Received: 22 December 2021 | Revised: 20 February 2022 | Accepted: 20 March 2022

DOI: 10.1002/brb3.2575

Brain and Behavior Open Access WILEY

ORIGINAL ARTICLE

Integrated analysis of single-cell RNA-seq dataset and bulk RNA-seq dataset constructs a prognostic model for predicting survival in human glioblastoma

Wenwen Lai[1,2] | Defu Li[1,2] | Jie Kuang[1] | Libin Deng[1,2] | Quqin Lu[1,2]

[1]Jiangxi Provincial Key Laboratory of Preventive Medicine, Nanchang University, Nanchang, China

[2]Department of Biostatistics and Epidemiology, School of Public Health, Nanchang University, Nanchang, China

Correspondence
Quqin Lu, Department of Biostatistics and Epidemiology, School of Public Health, Nanchang University, Nanchang, Jiangxi 330000, China.
E-mail: quqinlu@ncu.edu.cn

Funding information
National Natural Science Foundation of China, Grant/Award Number: 31860311

Abstract

Background: Glioblastoma (GBM) is the most common primary malignant brain tumor in adults. For patients with GBM, the median overall survival (OS) is 14.6 months and the 5-year survival rate is 7.2%. It is imperative to develop a reliable model to predict the survival probability in new GBM patients. To date, most prognostic models for predicting survival in GBM were constructed based on bulk RNA-seq dataset, which failed to accurately reflect the difference between tumor cores and peripheral regions, and thus show low predictive capability. An effective prognostic model is desperately needed in clinical practice.

Methods: We studied single-cell RNA-seq dataset and The Cancer Genome Atlas-glioblastoma multiforme (TCGA-GBM) dataset to identify differentially expressed genes (DEGs) that impact the OS of GBM patients. We then applied the least absolute shrinkage and selection operator (LASSO) Cox penalized regression analysis to determine the optimal genes to be included in our risk score prognostic model. Then, we used another dataset to test the accuracy of our risk score prognostic model.

Results: We identified 2128 DEGs from the single-cell RNA-seq dataset and 6461 DEGs from the bulk RNA-seq dataset. In addition, 896 DEGs associated with the OS of GBM patients were obtained. Five of these genes (*LITAF*, *MTHFD2*, *NRXN3*, *OSMR*, and *RUFY2*) were selected to generate a risk score prognostic model. Using training and validation datasets, we found that patients in the low-risk group showed better OS than those in the high-risk group. We validated our risk score model with the training and validating datasets and demonstrated that it can effectively predict the OS of GBM patients.

Conclusion: We constructed a novel prognostic model to predict survival in GBM patients by integrating a scRNA-seq dataset and a bulk RNA-seq dataset. Our findings may advance the development of new therapeutic targets and improve clinical outcomes for GBM patients.

KEYWORDS
glioblastoma, overall survival, prognostic model, single-cell RNA-seq, bulk RNA-seq

1 | INTRODUCTION

Glioblastoma (GBM) is the most common primary malignant brain tumor in adults, accounting for 48.6% of malignant tumors in the central nervous system and 14.5% of all tumors (Ostrom et al., 2020). The median overall survival (OS) time is around 14.6 months for patients diagnosed with GBM, with only a 5-year survival rate of 7.2% (Lynes et al., 2020; Ostrom et al., 2020). Currently, the main treatment measures for GBM include radiotherapy, chemotherapy, and surgical resection (Fabian et al., 2019). Unfortunately, little progress has been made toward prolonging survival in GBM despite considerable effort in improving treatments over the past decades (Alexander & Cloughesy, 2017). The OS of each GBM patient is a crucial factor in developing a personalized treatment plan. Therefore, it is imperative to develop a reliable tool to predict the survival probability for patients with newly diagnosed GBM.

With the advancement of high-throughput technologies, RNA sequencing (RNA-seq) from bulk tissue has become indispensable for transcriptome-wide analysis (Stark et al., 2019). Many public databases have been established, including The Cancer Genome Atlas (TCGA, https://cancergenome.nih.gov/) and Gene Expression Omnibus (GEO, https://www.ncbi.nlm.nih.gov/geo/). These databases enable scientists to investigate the relationship between the prognosis of diseases and gene expression profiles. Studies are in mounting numbers being reported to identify biomarkers associated with prognosis for GBM patients (Wang et al., 2019; Zhao et al., 2021; Zhou et al., 2021). However, the expression of genes obtained from bulk tissue does not reflect their expression in individual cells, which leads to the high heterogeneity of GBM being masked.

The development of single-cell RNA sequencing (scRNA-seq) in recent years has significantly expanded our knowledge about biological systems. As an emerging technology, scRNA-seq has been applied increasingly to explore extensive intratumoral heterogeneity (Kinker et al., 2020; Patel et al., 2014; Peng et al., 2019). Compared to calculating the average gene expression in all the cells, scRNA-seq allows the evaluation of gene expression at a single-cell resolution, which greatly compensates for the shortage of RNA-seq from bulk tissue (G. Chen et al., 2019). In addition, scRNA-seq analysis enables researchers to discover critical genes that are characteristic of cancer cells (Kulkarni et al., 2019). In this study, we studied a scRNA-seq dataset and a bulk RNA-seq dataset and integrated them to construct a novel prognostic model for predicting survival in GBM.

2 | MATERIALS AND METHODS

2.1 | Acquisition of bulk RNA-seq dataset and scRNA-seq dataset in GBM patients

We included the scRNA-seq dataset and three bulk RNA-seq datasets of human GBM samples in our study. We first obtained the gene expression dataset and related clinical information of GBM patients from The Cancer Genome Atlas-glioblastoma multiforme (TCGA-GBM) dataset. The gene expression dataset and the clinical information from GSE43378 were collected from the GEO database. In addition, the RNA sequencing dataset and corresponding clinical information that contained 693 samples (dataset ID: mRNAseq_693) were downloaded from Chinese Glioma Genome Atlas (CGGA; http://www.cgga.org.cn) database. We first performed data clean-up. We excluded cases that do not have follow-up time or survival status, as well as the ones that had clinical information but no corresponding RNA-seq data. According to the exclusion criteria, a total of 152 tumor samples and five normal controls in the TCGA-GBM dataset were enrolled in the study and selected as the training dataset, and a total of 50 tumor samples in the GSE43378 dataset and a total of 133 GBM samples in the mRNAseq_693 dataset were enrolled and selected as the validation datasets. The scRNA-seq dataset with a total of 3589 cells from four human primary GBM samples from the GSE84465 dataset was acquired from the GEO database. Among the 3589 cells, 2343 were from tumor cores and 1246 were from peripheral regions, with a reading depth of 10× genomics based on Illumina NextSeq 500.

