RNA processing genes characterize RNA splicing and further stratify lower-grade glioma

Rui-Chao Chai, …, Fan Wu, Yong-Zhi Wang

*JCI Insight.* 2019. [https://doi.org/10.1172/jci.insight.130591](https://doi.org/10.1172/jci.insight.130591).

**Graphical abstract**

Find the latest version:

[http://jci.me/130591/pdf](http://jci.me/130591/pdf)
RNA processing genes characterize RNA splicing and further stratify lower-grade glioma

Authors: Rui-Chao Chai\textsuperscript{1,2,3*}, Yi-Ming Li\textsuperscript{1,2,4*}, Ke-Nan Zhang\textsuperscript{1,2}, Yu-Zhou Chang\textsuperscript{4}, Yu-Qing Liu\textsuperscript{1,2}, Zheng Zhao\textsuperscript{1,2}, Zhi-Liang Wang\textsuperscript{1,2}, Yuan-Hao Chang\textsuperscript{1,2}, Guan-Zhang Li\textsuperscript{1,2}, Kuan-Yu Wang\textsuperscript{1,2}, Fan Wu\textsuperscript{1,2#}, Yong-Zhi Wang\textsuperscript{1,2,3,4#}

\textsuperscript{1}Department of Molecular Neuropathology, Beijing Neurosurgical Institute, Capital Medical University, Beijing, China;
\textsuperscript{2}Chinese Glioma Cooperative Group (CGCG), Beijing, China;
\textsuperscript{3}China National Clinical Research Center for Neurological Diseases;
\textsuperscript{4}Department of Neurosurgery, Beijing Tiantan Hospital, Capital Medical University, Beijing, China;

* Rui-Chao Chai and Yi-Ming Li contributed equally to this work.

# Corresponding authors:
1. Yong-Zhi Wang
Beijing Neurosurgical Institute, Beijing Tiantan Hospital, Capital Medical University,
No. 119 South Fourth Ring West Road,
Beijing 100070, China
Tel/Fax: +861067096794
E-mail: yongzhiwang_ng_bni@163.com

2. Fan Wu
Beijing Neurosurgical Institute, Beijing Tiantan Hospital, Capital Medical University,
No. 119 South Fourth Ring West Road,
Beijing 100070, China
Tel/Fax: +861067096794
E-mail: wufan0510284@163.com

Keywords: RNA processing, lower-grade gliomas, clinical outcomes, alternative splicing, \textit{TTF2}

Conflict of Interest: The authors declare no conflict of interest.
Abstract

BACKGROUND. Aberrant expression of RNA processing genes may drive the alternative RNA profile in lower-grade gliomas (LGGs). Thus, we aimed to further stratify LGGs based on the expression of RNA processing genes.

METHODS. This study included 446 LGGs from The Cancer Genome Atlas (TCGA, training set) and 171 LGGs from the Chinese Glioma Genome Atlas (CGGA, validation set). The least absolute shrinkage and selection operator (LASSO) Cox regression algorithm was conducted to develop a risk-signature. The receiver operating characteristic (ROC) curves and Kaplan–Meier curves were used to study the prognosis value of the risk-signature.

RESULTS. Among the tested 784 RNA processing genes, 276 were significantly correlated with the OS of LGGs. Further LASSO Cox regression identified a 19-gene risk-signature, whose risk score was also an independent prognosis factor (P<0.0001, multiplex Cox regression) in the validation dataset. The signature had better prognostic value than the traditional factors “age”, “grade” and “WHO 2016 classification” for 3- and 5-year survival both two datasets (AUCs > 85%). Importantly, the risk-signature could further stratify the survival of LGGs in specific subgroups of WHO 2016 classification. Furthermore, alternative splicing events for genes such as EGFR and FGFR were found to be associated with the risk score. mRNA expression levels for genes, which participated in cell proliferation and other processes, were significantly correlated to the risk score.

CONCLUSIONS. Our results highlight the role of RNA processing genes for further
stratifying the survival of patients with LGGs and provide insight into the alternative
splicing events underlying this role.

**FOUNDING.** Supported by the National Natural Science Foundation of China
(81773208, 81402052), the Beijing Nova Program (Z16110004916082), and the
National Key Research and Development Plan (2016YFC0902500).
Introduction

Diffuse glioma is the most frequent primary malignant brain tumors in adults (1, 2). Although intensive treatment with surgery, radiotherapy, and chemotherapy are conducted, nearly all gliomas relapse. The median overall survival (OS) of patients with glioblastoma (GBM, WHO grade IV) is only approximately 15 months (1, 3, 4). Patients with lower-grade diffuse glioma (LGG, WHO grade II/III) have a relatively favorable prognosis, but these tumors evolve and most relapse as therapy-resistant higher grade gliomas over time (5-7). LGGs can be classified as WHO grade II or III based on histological features or as oligodendroglioma with isocitrate dehydrogenase (IDH)-mutant and 1p/19q codeletion, astrocytoma with IDH-mutant, and astrocytoma with IDH-wildtype according to the WHO 2016 integrated analysis (1, 3, 8). However, this classification still does not fully reflect the rate of LGG evolution, and for the OS of patients with LGGs in a stratified subgroup is rather heterogeneous (7, 8). Thus, further stratification of LGGs using various methods (9-15), including RNA expression profiles (11, 14, 16-18), has attracted attention.

A dysregulated RNA expression profile is an important hallmark of tumors (19). RNA processing factors, such as RNA-binding proteins and RNA methylation regulators, are the main factors controlling the life-cycle of RNA in cells and thus play vital roles in numerous diseases and cancers by determining the final content and isoforms of mature RNA transcripts (6, 20-24). However, studies on RNA processing factors in gliomas have mainly focused on GBM, and the roles of RNA processing factors in stem cell maintenance, cell adhesion, cell migration, cell growth, colony formation, and invasiveness have been demonstrated in GBM (6, 20, 25-29). Recently, some RNA processing factors, such as HnRNPH1, PTB, SNRPI, and ELAVL, which are evaluated in the progression and prognosis of GBM, also showed prognostic value for all grades of diffuse gliomas (6), indicating their potential role in LGGs.

