Gene Biomarker Prediction in Glioma by Integrating scRNA-seq Data and the Gene Regulatory Network

Guimin Qin
Xidian University
Yuying Ma
Xidian University https://orcid.org/0000-0001-9714-2686
Yuhan Yang
Xidian University
Yu Yin
Xidian University
Xiyang Liu
Xidian University
Liming Wang (wanglm@mail.xidian.edu.cn)
Xidian University https://orcid.org/0000-0001-9596-8054

Technical advance

Keywords: glioma, single-cell gene expression profile, cell type, tumor marker genes

Posted Date: August 13th, 2020

DOI: https://doi.org/10.21203/rs.3.rs-49645/v1

Version of Record: A version of this preprint was published at BMC Medical Genomics on December 1st, 2021. See the published version at https://doi.org/10.1186/s12920-021-01115-6.

Abstract

Background: Although great efforts have been made to study the occurrence and development of glioma, the molecular mechanisms of glioma are still unclear. Single-cell sequencing technology provides a new perspective for researchers to explore the pathogens of tumors to further help make treatment and prognosis decisions for patients with tumors.

Methods: In this study, we proposed an algorithm framework to explore the molecular mechanisms of glioma by integrating single-cell gene expression profiles and gene regulatory relations. First, since there were great differences among malignant cells from different glioma samples, we analyzed the expression status of malignant cells for each sample, and then tumor consensus genes were identified by constructing and analyzing cell-specific networks. Second, to comprehensively analyze the characteristics of glioma, we integrated transcriptional regulatory relationships and consensus genes to construct a tumor-specific regulatory network. Third, we performed a hybrid clustering analysis to identify glioma cell types. Finally, candidate tumor marker genes were identified based on cell types and known glioma-related genes.

Results: We got six identified cell types using the method we proposed and for these cell types, we performed functional and biological pathway enrichment analyses. The candidate tumor marker genes were analyzed through survival analysis and verified using literature from PubMed.

Conclusions: The results showed that these candidate tumor marker genes were closely related to glioma and could provide clues for the diagnosis and prognosis of patients with glioma. In addition, we found that four of the candidate tumor marker genes (*NDUFS5*, *NDUFA1*, *NDUFA13*, and *NDUFB8*) belong to the NADH ubiquinone oxidoreductase subunit gene family, so we inferred that this gene family may be strongly related to glioma.

1 Background

Malignant tumors have a very large impact on human health due to their high mortality rate and high recurrence rate. There are many factors that affect tumorigenesis, including genetic variation, epigenetics, and external environmental influences. Glioma is the most common type of brain tissue tumor in complex diseases and accounts for approximately 40% of brain tissue tumors[1]. Therefore, it is of great importance to explore the molecular mechanisms of glioma, as these may help researchers develop glioma treatment strategies and drugs.

In recent years, many researchers have focused on the molecular mechanisms of glioma. Hu et al. constructed a coexpression network by calculating the differentially expressed genes (DEGs) between 971 glioma samples and 102 normal samples, and functional and pathway enrichment analyses indicated that the p53 signaling pathway and the pathway of neuroactive ligand-receptor interaction may play important roles in the progression of glioma, and three genes (*PUS7*, *EFR3B* and *NRCAM*) were potential biological agent landmarks[2]. Niu et al. used protein-protein interaction networks to screen key

DEGs and then applied machine learning methods to reveal the molecular mechanisms of glioma[3]. Wang et al. integrated gene interaction information into a weighted random survival forest method to perform an accurate survival prediction and to discover a survival biomarker for glioma[4]. Zhou et al. identified the glioma-specific protein interaction network based on bulk RNA-seq data and performed enrichment analysis to verify disease-specific molecular complexes[5]. Due to the complexity of glioma, more genetic markers need to be discovered.

Recent advances in microfluidic technology have made it possible to isolate a large number of cells, and single-cell sequencing (scRNA-seq) data analysis has become one of the most noteworthy technical fields in bioinformatics[6–8]. The resolution of scRNA-seq technology is accurate to a single cell, can resolve more subtle differences among cells and is widely used in biology, including development[9, 10], infectious diseases[11, 12], immunity[13, 14], neurology[15] and oncology[16–20]. Cell type identification and/or rare cell type prediction based on scRNA-seq data can deepen the understanding of tumors and analyze the process of tumor occurrence[21]. At present, many methods have been proposed to identify cell types. For example, Kiselev et al. proposed a method for consistent clustering of single cells[22]. Wang et al. proposed Single-cell Interpretation via Multi-kernel LeaRning (SIMLR), which is based on single-cell data and multicore learning similarity measures. They used downscaling and clustering to analyze cell types[23]. Kim et al. implemented a semi-supervised learning classification tool, scReClassfy, to fine tune cell type annotations generated using any method in single-cell sequence datasets[24]. Lin et al. adopted an implicit missing value processing method to reduce the impact of dropout values in scRNA-seq data and achieved rapid and accurate cell type identification[25]. Grun et al. designed the RaceID method to identify rare cell types in complex single-cell populations through k-means and outlier detection methods[26]. Most of these methods directly identified cell types based on single-cell gene expression data without integrating multi-omics data.