2.2 | The processing of GBM scRNA-seq dataset

We analyzed 2343 cells from tumor cores as follows. We used the Seurat package in R 4.0.0 to perform quality control, statistical analysis, and explore the scRNA-seq dataset (Gribov et al., 2010). We calculated the percentage of mitochondrial genes with the PercentageFeatureSet function and elucidated the relationship between the sequencing depth, the mitochondrial gene sequences, and total intracellular sequences through correlation analysis. We cleaned up data according to the following quality control criteria: first, genes detected in < 3 cells were omitted; second, cells with < 100 total detected genes were

excluded; third, cells with ≥ 5% mitochondria-expressed genes were discarded; and last, cells with nuclei gene counts < 200 or > 6000 were excluded. We normalized the gene expression of the remaining cells with the LogNormalize method, and we identified the top 1500 genes with highly variable features by variance analysis. We then performed principal component analysis (PCA) to identify significantly available dimensions with a p-value < .05 based on the expression of these genes (Ringnér, 2008). We next applied the t-distributed stochastic neighbor embedding (tSNE) algorithm to reduce dimensionality with 20 initial PCs and perform cluster classification analysis (Kobak & Berens, 2019). With the criteria of $\log_2$ [fold change (FC)] > 0.25 and an adjusted p-value < .05, marker genes in each cluster were obtained. Clusters were annotated through the "SingleR" package based on these marker genes (Aran et al., 2019).

Then, 1246 cells from peripheral regions were analyzed as described before, except that the cells with nuclei gene counts < 200 or > 4000 were excluded rather than cells with sequencing number < 200 or nuclei gene counts > 6000.

2.3 | The identification of DEGs from the scRNA-seq dataset and TCGA-GBM dataset

In the GBM scRNA-seq dataset, cancer cells were selected as representative tumor cores after annotation, and neurons were selected as representative peripheral regions. Then, the differentially expressed genes (DEGs) between cancer cells and neurons were identified by the "DEsingle" package (Miao et al., 2018). Genes with $|\log_2 FC| > 2$ and an adjusted p-value < .05 were considered DEGs. For the TCGA-GBM dataset, first, $\log_2$ transformation was employed to generate expression profiles, and then the genes between tumor samples and normal controls were used for differentially expressed analysis using the "Limma" package (Ritchie et al., 2015). Genes with $|\log_2 FC| > 1$ and p-value < .05 were defined as DEGs.

2.4 | Enrichment analysis of Gene Ontology functions and Kyoto Encyclopedia of Genes and Genomes pathways for DEGs

To investigate the biological implications of DEGs identified from both the GBM scRNA-seq and TCGA-GBM datasets, we performed an intersection of the datasets. We then conducted Gene Ontology (GO) function and Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway analysis using WebGestalt (WebGestalt: WEB-based GEne SeT AnaLysis Toolkit, RRID:SCR_006786). A gene set at p < .05 and false discovery rate (FDR) < 0.05 was considered to be significantly enriched.

2.5 | Analysis of DEGs associated with overall survival in GBM patients

First, univariate Cox proportional hazards regression analysis was used to assess the relationship between the expression of genes and the OS of patients in the TCGA-GBM dataset. Genes with hazard ratio (HR)≠1 and p < .05 were defined as genes associated with OS. Then, DEGs associated with OS were obtained by overlapping genes associated with OS and DEGs from both the GBM scRNA-seq and TCGA-GBM datasets.

2.6 | Prognostic model construction

The DEGs associated with OS were regarded as candidate genes for constructing a prognostic model. Then, we conducted least absolute shrinkage and selection operator (LASSO) Cox penalized regression analysis using the R package "glmnet" (Friedman et al., 2010), and the genes with nonzero coefficients were selected to establish a risk score prognostic model. Based on the results of LASSO Cox penalized regression analysis, we calculated the risk score for each GBM patient in the training dataset.

2.7 | Prognostic model validation

After constructing the risk score prognostic model, we used one independent dataset GSE43378 including 50 patients with complete OS information from GEO and the other independent dataset mRNAseq_693 including 133 patients with complete OS information from CGGA to validate the model, respectively. First, we performed time-dependent receiver operating characteristic (ROC) curve analysis to predict the 12-, 15-, and 18-month survival using the R package "survivalROC" (Lorent et al., 2014). Then, based on the median risk score, we divided GBM patients into high- and low-risk groups. We performed Kaplan–Meier survival analysis to determine the association between the risk score prognostic model and the OS of GBM patients. The significance of differences in survival between the two groups was determined by the log-rank test.

3 | RESULTS

3.1 | Identification of cancer cells and neurons in the GBM scRNA-seq dataset

Thirty-eight nonconforming cells were excluded, and 2305 cells were preserved for further analysis after quality control from tumor cores (Figure 1a). We performed correlation analysis and found that there appeared to be no correlation between sequencing depth and mitochondrial gene sequences (Figure 1b). However, there was a significant positive correlation between the sequencing depth and total intracellular sequences (r = 0.37, Figure 1c). We also found that among the total of 18,545 genes analyzed, 1500 had high variation and 17,045 had low intercellular variation (Figure 1d). For PCA analysis, we picked 20 principal components (PCs) that have a p-value < .05 for subsequent analysis (Figure 1e). Then, we applied the tSNE algorithm and successfully classified the cells from tumor cores into 13 separate

