Am J Transl Res 2020;12(4):1222-1238
www.ajtr.org /ISSN:1943-8141/AJTR0101960

Original Article

Identification of biomarkers for the transition from low-grade glioma to secondary glioblastoma by an integrated bioinformatic analysis

Liang Zhao*, Jiayue Zhang*, Zhiyuan Liu, Peng Zhao

*Department of Neurosurgery, The First Affiliated Hospital of Nanjing Medical University, Nanjing, Jiangsu, China. *Equal contributors.*

Received September 5, 2019; Accepted April 2, 2020; Epub April 15, 2020; Published April 30, 2020

Abstract: Secondary glioblastoma (sGBM) is a type of glioblastoma multiforme that evolves from low-grade glioma (LGG). However, the mechanism of this transition still remains poorly understood. In this study, we used weighted gene co-expression network analysis (WGCNA) on the gene expression profiles of glioma samples from the Chinese Glioma Genome Atlas (CGGA) database to identify key genetic module related to distinguish histological characteristics. Here, the brown module was highly correlated with histological characteristics and was selected as the hub module. By applying functional annotation analysis, we found that biological processes related to the cell-cycle and DNA-replication were enriched in the genes of the brown module. After constructing a protein-protein interaction (PPI) network, validation of differential gene expression, and survival analyses, we ultimately identified five hub genes: CCNB2 (Cyclin B2), KIF2C (Kinesin Family Member 2C), CDC20 (Cell Division Cycle 20), TPX2 (TPX2 Microtubule Nucleation Factor), and PLK1 (Polo Like Kinase 1). In addition, a computational risk model was developed for predicting the clinical outcomes of sGBM patients by combining gene expression levels. This gene signature was demonstrated to be an independent predictor of survival by univariate and multivariable Cox regression analysis. Finally, we used the Genomics of Drug Sensitivity in Cancer (GDSC) database to predict the responses of sGBM patients to routine chemotherapeutic drugs. Patients from the high-risk group were more sensitive to common chemotherapies during clinical treatment. Our findings based on comprehensive analyses might advance the understanding of sGBM transition and aid the development of novel biomarkers for diagnosing and predicting the survival of sGBM patients.

Keywords: Secondary glioblastoma, transition, bioinformatics analysis, weighted gene coexpression network analysis (WGCNA), molecular mechanism

Introduction

Glioma is the most common intracranial malignancy, and the prognosis remains poor in most cases. According to the World Health Organization (WHO) classification of central-nervous-system tumors [1], grade IV glioma-also known as glioblastoma multiforme (GBM)-is the most lethal and aggressive brain tumor [2]. GBM can be classified into two distinct subtypes. The *de novo* tumors without a prior malignant lesion can be classified as "primary GBM (pGBM)", whereas GBMs originating from low-grade glioma (LGG) are defined as "secondary GBM (sGBM)" [3]. Although sGBM shares certain histological similarities with pGBM, they differ in genetic and epigenetic aspects [3]. The phenotype of sGBM is often more aggressive, with significantly poorer clinical outcomes after developing from LGG. Accordingly, the median overall survival of sGBM patients (7.8 months) is much shorter than that of LGG patients (approximately seven years) [4, 5]. Despite intensive therapeutic methods, including surgical resection, chemotherapy and radiotherapy, the clinical efficacy of sGBM treatment still remains unsatisfactory [6]. Most studies on sGBM have mainly focused on exploring the biological differences between pGBM and sGBM [4, 7], and have rarely paid attention to the mechanisms of the transition from LGG to sGBM. Therefore, the changes in genetic profiles that accompany this conversion should be urgently clarified to aid the search for more

effective biomarkers and therapeutic targets for sGBM.

With the technological development of microarray and high-throughput sequencing methods, gene expression profiles have been widely used to identify potential key targets behind the vital molecular mechanisms for subsequent research. However, most studies have merely focused on seeking differentially expressed genes but ignored the interactions among them. Weighted gene co-expression network analysis (WGCNA) [8] and protein-protein interaction (PPI) network are powerful methods for exploring the correlations between gene clusters and clinical features. To date, the WGCNA algorithm has been widely used in studies of different diseases, especially various cancers [9]. The Chinese Glioma Genome Atlas (CGGA), a database consists of over 2000 samples from Chinese glioma cohorts, has provided a considerable amount of genomic and clinical data for glioma, offering a possibility to better understand the biology and pathology of this severe malignancy.

In the present study, we used systematic bioinformatic approaches to explore the potential diagnostic and prognostic targets of sGBM. A co-expression network was constructed and several key genes inside the hub module were identified. A risk-score model was built to evaluate the effect of these hub genes on the prognosis of sGBM patients. This study may improve our understanding of the genetic changes and potential mechanisms of the transition from LGG to sGBM, and may provide new ideas for the development of efficacious therapies for treating sGBM.

Material and methods

Data collection and preprocessing

The normalized gene-level RNA-sequencing, microarray data and clinical information of diffuse glioma samples ranging from WHO grade II to IV were downloaded from the CGGA database (http://www.cgga.org.cn). All recurrent LGG samples were eliminated before filtering appropriate samples. Only samples with a 'histology' valuation of LGG or sGBM were saved for further analysis. Accordingly, 142 LGG and 34 sGBM samples from the RNA-sequencing dataset were selected as the training set, and another independent dataset consisting of 151 LGG and 10 sGBM samples from the microarray gene expression profile was defined as the validation set. For the RNA-sequencing dataset, the fragments per kilobase million (FPKM) values were transformed into transcripts per kilobase million (TPM) values, which are more similar to those resulting from microarrays and more comparable between different samples [10]. All probes from the microarray data were re-annotated using the GENECODE29 GTF file to generate gene symbol names. All protein-coding genes from the two datasets were selected for subsequent analyses.

Clinical specimens

Archival paraffin-embedded LGG tissues (WHO grades II-III) were collected from eight patients who underwent surgery at the First Hospital of Nanjing Medical University. Four sGBM tissue samples were obtained from GBM patients who had previous pathohistological evidence of LGG. Written informed consent was obtained from all patients. This study was approved by the ethics committee of the First Hospital of Nanjing Medical University. All samples were collected under protocols approved by the institutional review boards of Nanjing Medical University. All tumor grades were evaluated by a pathologist according to the WHO classification criteria.

Construction of the co-expression network and module preservation analysis

The weighted co-expression network was constructed based on the RNA-sequencing dataset using the WGCNA software package [8]. Before applying the WGCNA algorithm, genes with insufficient abundance (TPM < 1) in more than 80% of all samples were removed to avoid noise and false correlations. This selection yielded 10,612 genes for the RNA-sequencing dataset. Since non-varying genes usually represent noise, the top 5000 genes according to median absolute deviation (MAD) were filtered for the following analysis. For the selected genes, a pairwise correlation matrix across all samples was calculated. Next, a soft threshold ($\beta = 11$; scale free $R^2 = 0.94$) was used to transform the correlation matrix into a signed weighted adjacency matrix. The adjacency matrix was then used to calculate the topological overlap matrix (TOM), which is a robust measure of net-

work connectivity. A cluster dendrogram, generated by average linkage hierarchical clustering of genes according to their topological overlap, was cut into modules with a minimum size of 30 using the dynamic tree-cutting function. Next, a cut-line (0.25) was applied for the module dendrogram, followed by the merging of similar modules. To assess the stability of each module identified in the microarray dataset, we conducted module preservation analysis using the module Preservation [11] method (nPermutations = 200) in the WGCNA package. In order to guarantee the independence and reproducibility of the process, duplicate samples that were present in both the microarray dataset and the RNA-sequencing dataset (five sGBM and 30 LGG patients) were excluded from the module preservation analysis. The preservation statistics, Zsummary and Median rank, were used to quantify the preservation of gene modules between two different datasets. A Zsummary > 10 indicated strong preservation and a module with a lower Median rank tended to exhibit stronger preservation statistics than a module with a higher Median rank.

Identification of clinically significant modules and functional annotation

The expression of a module eigengene (ME) was considered as representative of all genes in a given module. The correlations between MEs and clinical traits were calculated to identify the clinically significant modules. In addition, the gene significance (GS) was defined as the mediated *p*-value of each gene in the linear regression between gene expression values and clinical traits. Module significance (MS) was defined as the average absolute gene significance measured for all genes in a given module. To further clarify the mechanisms underlying the module genes, Gene Ontology (GO) term analysis and Kyoto Encyclopedia of Genes and Genomes (KEGG) functional pathway annotation were performed using the R package clusterProfiler [12]. Enriched GO terms and KEGG pathways were chosen according to the cutoff criterion of an adjusted *p*-value (FDR) < 0.05.

Identification and validation of hub genes

Hub genes are defined as a series of genes with high intramodular connectivity and are considered to play leading roles in the function of a module. Here, hub genes were filtered using the criteria of high intramodular connectivity (cor.geneModuleMembership > 0.8) and strong correlation with clinical traits inside a given module (cor.geneTraitSignificance > 0.5). The online STRING 11.0 database was used to construct a PPI network, which was then imported into Cytoscape software. The Maximal Clique Centrality (MCC) algorithm in the CytoHubba plugin [13] was used to explore important nodes in the network. The top-20 highest scoring genes were selected after the MCC process. Genes identified in both the co-expression network and PPI network were defined as candidate hub genes. The expression levels of candidate genes were validated in the CGGA RNA-sequencing and microarray datasets using boxplot. To further test the diagnostic accuracy of the candidate hub genes in distinguishing LGG and sGBM, the R package pROC [14] was used to visualize receiver-operator characteristic (ROC) curves and calculate the area under the curve (AUC).

Immunohistochemistry

Immunohistochemistry was performed on 4-μm-thick sections. Briefly, antigen retrieval was performed using sodium citrate buffer (pH 6.0), and endogenous peroxidase activity was blocked with 3% H_2O_2. Then, the appropriate primary antibodies were added and incubated overnight at 4°C. The secondary antibody was applied for 30 min at room temperature. Slides incubated with normal serum instead of the primary antibody were used as negative controls. The slides were counterstained with hematoxylin, dehydrated, and mounted. The primary antibodies were as follows: anti-CCNB2 (Proteintech, 21644-1-AP), anti-CDC20 (Proteintech, 10252-1-AP), anti-KIF2C (Proteintech, 28372-1-AP), anti-PLK1 (Abcam, ab17056), and anti-TPX2 (Abcam, ab32795). All the stained slides were scanned and captured using Pannoramic SCAN (3DHISTECH, Budapest, Hungary). The intensity of staining was assessed by measuring the ratio of the integrated optical density (IOD) to the area using Image-pro plus v6.0 software (Media Cybernetics Inc., Bethesda, MD, USA).

Construction of a prognostic signature based on hub genes

To construct the risk-score model of gene signatures for predicting the overall survival of sGBM patients, multivariate Cox proportional-

hazards regression was performed. The risk score for each sGBM patient was calculated as follows: risk score = $expr_{gene1} \times \beta_{gene1} + ex_{prgene2} \times \beta_{gene2} + ...+ expr_{gene\ n} \times \beta_{gene\ n}$. In this formula, β_{gene} is the regression coefficient calculated by the multivariate Cox proportional hazards regression model and $expr_{gene}$ represents the expression value of a given gene. To identify the specific gene signature that determines the patients' clinical outcomes, patients with risk scores above the median were defined as 'high-risk', and those with scores below the median were defined as 'low-risk'. Kaplan-Meier survival curves were visualized to compare the differences of prognosis between the high- and low-risk groups. Subsequently, time-dependent ROC analysis for overall survival was used to display the predictive capacity of the gene-signature-based risk model.