In addition, gene expression levels are affected by a variety of regulatory factors, and it is also crucial for the treatment and prevention of complex diseases to understand the disturbance of transcriptional regulatory relationships. In terms of regulatory mechanisms, a transcription factor (TF) is a key gene regulatory factor that mainly activates or inhibits gene expression during the transcriptional stage. TFs participate in many important cellular processes, such as cell proliferation and cell differentiation. These cellular processes may affect the development of many complex diseases, including tumors[27]. For example, Zhang et al. reconstructed a multilayer signaling network that contains pathways from intercellular ligand-receptor interactions, intracellular TFs and their target genes. In this way, they discovered a new multilayer network biomarker (MNB) that was indicated to be valuable for the prognosis and prediction of glioma patients[28].

To further analyze the molecular mechanisms of glioma, in this study, we identified multiple cell types and candidate tumor marker genes in glioma by integrating scRNA-seq data and transcriptional regulation pairs. Through gene enrichment analysis, survival analysis and PubMed analysis, our results showed that our method has an effective performance and provides clues for the diagnosis and prognosis of patients with glioma.

2 Methods

2.1 Materials

2.1.1 Single-cell gene expression data of glioma

To explore the molecular mechanisms of glioma, we downloaded the single-cell gene expression data with EXP0062 from the CancerSEA database[29]. The data contain a single-cell gene sequencing profile of 4044 tumor malignant cells, in which all malignant cells were derived from six glioma samples, and the tissue source of the samples was oligodendrocytes. The sample IDs are MGH36, MGH53, MGH54, MGH60, MGH93 and MGH97, respectively. The CancerSEA database uses methods such as copy number variation inference on the original single-cell data to ensure that all cells in the data set are tumor malignant cells.

2.1.2 Transcriptional regulation pairs

Gene transcriptional regulation pairs were collected from the HTRIdb[30] and TRRUST[31] databases. For HTRIdb, we collected 51871 regulation pairs, and for TRRUST, we collected 8427 regulation pairs. The regulation pairs were the pairs between TFs and the regulatory targets (TARGETs). TARGETs contain target genes and target TFs. Therefore, we divided the regulation pairs into TF-TF pairs and TF-gene pairs, according to whether a TARGET is a TF or gene. Finally, we obtained 952 TFs, 17600 target genes, 5694 TF-TF pairs and 53408 TF-gene pairs.

2.1.3 Known glioma-related genes

We collected known cancer-related genes from the Online Mendelian Inheritance in Man (OMIM)[32] and the Catalogue Of Somatic Mutations In Cancer (COSMIC)[33] databases. OMIM is an authoritative database focusing on the relationship between disease phenotypes and genotypes and contains cancer-related genes with high confidence. COSMIC is a comprehensive somatic mutation database that contains thousands of somatic mutation information related to cancer development. In addition, we obtained known cancer-related genes from Bailey's research results[34]. This research uses 26 different bioinformatics tools to analyze somatic mutations in a variety of cancers and provides services for cancer research. In total, we obtained 77 KGGs.

2.1.4 Bulk RNA-seq of gene expression data and clinical data from glioma

Bulk RNA-seq of gene expression data and clinical data from glioma were obtained from The Cancer Genome Atlas (TCGA)[35]. The clinical data contained overall survival (OS) data. To analyze the data more effectively, we retained samples that had a tissue type of oligodendrocytes only. In this way, the tissue type of the samples in bulk RNA-seq data was consistent with the tissue type of the samples in the single-cell gene expression data. In the end, we obtained 198 glioma samples. Then, to improve the data quality, we deleted genes whose expression values were less than 1 in more than half of the samples.

Finally, to mitigate the influence of different samples on the expression level and avoid the influence of overcapturing features with extreme values and outliers, the z-score was used to normalize the gene expression values in the bulk RNA-seq expression data.

2.2 Preprocessing of single-cell gene expression data

In the single-cell gene expression data, there are significant differences in tumor malignant cells among different patient samples. To comprehensively analyze tumor characteristics, we first explored the expression status of malignant cells in a single sample. Therefore, the original single-cell gene expression data were split according to the sample source to obtain multiple single-sample single-cell gene expression data.

Next, we cleaned the single-sample single-cell gene expression data from the perspective of cells and genes. First, the number of cells and genes were fitted to a normal distribution, and cells with significantly fewer expressed genes were deleted (FDR < 0.05). Then, the genes that had an expression value detected in at least 3 cells and had an average normalized expression value greater than 10^{-5} were retained. To effectively improve the signal-to-noise ratio, the genes affected by technical noise were ignored. We performed the M3Drop feature selection method[36] and obtained the feature genes of single-sample single-cell gene expression data with FDR < 0.01. Finally, we normalized each expression data with a logarithmic function of offset 1.