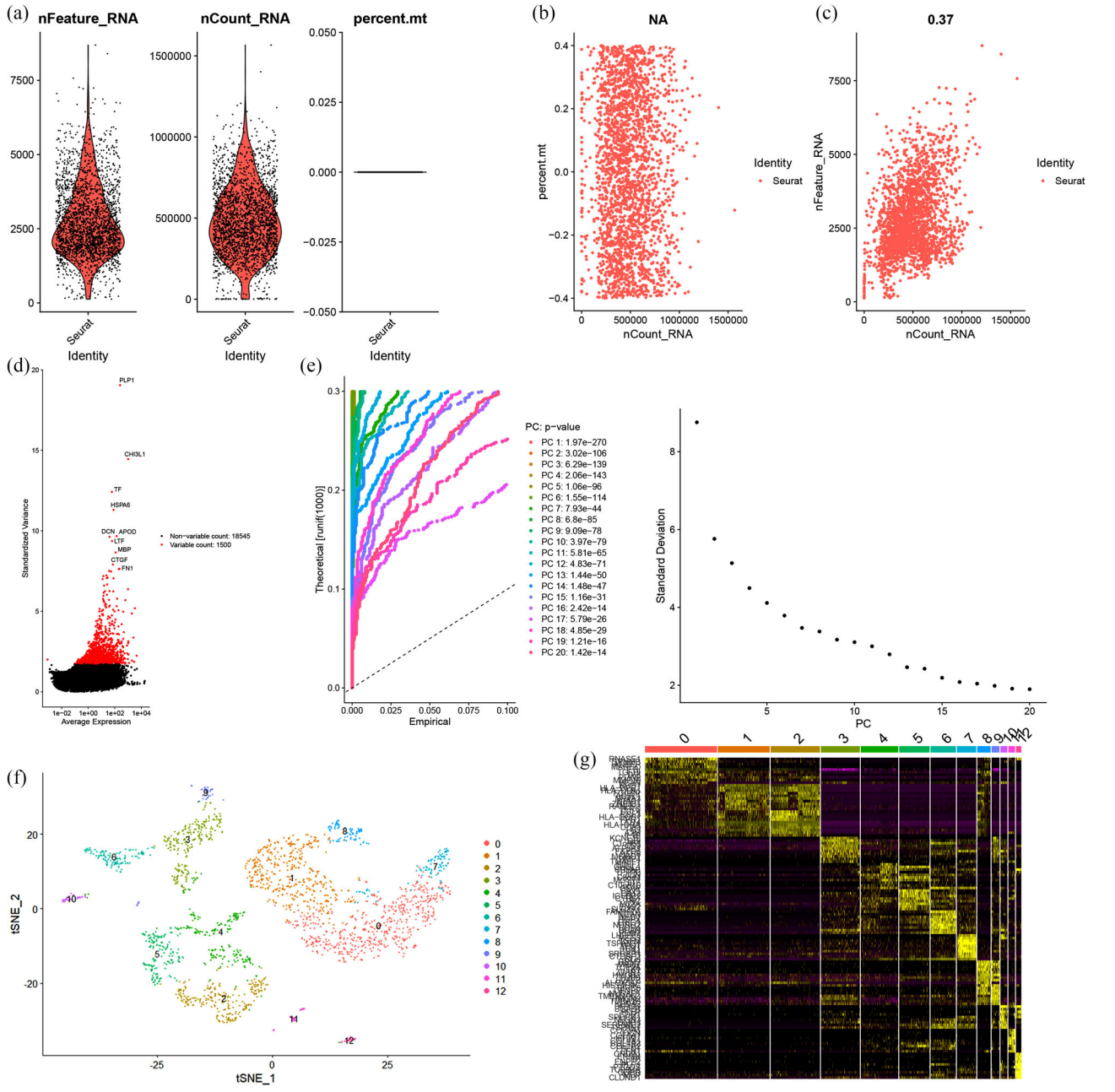

FIGURE 1 The processing of cells from tumor cores in the glioblastoma (GBM) scRNA-seq dataset. (a) Thirty-eight nonconforming cells were filtered out during quality control and normalization, and 2305 cells were screened for further analysis. (b) Correlation analysis of sequencing depth and mitochondrial gene sequences. (c) Correlation analysis of sequencing depth and total intracellular sequences. (d) Among the 18,545 genes analyzed, 17,045 showed low and 1500 showed high intercellular variation. (e) Twenty principal components (PCs) with significant differences were identified with $p < .05$. (f) Two thousand three hundred five cells were divided into 13 separate clusters. (g) Heatmap displaying the top 10 marker genes in each cluster

clusters (Figure 1f). We identified a total of 13,616 marker genes from all 13 clusters, and the top 10 marker genes from each cluster were presented in the heatmap (Figure 1g). We annotate clusters with singleR based on the expression of these marker genes (Figure 3a). We determined that the Clusters 0, 1, 7, and 8, containing 1176 cells, were macrophages; Clusters 2 and 4, containing 450 cells, were GBM cancer cells; Clusters 3, 5, 6, 9, 10, and 12, containing 637 cells, were astrocytes; and Cluster 11, containing 42 cells, was endothelial cells.

A total of 1193 cells were screened, and 53 nonconforming cells were excluded for further analysis after quality control from peripheral regions (Figure 2a). We found that though the sequencing depth did not have correlation with mitochondrial gene sequences (Figure 2b), it showed a significant positive correlation with total intracellular sequences ($R = 0.43$, Figure 2c). Among the 17,210 genes analyzed, 500 had high variation and 15,710 had low intercellular variation (Figure 2d). We executed PCA and selected 17 PCs with a p-value $< .05$