Considering that abnormal expression of RNA processing genes may drive changes in the dysregulated RNA expression profile in glioma, it is important to systematically examine the roles of RNA processing factors in LGGs.
RNA processing factors also govern the alternative splicing events (ASEs) of individual genes, which can affect several important factors in tumor initiation and progression (30). For instance, our recent study demonstrated the vital role of 14 exon skip of $MET$ in the malignant progression and targeted therapy of secondary GBM (31). The Cancer Genome Atlas (TCGA) SpliceSeq dataset lists seven ASE types (32). This makes it possible to link the dysregulation of RNA processing factors with aberrant ASEs in LGGs.

In this study, we systemically analyzed the expression profile of RNA processing genes and their prognostic values in 617 LGGs from TCGA (n = 446) and the Chinese Glioma Genome Atlas (CGGA; n = 171). We also profiled the ASEs underlying LGGs stratified by the risk signature built with RNA processing genes and identified the corresponding functions.
Results

Prognostic value of RNA processing genes and their biological function in LGGs

Our approach and workflow for the selection of RNA processing genes, development and validation of a prognostic risk signature, and the analysis of the ASEs and altered RNA expression profile that correlated to the signature are summarized in Figure 1. Of the 784 RNA processing genes tested, 276 were significantly correlated with the OS of LGGs (Supplementary Table 1). A total of 142 of the 276 genes had an HR >1 and were considered risk-associated, while the remaining 134 genes had HR <1 and were considered protection-associated. Risk-associated genes showing increasing expression (upper panel) and protection-associated genes showing decreasing expression (lower panel) were associated with increased malignancy in three subgroups according to the WHO 2016 classification (Figure 2A). Some of these genes also showed inconsistent expression levels within the same subgroups, suggesting the potential prognostic value of these genes within specific subgroups.

Since these 276 genes are a collection of genes that participated in any process involved in the conversion of one or more primary RNA transcripts into one or more mature RNA molecules. We used Gene Ontology (GO) analysis to study the more specific biological processes that these genes are enriched in, and the results indicated that these survival-associated genes were correlated with the terms such as rRNA processing, nuclear-transcribed mRNA catabolic process, nonsense-mediate decay, and mRNA splicing, via spliceosome (Figure 2B). We also further annotated the 142 risk-associated and 134 protection-associated genes through GO and KEGG analysis,
respectively (Supplementary Table 2 and 3). The results indicated that GO_BP term of mRNA splicing, via spliceosome and KEGG pathway of spliceosome were the most associated terms for the 142 risk-associated genes, and the GO_BP term of rRNA processing and KEGG pathway of ribosome were the most associated terms for the 134 protection-associated genes.

**Identification of a panel of 19 RNA processing genes as a risk signature in LGGs**

To easily and reliably stratify the outcomes of LGGs with RNA processing genes, we applied the least absolute shrinkage and selection operator (LASSO) Cox regression algorithm to the 276 genes in TCGA dataset (Supplementary Figure 1). A total of 19 genes were selected to build the risk signature, and the coefficients and normalized expression levels of these genes were used together to calculate risk scores for both the training dataset (TCGA dataset) and validation dataset (CGGA dataset) (Figure 2C and Supplementary Table 4).

We divided patients into high-risk and low-risk groups in various stratified glioma subtypes using their respective median risk score as a cutoff. We found that patients with low-risk scores had significantly longer OS than patients with high-risk scores in all subtypes evaluated in TGCA datasets: all LGGs (Figure 2D), oligodendroglioma with *IDH*-mutant and 1p/19q codeletion (Figure 2E), astrocytoma with *IDH*-mutant and *IDH*-wild-type (Figure 2F, Supplementary Figure 2A–B). Similar results were observed in the CGGA datasets (Figure 2G–I, Supplementary Figure 2C–D).

To investigate the prognostic value of the risk signature and other clinicopathological features, univariate and multivariate Cox regression analysis was
performed both in TCGA (training) and CGGA (validation) data set. The results revealed that high risk score was a prognostic factor in both datasets (P<0.0001), independent of grade, age, gender, WHO 2016 subgroups, IDH mutation, 1p19q codeletion, and MGMT methylation status (Table 1).

We also summarized the race composition in datasets used in this study, as a specific race may represent population-specific exposure and risk factor. In the TCGA dataset, 92.00% of cases are White, 3.36% of cases are Black and African American, 1.79% of cases are Asian, 0.22% of cases are American Indian or Alaska Native, and the rest 2.02% of cases unknown (Supplementary Figure 3A). All of the cases in the CGGA dataset are Asian. We found that patients with low risk scores had significantly longer OS than patients with high risk scores in White (Supplementary Figure 3B) and Black or African American (Supplementary Figure 3C) population of TGCA dataset. We cannot compare the OS of the 8 cases of Asian in the TCGA dataset, as all of them survived at the end of follow-up (Supplementary Figure 3D). However, the signature could stratify the survival of Asian population, since all the cases in the CGGA dataset are Asian.

Collectively, these results strongly support the ability of our risk signature to accurately stratify the prognosis of patients with definite subgroups based on WHO 2016 integrated diagnosis.