Gene set enrichment analysis

Gene set enrichment analysis (GSEA) was performed using GSEA software [15] from the Broad Institute. Differential gene analysis between high- and low-risk samples was performed using the limma package [16] and the calculated value of log2 (fold change) was used as the ranking metric. Collections from curated gene sets (C2) and hallmark gene sets (H) were downloaded from the Molecular Signatures Database (MSigDB) and were used as reference gene sets. Among them, H contains well-defined and representative biological states or processes, and C2 contains various curated canonical pathways, as well as genetic and chemical perturbations. While performing GSEA analysis, 1000 gene-set permutations were performed. The normalized enrichment score (NES) was calculated for each gene set. Enrichment *p*-values were adjusted using the Benjamini-Hochberg procedure to control the FDR [17].

Prediction of chemotherapeutic response

The chemotherapeutic response for each of the sGBM patients was predicted according to the public pharmacogenomic database, Genomics of Drug Sensitivity in Cancer (GDSC, www.cancerrxgene.org). The GDSC database contains data on a large collection of human cancer cell lines, anticancer compounds, and experimental data on drug sensitivity. The prediction of drug sensitivity (IC50) values was conducted using the R package "pRRophetic" [18], which uses a ridge regression model based on the GDSC cancer-cell-line expression profiles. Before processing, genes with low variety were removed and duplicate gene expression data was summarized as the mean value.

Statistical analysis

Kaplan-Meier survival curves were visualized to discover the differences of clinical outcomes between groups using the "survival" and "survminer" R packages. Two-sample Student's *t*-test and Wilcoxon test (Mann-Whitney test) were used to assess the significance of differences in gene expression levels and responses to chemotherapeutic drugs between two groups, respectively. The log-rank test was performed for survival analysis. For all hypothetical tests, a two-sided *p*-value < 0.05 was considered to indicate statistical significance. All statistical analyses were performed using R software (version 3.6.1, www.r-project.org).

Results

Construction of the co-expression network

The co-expression network was constructed using the CGGA RNA-sequencing dataset. Hierarchical clustering indicated that CGGA_1283 and CGGA_488 were outliers in these datasets, and they were excluded from subsequent study (Figure S1). Finally, the datasets of 33 sGBM patients and 141 LGG patients with complete clinical information were selected for WGCNA analysis (**Figure 1A**). The power of β = 11 (scale free R^2 = 0.94) was selected as the soft threshold parameter to ensure a scale-free network (**Figure 1B-E**). Specifically, 13 co-expression modules were identified after merging modules with similarities above 0.75 (**Figures 2A** and S2). The interaction relationship of the modules was analyzed by plotting a network heatmap (Figure S3). Each module showed significant independence from other modules.

Two methods were used to test the relationship between each module and the clinical information. Modules with a greater MS were considered to have a stronger connection with clinical features. The brown module had an R^2 of 0.64 (P = 4.1e-40; Figure S4), which was higher than that of other modules (**Figure 2C**). Additionally,

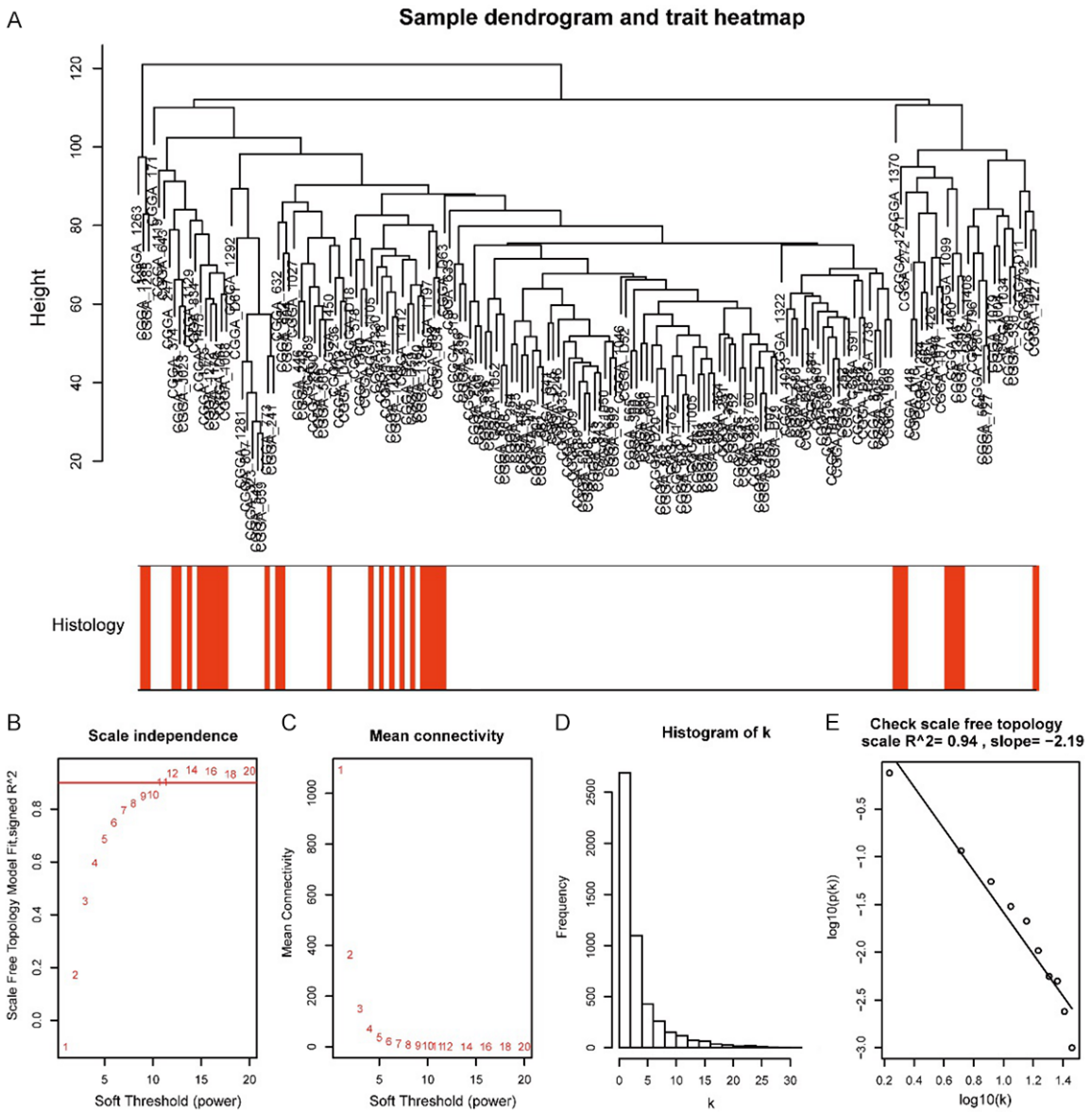


Figure 1. Clustering dendrogram of samples and selection of soft-threshold power. A. The clustering was based on the expression data from the CGGA RNA-sequencing dataset. The top 5000 genes with the highest MAD values were used for WGCNA analysis. The displayed colors correspond to the histological characteristics of samples. B. Analysis of the scale-free fit index for different soft-thresholding powers. C. Analysis of the mean connectivity for different soft-threshold powers. D. Histogram of connectivity distribution when β = 11. E. Linear model fitting of the R^2 index showing a good quality of fit (R^2 = -0.94).

the ME of the brown module showed the highest association with the histology (**Figure 2B**). Also, this module was robust and reproducible according to the module preservation analysis after comparing RNA-sequencing data with microarray gene expression data. The brown module could also be identified in another independent network with $Z_{summary}$ > 10 and relative low Median rank statistics (Figure S5). Consequently, this module was identified as the key module of interest for further analyses.

In order to provide an interpretation of the biological mechanisms underlying the impact of genes clustered in the brown module, GO functional and KEGG pathway enrichment analyses were performed on 336 genes of this module using the "clusterProfiler" R package (**Figure 3**). In the GO analysis, terms related to the cell cycle were most prominent, such as DNA replication (FDR = 5.433e-27), chromosomal segregation (FDR = 6.637e-22), sister chromatid segregation (FDR = 8.205e-20) and cell-cycle

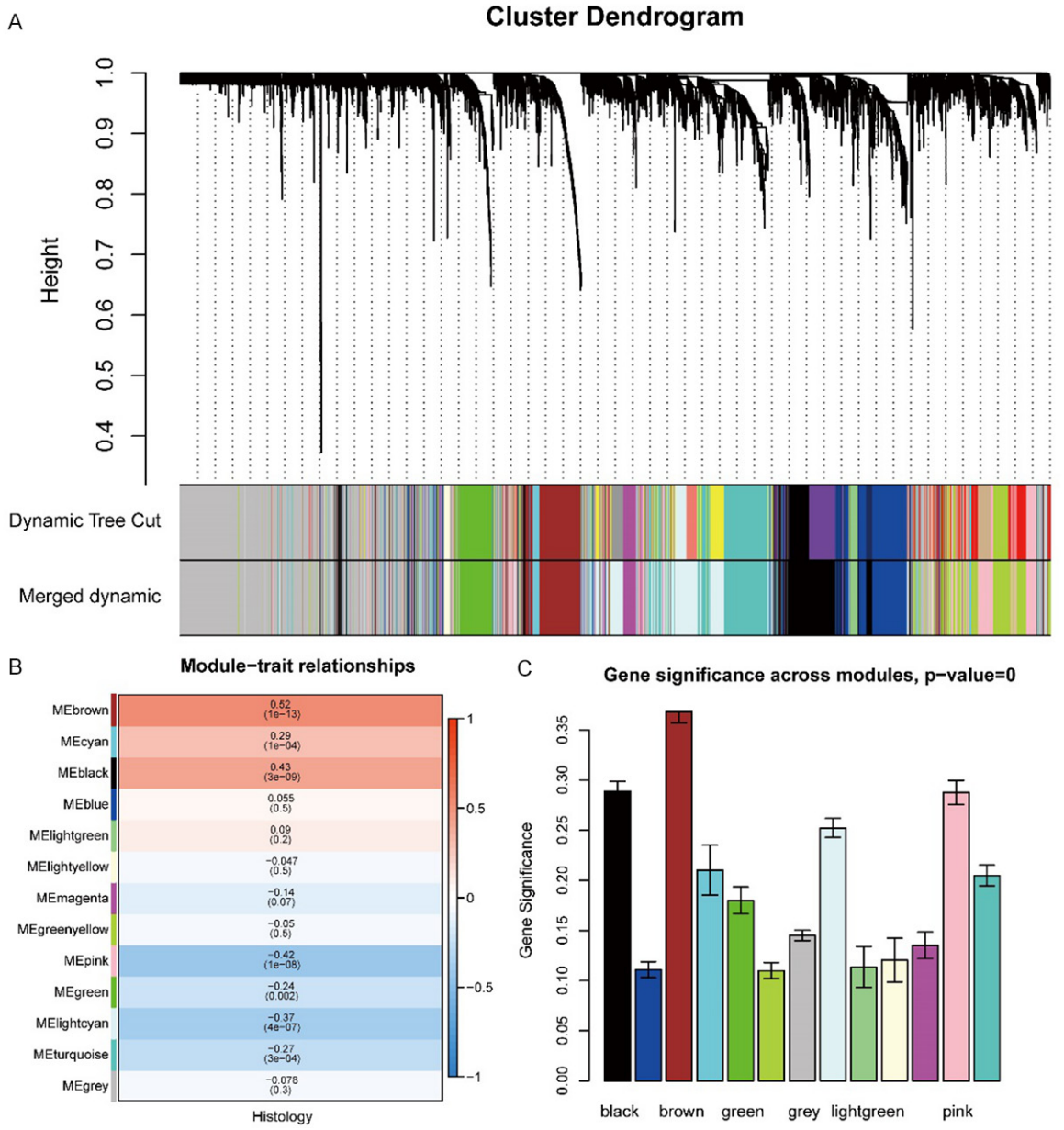


Figure 2. Identification of modules associated with the histopathological features of the samples. A. Clustering tree (dendrogram) of genes based on co-expression network analysis. Genes were clustered based on dissimilarity measure (1-TOM). Bars below correspond to modules of genes with high interconnectivity. B. Heatmap of correlations between the modules' eigengenes and histological characteristics of the samples. Each row corresponds to a specific module color. The upper number in each cell is the correlation coefficient of each module with histology, and the lower number is the *p*-value. Color is coded according to the correlation coefficient. C. Distribution of average gene significance and errors in the modules associated with the histological characteristics of the samples.