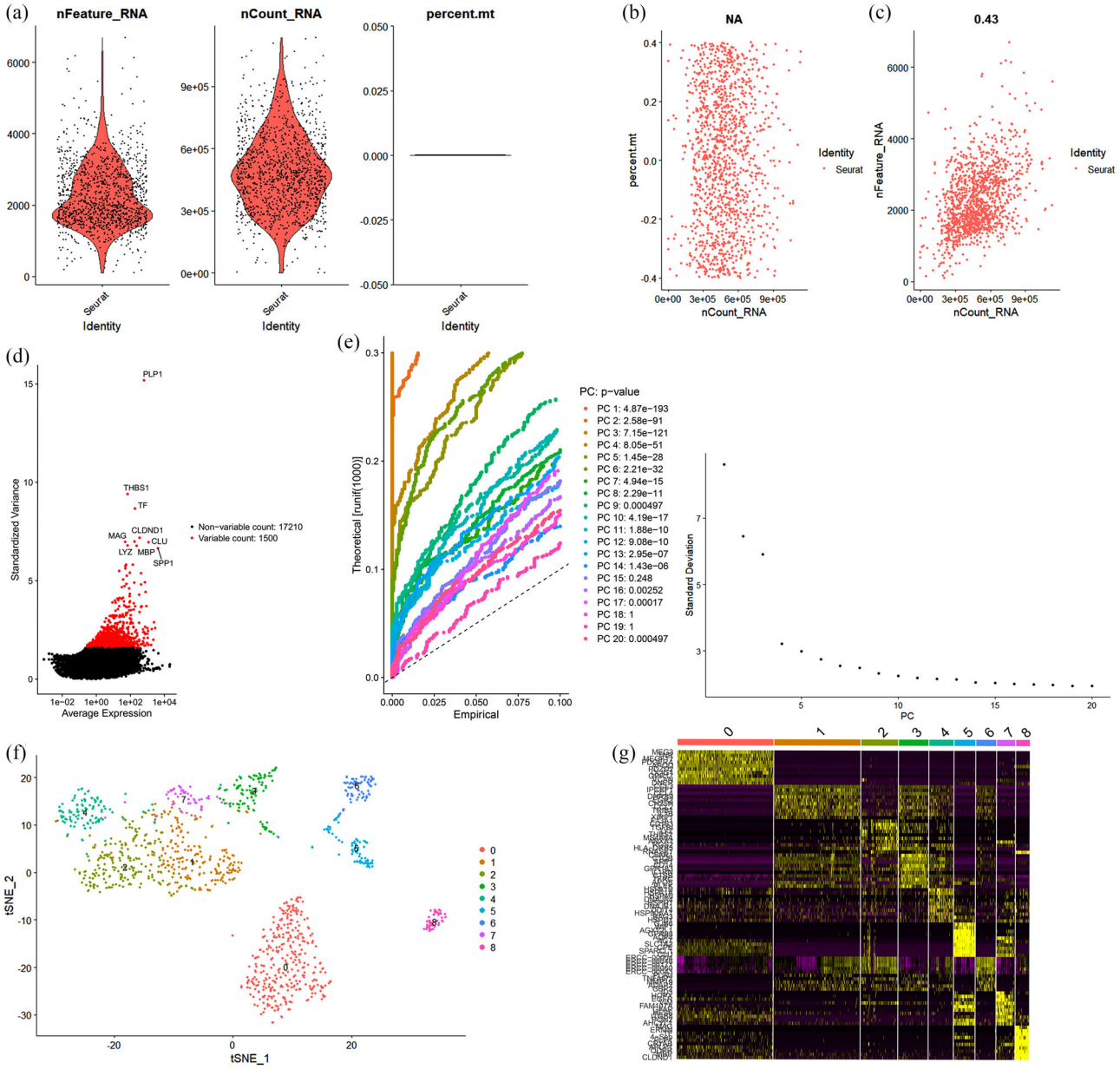

FIGURE 2 The processing of cells from peripheral regions in the glioblastoma (GBM) scRNA-seq dataset. (a) Fifty-three nonconforming cells were filtered out during quality control and normalization, and 1193 cells were screened for further analysis. (b) Correlation analysis of sequencing depth and mitochondrial gene sequences. (c) Correlation analysis of sequencing depth and total intracellular sequences. (d) Among the total of 17210 genes analyzed, 15710 showed low intercellular variation, while 1500 showed high variation. (e) Seventeen principal components (PCs) with significant differences were identified with $p < .05$. (f) A total of 1193 cells were divided into nine separate clusters. (g) Heatmap displaying the top 10 marker genes in each cluster

for subsequent analysis (Figure 2e). Then, we performed the tSNE algorithm and divided the cells from peripheral regions into nine separate clusters (Figure 2f). We have identified a total of 6748 marker genes from all nine clusters, and the top 10 marker genes from each cluster were laid out in the heatmap (Figure 2g). All clusters were annotated by singleR based on the expression of these marker genes (Figure 3b). Clusters 0, 5, and 6, containing 482 cells, were annotated as astrocytes; Clusters 1, 2, 4, and 7, containing 559 cells, were classified as macrophages; Cluster 3, containing 105 cells, was annotated as monocytes; and Cluster 8, containing 47 cells, was classified as neurons.

3.2 | DEGs from scRNA-seq dataset and TCGA-GBM dataset

A total of 450 cancer cells were selected as representative tumor cores, and 47 neurons were selected as representative peripheral regions in the GBM scRNA-seq dataset. We obtained 2128 DEGs at $|\log_2 FC| > 2$ and an adjusted p-value $< .05$ between these cancer cells and neurons. In addition, 6461 DEGs at $|\log2 FC| > 1$ and $p < .05$ were identified between 152 tumor samples and five normal controls in the TCGA-GBM dataset.

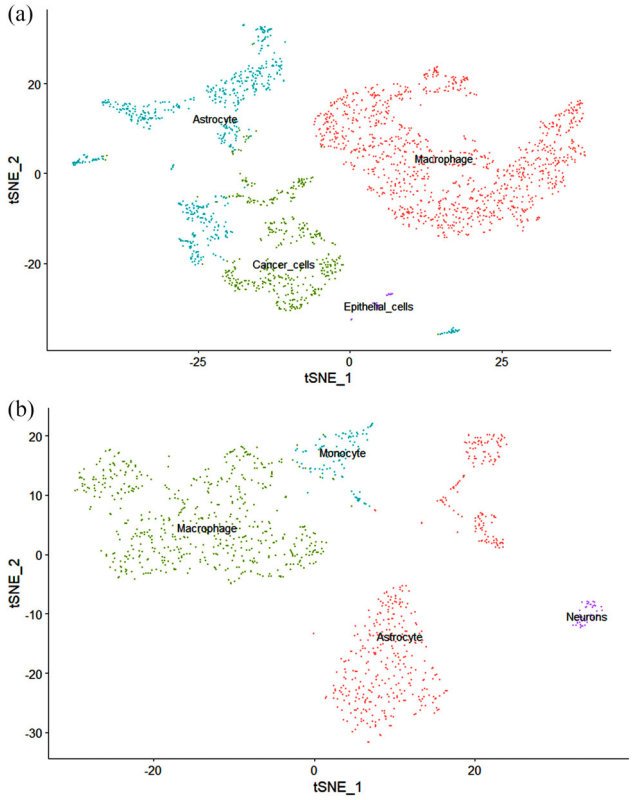

FIGURE 3 Cell annotation by singleR corresponding to the composition of the marker genes in each cluster. (a) All 13 clusters of cells from tumor cores in glioblastoma (GBM) scRNA-seq dataset were annotated. (b) All nine clusters of cells from peripheral regions in GBM scRNA-seq dataset were annotated

3.3 | Enrichment analysis of GO functions and KEGG pathways

There were 896 genes in the intersection of the identified DEGs from the GBM scRNA-seq and TCGA-GBM datasets. The KEGG analysis on the 896 genes suggested that they were enriched in signaling pathways such as the MAPK signaling pathway, the apelin signaling pathway, and pathways involved in circadian entrainment (Figure S1).

3.4 | DEGs associated with OS in GBM patients

We preliminarily identified 1418 genes that are linked to the OS of patients using univariate Cox proportional hazards regression analysis on the TCGA-GBM dataset. We found that 43 genes were at the intersection of genes associated with OS and DEGs from the GBM scRNA-seq dataset and TCGA-GBM dataset. These genes were chosen as candidate genes for constructing a prognostic model.

3.5 | Construction of the prognostic model

After LASSO Cox penalized regression analysis in the training dataset (Figure 4a,b), we constructed a five-gene (*LITAF*, *MTHFD2*, *NRXN3*, *OSMR*, and *RUFY2*)-based risk score prognostic model. The risk score = 0.01301 × expression of *LITAF* − 0.03406 × expression of *MTHFD2* + 0.04864 × expression of *NRXN3* + 0.09675 × expression of *OSMR* − 0.00038 × expression of *RUFY2*. We performed time-dependent ROC curve analysis to predict the 12-, 15-, and 18-month survival, and the area under curves (AUCs) for 12-, 15-, and 18-month OS were 0.728, 0.721, and 0.713, respectively (Figure 4c). In addition, the survival curve suggested that the high-risk group showed a worse prognosis, compared to the low-risk group, with $p < .001$ (Figure 4d).