The signature risk score is correlated with malignant clinicopathological features of LGGs
To determine whether the risk signature also reflects the malignant clinicopathological features of LGGs, we determined risk signature expression of 19 genes and clinicopathological features in LGGs with low and high risk in the form of a heatmap (Supplementary Figure 4A). Significant differences were observed between the high- and low-risk groups with respect to the WHO grade \((P < 0.001)\), \(IDH\) status \((P < 0.001)\), 1p/19q codeletion status \((P < 0.001)\), WHO 2016 subgroups \((P < 0.001)\), \(MGMT\) promoter methylation status \((P < 0.001)\), age \((P < 0.001)\), and \(CIC\) \((P < 0.001)\) (Supplementary Figure 4A and Supplementary Table 5). We also found that the risk score was significantly different in patients with LGGs stratified by WHO grade, subgroups according to WHO 2016 classification, \(IDH\) mutation status, 1p/19q codeletion status, \(MGMT\) promoter methylation status, and age (Supplementary Figure 4B–G). Moreover, the distributions of risk score between LGGs with and without \(TERT\) promoter mutation were highly dependent on the \(IDH\) mutant status (Supplementary Figure 4H). The risk score significantly increased in \(TERT\)-mutant LGGs with wild-type-\(IDH\) \((P < 0.01)\), while the risk score deceased in \(TERT\)-mutant LGGs with mutant-\(IDH\) \((P < 0.01)\). A similar distribution of risk score was observed in the CGGA dataset (all \(P\) values \(< 0.05\), except for \(TERT\) status, and \(P=0.07\) for \(TERT\) status, Supplementary Figure 5). These results indicated that the signature was associated with the malignant clinicopathological features of gliomas.

The receiver operating characteristic (ROC) curve analysis showed that the signature risk score had the best efficiency (compared with age, WHO 2016 subgroups and WHO grade) for predicting the 3-year and 5-year survival of patients with LGGs.
both in TCGA and CGGA datasets (Figure 3). Areas under the curve of risk score, age
WHO 2016 subgroups and WHO grade were 93.7%, 81.8%, 77.6%, 70.2% respectively
for 3 year’s survival in TCGA dataset; 88.1%, 78.8%, 74.8%, 68.4% respectively for 5
year’s survival in TCGA dataset; 85.9%, 64.4%, 76.1%, 78.6% respectively for 3 year’s
survival in CGGA dataset; 87.3%, 63.7%, 76.4%, 75.3% respectively for 5 year’s
survival in CGGA dataset.

The signature risk score is closely correlated with the alternative splicing events
RNA splicing activities are governed by RNA processing genes, and our findings
showed that worse survival-associated RNA processing genes were most enriched in
RNA splicing related activities. Thus, we also systematically characterized ASEs in
LGGs with different risk scores. In total, tens of thousands of seven ASE types,
including alternate acceptor site (AA), alternate donor site (AD), alternate promoter
(AP), alternate terminator (AT), exon skip (ES), mutually exclusive exons (ME), and
retained intron (RI), were detected in each LGG (Supplementary Figure 6A).
Furthermore, the proportion of differential ASE types in LGGs varied widely (from 0.5%
to approximately 35%). Although all LGGs shared similar patterns of ASE types (ES
and AP were the most frequently observed ASE types, while ME and RI were the least
frequently observed ASE types), the total number of detected differential ASEs
gradually increased along with the increasing risk score (Figure 4A), and the total
number of ASEs was significantly smaller in LGGs with a lower (1st quarter n=111) risk
score than with a higher (4th quarter n=111) risk score (Figure 4B).
Percent Spliced In (PSI) is the ratio of reads indicating the presence of a transcript element versus the total reads covering the event (32). We also analyzed the PSI levels for ASEs in all LGGs and LGGs with lower and higher risk scores. The results revealed that events with lower PSI levels (PSI \leq 0.2) and higher PSI levels (PSI > 0.8) constituted most types of ASEs in all LGGs and LGGs with lower and higher risk scores (Supplementary Figure 6B–H). These observations indicated that transcripts with high splice-out alternative exons or high splice-in alternative exons were the predominantly transcribed form of most genes.

We further identified differentially expressed RNA splicing genes ($P < 0.05$ and fold-change $\geq 2$ or $\leq 0.5$) and ASEs with significantly different PSI ($P < 0.05$ and fold-change $\geq 2$ or $\leq 0.5$) in LGGs with lower ($1^{\text{st}}$ quarter $n=111$) and higher ($4^{\text{th}}$ quarter $n=111$) risk scores (Figure 5A, Supplementary Table 6 and Supplementary Table 7). There were 257 ASEs for 247 genes with decreased PSI levels in LGGs with higher risk score and 604 ASEs for 527 genes with increased PSI levels in LGGs with higher risk scores (Supplementary Figure 7A). The proportion of AP type of ASEs dramatically increased (from 23.14% to 47.7%) in these ASEs with significantly different PSI. In LGGs with higher risk score, AP (45.1%) and ES (26.5%) showed the largest proportion in ASEs with decreased PSI, while AP (49.3%) and AT (25.0%) constituted a larger proportion of ASEs with increased PSI (Supplementary Figure 7B). This suggests that ES transcripts with high splice-out exons are more likely to be present in LGGs with higher risk scores and the AP type of ASEs appeared to be more relevant to the prognosis of LGGs.
We found that genes involved in the receptor tyrosine kinase signaling pathway (EGFR, FGFR1, and FGFR2), DNA damage response (C5orf45), RNA binding (CIRBP), transcription regulation (CREM), and brain disease development (MAPT) were differentially spliced between LGGs with lower and higher risk scores (Figure 5B). To further explore the function of alternative splicing in the malignancy of LGGs, GO analysis was performed for all genes that were differentially spliced in LGGs with lower and higher risk scores. In general, genes with differential PSI were mainly enriched in biological processes, such as “positive regulation of GTPase activity”, “SRP-dependent co-translational protein targeting to membrane”, “cytoskeleton”, “cytokinesis”, “Transmembrane receptor protein tyrosine kinase signaling pathway”, “cell adhesion”, and others (Figure 5C, Supplementary Figure 8). Our analysis indicated that differential ASEs participate in many cancer-related biological processes, particularly signal transduction pathways, suggesting that ASEs are a critical mechanism underlying the prognostic value of RNA processing genes in LGGs.