G1/S phase transition (FDR = 1.617e-16). KEGG pathway analysis showed that genes from this module were mostly enriched in cell cycle (FDR = 3.576e-18), DNA replication (FDR = 8.717e-12), human T-cell leukemia virus-1 infection (FDR = 1.333e-07), oocyte meiosis (FDR = 2.022e-07) and the p53 signaling pathway (FDR = 2.936e-07). These findings revealed that genes from the identified brown module are mainly involved in DNA replication, cell cycle regulation, and proliferation. It stands to reason that the relatively poor prognosis of sGBM is related to persistent proliferation during the transition from LGG to sGBM.

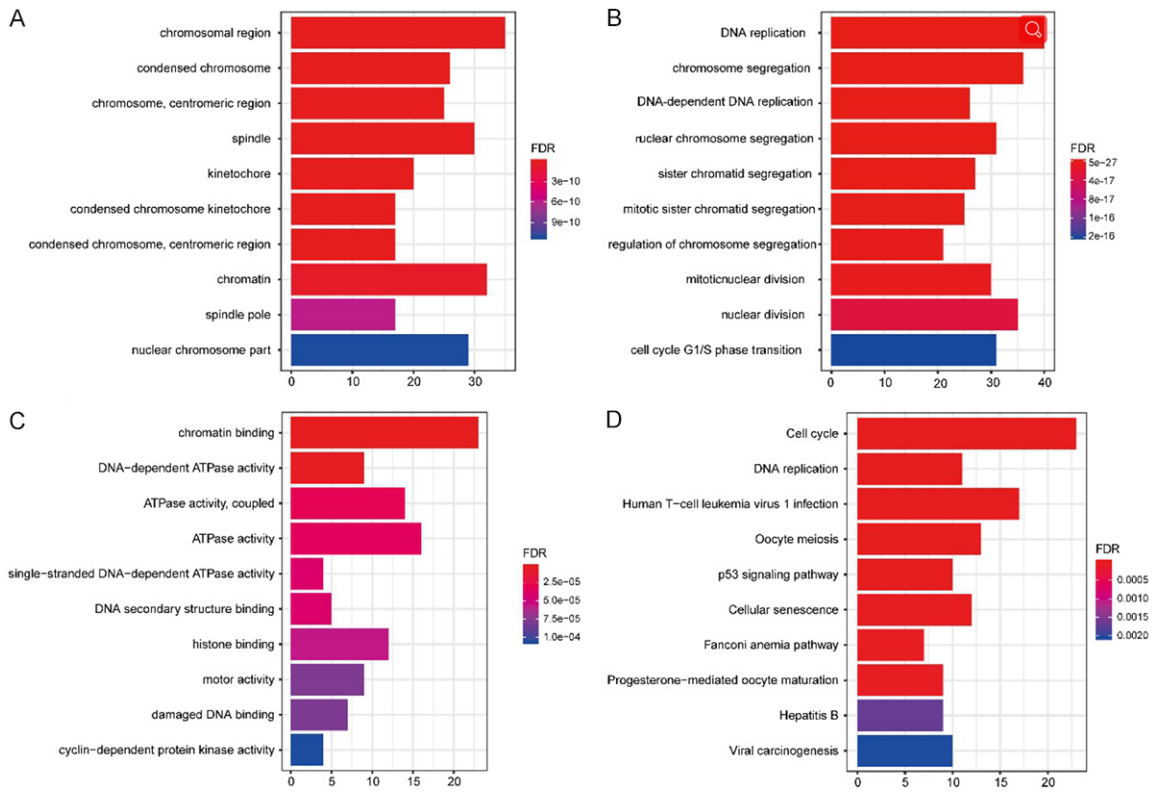


Figure 3. GO functional and KEGG pathway enrichment analysis of genes in the brown module. A. Top 10 significantly enriched biological process annotations. B. Top 10 significantly enriched cellular component annotations. C. Top 10 significantly enriched molecular function annotations. D. Top 10 significantly enriched KEGG pathways. The x-axis represents the number of genes in the corresponding gene term and the y-axis shows the gene terms. The adjusted *p*-value of each term is colored according to the legend.

Identification of hub genes in the brown module

Inspired by these findings, we further evaluated how these genes of the brown module may drive tumor development. Based on the cut-off criteria (cor.geneModuleMembership > 0.8 and cor.geneTraitSignificance > 0.5), 22 genes with high connectivity in the brown module were identified as primary hub genes. Subsequently, we constructed a network of PPI for these 22 genes according to the STRING database using Cytoscape software, and the MCC score of each node was calculated using the CytoHubba algorithm. Finally, the top-20 genes ranked by MCC scores were filtered, and the PPI network was plotted in Cytoscape using genes filtered by the two methods. Seven genes were eventually identified as potential hub genes (CCNB2, KIF2C, KIF20A, CDC20, TPX2, CCNB1 and PLK1) (**Figures 4A** and S6).

Next, we further explored the differences of expression levels between the sGBM and LGG samples, as well as the diagnostic/prognostic value of the identified potential hub genes. We used boxplots to show the relationships between the hub genes and clinical features in the RNA-sequencing and microarray datasets. The expression levels of hub genes were significantly higher in the sGBM samples than in LGG samples (**Figure 4B**). To evaluate the performance of the gene signature in distinguishing between sGBM and LGG samples, we used ROC curves to measure the true-positive rates against the false-positive rates at various expression levels of hub genes. All tested genes consistently showed a satisfactory performance in two independent datasets. Considering the limited sample size of the microarray dataset, Kaplan-Meier overall survival curves were further compared for patients with high versus low expression levels of individual hub genes in the RNA-sequencing dataset. Comparison of survival curves indicated that five of the seven genes were significantly correlated with the survival time in the RNA-sequencing dataset, with hazard ratios (HRs) ranging from

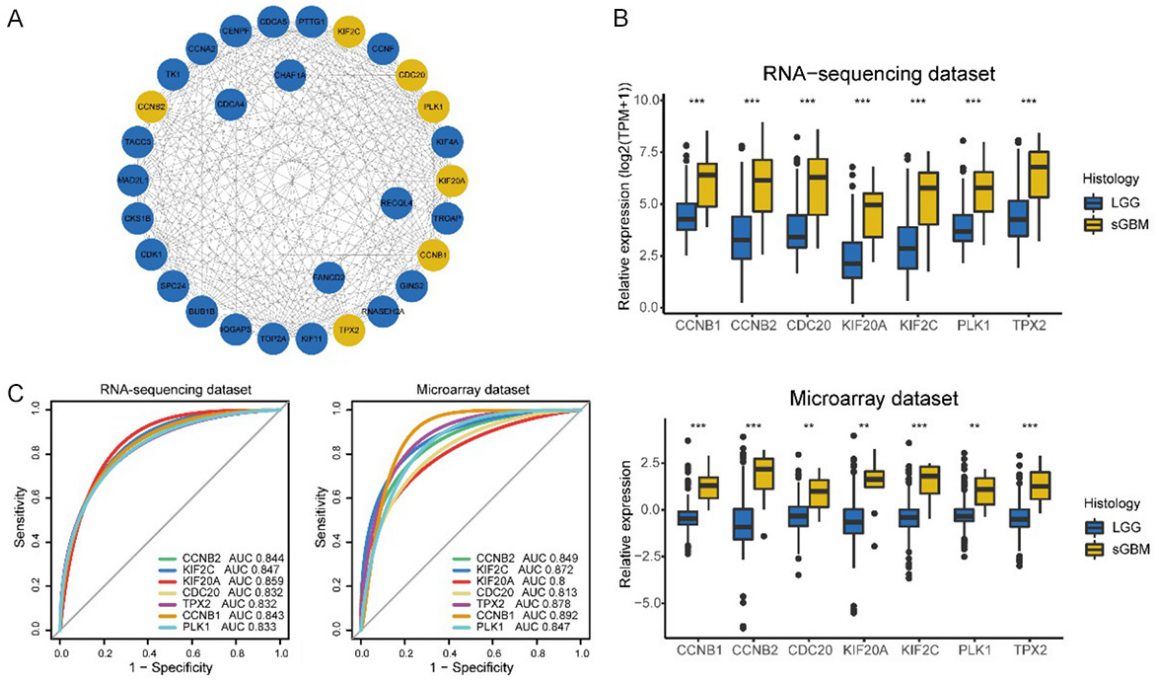


Figure 4. Detection and validation of candidate hub genes. A. PPI network of 22 genes with high connectivity and top-20 MCC genes in the brown module. Nodes colored in yellow are candidate genes identified both in the co-expression network and the PPI network. B. Boxplots of the expression levels of candidate genes in LGG and sGBM samples in the CGGA RNA-sequencing and microarray datasets. **; P < 0.01, ***; P < 0.001. Two-tailed Student's *t*-test was used to evaluate the statistical significance of differences. C. ROC curves measuring the predictive value of each candidate gene in the CGGA database. The X-axis shows the false-positive rate, shown as "1-Specifcity". The Y-axis indicates the true positive rate, shown as "Sensitivity".

2.89 to 4.73 (**Figure 5**). These *in silico* analysis results indicated that the five genes may serve as effective indicators for the diagnosis and prognosis of sGBM patients.

Validation of hub genes in clinical specimens

In order to confirm that the five identified hub genes were also overexpressed at the protein level, we carried out IHC staining and measured the expression levels of target proteins in LGG (n = 8) and sGBM (n = 4) samples from our institution. Consistent with the results of the bioinformatic analyses, all five proteins were significantly upregulated in sGBM samples compared with the LGG samples (**Figure 6**). Taken together, we concluded that the five hub genes (CCNB2, KIF2C, CDC20, TPX2 and PLK1) play an important role in promoting the transition of LGG to sGBM, and selected them as bona fide hub genes for subsequent analyses.

Construction and validation of a sGBM risk signature based on the identified hub genes

To assess the prognostic value of the five identified hub genes, we constructed a prognostic gene signature by integrating the expression levels of these genes and the corresponding regression coefficients. The risk score was calculated for each patient using the following formula: risk score = (-1.0519) expression value of CCNB2 + 2.2897 expression value of KIF2C + 1.628 expression value of CDC20 + (-1.6286) expression value of TPX2 + (-0.8481) expression value of PLK1.