After preprocessing, we obtained a total of 6 single-sample single-cell gene expression data. In MGH36, there were 694 cells and 4608 feature genes. In MGH53, there were 726 cells and 4126 feature genes. In MGH54, there were 1174 cells and 4732 feature genes. In MGH60, there were 428 cells and 3609 feature genes. In MGH93, there were 440 cells and 3879 feature genes. In MGH97, there were 582 cells and 4113 feature genes.

2.3 Identification of tumor consensus genes

Differences among samples of tumor malignant cells may affect the identification of genetic markers, so we first analyzed each single-sample single-cell gene expression data. First, we performed PCA on each single-sample single-cell gene expression data to determine the appropriate principal components. To prevent excessive capture of certain genes with large values, the z-score was used to normalize gene expression data. In addition, the criteria for determining the number of principal components were as follows: (1) the cumulative contribution rate was greater than 90%, and (2) the difference between two consecutive principal components was less than 0.1%. We used the minimal number in the numbers obtained from condition (1) and condition (2) as the final number of principal components.

Then, we adopted the idea of the k-nearest neighbors to construct a cell-specific network within a single sample. The Euclidean distance was used to calculate the distance between all cell pairs. The k-nearest neighbor relationships were determined for each cell, and the similarity between the two cells was calculated by Jaccard coefficients. Next, Louvain clustering[37] was used to achieve the initial division of cells in a single sample. The results of the initial division helped to analyze the expression status of

malignant cells in a single sample. We constructed a cell-specific network with k as 20 and used the Seurat package[38] to complete the clustering process. According to each cell cluster in the initial division of each sample, the cells were divided into the cells belonging to the cell cluster and the remaining cells. We then performed the Limma package[39] to calculate the DEGs for each sample (lg2 | FC |> 1, p-value < 0.05).

If a gene was a differential gene in multiple samples, the gene reflected the coexpression pattern among samples to some extent. We selected tumor consensus genes by screening the genes that were differentially expressed in at least 30% of the samples.

2.4 Identification of tumor cell types

Each TF in the specific regulatory network was used to form its corresponding regulation meta module (RMM). Each RMM included all target genes and other TFs directly regulated by the core TF. Then, based on the entire single-cell gene expression data that contained all cells from different samples, the RMM was regarded as a new feature of malignant cells to construct a specific regulation expression matrix, in which the feature value was calculated by the Cell Score Method[40]. Then, hybrid clustering was used to identify the glioma cell types based on the matrix. The canopy clustering algorithm[41] was first performed on all malignant cells to provide the k value and initial clustering center for k-means clustering. Then, k-means clustering was used to identify cell types. The Calinski-Harabaz (CH) coefficient was used to tune the parameters of the hybrid clustering, and the larger the CH value was, the better the clustering results.

In the identified cell types, we combined the entire single-cell gene expression data after M3Drop feature selection and divided all cells into two groups for each cell type: the cells that belonged to the cell type and the remaining cells. The ROTS method[42] was performed to obtain the marker genes for each cell type.

2.5 Gene set enrichment analysis

To further analyze the functional characteristics of cell types, GO functional enrichment and pathway enrichment analyses were conducted for marker genes from these cell types. We applied the Metascape tool[43] for enrichment analysis, which mainly provided five forms of gene annotations, including GO biological processes, Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway, Reactome pathway database, Reactome canonical pathways and CORUM.

3 Results

3.1 Overview of the computational framework

We proposed a computational framework, which consisted of four steps (Fig. 1), to gain insight into the molecular mechanisms of glioma.

Step 1. Preprocessing of single-cell gene expression data. The original single-cell gene expression data were split according to the sample ID. Then, we preprocessed the gene expression data through data cleaning, feature selection and standardization.

Step 2. Identification of tumor consensus genes. For each single-sample single-cell gene expression data, we explored the gene expression patterns of all malignant cells through principal component analysis (PCA), cell-specific network construction, and differential gene identification. Then, based on the overlapping degree of the differential genes among samples, tumor consensus genes were identified.

Step 3. Identification of tumor cell types. We combined the gene expression profiles of each sample and integrated transcriptional regulatory pairs. As a result, a specific regulatory network was built based on tumor consensus genes and feed forward loops (FFLs). Finally, the single-cell specific regulatory expression matrix was constructed, and a hybrid clustering method was used to obtain the cell types of glioma.

Step 4. Identification of candidate tumor marker genes. The marker genes of the cell types were regarded as candidate genes, and then the tumor eigenvector was calculated by known glioma-related genes (KGGs). Finally, the tumor marker genes were identified according to the degree of correlation between the candidate genes and the tumor eigenvector.

3.2 Differential analysis of single-cell gene expression data among samples

We performed data cleaning and feature selection on the single-cell gene expression data, and then t-distributed stochastic neighbor embedding (TSNE)[44] was used to visualize the malignant cell clusters.

Each point in Fig. 2 represents a cell, and each color represents a tumor sample. All of the tumor malignant cells were clustered according to their tumor sample source, and there were almost no mixing results of multiple tumor sample cells, which was consistent with previous studies[40, 45, 46]. Thus, there were considerably significant differences in malignant cells among samples, which inspired us to conduct our analysis at the sample level.