We further investigated the potential regulatory networks between the significantly changed 28 RNA splicing genes and 862 ASEs (see Methods). Finally, a network of the 28 differential RNA splicing genes and 221 of their significantly correlated differential ASEs was constructed (Figure 6A and Supplementary Table 8). The 28 RNA splicing genes regulated distinct amounts of ASEs ranging from 4 to 104 (Figure 6B). For RNA splicing genes whose expression was increased in LGGs with higher risk scores, TTF2 and MBNL3 were found to regulate a larger number of ASEs, while GPATCH1 and DHX15 regulated fewer ASEs (Figure 6B upper panel). In RNA splicing genes whose
expression was decreased in LGGs with higher risk scores, *POLR2F* and *PQBP1*
regulated a larger number of ASEs, while *GPKOW* and *SAP18* regulated fewer ASEs
(Figure 6B lower panel). The proportion of AP type of ASE was still the largest one in
these 221 ASEs.

**The function analysis of genes whose mRNA expression are correlated with the
signature risk score in LGGs**

Since RNA processing genes are the main factors controlling the life-cycle of RNA in
cells, we also evaluate the RNA expression profile influenced by the differentially
expressed RNA processing genes. We identified genes positively (Pearson
coefficient >0.5 and Bonferroni corrected *P* < 0.01) or negatively (Pearson coefficient
< -0.5, and Bonferroni corrected *P* < 0.01) correlated with the risk score and then
annotated their functions using GO analysis for biological processes and KEGG
pathways analysis (Figure 7A and B). Overall, the results indicated that the risk scores
positively correlated genes were mainly enriched in regulation of chromosome division
and DNA stability, cell division, DNA replication, cell cycle, DNA repair, angiogenesis
and other malignancy-related biological processes (Figure 7A). The corresponding
pathways could also be observed in the Kyoto Encyclopedia of Genes and Genomes
(*KEGG*) pathway analysis (Figure 7B). Suggesting that the altered expression of RNA
processing genes may be associated with the increased expression of genes enriched in
these processes. Interesting, we observed that the signature risk scores negatively
correlated genes were mainly involved in the GO processes of translation and rRNA
processing, and the KEGG pathway of ribosome (Figure 7A and B). This finding is consistent with the finding that the GO_BP term of rRNA processing and KEGG pathway of ribosome are the most associated terms for the 134 protection-associated RNA processing genes (Supplementary Table 2).

Furthermore, Gene Set Enrichment Analysis (GSEA) revealed that the hallmarks of malignant tumors, including epithelial-mesenchymal transition, angiogenesis, G2M checkpoint, mitotic spindle, inflammatory response, complement, PI3K-AKT-MTOR signaling, and KRAS signaling (Figure 8A–H), as significantly enriched in LGGs with risk scores higher than median risk score. These findings indicate that the risk score of the RNA processing signature reflects the expression alterations of genes involved in malignant biological processes, signaling pathways, and malignant hallmarks in LGGs.

**RNA expression profile influenced by knock-down of mRNA expression of TTF2**

To study whether the change of RNA processing gene is a driver or merely correlative to the dysregulated RNA profile. We selected three specific siRNA to knock-down the mRNA expression of TTF2 genes, this gene is risk-associated in the signature (with highest HR), and it also included in the GO term of RNA splicing. The results indicated that all three siRNA could significantly downregulate the mRNA expression of TTF2 in glioma cell line LN229 cells, and TTF2 siRNA2 appeared to be the most effective siRNA (Figure 9A).

We collected whole transcriptome data of LN229 cells at 48 hr after transfection of TTF2 siRNA2 (n=3) and scramble siRNA (n=3), respectively. Compared with scrambled siRNA group, there were 2676 genes with increased expression and 2710
genes with decreased expression in the group of TTF2 siRNA, and the mRNA expression of TTF2 gene was also significantly decreased (Figure 9B, Supplementary Table 9). We found that the 2710 genes with decreased expression after knock-down mRNA of TTF2 gene were significantly enriched in the biological processes of chromosome segregation, cell division, cell cycle, mRNA metabolic, and RNA splicing (Figure 9C, Supplementary Table 10), and the KEGG pathway analysis showed these genes also were enriched in the pathways of “cell cycle”, “spliceosome”, “DNA replication”, “RNA transport”, “proteasome”, “mismatch repair”, and others (Figure 9D, Supplementary Table 11). This finding was largely consistent with the result of function analysis of genes whose mRNA expressions were positively correlated to the risk signature, indicating that changes of RNA processing gene TTF2 maybe a driver of these dysregulated RNA profiles in glioma.

As mentioned above, functions of genes whose mRNA expressions were negatively correlated to the risk signature mainly enrich in the biological processes of “translation” and “SRP-dependent co-translational protein targeting to membrane”, the KEGG pathways of “Ribosome”. We also noticed that functions of the 2676 up-regulated genes were mainly involved in the protein processing activities, such as the biological processes of “extracellular matrix organization”, “response to endoplasmic reticulum stress” and others, the KEGG pathways of “lysosome”, “protein processing in endoplasmic reticulum” and others (Figure 9C and D, Supplementary Table 12 and 13).”
Discussion

In this study, we found that the general expression pattern of RNA processing genes is correlated with the malignancy features of LGGs and identified RNA processing genes significantly associated with the prognosis of LGGs. We further built a risk signature that not only further predicted the prognosis of stratified LGGs but also could perfectly reflect the malignancy-correlated clinicopathological features, biological processes, key signaling pathways, and hallmarks. Moreover, we systemically analyzed ASEs underlying LGGs with lower and higher risk scores and identified the corresponding functions. Our results highlight the role of RNA processing genes in further stratifying the survival of patients with LGGs, and the developed signature shows a potential to become an effective supplement for the integrated diagnosis criteria of WHO 2016 in further stratifying the prognosis of LGGs. Moreover, we also reveal the associated ASEs and biological processing that were associated with the risk scores of the developed signature.