Patients were subdivided into a high- and a low-risk group based on the median risk score of 0.4996. The risk score distribution, survival status, and expression profile of the hub genes are shown in **Figure 7A**. Unsurprisingly, more surviving patients were found in the low-risk group than in the high-risk group. Heatmap depicting the expression patterns of the hub genes in the two distinct risk groups showed that patients with high risk scores were characterized by upregulation of all hub genes, whereas hub genes were downregulated in low-risk patients. A Kaplan-Meier curve was plotted to analyze the distribution and correlation of patient risk scores with corresponding statuses (**Figure 7B**). The HR value of the high versus the low-risk group was 5.37 for overall survival (OS;

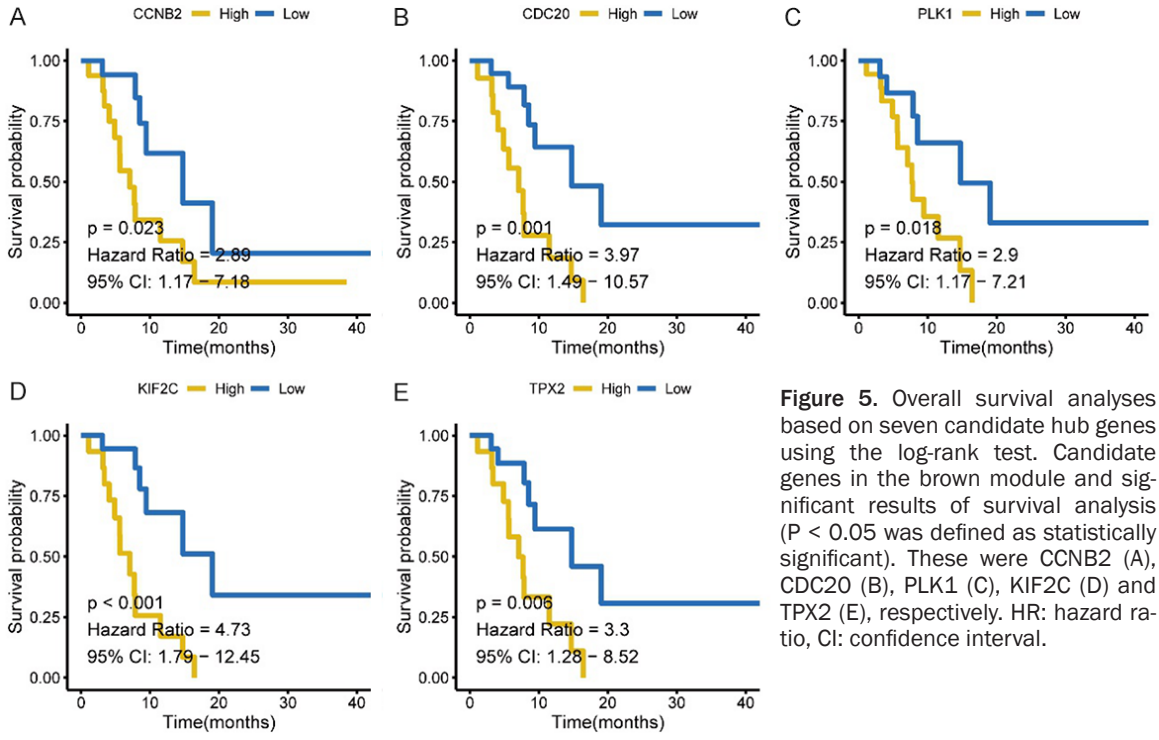


Figure 5. Overall survival analyses based on seven candidate hub genes using the log-rank test. Candidate genes in the brown module and significant results of survival analysis ($P < 0.05$ was defined as statistically significant). These were CCNB2 (A), CDC20 (B), PLK1 (C), KIF2C (D) and TPX2 (E), respectively. HR: hazard ratio, CI: confidence interval.

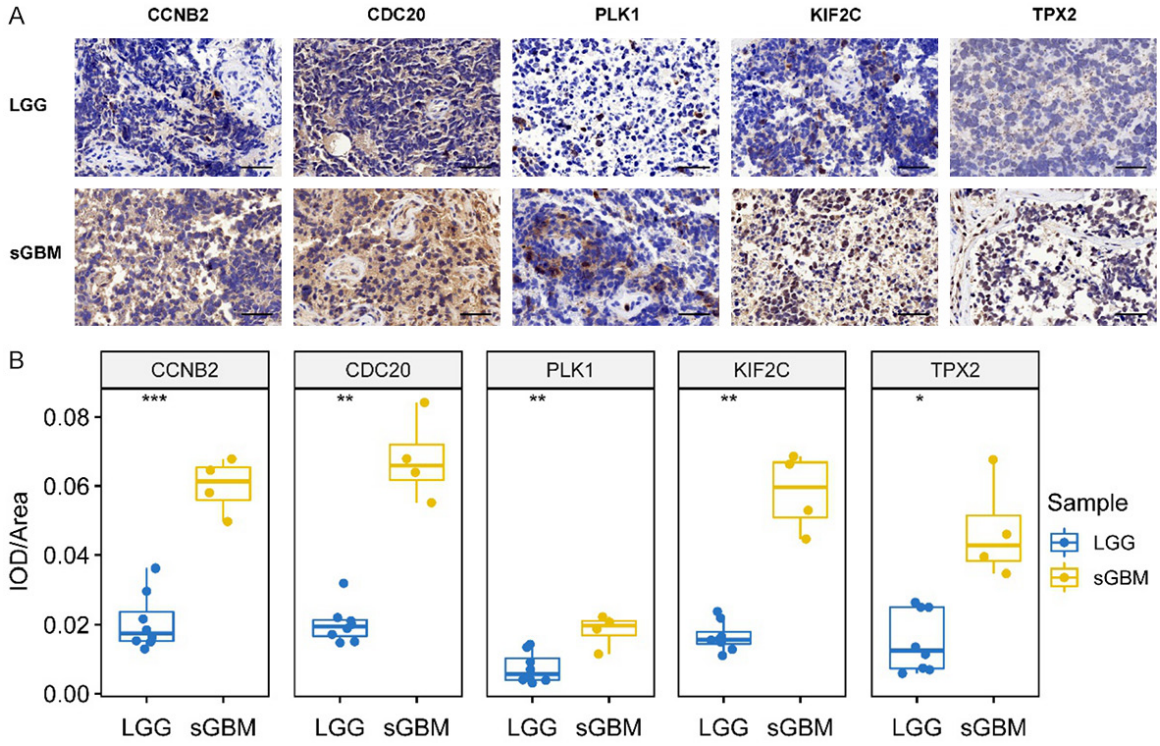


Figure 6. Protein expression levels of five hub genes in LGG and sGBM tissues. A. Representative photographs of IHC staining for CCNB2, CDC20, PLK1, KIF2C, and TPX2 in clinical human samples of LGG and sGBM. Magnification, × 200. Scale bar = 50 μm. B. Quantification of IHC staining intensity for each protein in the specimens. The integrated optical density (IOD) and the area were quantified using Image-Pro Plus software. Significance tested by Student's *t*-test, *; $P < 0.05$, **; $P < 0.01$, ***; $P < 0.001$.

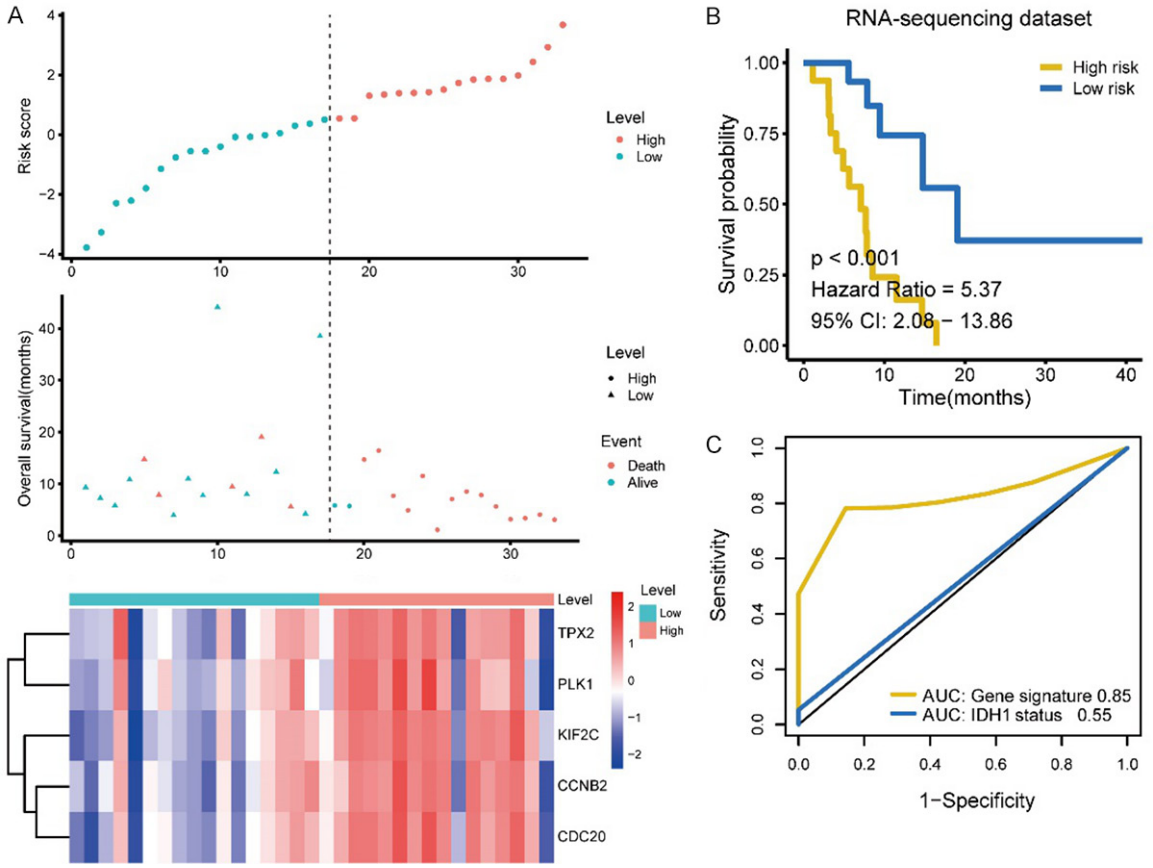


Figure 7. Construction of the risk signature for sGBM based on five hub genes. The patients were classified into high- and low-risk groups based on the median value of the risk scores. A. Risk score distribution, survival status of sGBM patients and expression heatmap of the five hub genes. Red indicates a high expression level of a given gene, whereas blue indicates a low expression level. B. Kaplan-Meier survival curves for overall survival in the high- and low-risk group. C. Time-dependent ROC curves for predicting one-year survival of sGBM patients in the RNA-sequencing dataset based on the signature and IDH1 status. The AUC of the gene signature was 0.85, and the AUC of the IDH1 mutation was 0.55.

P < 0.001, 95% confidence interval (CI) = 2.08-13.86). Thus, the patients with low risk scores had significantly longer survival times than those with high scores. Considering that the median survival time of these sGBM patients was approximately eight months, we evaluated the accuracy of this risk model in predicting the one-year survival status. The isocitrate dehydrogenase 1 (IDH1) mutation status has been previously demonstrated to be strongly correlated with increased OS and is regarded as a molecular predictor in the prognosis of sGBM patients [19]. The AUC for the risk model was 0.85, which was much higher than that of the IDH1 mutation status (0.55), suggesting that the risk model outperforms currently known biomarker in accurately predicting the survival of sGBM patients (**Figure 7C**).

To confirm whether the risk model can be used as an independent prognostic tool for sGBM, univariate and multivariate Cox regression analyses were applied to the RNA-sequencing dataset according to clinicopathologic features, including age, gender, chemotherapy, radiotherapy and IDH1 mutation status. The risk score signature was adjusted using prognostic information of clinical factors that had been deemed statistically significant in univariate analysis (P < 0.05). The HRs for the signature in the univariate and multivariate analyses were 6.53 (P < 0.001, 95% CI: 2.121-20.109) and 5.167 (P = 0.015, 95% CI: 1.368-19.52), respectively (**Table 1**). Thus, the risk score signature based on the hub genes can be used independently to predict the overall survival of sGBM patients.

