860. https://doi.org/10.1214/08-AOAS169
[35] ISHWARAN H. Variable importance in binary regression trees and forests. Electron J Statist 2007; 1: 519–537. https://doi.org/10.1214/07-EJS039
[36] VILLANUEVA A, PORTELA A, SAYOLS S, BATTISTON C, HOSHIDA Y et al. DNA methylation-based prognosis and epidrivers in hepatocellular carcinoma. Hepatology 2015; 61: 1945–1956. https://doi.org/10.1002/hep.27732
[37] DWEEP H, GRETZ N. miRWalk2. 0: a comprehensive atlas of microRNA-target interactions. Nat Methods 2015; 12: 697. https://doi.org/10.1038/nmeth.3485
[38] LI J-H, LIU S, ZHOU H, QU L-H, YANG J-H. starBase v2.0: decoding miRNA-ceRNA, miRNA-ncRNA and protein–RNA interaction networks from large-scale CLIP-Seq data. Nucleic Acids Res 2014; 42: D92–D97. https://doi.org/10.1093/nar/gkt1248
[39] DAS S, GHOSAL S, SEN R, CHAKRABARTI J. lnCeDB: database of human long noncoding RNA acting as competing endogenous RNA. PLoS One 2014; 9: e98965. https://doi.org/10.1371/journal.pone.0098965
[40] YU G, WANG LG, HAN Y, HE QY. clusterProfiler: an R package for comparing biological themes among gene clusters. Omics 2012; 16: 284–287. https://doi.org/10.1089/omi.2011.0118
[41] DIETRICH S, FLOEGEL A, TROLL M, KUHN T, RATHMANN W et al. Random Survival Forest in practice: a method for modelling complex metabolomics data in time to event analysis. Int J Epidemiol 2016; 45: 1406–1420. https://doi.org/10.1093/ije/dyw145
[42] DATEMA FR, MOYA A, KRAUSE P, BACK T, WILLMES L et al. Novel head and neck cancer survival analysis approach: random survival forests versus Cox proportional hazards regression. Head Neck 2012; 34: 50–58. https://doi.org/10.1002/hed.21698
[43] SEVER R, BRUGGE JS. Signal transduction in cancer. Cold Spring Harbor perspectives in medicine 2015; 5. https://doi.org/10.1101/cshperspect.a00609
[44] ENGLAND B, HUANG T, KARSY M. Current understanding of the role and targeting of tumor suppressor p53 in glioblastoma multiforme. Tumour Biol 2013; 34: 2063–2074. https://doi.org/10.1007/s13277-013-0871-3
[45] VILLALONGA-PLANELLS R, COLL-MULET L, MARTINEZ-SOLER F, CASTANO E, ACEBES JJ et al. Activation of p53 by nutlin-3a induces apoptosis and cellular senescence in human glioblastoma multiforme. PLoS One 2011; 6: e18588. https://doi.org/10.1371/journal.pone.0018588
[46] COSTA B, BENDINELLI S, GABELLONI P, DA POZZO E, DANIELE S et al. Human glioblastoma multiforme: p53 reactivation by a novel MDM2 inhibitor. PLoS One 2013; 8: e72281. https://doi.org/10.1371/journal.pone.0072281
[47] HENNESSY BT, SMITH DL, RAM PT, LU Y, MILLS GB. Exploiting the PI3K/AKT pathway for cancer drug discovery. Nat Rev Drug Discov 2005; 4: 988–1004. https://doi.org/10.1038/nrd1902
[48] LI X, WU C, CHEN N, GU H, YEN A et al. PI3K/Akt/mTOR signaling pathway and targeted therapy for glioblastoma. Oncotarget 2016; 7: 33440–33450. https://doi.org/10.18632/oncotarget.7961
[49] WESTHOFF MA, KARPEL-MASSLER G, BRUHL O, ENZENMULLER S, LA FERLA-BRUHL K et al. A critical evaluation of PI3K inhibition in Glioblastoma and Neuroblastoma therapy. Mol Cell Ther 2014; 2: 32. https://doi.org/10.1186/2052-8426-2-32
[50] SAJI M, RINGEL MD. The PI3K-Akt-mTOR pathway in initiation and progression of thyroid tumors. Mol Cell Endocrinol 2010; 321: 20–28. https://doi.org/10.1016/j.mce.2009.10.016
[51] BAGHERI-YARMAND R, MANDAL M, TALUDKER AH, WANG RA, VADLAMUDI RK et al. Etk/Bmx tyrosine kinase activates Pak1 and regulates tumorigenicity of breast cancer cells. J Biol Chem 2001; 276: 29403–29409. https://doi.org/10.1074/jbc.M103129200
[52] MAZUMDAR A, KUMAR R. Estrogen regulation of Pak1 and FKHR pathways in breast cancer cells. FEBS Lett 2003; 535: 6–10. https://doi.org/10.1016/S0014-5793(02)03846-2
[53] ZHU G, WANG Y, HUANG B, LIANG J, DING Y et al. A Rac1/PAK1 cascade controls beta-catenin activation in colon cancer cells. Oncogene 2012; 31: 1001–1012. https://doi.org/10.1038/onc.2011.294
[54] CARTER JH, DOUGLASS LE, DEDDENS JA, COLLIGAN BM, BHATT TR et al. Pak-1 expression increases with progression of colorectal carcinomas to metastasis. Clin Cancer Res 2004; 10: 3448–3456. https://doi.org/10.1158/1078-0432.CCR-03-0210
[55] YANG G, ZHANG X, SHI J. MiR-98 inhibits cell proliferation and invasion of non-small cell carcinoma lung cancer by targeting PAK1. Int J Clin Exp Med 2015; 8: 20135–20145.
[56] RETTIG M, TRINIDAD K, PEZESHKPOUR G, FROST P, SHARMA S et al. PAK1 kinase promotes cell motility and invasiveness through CRK-II serine phosphorylation in non-small cell lung cancer cells. PLoS One 2012; 7: e42012. https://doi.org/10.1371/journal.pone.0042012

[57] SHRESTHA Y, SCHAFER EJ, BOEHM JS, THOMAS SR, HE F et al. PAK1 is a breast cancer oncogene that coordinately activates MAPK and MET signaling. Oncogene 2012; 31: 3397–3408. https://doi.org/10.1038/onc.2011.515

[58] CROWDER KM, GUNTHER JM, JONES TA, HALE BD, ZHANG HZ et al. Abnormal neurotransmission in mice lacking synaptic vesicle protein 2A (SV2A). Proc Natl Acad Sci U S A 1999; 96: 15268–15273. https://doi.org/10.1073/pnas.96.26.15268

[59] JANZ R, GODA Y, GEPPERT M, MISSLER M, SUDHOF TC. SV2A and SV2B function as redundant Ca2+ regulators in neurotransmitter release. Neuron 1999; 24: 1003–1016. https://doi.org/10.1016/S0896-6273(00)81046-6

[60] CHIU YC, WANG LJ, LU TP, HSIAO TH, CHUANG EY et al. Differential correlation analysis of glioblastoma reveals immune ceRNA interactions predictive of patient survival. BMC Bioinformatics 2017; 18: 132. https://doi.org/10.1186/s12859-017-1557-4

[61] DE GROOT M, ARONICA E, HEIMANS JJ, REIJNEVELD JC. Synaptic vesicle protein 2A predicts response to levetiracetam in patients with glioma. Neurology 2011; 77: 532–539. https://doi.org/10.1212/WNL.0b013e318228c110