Genetic alterations, such as IDH mutation and 1p/19q codeletion, have been included in the integrated diagnosis criteria of WHO 2016 (3). Very large differences in prognosis exist in LGGs within the definite WHO 2016 subgroup (16). Several important biomarkers have been identified to further stratify LGGs, including genetic alterations in CDKN2A/B in astrocytoma with mutant-IDH (11, 33); EGFR amplification, chromosome 7 gain and chromosome 10 loss, and TERT promoter mutation in astrocytoma with wild-type-IDH (11). Importantly, “the EGFR amplification, or chromosome 7 gain and chromosome 10 loss, or TERT promoter
mutation” had been recommended as the diagnostic criteria for “Diffuse astrocytic gayla, IDH-wild-type, with molecular features of glioblastoma, WHO grade IV” by the Consortium to Inform Molecular and Practical Approaches to Central Nervous System Tumor Taxonomy (34). Thus, further stratification of LGGs with definite IDH and 1p/19q status is urgently needed.

Apart from these genetic alteration markers, RNA expression is also valuable for predicting the outcomes of tumors with low rates of mutation, and the RNA expression profile had become an indispensable element for classifying medulloblastoma (35). Compared to GBM, LGGs have relative lower mutation rates and the RNA expression of a set of genes was suggested to be useful for predicting the prognosis of patients with 1p19q co-deletion diffuse glioma (16). In this study, we confirmed the prognostic value of a signature built with 19 RNA processing genes in each stratified subgroup of LGGs (Figure 2D–I). Compared with traditional stratifying factors (age, WHO grade, subgroups of WHO 2016 classification), The risk score of this signature also had the best predictive efficiency on both the 3-year and 5-year survival (Figure 3). Though the risk score of the signature is highly associated with traditional clinicopathological features, we confirmed that the risk score is an independent prognosis factor in both the training and validation dataset (Table 1). A previous study identified 104 key genes that were prognostic for LGGs. However, the average area under the ROC curve (AUC) values ranged from 0.7 to 0.8 (16), which was lower than the performance of our signature (AUC > 0.85). These data indicate that the RNA processing gene signature is a powerful tool for further predicting prognosis of LGGs, which were stratified by
isocitrate dehydrogenase and 1p/19q status based on the 2016 World Health Organization classification guidelines

The ASEs of individual genes which can affect several important players in tumor initiation and progression (6, 30, 36-38). For instance, delta isoform of Max could promote glioma cell proliferation by enhancing functions of MYC in GBM harboring EGFRvIII mutation (39); two of CD97 isoforms, EGF (1,2,5) and EGF (1,2,3,5) could promote growth, migration, metastasis, and angiogenesis in GBM (40); and an aberrant splicing of cyclin-dependent kinase-associated protein phosphatase KAP increases proliferation and migration in glioblastoma (30). Currently, genome-wide analyses of exon expression arrays have begun to reveal the roles of ASEs correlated with the progression and prognosis of GBM (41). However, the role of alterations of ASEs in LGGs is not well-understood. Our study not only reveals the landscape of ASEs in LGGs, but also identified 861 ASEs correlated with the prognosis of LGGs (Supplementary Table 5), although more specific studies are needed to thoroughly determine the functions of these identified ASEs. Moreover, we identified RNA splicing genes correlated with differential ASEs in LGGs with distinct prognosis.

RNA plays crucial roles in biological functions in cells not only by passing genetic information from DNA to protein but also by regulating various biological processes (42). Dysregulation of RNA profiles, including hypoxia-associated gene sets (43), vascular gene sets (44), WNT/beta-catenin pathway-related genes (45), genes in the NF-κB signaling pathway (46), micro-RNA (18), and long non-coding RNA (47) have been shown to play crucial roles in the malignant progression and prognosis of
LGGs. Here, we highlight the stratification ability of RNA processing genes in LGGs. Additionally, we found that genes whose expression was significantly correlated with the risk score of the signature built with 19 RNA processing genes were enriched in the biological processes/pathways of cell cycle, cell division, DNA replication and repair, angiogenesis, cell proliferation, and pathways in cancer (Figure 6A). Among the 19 genes included in the signature, *RPL3* and *BICD1* have been suggested to be associated with temozolomide resistant in glioblastoma (48, 49); and *ELAVL1* was reported participated in the cell proliferation, migration and glioma stem-like cell maintenance (50, 51). RNA expression profile and RNA fate are highly dependent on RNA processing factors responsible for precise temporal and spatial coordinating gene expression (6, 20). In addition, our finding also confirmed that knock-down the expression of *TTF2* genes could alter the expression of genes involved in the activities of the cell cycle, RNA splicing, DNA replication, mismatch repair, extracellular matrix organization, and others. All of these indicate that abnormal expression of RNA processing genes is potential drivers of the previously reported dysregulated RNA profiles in LGGs.