Table 1. Uni- and multivariate analyses of the relationship of clinicopathological characteristics, IDH1 mutation status, and the hub gene signature with overall survival in the CGGA sGBM RNA-sequencing cohort

Variable	Univariate analysis		Multivariate analysis	
	HR (95% CI)	*P*	HR (95% CI)	*P*
Gender Male vs. female	1.332 (0.496-3.573)	0.57		
Age ≥ median vs. < median	0.709 (0.276-1.817)	0.474	1.869 (0.579-6.035)	0.295
Radiotherapy Yes vs. no	0.267 (0.075-0.957)	***0.043***	0.264 (0.062-1.124)	0.072
Chemotherapy Yes vs. no	0.13 (0.039-0.435)	***< 0.001***	0.128 (0.036-0.451)	***0.001***
IDH1 status Mutation vs. wild-type	0.707 (0.264-1.892)	0.49		
Risk score High vs. low	6.53 (2.121-20.109)	***0.001***	5.167 (1.368-19.52)	***0.015***

Note: The multivariate analysis used stepwise addition and removal of age and covariates associated with survival in the univariate models (P < 0.05). The final models only include covariates significantly associated with survival (Wald statistic, P < 0.05). Bolded italics indicate statistically significant values (P < 0.05).

Functional analysis of the risk score signature

To investigate the potential biological characteristics and processes related to the hub gene signature, we performed GSEA analysis, which is a computational method that assesses whether an a priori defined set of genes shows statistically significant and concordant differences between two groups. The results showed that signaling pathways related to the cell-cycle and epithelial-mesenchymal transition (EMT) were consistently and significantly upregulated in the high-risk samples. These include pathways implicated in the regulation of the G2/M checkpoint, E2F transcription factors, cell cycle, and elevation of the EMT (**Figure 8A**). By contrast, the low-risk group showed high enrichment for the p53 pathway, the inflammatory response, and downregulation of glioma stemness and tumor metastasis (**Figure 8B**). The "lein_oligodendrocyte_markers" and "lein_astrocyte_markers" are large gene sets comprising many oligodendrocyte and astrocyte markers, and their high enrichment in the low-risk group suggested that malignant tissues derived from these sGBM patients are more similar to those of LGG (diffuse astrocytoma, oligodendroglioma) patients at the genetic level (**Figure 8B**). In recent years, clinically relevant molecular subtypes of GBM were defined by integrated genomic analysis. Verhaak et al. used unified transcriptomic and genomic dimensions to stratify GBM into four distinct subtypes: Proneural, Neural, Classical and Mesenchymal. The Neural subtype is typified by the expression of neuronal markers, and often indicates a more favorable prognosis compared with Mesenchymal and Classical subtypes. Intriguingly, we noticed that the "verhaak_glioblastoma_neural" gene term was enriched in the low-risk group, which might partially explain the difference in prognosis between the high- and low-risk patients in this study.

Greater sensitivity to chemotherapy in the high-risk group

Considering that chemotherapy is still the conventional approach in the clinical treatment of glioma, we attempted to evaluate the responses of sGBM patients from different risk groups to common drugs. Temozolomide (TMZ)-based chemotherapy is a standard strategy for glioma, and cisplatin is one of the most effective drugs for adjuvant therapy [20]. We managed to predict the responses of the two patient groups to these chemotherapy drugs by ridge regression, using GDSC nervous system cell lines gene expression data as the training set. A 10-fold cross-validation was applied to ensure the accuracy of this prediction. We found a significant difference in the estimated IC50 values of the two drugs between the high- and low-risk groups (**Figure 8C**). Hence, we concluded that the high-risk group may be more sensitive to common chemotherapies during clinical treatment (P < 0.001 for TMZ, P < 0.001 for cisplatin).

Discussion

Secondary glioblastoma is a type of grade IV glioma that originates from LGG. Because it is a rare disease, insufficient attention has been paid to the potential mechanisms of the pro-

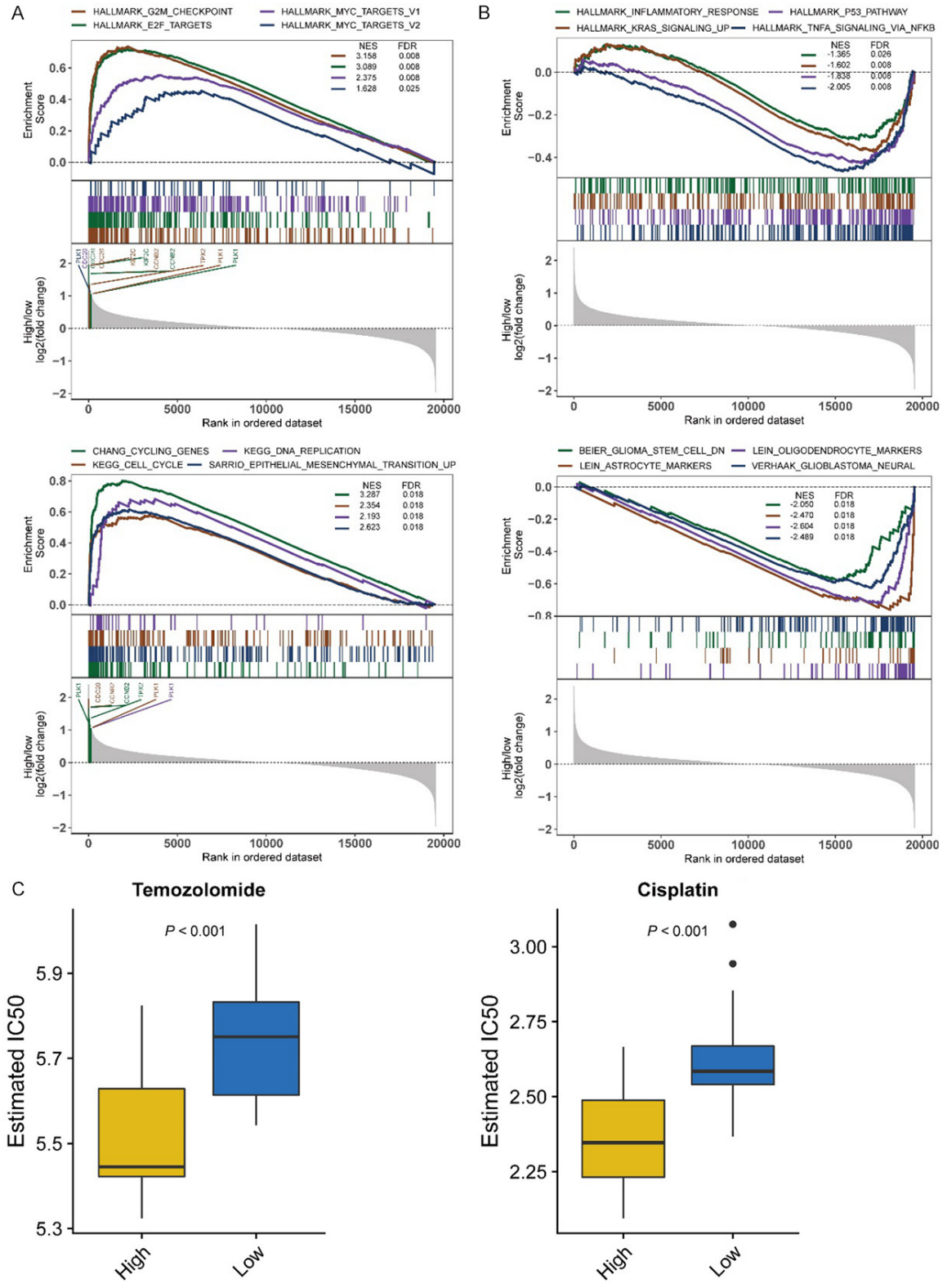


Figure 8. Differentially regulated pathways and predicted responses to chemotherapy in the high- and low-risk group. All transcripts were ranked according to the log2 (fold change) value derived from differential gene expression analysis between the two groups. Various pathways enriched in the high- (A) and low-risk group (B) were plotted. The NES and FDR value of each term were shown. (C) Boxplots of the estimated IC50 values of TMZ and cisplatin for tumor cells from two groups. Wilcoxon test (Mann-Whitney test) was used for comparison across groups.

gression from LGG to sGBM. Although various chemotherapies and radiotherapies have been applied in clinical treatment, little improvement has been documented in sGBM patients. TMZ, the first-line chemotherapy for GBM, functions by inducing DNA damage, which is unfortunately accompanied by the development of chemoresistance and various side effects. Almost all patients will suffer from tumor recurrence after TMZ treatment, and the recurrent lesion will be resistant to TMZ [21]. Therefore, there is an urgent need to explore the genetic alterations and unknown biological processes behind this conversion, as well as identify new therapeutic targets for clinical treatment. By using integrated bioinformatic methods, we successfully identified five key genes strongly associated with the patients' pathological information and developed an effective computational tool to predict the risk of sGBM patients. This study may offer other researchers useful novel ideas to improve our understanding of the initiation of sGBM.

In order to ensure the robustness of our conclusion, we used both RNA-sequencing and microarray gene expression data from the CGGA database for the whole study. WGCNA, a powerful algorithm for the mining of key genetic modules behind different phenotypes of interest, has been widely used in cancer research. We did not filter the differentially expressed genes between LGG and sGBM samples before performing WGCNA, because WGCNA is an unsupervised analysis method that seeks for clusters of genes based on expression profiles. Accordingly, pre-filtering genes will yield a set of correlated genes that will result in highly correlated modules. Moreover, the clustering method inside WGCNA mainly focuses on sub-classification of expression profiles based on similar biological processes rather than geometric distance. According to the correlation between modules and the pathohistological data of glioma patients, the brown module was selected as the preliminary module. Preservation analysis using the microarray dataset showed that the module has sufficiently high stability.

Molecular function terms of the GO analysis describe activities that occur at the molecular level [22]. The genes of the brown module from this study were significantly enriched in several categories, including chromatin binding, DNA-dependent ATPase activity, ATPase activity, histone binding, and cyclin-dependent protein kinase activity. Chromatin binding can affect the local chromatin structure and regulate gene transcription [23]. Cyclin-dependent protein kinases are a family of regulatory kinases that are necessary for cell cycle progression [24]. Several signatures related to biological process, such as DNA replication, chromosome segregation, mitotic nuclear division, and cell cycle G1/S phase transition, were found to be enriched in the brown module, indicating that the main changes of the progression from LGG to sGBM were relevant to the cell cycle and DNA replication pathways. KEGG pathway analysis further supported these results. All the identified gene terms basically mediate cell proliferation, tumorigenesis, and oncogenic activities. Our study therefore indicated that dysregulation of the cell cycle and DNA replication processes may be responsible for the sGBM transition.

We adopted strict criteria to ensure the precision of hub gene selection. CytoHubba is a Cytoscape plugin that can be used to rank nodes in a network according to their network features. MCC has been proved to have the best performance among all methods built into CytoHubba for predicting key genes in ranked nodes. Combined with filtering high-connectivity genes in WGCNA, seven genes were selected as hub genes. Nevertheless, KIF20A and CCNB1 were eliminated because of lacking significant prognostic value. ROC curve analysis showed that these five genes may serve as potential biomarkers for the diagnosis of sGBM with relatively high sensitivity and specificity. The expression levels of these hub genes were tested in two databases and protein levels were further validated in clinical samples. The overexpression and unfavorable prognostic value of these genes in sGBM patients may explain our hypothesis.