3.3 Analysis of consensus genes

For each sample, the k-nearest neighbors and Jaccard coefficient were used to construct a cell-specific network, then Louvain clustering[37] was used to obtain the initial division of all cells, and the differentially expressed genes (DEGs) were calculated (Table 1). Finally, consensus genes were identified based on the overlapping degree of differential genes among different samples. In this paper, a total of 1123 tumor consensus genes were conservatively conserved by screening differential genes that were present in at least two samples.

Table 1
Cell-specific network of single-sample

Sample ID	Cell-specific network		Initial division	Num. of DEGs
	Num. of nodes	Num. of edges		
MGH36	694	20452	6	1206
MGH53	726	21324	8	661
MGH54	1174	35915	8	924
MGH60	428	15271	5	545
MGH93	440	12695	5	443
MGH97	582	19057	6	607

To show that the overlapping degree of DEGs could describe the co-expression patterns of genes among different samples, we counted the overlapping degree of the DEGs in a certain sample and the other samples (Fig. 3). The x-axis represents the sample source of all malignant cells. For each sample, the DEGs that overlapped with the other 5 samples were calculated, respectively, and the y-axis represents the proportion of the overlapping DEGs in all the DEGs of the sample. Figure 3 shows that the highest percentage of overlap was 75%, the lowest percentage of overlap was 28%, and more than half of the overlap percentages were from 30–50%. The analysis results showed that the DEGs in different samples had a high degree of consistency, which further showed that the consensus genes reflected the common gene expression patterns in different samples.

3.4 Specific regulatory network analysis

TF-TF pairs and TF-gene pairs were obtained from the Human Transcriptional Regulation Interactions database (HTRIdb)[30] and Transcriptional Regulatory Relationships Unraveled by Sentence-based Text mining (TRRUST)[31], and we constricted the target genes as tumor consensus genes. To improve the specificity of the regulatory network, the entire single-cell gene expression data containing all cells from all samples were used to adjust the network links. We first reassigned the missing values in the entire single-cell gene expression data using the method proposed by Venteicher et al.[47]. The new value of $E_{i,j}$ was proportional to the expected expression of gene *i* in cell *j*, which was calculated by the average expression of gene *i* and the complexity of cell *j* (the number of detected genes). Then, we calculated the Euclidean distance for each regulation pair based on the entire expression data, and the maximum and minimum normalization was used to shrink the range of distance. We then calculated the similarity of the regulation pairs based on the 1-distance and filtered the pairs whose similarity was less than 0.6. In addition, a feed forward loop (FFL) is an important building block of regulatory mechanisms and is related to the development of tumors, in which one TF *M* regulates another TF *N*, and *M* and *N* jointly regulate their target gene *G*. Therefore, we identified FFLs in the regulatory network to construct the final specific regulatory network.

Ultimately, the specific regulatory network consisted of 121 TFs, 439 target genes and 2081 regulatory pairs. There were two categories of edges in the specific regulatory network: 394 TF-TF pairs and 1687 TF-target gene pairs. Each edge that corresponded to two nodes in the network had a tumor-specific regulation relationship, and the similarity of the edge represented the degree of regulation between the two nodes.

ETS1 was the node with the highest degree in the specific regulatory network. *ETS1* is a protein-coding TF that can act as an activator or inhibitor of multiple genes in a variety of different cellular environments. Moreover, annotations of Gene Ontology (GO) related to *ETS1* indicate that the gene participates in various biological functions, such as cell senescence, apoptosis, and cell development, and plays an important role in the occurrence of diseases. *ETS1* upregulates the expression of the integrin α5 subunit and mediates intracellular signal transduction and invasion processes, leading to the occurrence of malignant glioma[48].

3.5 Cell type identification

We identified 121 regulation meta modules (RMMs) in the specific regulatory network, and then the RMMs were considered as single-cell features to obtain the specific regulation expression matrix. Next, we used hybrid clustering to identify the cell types, and reproducibility-optimized test statistic (ROTS) method[42] was used to identify the marker genes of different cell types. The process of cell type identification fully considered the differences among tumor samples from malignant cells and the effect of transcriptional regulatory mechanisms on gene expression profiles. The resulting 6 cell types identified are shown in Table 2 and were named cell types A to F.

Table 2
Results of cell types identification of glioma

cell type	A	B	C	D	E	F
Num. of cells	136	983	186	214	238	2287
Num. of markers genes in cell type	229	29	11	388	4	49

To further analyze the functional and biological significance of different cell types, we performed enrichment analysis for marker genes in each cell type. Enrichment analysis was used as the priori knowledge such as gene annotation to classify a group of genes, and the classification results could help explore whether these genes had certain functions in common and understand the role of genes in life activities. In this study, the Metascape tool[43] was used for analysis.