3.6 | Validation of the prognostic model

We used the dataset GSE43378 including 50 patients with complete OS information from GEO and the dataset mRNAseq_693 including 133 patients with complete OS information from CGGA as the external validation datasets to evaluate the robustness and effectiveness of our risk score prognostic model. We also performed time-dependent ROC curve analysis to predict the 12-, 15-, and 18-month survival. The AUCs for 12-, 15-, and 18-month OS were 0.645, 0.701, and 0.733 in GSE43378, respectively (Figure 4e). The AUCs for 12-, 15-, and 18-month OS were 0.616, 0.634, and 0.622 in mRNAseq_693, respectively (Figure 4g). Additionally, in Figure 4f,h, the survival curve indicated that the high-risk group presented a worse prognosis than the low-risk group ($p < .001$) in these two validation datasets. In summary, these results indicate the effective predictive capability of the risk score prognostic model constructed by integrated analysis of scRNA-seq and bulk RNA-seq datasets.

4 | DISCUSSION

Over years, there are increasing studies using public databases to predict survival in GBM. However, the DEGs were identified from bulk RNA-seq in most studies, which failed to accurately reflect the difference between tumor cores and peripheral regions, thus weakening the predictive ability of the models (P. F. Chen et al., 2019; Xu et al., 2020; Zhang et al., 2020). On the other hand, analysis using the scRNA-seq data could resolve gene expression at single-cell resolution, allowing the classification and annotation of their expression in a cell-type- or tissue-specific manner. Neurons are regarded as the original cells of cancer cells in GBM (Friedmann-Morvinski et al., 2012; Vescovi et al., 2006). In this study, we first obtained candidates DEGs from the scRNA-seq dataset GSE84465, and then combined them with TCGA-GBM to acquire the DEGs associated with OS. Finally, we constructed a risk score prognostic model and validated the model with the external dataset GSE43378. Figure 5 shows the flow of the study.

The results of KEGG analysis showed that DEGs were enriched in signaling pathways such as the MAPK pathway, the apelin pathway, and pathways involved in circadian entrainment. A report suggested that the MAPK signaling pathway plays an important role in GBM development and malignant progression through promoting GBM cell tumorigenicity (X. Chen et al., 2020). Another study also demonstrated that

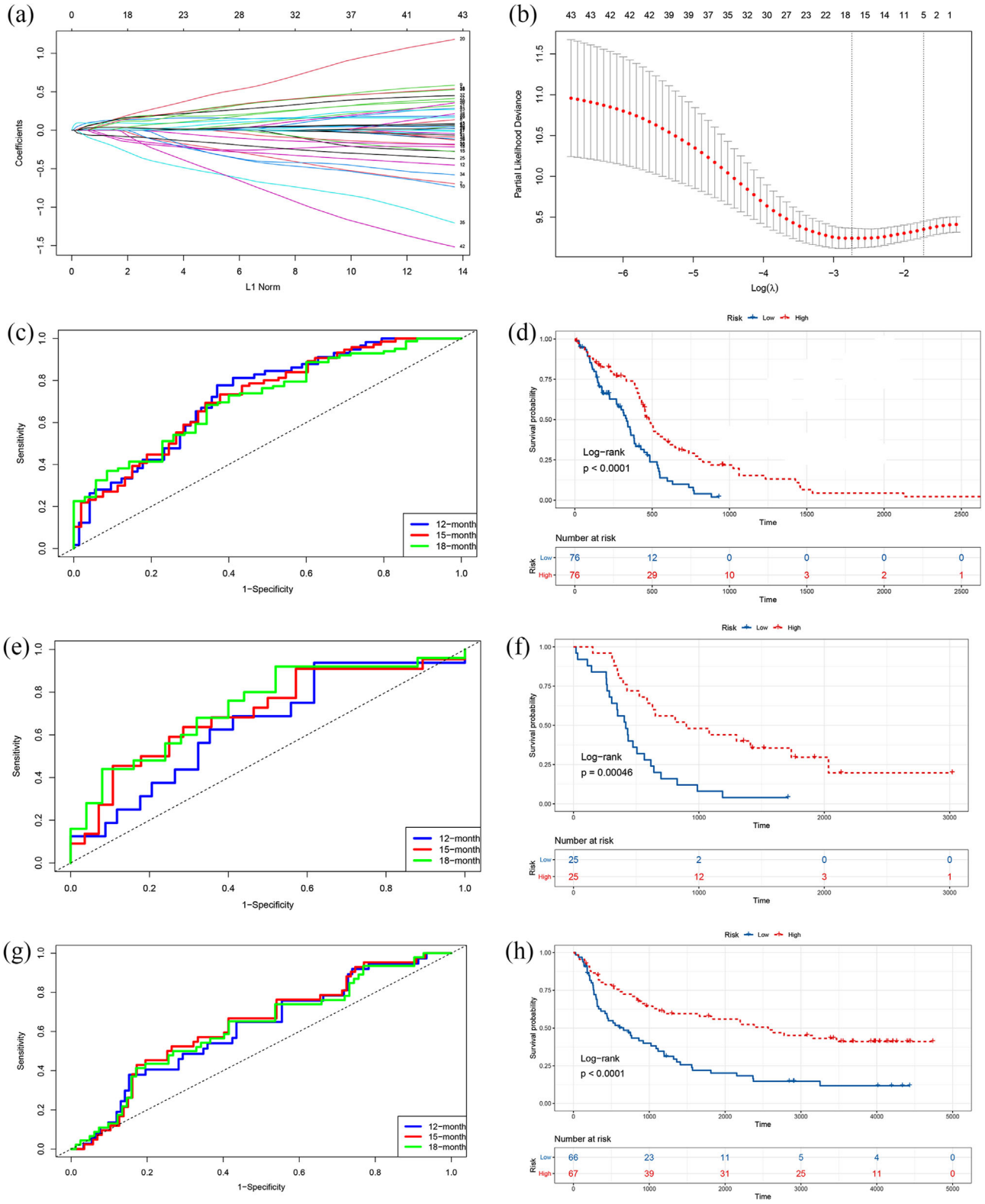

FIGURE 4 (a) least absolute shrinkage and selection operator (LASSO) coefficient profiles of the differentially expressed genes (DEGs) associated with the overall survival (OS) of glioblastoma (GBM) patients. (b) Partial likelihood deviance plotted versus log(lambda). The vertical dotted line indicates the lambda value with the minimum error and the largest lambda value where the deviance is within one SE of the minimum. (c) The receiver operating characteristic (ROC) curves for the risk score model in the training dataset. (d) The OS of patients in the five-gene risk score model low- and high-risk groups in the training dataset. (e) The ROC curves for the risk score model in the GSE43378. (f) The OS of patients in the five-gene risk score model low- and high-risk groups in the GSE43378. (g) The ROC curves for the risk score model in the mRNAseq_693. (h) The OS of patients in the five-gene risk score model low- and high-risk groups in the mRNAseq_693