Multiple studies underscored the tumor-driver roles of the hub genes identified in this study. CCNB2, a member of the mitotic cyclin family, is essential for cyclin-dependent kinase 1 (CDK1) activation and is expressed in most cells with high mitotic activities [25, 26]. CCNB2 was reported to play an important role in regulating the G2/M transition [27]. Overexpression of CCNB2 is an unfavorable prognostic factor in

multiple human cancers, including breast carcinoma, gastric cancer, non-small cell lung cancer, and colorectal adenocarcinoma [28-31]. CDC20 is a mitotic regulator that functions by co-activating the anaphase-promoting complex (APC) E3 ubiquitin ligase [32]. In GBM, CDC20-APC can control the cell cycle and contribute to the maintenance of the invasiveness and self-renewal abilities of GBM stem-like cells by interacting with SRY-Box 2 (SOX2) [33]. Increased expression of CDC20 was observed and verified to be associated with clinicopathological parameters in a variety of human cancers [34-37]. Recently, small-molecule inhibitors targeting CDC20 were developed and found to have considerable anti-tumor effects. Examples include tosyl-l-arginine methyl ester (TAME), APC inhibitor (Apcin), withaferin A, and genistein [38-41]. PLK1 has been found to control mitotic processes, including centrosome maturation, chromosome segregation, spindle assembly, and cytokinesis [42]. By phosphorylating cell division cycle-25 (CDC25), cyclin B/cdc2 kinase is activated and in turn contributes to cell proliferation. At the same time, PLK1 can inhibit the activation of checkpoint kinase 1 (Chk1) and checkpoint kinase 2 (Chk2) to further prevent DNA damage repair [43]. Deficiency of mismatch repair (MMR) genes has been identified as a vital mechanism of the recurrence of GBM and acquisition of TMZ resistance [44]. Selective PLK1 inhibitors, such as volasertib, have been proved to inhibit the proliferation of GBM cells and suppress the development of MMR-deficient TMZ resistant GBM tumors [45]. KIF2C, also known as mitotic centromere-associated kinesin (MCAK), is regulated by aurora kinase-B and modulates cell cycle progression by participating in chromosome segregation and microtubule depolymerization [46]. The expression level of KIF2C escalates with increasing glioma grades, and is correlated with a poor prognosis in glioma patients [47]. TPX2 binds to the Thr288 residue in the catalytic domain of Aurora-A, preventing Aurora-A from phosphorylation and further achieve complete activation during mitosis [48]. Considering the overexpression and unfavorable prognostic correlation of TPX2 in various human cancer types, it has been regarded as an oncogene and a promising therapeutic target [49-51].

Although each single gene showed satisfactory efficiency in the diagnosis and prognosis of sGBM patients, the robustness of results is still a major concern. In this study, we successfully developed a computational risk model by combining all these key genes. Integrated analysis indicated that a combination of the five genes can serve as an indicator for predicting the clinical outcomes of sGBM patients. Somatic mutations of the IDH1 gene in GBM have been found to be associated with better overall survival. It has been reported that the frequency of IDH1 mutation in sGBM and LGG is much higher than in pGBM [52]. Moreover, patients with IDH1 mutation often show a relatively better response to TMZ-based chemotherapy, which indicates that sGBM patients can better benefit from TMZ application than pGBM patients [19]. In our study, ROC curve analysis showed that the AUC of this hub-gene-based signature for the prediction of 1-year survival was much higher than that of IDH1 mutation, which means that this risk model is superior to IDH1 status in predicting the clinical outcomes of sGBM. Given the scarcity of effective biomarkers for sGBM, this gene signature is potentially a useful alternative indicator for the clinical management of sGBM patients.

GBM is comprised of heterogeneous cells and manifests in distinct phenotypes and clinical outcomes [53]. An interesting finding of the present study was that genetic markers of oligodendrocytes and astrocytes were enriched in the low-risk group. According to the WHO classification of tumors of the CNS, low-grade diffuse glioma can be classified as oligodendroglioma, astrocytoma, and mixed oligoastrocytoma [54]. Different types of diffuse glioma may originate from different cell lineages. Traditionally, oligodendroglioma and astrocytoma are believed to respectively originate from oligodendrocytes and astrocytes in the normal brain. Hence, the high enrichment of these brain cell markers indicated that tumor tissues derived from the low-risk patients were well-differential and more similar to LGG at the genetic level. The Neural subtype of GBM is characterized by the expression of neuronal markers, including NEFL, GABRA1, and SYT1, as well as a strong association with oligodendrocytic and astrocytic differentiation [55]. Therefore, the longer overall survival time of low-risk sGBM patients probably results from distinct genetic characteristics.

Due to the pressing need to discover potential new anti-cancer drugs, an increasing number of drug databases has been established. GDSC is the largest public database containing information on the drug sensitivity of cancer cell lines and molecular markers of drug response based on large genomic data. Here, we built statistical models based on gene expression and drug sensitivity data of nervous system cancer cell lines. Cisplatin is a well-known chemotherapeutic drug that has been widely used for the treatment of numerous different human cancers [56]. It can induce DNA damage and apoptosis in cancer cells and is an effective regimen for LGG in children [57]. In this study, the IC50 of TMZ and cisplatin predicted using the GDSC dataset indicated that tumors in the high-risk group may be more sensitive to chemotherapy. This finding may provide guidance for deciding which treatment plan is appropriate for sGBM patients with different risk scores.

Finally, we are aware that there are also some limitations in this study. Considering the extremely low incidence of sGBM, it is difficult to collect enough tumor samples and primary sGBM cells for experimental validation. Considering that this study was mainly focused on data mining and analysis, further experiments are needed to confirm the findings of our work. For similar reasons, the CGGA database contains a limited number of sGBM samples, which may limit the power of statistical analyses. This will remain a significant challenge until a large-scale genomic database becomes available.

In summary, our analyses uncovered several key genes that might play critical roles in the progression from LGG to sGBM. These genes may serve as potential diagnostic and prognostic biomarkers, as well as possible therapeutic targets for sGBM. Most importantly, a risk model based on hub genes revealed the intrinsic mechanisms of this transition, underscoring the value of this model in predicting the clinical outcomes and directing clinical decisions in sGBM cases.

Acknowledgements

This work was supported by the National Natural Science Foundation of China (Grant No. 81673210). We acknowledge the work of all staff involved in the establishment of the Chinese Glioma Genome Atlas (CGGA) database.

Disclosure of conflict of interest

None.

Address correspondence to: Peng Zhao, Department of Neurosurgery, The First Affiliated Hospital of Nanjing Medical University, Nanjing, Jiangsu, China. Tel: +86-13655171913; E-mail: zhaopeng@njmu.edu.cn

References

[1] Louis DN, Perry A, Reifenberger G, von Deimling A, Figarella-Branger D, Cavenee WK, Ohgaki H, Wiestler OD, Kleihues P and Ellison DW. The 2016 World Health Organization classification of tumors of the central nervous system: a summary. Acta Neuropathol 2016; 131: 803-820.

[2] Riemenschneider M and Reifenberger G. Molecular neuropathology of gliomas. Int J Mol Sci 2009; 10: 184-212.

[3] Ohgaki H and Kleihues P. The definition of primary and secondary glioblastoma. Clin Cancer Res 2013; 19: 764-772.

[4] Ohgaki H and Kleihues P. Genetic pathways to primary and secondary glioblastoma. Am J Pathol 2007; 170: 1445-1453.

[5] Claus EB, Walsh KM, Wiencke JK, Molinaro AM, Wiemels JL, Schildkraut JM, Bondy ML, Berger M, Jenkins R and Wrensch M. Survival and low-grade glioma: the emergence of genetic information. Neurosurg Focus 2015; 38: E6.

[6] Jansen M, Yip S and Louis DN. Molecular pathology in adult gliomas: diagnostic, prognostic, and predictive markers. Lancet Neurol 2010; 9: 717-726.

[7] Tso CL, Freije WA, Day A, Chen Z, Merriman B, Perlina A, Lee Y, Dia EQ, Yoshimoto K and Mischel PS. Distinct transcription profiles of primary and secondary glioblastoma subgroups. Cancer Res 2006; 66: 159-167.

[8] Langfelder P and Horvath S. WGCNA: an R package for weighted correlation network analysis. BMC Bioinformatics 2008; 9: 559.

[9] Luo Z, Wang W, Li F, Songyang Z, Feng X, Xin C, Dai Z and Xiong Y. Pan-cancer analysis identifies telomerase-associated signatures and cancer subtypes. Mol Cancer 2019; 18: 106.

[10] Wagner GP, Kin K and Lynch VJ. Measurement of mRNA abundance using RNA-seq data: RPKM measure is inconsistent among samples. Theory Biosci 2012; 131: 281-285.

[11] Langfelder P, Luo R, Oldham MC and Horvath S. Is my network module preserved and reproducible? PLoS Comput Biol 2011; 7: e1001057.

[12] Yu G, Wang LG, Han Y and He QY. ClusterProfiler: an R package for comparing biological

themes among gene clusters. Omics 2012; 16: 284-287.
[13] Chin CH, Chen SH, Wu HH, Ho CW, Ko MT and Lin CY. cytoHubba: identifying hub objects and sub-networks from complex interactome. BMC Syst Biol 2014; 8 Suppl 4: S11.
[14] Robin X, Turck N, Hainard A, Tiberti N, Lisacek F, Sanchez JC and Müller M. pROC: an open-source package for R and S+ to analyze and compare ROC curves. BMC Bioinformatics 2011; 12: 77.
[15] Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, Paulovich A, Pomeroy SL, Golub TR and Lander ES. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc Natl Acad Sci U S A 2005; 102: 15545-15550.
[16] Ritchie ME, Phipson B, Wu D, Hu Y, Law CW, Shi W and Smyth GK. limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Res 2015; 43: e47.
[17] Benjamini Y and Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. J R Stat Soc Series B Stat Methodol 1995; 57: 289-300.
[18] Geeleher P, Cox N and Huang RS. pRRophetic: an R package for prediction of clinical chemotherapeutic response from tumor gene expression levels. PLoS One 2014; 9: e107468.
[19] SongTao Q, Lei Y, Si G, YanQing D, HuiXia H, XueLin Z, LanXiao W and Fei Y. IDH mutations predict longer survival and response to temozolomide in secondary glioblastoma. Cancer Sci 2012; 103: 269-73.
[20] Van Den Bent MJ, Hegi ME and Stupp R. Recent developments in the use of chemotherapy in brain tumours. Eur J Cancer 2006; 42: 582-8.
[21] Wang J, Cazzato E, Ladewig E, Frattini V, Rosenbloom DI, Zairis S, Abate F, Liu Z, Elliott O and Shin YJ. Clonal evolution of glioblastoma under therapy. Nat Genet 2016; 48: 768-76.
[22] Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, Harris MA, Hill DP, Issel-Tarver L, Kasarskis A, Lewis S, Matese JC, Richardson JE, Ringwald M, Rubin GM and Sherlock G. Gene ontology: tool for the unification of biology. Nat Genet 2000; 25: 25-9.
[23] Baylin SB. DNA methylation and gene silencing in cancer. Nature Clinical Practice Oncology 2005; 2: S4.
[24] Malumbres M. Cyclin-dependent kinases. Genome Biology 2014; 15: 122.
[25] Jackman M, Firth M and Pines J. Human cyclins B1 and B2 are localized to strikingly different structures: B1 to microtubules, B2 primarily to the Golgi apparatus. EMBO J 1995; 14: 1646-54.
[26] Brandeis M, Rosewell I, Carrington M, Crompton T, Jacobs MA, Kirk J, Gannon J and Hunt T. Cyclin B2-null mice develop normally and are fertile whereas cyclin B1-null mice die in utero. Proc Natl Acad Sci U S A 1998; 95: 4344-4349.
[27] Liu JH, Wei S, Burnette PK, Gamero AM, Hutton M and Djeu JY. Functional association of TGF-β receptor II with cyclin B. Oncogene 1999; 18: 269-275.
[28] Shubbar E, Kovács A, Hajizadeh S, Parris TZ, Nemes S, Gunnarsdóttir K, Einbeigi Z, Karlsson P and Helou K. Elevated cyclin B2 expression in invasive breast carcinoma is associated with unfavorable clinical outcome. BMC Cancer 2013; 13: 1.
[29] Shi Q, Wang W, Jia Z, Chen P, Ma K and Zhou C. ISL1, a novel regulator of CCNB1, CCNB2 and c-MYC genes, promotes gastric cancer cell proliferation and tumor growth. Oncotarget 2016; 7: 36489-36500.
[30] Qian X, Song X, He Y, Yang Z, Sun T, Wang J, Zhu G, Xing W and You C. CCNB2 overexpression is a poor prognostic biomarker in Chinese NSCLC patients. Biomed Pharmacother 2015; 74: 222-227.
[31] Park SH, Yu GR, Kim WH, Moon WS, Kim JH and Kim DG. NF-Y-dependent cyclin B2 expression in colorectal adenocarcinoma. Clin Cancer Res 2007; 13: 858-867.
[32] Peters JM. The anaphase promoting complex/cyclosome: a machine designed to destroy. Nat Rev Mol Cell Biol 2006; 7: 644-56.
[33] Mao DD, Gujar AD, Mahlokozera T, Chen I, Pan Y, Luo J, Brost T, Thompson EA, Turski A and Leuthardt EC. A CDC20-APC/SOX2 signaling axis regulates human glioblastoma stem-like cells. Cell Rep 2015; 11: 1809-1821.
[34] Chang DZ, Ma Y, Ji B, Liu Y, Hwu P, Abbruzzese JL, Logsdon C and Wang H. Increased CDC20 expression is associated with pancreatic ductal adenocarcinoma differentiation and progression. J Hematol Oncol 2012; 5: 15.
[35] Karra H, Repo H, Ahonen I, Löyttyniemi E, Pitkänen R, Lintunen M, Kuopio T, Söderström M and Kronqvist P. Cdc20 and securin overexpression predict short-term breast cancer survival. Br J Cancer 2014; 110: 2905-13.
[36] Kato T, Daigo Y, Aragaki M, Ishikawa K, Sato M and Kaji M. Overexpression of CDC20 predicts poor prognosis in primary non-small cell lung cancer patients. J Surg Oncol 2012; 106: 423-430.
[37] Li J, Gao JZ, Du JL, Huang ZX and Wei LX. Increased CDC20 expression is associated with development and progression of hepatocellular carcinoma. Int J Oncol 2014; 45: 1547-1555.