Figures 4 and 5 show the enrichment analysis results for cell types A and D, respectively. The more depth the color of the bar is, the greater the enrichment of the gene. For cell type A, the genes were mainly enriched in biological functions such as glial cell differentiation, the ERBB4 signaling pathway, and nervous system development. Among them, ERBB4 belongs to the ERBB receptor family and plays an important role in the development of the nervous system, and the ERBB growth factor receptor is

considered to be a key signaling pathway for many human tumors, including glioma[49]. For cell type D, the genes were mainly enriched in a number of biological functions related to cellular respiration, including aerobic respiration and the negative regulation of respiration involving inflammation. Hypoxia could lead to increased aggressiveness of tumors, and tumor growth, metastasis and resistance to drug treatment greatly improved in the hypoxic microenvironment. There was also some evidence that the hypoxic response plays a key role in the behavior of glioma cells, which is very important for personalized treatment of patients with glioma[50].

The four most enriched entries of cell types B, C and F are shown in Table 3. Table 3 shows that cell type B was mainly involved in a variety of cellular metabolic activities; cell type C was mainly related to apoptosis, inhibition of cell growth and other biological functions; and cell type F was associated with biological functions related to multiple ribosomal proteins. Cell metabolism, apoptosis, and disturbance of ribosomal proteins could cause many complex diseases, including cancer. In addition, since only 4 marker genes were found in this cell type, there were no related enrichment items. Howerer, two of these genes are known cancer-related genes, indicating that cell type E may also be related to the development of glioma.

Table 3
Enrichment analysis results of gene sets in cell type B, C, and F

cell type	enriched item	function	log10(p)
cell type B	R-HSA-71291	Metabolism of amino acids and derivatives	-5.3
cell type B	R-HSA-69206	G1/S Transition	-3.3
cell type B	GO:0010565	regulation of cellular ketone metabolic process	-2.9
cell type B	GO:0032787	monocarboxylic acid metabolic process	-2.1
cell type C	GO:0072331	signal transduction by p53 class mediator	-3.7
cell type C	GO:0097193	intrinsic apoptotic signaling pathway	-3.6
cell type C	GO:0071363	cellular response to growth factor stimulus	-2.5
cell type C	GO:0080135	regulation of cellular response to stress	-2.4
cell type F	R-HSA-72689	Formation of a pool of free 40S subunits	-35.0
cell type F	R-HSA-72695	Formation of the ternary complex, and subsequently, the 43S complex	-15.1
cell type F	CORUM:5380	TRBP containing complex (DICER, RPL7A, EIF6, MOV10 and subunits of the 60S ribosomal particle)	-8.8
cell type F	GO:0042255	ribosome assembly	-5.1

Metascape enrichment analysis indicated that each cell type had unique functionality, and the marker genes may be closely related to the occurrence and treatment of glioma.

3.6 Candidate tumor marker gene analysis of glioma

Assuming that KGGs are specifically expressed in glioma, we used the first principal component method to calculate the tumor feature vector (TEV) for all KGGs based on bulk RNA-seq gene expression data.

TEV was a linear combination of all KGG expression vectors, which could represent the expression level of all KGGs. In addition, marker genes of cell types reflected the biological function of glioma and were likely to be the causative molecules of glioma. Therefore, we took all of these genes as candidate genes and calculated the Pearson correlation coefficient (PCC) between the candidate gene and TEV. The greater the absolute PCC value is, the stronger the relationship between the candidate gene and glioma. The absolute PCC value was used as the correlation between the candidate gene and TEV. We analyzed the top 20 genes in detail and defined them as candidate tumor marker genes of glioma (Table 4). Statistical results showed that the correlations between two candidate tumor marker genes (*ATP6V0B* and *GUK1*) were extremely strong, with correlations greater than 0.8. The correlations of the remaining 18 candidate tumor marker genes were strong (between 0.6 and 0.8).

Table 4
Candidate tumor marker genes of glioma

Ranking	Candidate tumor marker genes	Correlation	PubMed ID
1	ATP6V0B	0.825678	-
2	GUK1	0.816174	11156382
3	MRPL20	0.774609	-
4	RAB30	0.718324	24080485
5	NDUFS5	0.712145	31747975
6	TMEM160	0.710739	-
7	ZNF195	0.70465	-
8	DDRGK1	0.69812	-
9	ARF5	0.691213	-
10	MRPL41	0.690842	28351326
11	NDUFA1	0.685764	29211022
12	GOLIM4	0.683062	-
13	COX5B	0.682042	29180880
14	NDUFA13	0.681426	31747975
15	NDUFB8	0.675849	29928884
16	NCOA4	0.673071	-
17	ARL2	0.659971	29843637
18	SDHB	0.645852	29890994
19	ROGDI	0.645402	-
20	BMPR1A	0.643674	26683138
-: No supported publiccations were found in PubMed			

Additionally, 11 out of the 20 candidate tumor marker genes were confirmed by relevant medical literature that they had a direct or indirect relationship with glioma (Table 4), which showed that the identified candidate tumor marker genes were reliable to a certain extent. In addition, we found that four (*NDUFS5*, *NDUFA1*, *NDUFA13*, and *NDUFB8*) of the candidate tumor marker genes belong to the NADH ubiquinone oxidoreductase subunit gene family. This gene family plays a key role in the transfer of NADH to the respiratory chain. NADH is the reduced state of nicotinamide adenine dinucleotide and is mainly involved in the metabolism of matter and energy in cells, which plays a key role in maintaining cell growth and

differentiation. The studies of Yuan et al.[51] and Trinh et al.[52] showed that NADH is regarded as a new marker to classify glioma cancer cells. Therefore, we inferred that the NADH ubiquinone oxidoreductase subunit gene family may be closely related to glioma.