In summary, our study highlights the prognostic value of RNA processing genes in LGGs and revealed a signature with 19 RNA processing genes for further stratifying the outcomes of LGGs with definite WHO subgroups. Clinicopathological features, ASEs, biological processes, signaling pathways, and hallmarks of tumors correlated with the risk signature were also identified. These results provide fundamental information for understanding the roles of RNA processing and indicate the potential
clinical implications of RNA processing genes in LGGs.
Methods

Patients

A total of 457 LGG samples were collected from the TCGA publication (52). Of these TCGA samples, we excluded 11 samples that do not contain RNA sequencing data or molecular pathological information or useful OS information, and the other 446 LGG samples with RNA-seq transcriptome data and corresponding clinical and molecular pathological information were obtained from TCGA (http://cancergenome.nih.gov/) and used for systematic analysis. Another 171 LGG samples with RNA-seq transcriptome and corresponding clinicopathological information form CGGA (www.cgga.org.cn) were used to validate the performance of the risk signature. Clinicopathological information for the TCGA and CGGA datasets was summarized in Supplementary Table 14.

Selection of RNA processing genes

We first collated a list of 835 genes that participated in any process involved in the conversion of one or more primary RNA transcripts into one or more mature RNA molecules (http://amigo.geneontology.org/amigo/term/GO:0006396), and then we used the 784 genes with available RNA expression data in TCGA datasets for further analysis.

Identification of the risk signature

We performed univariate Cox regression analyses of the expression of RNA pressing genes to identify genes significantly correlated with the prognosis of patients with LGGs in TCGA dataset. Next, we employed the LASSO Cox regression algorithm to build an optimal risk signature with the minimum number of genes (53). Finally, a set
of genes and their coefficients were determined by minimum criteria, which involves selecting the best penalty parameter $\lambda$ associated with the smallest 10-fold cross validation within the training set. The risk score for the gene signature was calculated using the formula:

$$\text{Risk score} = \sum_{i=1}^{n} \text{Coef}_i \cdot x_i$$

where $\text{Coef}$ is the coefficient and $x_i$ is the z-score-transformed relative expression value of each selected gene.

Patients were divided into “high-risk” and “low-risk” groups using the median risk score as the cutoff value. Patients could also be compared between groups with lower risk score (1st quarter) and higher risk score (4th quarter).

**Bioinformatic analysis**

GO and KEGG pathway enrichment analyses with the Database for Annotation, Visualization, and Integrated Discovery (http://david.abcc.ncifcrf.gov/home.jsp) to functionally annotate genes with prognosis value in LGGs, significantly correlated to the risk score, and differentially spliced between patients with lower (1st quarter) and higher (4th quarter) risk score. GSEA was performed to investigate the functions of genes that significantly correlated to the risk score.

**Construction of regulation networks between RNA splicing genes and ASEs in LGGs**

The RNA splicing data were collected from the online database (http://bioinformatics.mdanderson.org/TCGASpliceSeq). The ASEs could be quantified by the PSI value, which represents the ratio of included transcript reads in
total transcript reads forms, and the detailed information on the PSI calculation as reported study (32).

RNA splicing genes (GO: 0008380) showing significant changes in expression levels were predicted to be correlated with the differential PSI level of ASEs between LGGs with lower (1st quarter) and higher (4th quarter) risk scores. We calculated the Pearson correlation for each RNA splicing gene-ASE pair, RNA splicing gene-ASE pair with correlation coefficients greater than 0.4 or less than -0.4, and the corresponding $P$ value less than 0.05 were considered significantly correlated (Bonferroni correction was performed to adjust the $P$ value for multiple comparisons). The predicted regulatory network was visualized with Cytoscape (http://www.cytoscape.org).

**Cell lines**

Human glioma cell line LN229 cell was purchased from the ATCC (Manassas, VA). We have performed the reauthentication by short tandem repeat analysis in July 2017 in our laboratory. Cells were routinely cultured in Dulbecco’s modified Eagle’s medium supplemented with 10% fetal bovine serum (HyClone, Logan, Utah), 100 units/ml penicillin and 100 mg/ml streptomycin (Invitrogen, Carlsbad, CA) at 37°C in a humidified atmosphere of 5% CO$_2$ as our previously reported study (7).

**Small interference RNA (siRNA), quantitative PCR**

The scrambled small interfering RNA (siRNA) and KIF2C specific siRNA were synthesized from the GenePharma Corporation (Suzhou, Jiangsu, China). The sequences for TTF2 siRNA as follow: TTF2 siRNA1 5’-GGA CCU CAG GUA AUG CUA ATT-3’; TTF2 siRNA2 5’- CCA AGA UCA CGU UCA UGC ATT-3’; TTF2
siRNA3 5’- GGA AAG AGC UUC UAC GUG UTT-3’.

The quantitative PCR was performed by the SYBR® Select Master Mix (Thermo Fisher, Waltham, MA) in ABI 7500 (Thermo Fisher). The sequences for the primers of TTF2 were 5’-ATT TAC CGA GTA GGG CAG CA-3’ (forward primer) and 5’- GGA CTC TGA GGT CAG CCA AG-3’ (reverse primer). The sequences for the primers of glyceraldehyde 3-phosphate dehydrogenase were 5’-GGT GGT CTC CTC TGA CTT CAA CA-3’ (forward primer) and 5’-GTT GCT GTA GCC AAA TTC GTT GT-3’ (reverse primer).

Transcriptome sequencing

The total RNA of LN229 cells were extracted by the Trizol reagent (Thermo Fisher, Waltham, MA). Then the RNA sequencing library and sequencing were finished by the Novogene (Beijing, China) using Illumina Novaseq 6000 System (Illumina, San Diego, CA, USA). After quality control and data mapping, quantitative analysis of gene expression was finished by the package of subread. Differential analysis of gene expression was performed by the DESeq2 (54). The quantitative RNA expression data of genes in LN229 with or without TTF2 knock-down had been uploaded to the figshare website (https://figshare.com/), and the DOI is 10.6084/m9.figshare.9164399.