[38] Zeng X, Sigoillot F, Gaur S, Choi S, Pfaff KL, Oh DC, Hathaway N, Dimova N, Cuny GD and King RW. Pharmacologic inhibition of the anaphase-promoting complex induces a spindle checkpoint-dependent mitotic arrest in the absence of spindle damage. Cancer Cell 2010; 18: 382-395.

[39] Sackton KL, Dimova N, Zeng X, Tian W, Zhang M, Sackton TB, Meaders J, Pfaff KL, Sigoillot F and Yu H. Synergistic blockade of mitotic exit by two chemical inhibitors of the APC/C. Nature 2014; 514: 646-9.

[40] Das T, Roy KS, Chakrabarti T, Mukhopadhyay S and Roychoudhury S. Withaferin A modulates the Spindle Assembly Checkpoint by degradation of Mad2-Cdc20 complex in colorectal cancer cell lines. Biochem Pharmacol 2014; 91: 31-39.

[41] Zhang L, Yang B, Zhou K, Li H, Li D, Gao H, Zhang T, Wei D, Li Z and Diao Y. Potential therapeutic mechanism of genistein in breast cancer involves inhibition of cell cycle regulation. Mol Med Rep 2015; 11: 1820-1826.

[42] Gutteridge RE, Ndiaye MA, Liu X and Ahmad N. Plk1 inhibitors in cancer therapy: from laboratory to clinics. Mol Cancer Ther 2016; 15: 1427-35.

[43] Shaltiel IA, Krenning L, Bruinsma W and Medema RH. The same, only different-DNA damage checkpoints and their reversal throughout the cell cycle. J Cell Sci 2015; 128: 607-620.

[44] Yip S, Miao J, Cahill DP, Iafrate AJ, Aldape K, Nutt CL and Louis DN. MSH6 mutations arise in glioblastomas during temozolomide therapy and mediate temozolomide resistance. Clin Cancer Res 2009; 15: 4622-9.

[45] Higuchi F, Fink AL, Kiyokawa J, Miller JJ, Koerner MV, Cahill DP and Wakimoto H. PLK1 inhibition targets Myc-activated malignant glioma cells irrespective of mismatch repair deficiency-mediated acquired resistance to temozolomide. Mol Cancer Ther 2018; 17: 2551-2563.

[46] Lan W, Zhang X, Kline-Smith SL, Rosasco SE, Barrett-Wilt GA, Shabanowitz J, Hunt DF, Walczak CE and Stukenberg PT. Aurora B phosphorylates centromeric MCAK and regulates its localization and microtubule depolymerization activity. Curr Biol 2004; 14: 273-286.

[47] Bie L, Zhao G, Wang YP and Zhang B. Kinesin family member 2C (KIF2C/MCAK) is a novel marker for prognosis in human gliomas. Clin Neurol Neurosurg 2012; 114: 356-360.

[48] Asteriti IA, Rensen WM, Lindon C, Lavia P and Guarguaglini G. The Aurora-A/TPX2 complex: a novel oncogenic holoenzyme? Biochim Biophys Acta 2010; 1806: 230-239.

[49] Li B, Qi XQ, Chen X, Huang X, Liu GY, Chen HR, Huang CG, Luo C and Lu YC. Expression of targeting protein for Xenopus kinesin-like protein 2 is associated with progression of human malignant astrocytoma. Brain Res 2010; 1352: 200-207.

[50] Li F, Su M, Zhao H, Xie W, Cao S, Xu Y, Chen W, Wang L, Hou L and Tan W. HnRNP-F promotes cell proliferation by regulating TPX2 in bladder cancer. Am J Transl Res 2019; 11: 7035-7048.

[51] Sillars-Hardebol AH, Carvalho B, Tijssen M, Beliën JA, de Wit M, Delis-van Diemen PM, Pontén F, van de Wiel MA, Fijneman RJ and Meijer GA. TPX2 and AURKA promote 20q amplicon-driven colorectal adenoma to carcinoma progression. Gut 2012; 61: 1568-1575.

[52] Nobusawa S, Watanabe T, Kleihues P and Ohgaki H. IDH1 mutations as molecular signature and predictive factor of secondary glioblastomas. Clin Cancer Res 2009; 15: 6002-6007.

[53] Soeda A, Hara A, Kunisada T, Yoshimura SI, Iwama T and Park DM. The evidence of glioblastoma heterogeneity. Sci Rep 2015; 5: 7979.

[54] Louis DN, Ohgaki H, Wiestler OD, Cavenee WK, Burger PC, Jouvet A, Scheithauer BW and Kleihues P. The 2007 WHO classification of tumours of the central nervous system. Acta Neuropathol 2007; 114: 97-109.

[55] Verhaak RG, Hoadley KA, Purdom E, Wang V, Qi Y, Wilkerson MD, Miller CR, Ding L, Golub T, Mesirov JP, Alexe G, Lawrence M, O'Kelly M, Tamayo P, Weir BA, Gabriel S, Winckler W, Gupta S, Jakkula L, Feiler HS, Hodgson JG, James CD, Sarkaria JN, Brennan C, Kahn A, Spellman PT, Wilson RK, Speed TP, Gray JW, Meyerson M, Getz G, Perou CM and Hayes DN; Cancer Genome Atlas Research Network. Integrated genomic analysis identifies clinically relevant subtypes of glioblastoma characterized by abnormalities in PDGFRA, IDH1, EGFR, and NF1. Cancer Cell 2010; 17: 98-110.

[56] Dasari S and Tchounwou PB. Cisplatin in cancer therapy: molecular mechanisms of action. Eur J Pharmacol 2014; 740: 364-378.

[57] Massimino M, Spreafico F, Cefalo G, Riccardi R, Tesoro-Tess JD, Gandola L, Riva D, Ruggiero A, Valentini L and Mazza E. High response rate to cisplatin/etoposide regimen in childhood low-grade glioma. J Clin Oncol 2002; 20: 4209-4216.

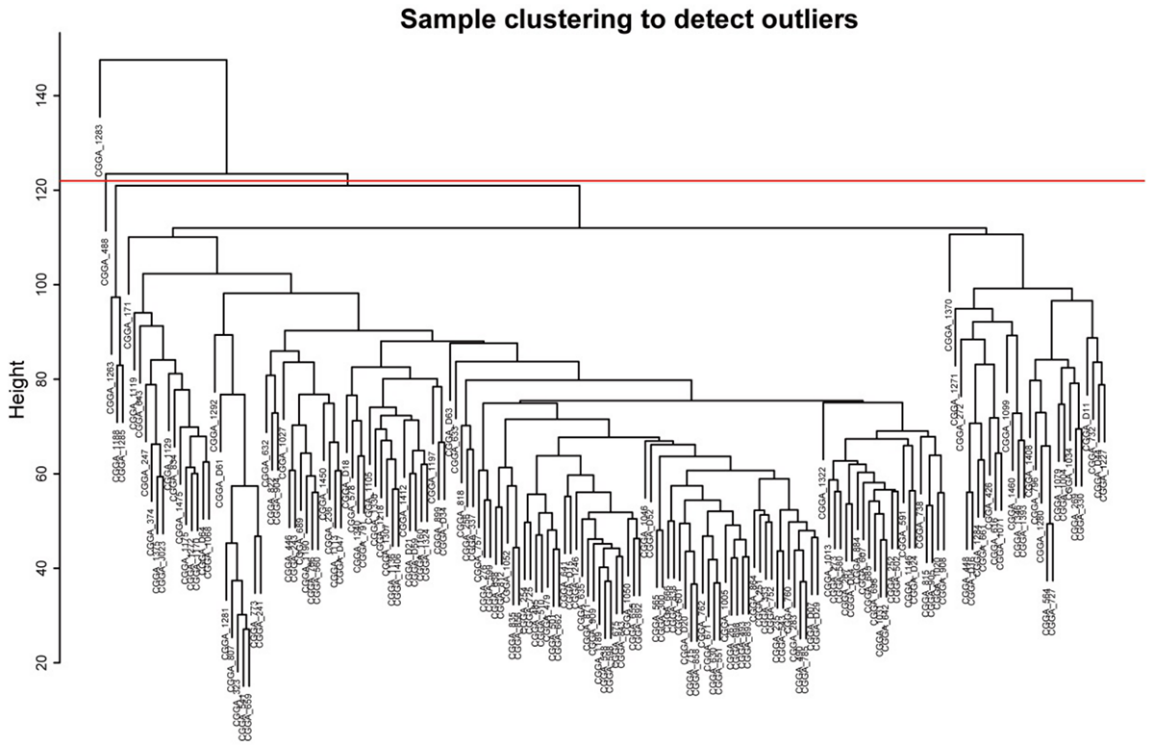


Figure S1. Clustering analysis of all samples in CGGA RNA-sequencing dataset. The top 5,000 genes with the highest median absolute deviation values of glioma samples were used. The red line was plotted to distinguish the outlier samples. Two outliner samples, CGGA_1283 and CGGA_488, were figured out.