3.7 Survival analysis of candidate tumor marker genes of glioma

To further explore the effect of the expression level of candidate tumor marker genes on the prognosis of gliomas, overall survival (OS) data were used for survival analysis. Specifically, we used specific survival time for the Kaplan-Meier (KM) survival curve analysis. Figure 6 shows the results of the KM survival analysis of the most highly correlated tumor marker gene (*ATP6V0B*). Glioma samples were divided into a high expression group (expression_level = 1) and a low expression group (expression_level = 0), according to the median expression value of the candidate tumor marker gene in bulk RNA-seq profiles. The red curve and the blue curve represent the survival curves of the low and high expression sample groups, respectively. Analysis of the downward trend of the KM curve in Fig. 6 revealed that the interval between the survival curves of the high and low expression samples of ATP6V0B was quite obvious, and the samples with high expression exhibited the worse prognosis.

For each candidate tumor marker gene, we divided samples into a high expression group and a low expression group and analyzed the downward trend of the KM survival analysis between the two curves. The interval between curves of the two groups clearly indicated that the gene expression level affected the survival of patients with glioma. The survival curve of the high expression sample group was below that of the low expression sample group, indicating that the patients with high gene expression had a worse prognosis, whereas the patients with low gene expression of the gene had a worse prognosis. Table 5 summarizes the KM survival analysis results of candidate tumor marker genes. There were 5 genes (*ATP6V0B*, *MRPL20*, *NDUFS5*, *DDRGK1*, and *SDHB*) with high expression levels, and the prognosis of the patient was worse. In addition, there were 5 genes (*GUK1*, *ZNF195*, *MRPL41*, *NDUFA13*, and *BMPR1A*) in which the expression level was low, and the prognosis of the patient was worse.

Table 5
KM survival analysis results of candidate tumor marker genes

Expression level	Candiate tumor marker genes
High expression	*ATP6V0B MRPL20 NDUFS5 DDRGK1 SDHB*
Low expression	*GUK1 ZNF195 MRPL41 NDUFA13 BMPR1A*

The experimental results of survival analysis showed that the identified candidate tumor marker genes were reliable and had a strong correlation with glioma, and these genes could provide clues for the diagnosis and treatment of patients with gliomas and further help to understand the molecular mechanisms of glioma.

4 Discussion

Due to the batch effect or other factors, the individual differences in malignant cells among tumor samples are strong. The identification of cell types can essentially be regarded as an unsupervised clustering process. If cluster analysis of malignant cells based on single-cell expression profiles is performed directly, malignant cells from the same individual will often cluster together, and the clustering results may not reflect the tumor cell types. However, there will be consistency among different samples of the tumor. To identify the genes that were coexpressed in the tumor and the genes that expressed heterogeneity in different tumor samples, we analyzed the cell-specific network from the single-sample level to identify tumor consensus genes and then combined the transcriptional regulatory pairs with multisample cell expression data to identify glioma cell types. Since the marker genes of the cell types have a strong correlation with glioma, we used the correlation assessment method to predict the candidate tumor marker genes that are highly related to glioma. From the analysis results, we concluded that the identified candidate tumor marker genes had a strong correlation with glioma. The research results may be helpful for the diagnosis and treatment of patients with glioma, but these predicted candidate tumor marker genes should be verified by further biological experiments. Of course, there are some problems in our study. For example, the results were heavily affected by the input gene expression data and noise in the data. In the future, we will integrate more omics data to perform further analyses, such as DNA methylation, noncoding RNA regulation, and protein interactions.

5 Conclusion

In this study, we have proposed a new framework for identifying candidate tumor marker genes based on single-cell gene expression profiles and transcriptional regulation pairs. The framework mainly contains four steps: preprocessing of single-cell gene expression data, identification of tumor consensus genes, identification of tumor cell types, and identification of candidate tumor marker genes. We have shown the framework's performance by exploring the molecular mechanisms of glioma. For glioma, 6 cell types and 20 candidate tumor marker genes were identified. The Metascape enrichment analysis showed that the cell types had significant functionality, and the analysis of candidate tumor marker genes showed that it had a strong correlation with glioma. In addition, recent relevant studies have also shown that some candidate tumor marker genes were recognized as targets of glioma, and 4 genes (*NDUFS5*, *NDUFA1*, *NDUFA13*, and *NDUFB8*) of the candidate tumor marker genes belonged to the NADH ubiquinone oxidoreductase subunit gene family, indicating that this gene family may have a strong correlation with glioma. These findings contributed to the clinical diagnosis, therapeutic drug development, and pathological mechanisms of glioma.