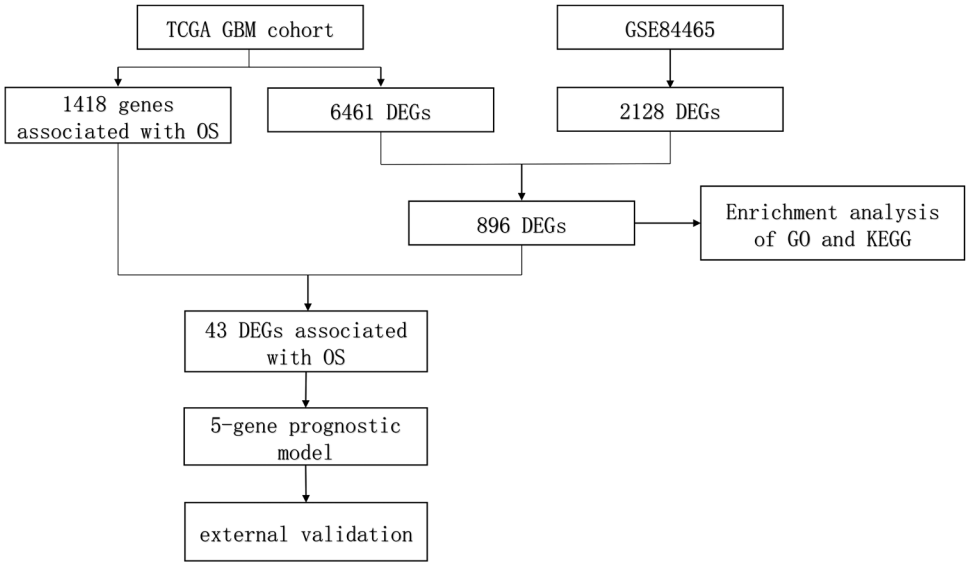

FIGURE 5 The process of constructing the five-gene risk score model. First, 2128 differentially expressed genes (DEGs) and 6461 DEGs were identified from the GSE84465 and TCGA-GBM datasets by differential expression analysis, respectively. In addition, using univariate Cox proportional hazards regression analysis, 1418 genes were identified to be associated with overall survival (OS) in glioblastoma (GBM) patients. Then, 43 DEGs associated with OS in GBM patients were obtained. Next, least absolute shrinkage and selection operator (LASSO) Cox penalized regression analysis was applied to construct a gene risk score model for prognosis prediction. Subsequently, the gene risk score model was constructed based on the five genes (*LITAF*, *MTHFD2*, *NRXN3*, *OSMR*, and *RUFY2*). Finally, the five-gene risk score model was validated using validation datasets

the apelin signaling pathway controls GBM angiogenesis and invasion (Mastrella et al., 2019). Our risk score model suggests that the five gene (*LITAF*, *MTHFD2*, *NRXN3*, *OSMR*, and *RUFY2*) might affect the OS of GBM patients through these pathways.

In the training dataset, we constructed a five-gene (*LITAF*, *MTHFD2*, *NRXN3*, *OSMR*, and *RUFY2*)-based risk score prognostic model. The risk score = 0.01301 × expression of *LITAF* − 0.03406 × expression of *MTHFD2* + 0.04864 × expression of *NRXN3* + 0.09675 × expression of *OSMR* − 0.00038 × expression of *RUFY2*. *LITAF* was identified as a transcription factor which activates the proinflammatory cytokine transcription in macrophages upon response to lipopolysaccharide. Recently, a study demonstrated that the *LITAF* expression is decreased in glioma tissues, which likely enhances the radiosensitivity of glioma cells through upregulating the FoxO1 pathway (Huang et al., 2019). *MTHFD2* is broadly required for cancer cell proliferation and viability as a metabolic enzyme and was overexpressed around the tumor regions with poor nutrient access in GBM patients, and the suppression of *MTHFD2* could cause cancer cell death (Tanaka et al., 2021). *NRXN3* belongs to a family of highly polymorphic neuronal-specific cell surface proteins, and it was reported to promote glioma cell proliferation and migration under the regulation of Fox Q1 (Sun et al., 2013). Our results not only verified the relationship between *NRXN3* and GBM but also demonstrated the effectiveness of using *NRXN3* as an important indicator for prognosis prediction in GBM patients. *OSMR* is a member of the interleukin-6 receptor family, and a previous study reported that it regulates GBM tumor growth through orchestrating a feed-forward signaling mechanism with EGFRvIII and *STAT3* to promote tumorigenesis (Jahani-Asl et al., 2016). Furthermore, another study showed that *OSMR* conferred resistance to ionizing radiation via regulation of oxidative phosphorylation and that loss of *OSMR* sensitized GBM tumors to ionizing radiation therapy (Sharanek et al., 2020). However, the function of *RUFY2* remains unknown.

In the validation dataset GSE43378, the AUCs for 12-, 15-, and 18-month OS were 0.645, 0.701, and 0.733, respectively. MRNAseq_693 downloaded from CGGA, whose all GBM samples were Chinese patients, was selected as the other validation dataset. The AUCs for 12-, 15-, and 18-month OS were 0.616, 0.634, and 0.622 in mRNAseq_693, respectively. In addition, the survival curve suggested that the high-risk group exhibited a worse prognosis than the low-risk group ($p < .001$). These results indicate the predictive capability of the risk score prognostic model. Unlike the previous prognostic signature, our risk score prognostic model was constructed by integrating the scRNA-seq dataset and bulk RNA-seq dataset. This finding could reflect the effect of these genes on GBM patient prognosis. Based on the risk score, the survival probabilities of an individual can be queried based on the level of the five genes (*LITAF*, *MTHFD2*, *NRXN3*, *OSMR*, and *RUFY2*). For patients with high risk of progression to severe conditions, it is important for them to receive adequate attention and care during treatments; therefore, our model will be a valuable tool to provide a good reference for clinicians.

In summary, we constructed a novel prognostic model to predict the survival in GBM patients though integrative analysis of a scRNA-seq dataset and a bulk RNA-seq dataset. Our findings have the potential to advance the development of new therapeutics for the treatment of GBM and improve the clinical outcomes for GBM patients. However, there are some limitations in our study. First, clinical characteristics

were not taken into account in our model. In the future, we can improve our model by integrating clinical characteristics from the analysis of patients with more comprehensive clinical information. In addition, our model was constructed and validated using public databases, and it would also be helpful to validate the model with private clinical or experimental datasets in the further.