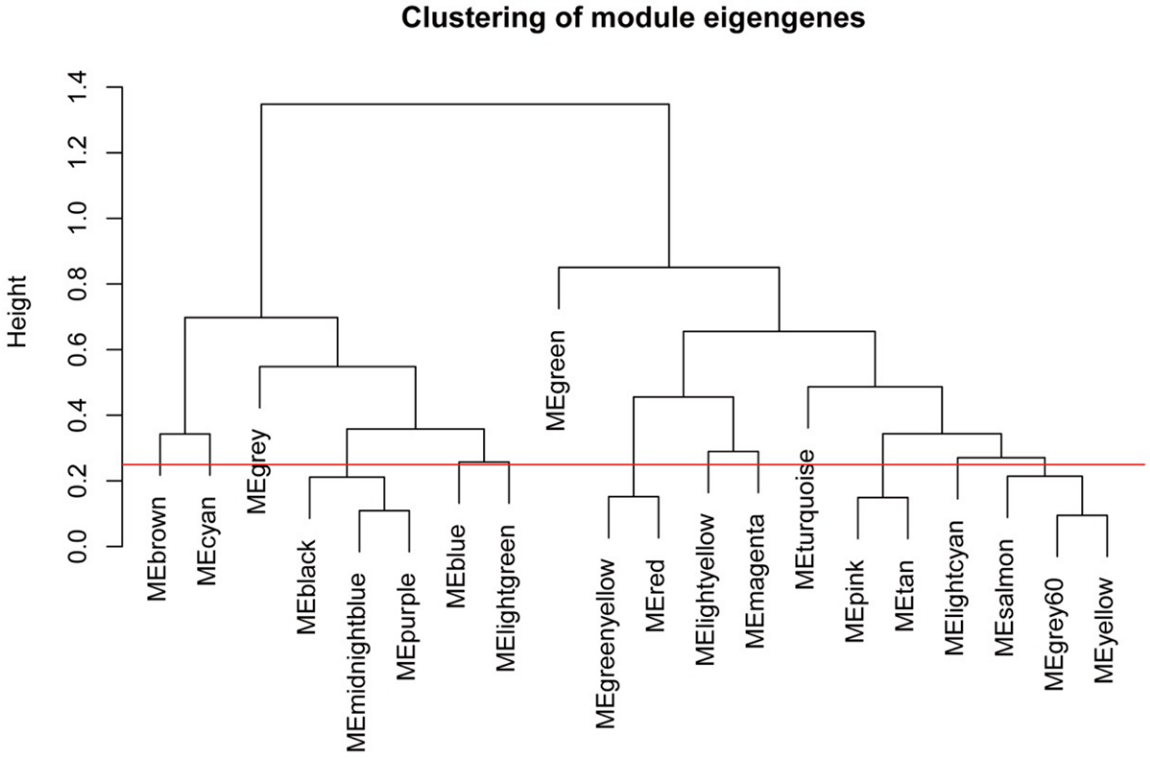


Figure S2. Visualization of the eigengene network representing the relationships among the modules. Hierarchical clustering dendrogram of the eigengenes based on the dissimilarity diss. The horizontal line (red) represents the threshold (0.25) for identifing the meta-modules.

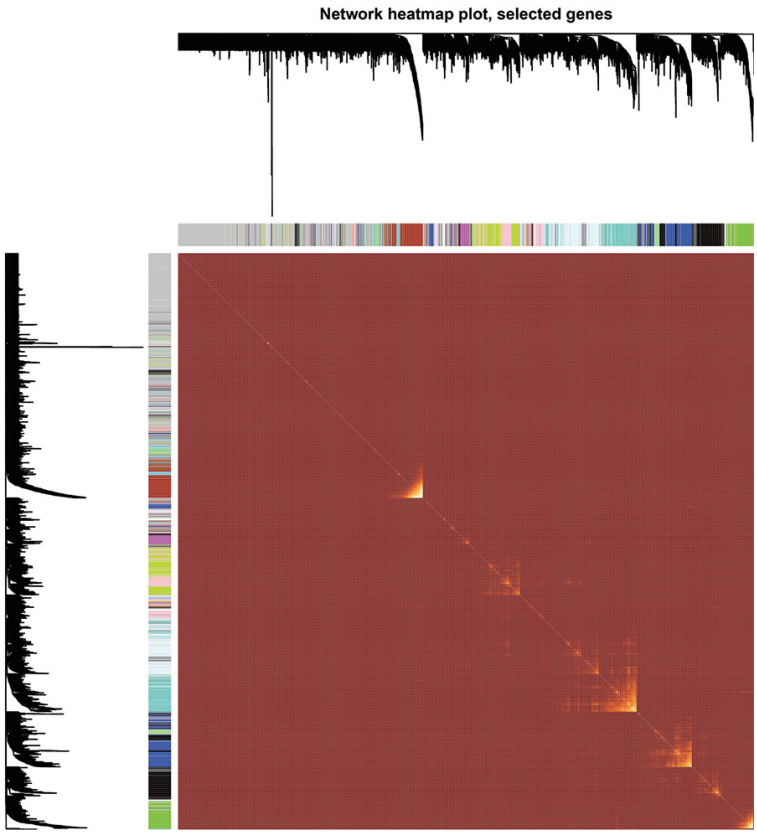


Figure S3. Heatmap of the topological overlap within the gene network. Each row and column represents a gene. Different colors on the x-axis and y-axis correspond to different modules. Bright color represents high topological overlap while darker color denoting low topological overlap.

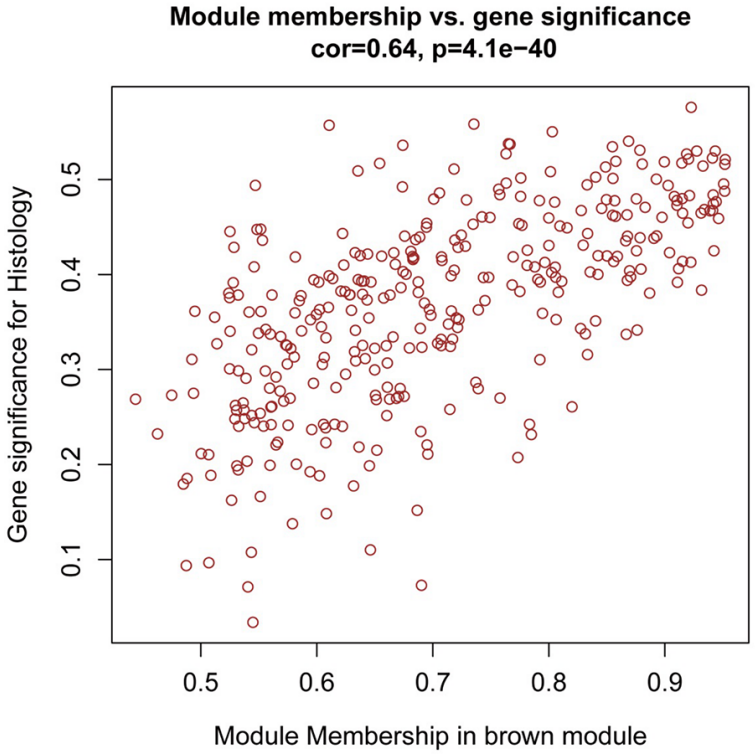


Figure S4. Scatterplot of gene significance (GS) for histology versus module membership (MM) in the brown module. The correlation coefficient and *p*-value were calculated using Pearson correlation analysis.

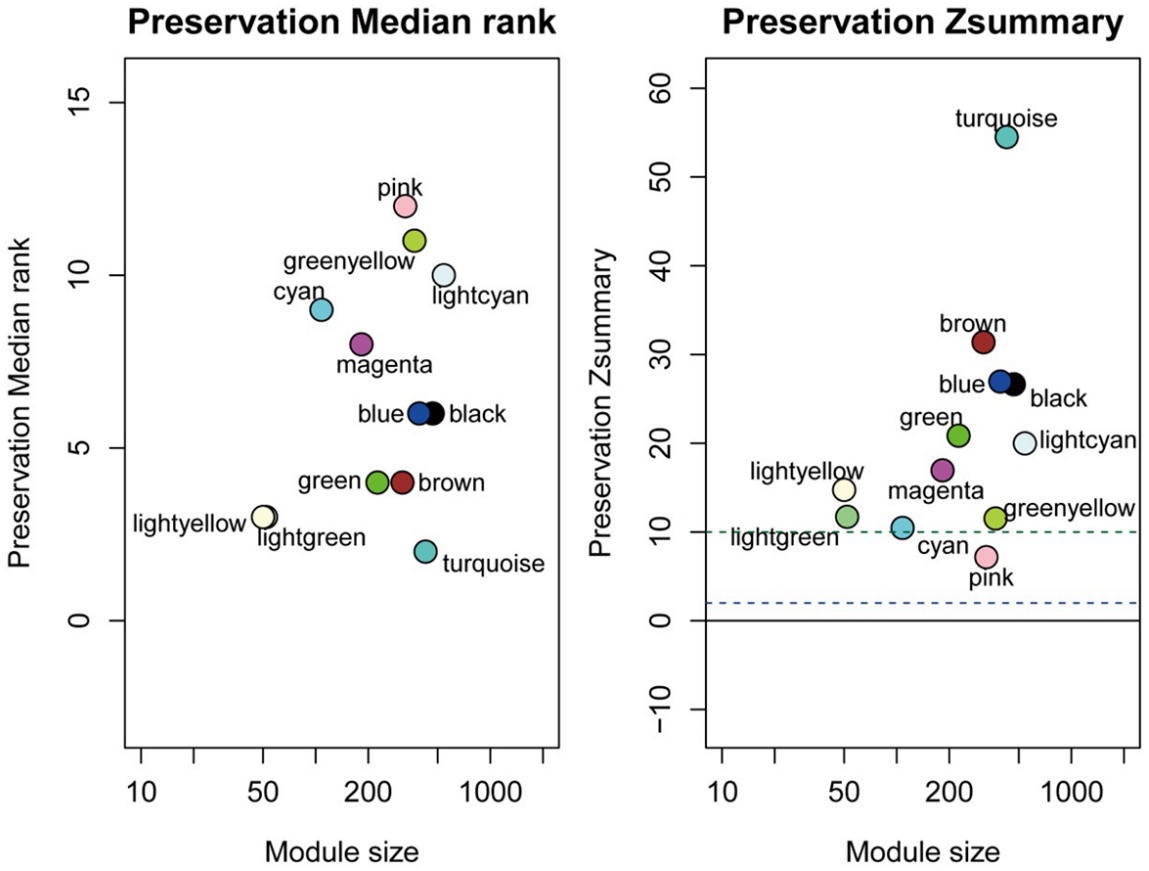
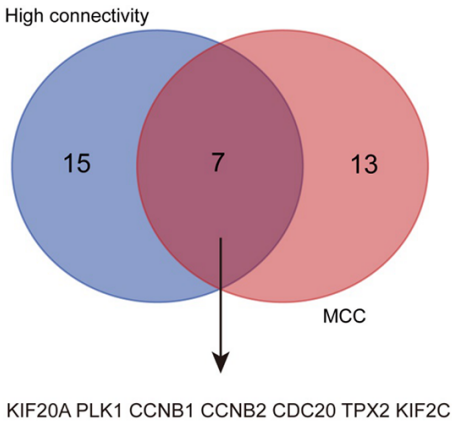


Figure S5. Preservation analysis of modules using another independent cohort (microarray dataset). The left panel shows the composite statistic preservation median rank. Modules with high median ranks usually indicate low preservation. The right panel shows the composite statistic median rank versus module size. Zsummary > 10 means strong evidence that the module is preserved. The dashed blue and green lines represent the thresholds Z = 2 and Z = 10, respectively.

A



B

High connectivity	MCC
CCNB2	CDK1
KIF2C	BUB1B
KIF20A	KIF11
CDC20	CCNA2
TPX2	TOP2A
SPC24	KIF20A
TROAP	CCNB1
TK1	CCNB2
IQGAP3	KIF2C
TACC3	CDC20
CCNB1	TPX2
PLK1	CENPF
IGIP	KIF4A
RECQL4	MAD2L1
TONSL	TRIP13
CCNF	CDC6
LMNB2	CDCA5
RNASEH2A	PLK1
CDCA4	GINS2
CHAF1A	PTTG1
CKS1B	
FANCD2	

Figure S6. Identification of hub genes in the brown module. A. Overlap of hub genes in the brown module identified by accessing gene connectivity and performing MCC algorithm, respectively. Intersected genes were regarded as hub genes for the subsequent analysis. B. List of detailed gene names obtained from these two distinct methods.