Abbreviations

KGG
known glioma-related gene
TF
transcription factor

FFL
feed forward loop
RMM
regulation meta module
PCC
Pearson correlation coefficient
OS
overall survival

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Availability of data and materials

The datasets generated during and/or analysed during the current study are available as follow: Single-cell gene expression data of glioma is available in CancerSEA, [https://doi.org/10.1093/nar/gky939][29]. Transcriptional regulation pairs are available in HTRIdb, [https://doi.org/10.1186/1471-2164-13-405][30], and TRRUST, [https://doi.org/10.1093/nar/gkx1013][31]. Known cancer-related genes are available in OMIM, [https://doi.org/10.1093/nar/30.1.52][32], COSMIC, [https://doi.org/10.1093/nar/gku1075][33], and Bailey's research, [https://doi.org/10.1016/j.cell.2018.02.060][34]. Bulk RNA-seq and clinical data from glioma is available in TCGA, [https://doi.org/10.1038/ng.2764][35].

Competing interests

The authors declare that they have no competing interests.

Funding

This study was founded by the Natural Science Foundation of Shaanxi Province (No. 2017JM6038) , the National Key Research and Development Program of China (No.2018YFC0116500), the Natural Science Basic Research Program of Shaanxi (Program No.2018JQ6047) and the Fundamental Research Funds for the Central Universities (Program No.JB181004).

Authors' contributions

GQ, YM, and LW designed the analysis pipeline. YM and GQ preprocessed the data, and performed the analysis to generate presented results. GQ, YM, YH, YY, LW, and XL prepared, revised and discussed the manuscript. All authors read and approved the final version of the manuscript.

Acknowledgement

Not applicable.

References

1. Wakabayashi, TJRsCn. Clinical trial updates for malignant brain tumors. 2011, 51(11):853–856.
2. Hu G, Wang R, Wei B, Wang L, Yang Q, Kong D, Du C. Prognostic Markers Identification in Glioma by Gene Expression Profile Analysis. J Comput Biol. 2019;27(1):81–90.
3. Niu B, Liang C, Lu Y, Zhao M, Chen Q, Zhang Y, Zheng L, Chou K-C. Glioma stages prediction based on machine learning algorithm combined with protein-protein interaction networks. Genomics. 2020;112(1):837–47.
4. Wang W, Liu W. Integration of gene interaction information into a reweighted random survival forest approach for accurate survival prediction and survival biomarker discovery. Sci Rep. 2018;8(1):13202–2.
5. Zhou C, Teng WJ, Zhuang J, Liu HL, Tang SF, Cao XJ, Qin BN, Wang CC, Sun CG. Analysis of the gene-protein interaction network in glioma. Genet Mol Res. 2015;14(4):14196–206.
6. De Souza N. Single-cell methods. Nat Methods. 2012;9(1):35.
7. Pennisi E. Single-cell sequencing tackles basic and biomedical questions. Science. 2012;336(6084):976–7.
8. Chi KR. Singled out for sequencing. Nat Methods. 2014;11(1):13–7.
9. Xue Z, Huang K, Cai C, Cai L, Jiang C-y, Feng Y, Liu Z, Zeng Q, Cheng L, Sun YE, et al. Genetic programs in human and mouse early embryos revealed by single-cell RNA sequencing. Nature. 2013;500(7464):593–7.
10. Hu Y, Hase T, Li HP, Prabhakar S, Kitano H, Ng SK, Ghosh S, Wee LJK. A machine learning approach for the identification of key markers involved in brain development from single-cell transcriptomic data. BMC Genom. 2016;17(13):1025.
11. Avraham R, Haseley N, Brown D, Penaranda C, Jijon Humberto B, Trombetta John J, Satija R, Shalek Alex K, Xavier Ramnik J, Regev A, et al. Pathogen cell-to-cell variability drives heterogeneity in host immune responses. Cell. 2015;162(6):1309–21.
12. Bossel Ben-Moshe N, Hen-Avivi S, Levitin N, Yehezkel D, Oosting M, Joosten LAB, Netea MG, Avraham R. Predicting bacterial infection outcomes using single cell RNA-sequencing analysis of human

immune cells. Nat Commun. 2019;10(1):3266.