Statistics

Patients were divided into “high-risk” and “low-risk” groups using the median risk score as the cutoff value in both the training and validation datasets. A nonparametric test was used to compare the distribution of age between the two risk groups, and Chi-square tests were used to compare the distribution of other clinicopathological features.
One-way ANOVA test was performed to compare the risk scores in patients grouped by WHO 2016 subgroup or combination of TERT promoter and IDH status. Two tailed Student’s t test was performed to compare the risk scores in patients grouped by other clinical or molecular-pathological characteristics, and a P value less than 0.05 was considered significant.

The prediction efficiency of the signature risk scores, age, WHO grade, and subgroups according to WHO 2016 classification for 3-year survival and 5-year survival were examined using ROC curves. Univariate and multivariate Cox regression analysis was performed to determine the prognostic value of the risk score and various clinical and molecular-pathological characteristics.

The Kaplan–Meier method with a two-sided log-rank test was used to compare the OS of patients in the different RNA processing clusters or in the high- and low-risk groups. All statistical analyses were conducted using R v3.4.1 (https://www.r-project.org/), SPSS 16.0 (SPSS, Inc., Chicago, IL, USA) and Prism 7 (GraphPad Software, Inc., La Jolla, CA, USA).

Study approval

This retrospective study was approved and reviewed by the institutional review board of Beijing Tiantan Hospital, and the informed consent was waived.

Author Contributions

R.C.C. and Y.M.L.: study design, data analysis and manuscript drafting; R.C.C and Y.Z.C performed the cell biological experiments and analyzed transcriptional data.

Acknowledgments

This study was supported by the National Natural Science Foundation of China (81773208, 81402052), the Beijing Nova Program (Z16110004916082), and the National Key Research and Development Plan (2016YFC0902500).

Abbreviations:

The abbreviations used in the manuscript are listed in the Supplementary Table 15.
References:


Figure legends

1. Figure 1. The workflow of the current study.
Figure 2. Prognosis-associated RNA processing genes’ expression profile in LGGs.

(A) Heatmap showing the expression pattern of the 276 RNA processing genes associated with patient survival. (B) GO biological process terms enriched among the 276 survival-associated RNA processing genes. (C) The 19 genes included in the signature, their hazard ratios (HR), and 95% confidence intervals (CI) by univariate Cox regression analysis, and the coefficients by multivariate Cox regression analysis using LASSO. (D–I) Kaplan–Meier overall survival (OS) curves for patients from TCGA (D–F) and CGGA (G–I) datasets with total lower grade glioma (D and G), oligodendroglioma with IDH-mutant and 1p/19q codeletion (E and H), Astrocytoma
with $IDH$-mutant or $IDH$-wild-type (F and I). All of the sample numbers (n) and P values are labeled in the figure.
Figure 3. Comparisons of survival predictive efficiencies between the risk score and clinicopathological characteristics.

ROC curves showed the predictive efficiencies of risk scores, age, subgroups according to WHO 2016 classification, and WHO grade on 3-year and 5-year survival in TCGA (A, n=143 and B, n=111) and CGGA (C, n=121 and D, n=84) datasets.
Figure 4. Alternative splicing profile analysis in LGGs with lower or higher risk scores.

(A) Differentially spliced events in 446 LGGs patients with increased risk score. Bars indicate the proportion of each ASE type. Points indicate the number of differentially spliced events in each patient. The IDH status and 1p/19q codeletion status of these LGGs are also presented. (B) The absolute numbers of all alternative spliced events were compared in LGGs with lower (n=111) or higher risk (n=111) score. ****P < 0.0001.
**Figure 5.** Differential ASE in LGGs with lower or higher risk scores.

(A) Heatmaps showing the expression level of RNA splicing genes (upper panel) and PSI of ASEs with significant differences between LGG groups with lower and higher risk score. (B) Representative ASEs with differential PSI between LGG groups with lower and higher risk score. (C) GO_BP terms of spliced genes with differential PSI between LGGs with lower and higher risk scores.
Figure 6. The ASE networks and RNA splicing genes.

(A) Labeled circles in the center represent RNA splicing genes. Red circles indicate up-regulated RNA splicing genes in LGGs with higher risk score, while green circles indicate down-regulated splicing genes. Colored circles connected to splicing genes by red or blue lines are distinct types of differential ASEs. The red connecting lines represent positive correlation, while blue connecting lines represent negative correlation. (B) The numbers of ASE significantly correlated to up-regulated (upper panel) or down-regulated (lower panel) RNA splicing genes are summarized.
Figure 7. Functions of genes correlated with risk scores.

(A–B) Functional analysis of genes positively (red bar chart) or negatively (green bar chart) correlated with the risk score using GO terms of biological processes (A) and KEGG pathway (B).
Figure 8. GSEA analysis of genes correlated with the risk scores.

(A–H) GSEA revealed the hallmarks of malignant tumors positively correlated with LGGs with high risk scores.
Figure 9. Transcriptome changes of LN229 cells after knock-down mRNA expression of TTF2.