Figure S1: GO functional and KEGG pathway analysis. (a) Summary of the differentially expressed genes and GO pathway enrichment. Red, blue, and green bars represent the biological process, cellular component, and molecular function categories, respectively. The height of the bar represents the number of differentially expressed genes observed in each category. (b) The top 10 pathways involving the differentially expressed genes.

ACKNOWLEDGMENT

This work was supported by the National Natural Science Foundation of China (31860311 to Quqin Lu).

CONFLICT OF INTEREST

The authors declare no conflict of interest.

AUTHOR CONTRIBUTIONS

Methodology, software, validation, data curation, and writing-original draft: Wenwen Lai. *Methodology, formal analysis, and visualization*: Defu Li. *Software and Data curation*: Jie Kuang. *Supervision, methodology, and formal analysis*: Libin Deng. *Conceptualization, project administration, and funding acquisition*: Quqin Lu.

DATA AVAILABILITY STATEMENT

Data were downloaded from the TCGA, GEO, and CGGA website.

PEER REVIEW

The peer review history for this article is available at https://publons.com/publon/10.1002/brb3.2575.

ORCID

Quqin Lu https://orcid.org/0000-0003-0774-593X

REFERENCES

Alexander, B. M., & Cloughesy, T. F. (2017). Adult glioblastoma. *Journal of Clinical Oncology*, *35*(21), 2402–2409. https://doi.org/10.1200/JCO.2017

Aran, D., Looney, A. P., Liu, L., Wu, E., Fong, V., Hsu, A., Chak, S., Naikawadi, R. P., Wolters, P. J., Abate, A. R., Butte, A. J., & Bhattacharya, M. (2019). Reference-based analysis of lung single-cell sequencing reveals a transitional profibrotic macrophage. *Nature Immunology*, *20*(2), 163–172. https://doi.org/10.1038/s41590-018-0276-y

Chen, G., Ning, B., & Shi, T. (2019). Single-Cell RNA-Seq technologies and related computational data analysis. *Frontiers in Genetics*, *10*, 317. https://doi.org/10.3389/fgene.2019.00317

Chen, P.-F., Li, Q.-H., Zeng, L.-R., Yang, X.-Y., Peng, P.-L., He, J.-H., & Fan, B. (2019). A 4-gene prognostic signature predicting survival in hepatocellular carcinoma. *Journal of Cellular Biochemistry*, *120*(6), 9117–9124. https://doi.org/10.1002/jcb.28187

Chen, X., Hao, A., Li, X., Ye, K., Zhao, C., Yang, H., Ma, H., Hu, L., Zhao, Z., Hu, L., Ye, F., Sun, Q., Zhang, H., Wang, H., Yao, X., & Fang, Z. (2020). Activation of JNK and p38 MAPK Mediated by ZDHHC17 drives glioblastoma multiforme development and malignant progression. *Theranostics*, *10*(3), 998–1015. https://doi.org/10.7150/thno.40076

Fabian, D., Guillermo Prieto Eibl, M., Alnahhas, I., Sebastian, N., Giglio, P., Puduvalli, V., Gonzalez, J., & Palmer, J. (2019). Treatment of glioblastoma (GBM) with the addition of tumor-treating fields (TTF): A review. *Cancers (Basel)*, *11*(2), 174, https://doi.org/10.3390/cancers11020174

Friedman, J., Hastie, T., & Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, *33*(1), 1–22. https://doi.org/10.18637/jss.v033.i01

Friedmann-Morvinski, D., Bushong, E. A., Ke, E., Soda, Y., Marumoto, T., Singer, O., Ellisman, M. H., & Verma, I. M. (2012). Dedifferentiation of neurons and astrocytes by oncogenes can induce gliomas in mice. *Science*, *338*(6110), 1080–1084. https://doi.org/10.1126/science.1226929

Gribov, A., Sill, M., Lück, S., Rücker, F., Döhner, K., Bullinger, L., Benner, A., & Unwin, A. (2010). SEURAT: Visual analytics for the integrated analysis of microarray data. *BMC Medical Genomics*, *3*, 21. https://doi.org/10.1186/1755-8794-3-21

Huang, C., Chen, D., Zhu, H., Lv, S., Li, Q., & Li, G. (2019). LITAF enhances radiosensitivity of human glioma cells via the FoxO1 pathway. *Cellular and Molecular Neurobiology*, *39*(6), 871–882. https://doi.org/10.1007/s10571-019-00686-4

Jahani-Asl, A., Yin, H., Soleimani, V. D., Haque, T., Luchman, H. A., Chang, N. C., Sincennes, M.-C., Puram, S. V., Scott, A. M., Lorimer, I. A. J., Perkins, T. J., Ligon, K. L., Weiss, S., Rudnicki, M. A., & Bonni, A. (2016). Control of glioblastoma tumorigenesis by feed-forward cytokine signaling. *Nature Neuroscience*, *19*(6), 798–806. https://doi.org/10.1038/nn.4295

Kinker, G. S., Greenwald, A. C., Tal, R., Orlova, Z., Cuoco, M S., Mcfarland, J M., Warren, A., Rodman, C., Roth, J. A., Bender, S. A., Kumar, B., Rocco, J W., Fernandes, P. A. C. M., Mader, C. C., Keren-Shaul, H., Plotnikov, A., Barr, H., Tsherniak, A., Rozenblatt-Rosen, O., … Tirosh, I. (2020). Pan-cancer single-cell RNA-seq identifies recurring programs of cellular heterogeneity. *Nature Genetics*, *52*(11), 1208–1218. https://doi.org/10.1038/s41588-020-00726-6

Kobak, D., & Berens, P. (2019). The art of using t-SNE for single-cell transcriptomics. *Nature Communication*, *10*(1), 5416. https://doi.org/10.1038/s41467-019-13056-x

Kulkarni, A., Anderson, A. G., Merullo, D. P., & Konopka, G. (2019). Beyond bulk: A review of single cell transcriptomics methodologies and applications. *Current Opinion in Biotechnology*, *58*, 129–136. https://doi.org/10.1016/j.copbio.2019.03.001

Lorent, M., Giral, M., & Foucher, Y. (2014). Net time-dependent ROC curves: A solution for evaluating the accuracy of a marker to predict disease-related mortality. *Statistics in Medicine*, *33*(14), 2379–2389. https://doi.org/10.1002/sim.6079

Lynes, J. P., Nwankwo, A. K., Sur, H. P., Sanchez, V. E., Sarpong, K. A., Ariyo, O. I., Dominah, G. A., & Nduom, E. K. (2020). Biomarkers for immunotherapy for treatment of glioblastoma. *Journal for ImmunoTherapy of Cancer*, *8*(1), e000348. https://doi.org/10.1136/jitc-2019-000348