(A) the mRNA expression changes of TTF2 at 48 hr after transfection of TTF2 specific siRNA. ****P<0.0001, n=3. (B) The volcano figure of deferentially expression genes between TTF2 siRNA (n=3) and Scramble siRNA (n=3). (C–D) Functional analysis of genes down-regulated (red bar chart) or up-regulated (green bar chart) after knock-down mRNA expression of TTF2 using GO terms of biological processes (A) and KEGG pathway (B).
<table>
<thead>
<tr>
<th></th>
<th>Univariate Cox Regression</th>
<th>Multivariate Cox Regression</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Univariate HR (95% CI)</td>
<td>Multivariate HR (95% CI)</td>
</tr>
<tr>
<td></td>
<td>P-value</td>
<td>HR</td>
</tr>
<tr>
<td>Grade (II vs III)</td>
<td>3.86E-06</td>
<td>0.306</td>
</tr>
<tr>
<td>Age (Age&lt;40 vs Age≥40)</td>
<td>2.31E-12</td>
<td>1.068</td>
</tr>
<tr>
<td>Gender (female vs male)</td>
<td>8.78E-01</td>
<td>1.036</td>
</tr>
<tr>
<td>IDH (mutant vs wildtype)</td>
<td>4.49E-14</td>
<td>0.161</td>
</tr>
<tr>
<td>1p/19q (codel vs non codel)</td>
<td>2.95E-03</td>
<td>0.425</td>
</tr>
<tr>
<td>The TCGA dataset</td>
<td></td>
<td></td>
</tr>
<tr>
<td>WHO 2016 subgroups</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Oligo, 1p/19q codel, IDH-mutant</td>
<td>2.71E-01</td>
<td>1.412</td>
</tr>
<tr>
<td>Astro, IDH-mutant</td>
<td>2.04E-10</td>
<td>7.754</td>
</tr>
<tr>
<td>Astro, IDH-wildtype</td>
<td>1.52E-03</td>
<td>0.444</td>
</tr>
<tr>
<td>MGMT (methylation vs unmethylation)</td>
<td>4.85E-26</td>
<td>4.7</td>
</tr>
<tr>
<td>Risk score (high vs low)</td>
<td>9.97E-09</td>
<td>0.166</td>
</tr>
<tr>
<td>Grade (II vs III)</td>
<td>9.21E-04</td>
<td>1.048</td>
</tr>
<tr>
<td>Increasing age</td>
<td>7.83E-01</td>
<td>0.923</td>
</tr>
<tr>
<td>Gender (female vs male)</td>
<td>2.36E-06</td>
<td>0.253</td>
</tr>
<tr>
<td>IDH (mutant vs wildtype)</td>
<td>8.08E-03</td>
<td>0.147</td>
</tr>
<tr>
<td>1p/19q (codel vs non codel)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>The CGGA dataset</td>
<td></td>
<td></td>
</tr>
<tr>
<td>WHO 2016 subgroups</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Oligo, 1p/19q codel, IDH-mutant</td>
<td>4.41E-02</td>
<td>4.426</td>
</tr>
<tr>
<td>Astro, IDH-mutant</td>
<td>2.19E-04</td>
<td>15.542</td>
</tr>
<tr>
<td>Astro, IDH-wildtype</td>
<td>5.30E-03</td>
<td>0.249</td>
</tr>
<tr>
<td>MGMT (methylation vs unmethylation)</td>
<td>1.13E-13</td>
<td>11.578</td>
</tr>
</tbody>
</table>

IDH: isocitrate dehydrogenase; codel: codeletion; Oligo: oligodendrocytoma; Astro: astrocytoma.
Supplementary Figures:

Supplementary Figure 1. Ten-fold cross validation for tuning parameter selection in the LASSO model. LASSO coefficient profiles of the 1203 1p19q genes. The minimum criteria are indicated by the dashed vertical line (left).
Supplementary Figure 2. Prognostic value of the risk signature in astrocytoma. Kaplan–Meier overall survival (OS) curves for patients from TCGA (A-B) and CGGA (C-D) datasets with astrocytoma with IDH-mutant or IDH-wildtype. The P value and sample numbers (n) are labeled on the figure.
Supplementary Figure 3. Prognostic value of the risk signature in different races of TCGA dataset.

(A) The composition of races in TCGA dataset. (B-D) Kaplan–Meier overall survival (OS) curves for White (B), Black or African American (C), and Asian population in TCGA (D), respectively. The P value and sample numbers (n) are labeled on the figure.
Supplementary Figure 4. Distribution of risk scores in patients stratified by clinicopathological and gene expression characteristics. (A) Heatmap showing the expression level of the 19 signature genes in low-risk and high-risk LGGs. Expression levels are indicated by the color bar to the right ranging from blue (no/low) to red (high) expression. The distribution of clinicopathological features was also compared between the low- and high-risk groups. (B–H) Distribution of risk scores for patients in TCGA dataset stratified by WHO grade (B) WHO 2016 subgroup (C), IDH status (D), 1p/19q codeletion status (E), MGMT promoter methylation status (F), age (G), and TERT promoter (H). **P < 0.01, ***P < 0.001, and ****P < 0.0001.
Supplementary Figure 5. Distribution of risk score in LGGs stratified by different pathological features in CGGA dataset. (A–G) Distribution of risk scores for patients in the CGGA dataset stratified by WHO grade (A), IDH status (B), 1p/19q codeletion status (C), MGMT promoter methylation status (D), age (E), WHO 2016 subgroup (F), and TERT promoter (G). *P < 0.05, **P < 0.01, ***P < 0.001, and ****P < 0.0001.
Supplementary Figure 6. Distribution of percent spliced in (PSI) levels for ASEs in LGGs. (A) Fraction of events of distinct PSI levels in total LGGs, LGGs with lower and higher risk score. (B–H) Numbers of events of distinct PSI in total LGGs (n = 446), LGGs with lower (n = 111), and higher (n = 111) risk score. AA: alternate acceptor site; AD: alternate donor site; AP: alternate promoter; AT: alternate terminator; ES: exon skip; ME: mutually exclusive exons; RI: retained intron.
Supplementary Figure 7. General information for ASE with differential PSI between LGG with lower and higher risk scores. (A–B) Total number of changed spliced genes and ASEs (A) and the fraction of changed ASE types (B) between LGG groups with lower and higher risk scores are summarized.
Supplementary Figure 8. Functions of spliced genes with differentially PSI between gliomas with lower (n=111) and higher (n=111) risk scores. (A–B) functional analysis of spliced genes with decreased PSI (A) or increased PSI (B) in LGGs with higher risk score.