Mastrella, G., Hou, M., Li, M., Stoecklein, V. M., Zdouc, N., Volmar, M. N. M., Miletic, H., Reinhard, S., Herold-Mende, C. C., Kleber, S., Eisenhut, K., Gargiulo, G., Synowitz, M., Vescovi, A. L., Harter, P. N., Penninger, J. M., Wagner, E., Mittelbronn, M., Bjerkvig, R., … Kälin, R E. (2019). Targeting APLN/APLNR improves antiangiogenic efficiency and blunts proinvasive side Effects of VEGFA/VEGFR2 blockade in glioblastoma. *Cancer Research*, *79*(9), 2298–2313. https://doi.org/10.1158/0008-5472.CAN-18-0881

Miao, Z., Deng, K., Wang, X., & Zhang, X. (2018). DEsingle for detecting three types of differential expression in single-cell RNA-seq data. *Bioinformatics*, *34*(18), 3223–3224. https://doi.org/10.1093/bioinformatics/bty332

Ostrom, Q. T., Patil, N., Cioffi, G., Waite, K., Kruchko, C., & Barnholtz-Sloan, J. S. (2020). CBTRUS statistical report: Primary brain and other central nervous system tumors diagnosed in the United States in 2013–2017. *Neuro-Oncology*, *22*(12), iv1–iv96. https://doi.org/10.1093/neuonc/noaa200

Patel, A. P., Tirosh, I., Trombetta, J. J., Shalek, A. K., Gillespie, S. M., Wakimoto, H., Cahill, D. P., Nahed, B. V., Curry, W. T., Martuza, R. L., Louis, D. N., Rozenblatt-Rosen, O., Suvà, M. L., Regev, A., & Bernstein, B. E. (2014). Single-cell RNA-seq highlights intratumoral heterogeneity in primary glioblastoma. *Science*, *344*(6190), 1396–1401. https://doi.org/10.1126/science.1254257

Peng, J., Sun, B.-F., Chen, C.-Y., Zhou, J.-Y., Chen, Y.-S., Chen, H., Liu, L., Huang, D., Jiang, J., Cui, G.-S., Yang, Y., Wang, W., Guo, D., Dai, M., Guo, J., Zhang, T., Liao, Q., Liu, Y., Zhao, Y.-L., … Wu, W. (2019). Single-cell RNA-seq highlights intra-tumoral heterogeneity and malignant progression in pancreatic ductal adenocarcinoma. *Cell Research*, *29*(9), 725–738. https://doi.org/10.1038/s41422-019-0195-y

Ringnér, M. (2008). What is principal component analysis? *Nature Biotechnology*, *26*(3), 303–304. https://doi.org/10.1038/nbt0308-303

Ritchie, M. E., Phipson, B., Wu, D., Hu, Y., Law, C. W., Shi, W., & Smyth, G. K. (2015). Limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Research*, *43*(7), e47. https://doi.org/10.1093/nar/gkv007

Sharanek, A., Burban, A., Laaper, M., Heckel, E., Joyal, J.-S., Soleimani, V. D., & Jahani-Asl, A. (2020). OSMR controls glioma stem cell respiration and confers resistance of glioblastoma to ionizing radiation. *Nature Communication*, *11*(1), 4116. https://doi.org/10.1038/s41467-020-17885-z

Stark, R., Grzelak, M., & Hadfield, J. (2019). RNA sequencing: The teenage years. *Nature Reviews Genetics*, *20*(11), 631–656. https://doi.org/10.1038/s41576-019-0150-2

Sun, H.-T., Cheng, S.-X., Tu, Y., Li, X.-H., & Zhang, S. (2013). FoxQ1 promotes glioma cells proliferation and migration by regulating NRXN3 expression. *PLoS One*, *8*(1), e55693. https://doi.org/10.1371/journal.pone.0055693

Tanaka, K., Sasayama, T., Nagashima, H., Irino, Y., Takahashi, M., Izumi, Y., Uno, T., Satoh, N., Kitta, A., Kyotani, K., Fujita, Y., Hashiguchi, M., Nakai, T., Kohta, M., Uozumi, Y., Shinohara, M., Hosoda, K., Bamba, T., & Kohmura, E. (2021). Glioma cells require one-carbon metabolism to survive glutamine starvation. *Acta Neuropathologica Communications*, *9*(1), 16. https://doi.org/10.1186/s40478-020-01114-1

Vescovi, A L., Galli, R., & Reynolds, B A. (2006). Brain tumour stem cells. *Nature Reviews Cancer*, *6*(6), 425–436. https://doi.org/10.1038/nrc1889

Wang, L., Yan, Z., He, X., Zhang, C., Yu, H., & Lu, Q. (2019). A 5-gene prognostic nomogram predicting survival probability of glioblastoma patients. *Brain and Behavior*, *9*(4), e01258. https://doi.org/10.1002/brb3.1258

Xu, Y., Wu, G., Li, J., Li, J., Ruan, N., Ma, L., Han, X., Wei, Y., Li, L., Zhang, H., Chen, Y., & Xia, Q. (2020). Screening and identification of key biomarkers for bladder cancer: A study based on TCGA and GEO data. *BioMed Research International*, *1–20*, 8283401. https://doi.org/10.1155/2020/8283401

Zhang, Y., Yang, X., Zhu, X.-L., Hao, J.-Q., Bai, H., Xiao, Y.-C., Wang, Z.-Z., Hao, C.-Y., & Duan, H.-B. (2020). Bioinformatics analysis of potential core genes for glioblastoma. *Bioscience Reports*, *40*(7), BSR20201625. https://doi.org/10.1042/BSR20201625

Zhao, B., Wang, Y., Wang, Y., Chen, W., Liu, P. H., Kong, Z., Dai, C., Wang, Y., & Ma, W. (2021). Systematic identification, development, and validation of prognostic biomarkers involving the tumor-immune microenvironment for glioblastoma. *Journal of Cellular Physiology*, *236*(1), 507–522. https://doi.org/10.1002/jcp.29878

Zhou, J., Guo, H., Liu, L., Hao, S., Guo, Z., Zhang, F., Gao, Y., Wang, Z., & Zhang, W. (2021). Construction of co-expression modules related to survival by WGCNA and identification of potential prognostic biomarkers in glioblastoma. *Journal of Cellular and Molecular Medicine*, *25*(3), 1633–1644. https://doi.org/10.1111/jcmm.16264

SUPPORTING INFORMATION

Additional supporting information may be found in the online version of the article at the publisher's website.

How to cite this article: Lai, W., Li, D., Kuang, J., Deng, L., & Lu, Q. (2022). Integrated analysis of single-cell RNA-seq dataset and bulk RNA-seq dataset constructs a prognostic model for predicting survival in human glioblastoma. *Brain and Behavior*, *12*, e2575. https://doi.org/10.1002/brb3.2575