13. Stephenson W, Donlin LT, Butler A, Rozo C, Bracken B, Rashidfarrokhi A, Goodman SM, Ivashkiv LB, Bykerk VP, Orange DE, et al. Single-cell RNA-seq of rheumatoid arthritis synovial tissue using low-cost microfluidic instrumentation. Nat Commun. 2018;9(1):791.
14. Samir J, Rizzetto S, Gupta M, Luciani F. Exploring and analysing single cell multi-omics data with VDJView. BMC Med Genomics. 2020;13(1):29.
15. Raj B, Wagner DE, McKenna A, Pandey S, Klein AM, Shendure J, Gagnon JA, Schier AF. Simultaneous single-cell profiling of lineages and cell types in the vertebrate brain. Nat Biotechnol. 2018;36(5):442–50.
16. Olmos D, Arkenau HT, Ang JE, Ledaki I, Attard G, Carden CP, Reid AHM, A'Hern R, Fong PC, Oomen NB, et al. Circulating tumour cell (CTC) counts as intermediate end points in castration-resistant prostate cancer (CRPC): a single-centre experience. Ann Oncol. 2008;20(1):27–33.
17. Levitin HM, Yuan J, Sims PA. Single-Cell Transcriptomic Analysis of Tumor Heterogeneity. Trends in Cancer. 2018;4(4):264–8.
18. Jerby-Arnon L, Shah P, Cuoco MS, Rodman C, Su M-J, Melms JC, Leeson R, Kanodia A, Mei S, Lin J-R, et al. A cancer cell program promotes T cell exclusion and resistance to checkpoint blockade. Cell. 2018;175(4):984–97.
19. Subramanian A, Schwartz R. Reference-free inference of tumor phylogenies from single-cell sequencing data. BMC Genom. 2016;17(1):348.
20. Gan YL, Li N, Zou GB, Xin YC, Guan JH. Identification of cancer subtypes from single-cell RNA-seq data using a consensus clustering method. *BMC Med Genomics* 2018, 11.
21. Baslan T, Hicks J. Unravelling biology and shifting paradigms in cancer with single-cell sequencing. Nat Rev Cancer. 2017;17(9):557–69.
22. Kiselev VY, Kirschner K, Schaub MT, Andrews T, Yiu A, Chandra T, Natarajan KN, Reik W, Barahona M, Green AR, et al. SC3: consensus clustering of single-cell RNA-seq data. Nat Methods. 2017;14(5):483–6.
23. Wang B, Zhu J, Pierson E, Ramazzotti D, Batzoglou S. Visualization and analysis of single-cell RNA-seq data by kernel-based similarity learning. Nat Methods. 2017;14(4):414–6.
24. Kim T, Lo K, Geddes TA, Kim HJ, Yang JYH, Yang P. scReClassify: post hoc cell type classification of single-cell rNA-seq data. BMC Genom. 2019;20(9):913.
25. Lin P, Troup M, Ho JWK. CIDR: Ultrafast and accurate clustering through imputation for single-cell RNA-seq data. In: *Genome Biol*. vol. 18; 2017: 59.
26. Grun D, Lyubimova A, Kester L, Wiebrands K, Basak O, Sasaki N, Clevers H, van Oudenaarden A. Single-cell messenger RNA sequencing reveals rare intestinal cell types. Nature. 2015;525(7568):251–5.
27. Hobert O. Gene Regulation by Transcription Factors and MicroRNAs. Science. 2008;319:1785–6.

28. Zhang J, Guan M, Wang Q, Zhang J, Zhou T, Sun X. Single-cell transcriptome-based multilayer network biomarker for predicting prognosis and therapeutic response of gliomas. Brief Bioinform. 2020;21(3):1080–97.
29. Yuan H, Yan M, Zhang G, Liu W, Deng C, Liao G, Xu L, Luo T, Yan H, Long Z, et al. CancerSEA: a cancer single-cell state atlas. Nucleic Acids Res. 2019;47(D1):D900–8. https://doi.org/10.1093/nar/gky939.
30. Bovolenta LA, Acencio ML, Lemke N. HTRIdb: an open-access database for experimentally verified human transcriptional regulation interactions. BMC Genom. 2012;13(1):405. https://doi.org/10.1186/1471-2164-13-405.
31. Han H, Cho J-W, Lee S, Yun A, Kim H, Bae D, Yang S, Kim CY, Lee M, Kim E, et al. TRRUST v2: an expanded reference database of human and mouse transcriptional regulatory interactions. Nucleic Acids Res. 2018;46(D1):D380–6. https://doi.org/10.1093/nar/gkx1013.
32. Hamosh A, Scott AF, Amberger JS, Bocchini CA, McKusick VA. Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. Nucleic Acids Res. 2005;33(Database issue):D514–7. https://doi.org/10.1093/nar/gki033.
33. Forbes SA, Beare D, Gunasekaran P, Leung K, Bindal N, Boutselakis H, Ding M, Bamford S, Cole C, Ward S, et al. COSMIC: exploring the world's knowledge of somatic mutations in human cancer. Nucleic Acids Res. 2014;43(D1):D805–11. https://doi.org/10.1093/nar/gku1075.
34. Bailey MH, Tokheim C, Porta-Pardo E, Sengupta S, Bertrand D, Weerasinghe A, Colaprico A, Wendl MC, Kim J, Reardon B, et al. Comprehensive characterization of cancer driver genes and mutations. Cell. 2018;173(2):371–85. https://doi.org/10.1016/j.cell.2018.02.060.
35. Cancer Genome Atlas Research N. Weinstein JN, Collisson EA, Mills GB, Shaw KRM, Ozenberger BA, Ellrott K, Shmulevich I, Sander C, Stuart JM. The Cancer Genome Atlas Pan-Cancer analysis project. Nat Genet 2013, 45(10):1113–20. https://doi.org/10.1038/ng.2764.
36. Andrews TS, Hemberg M. M3Drop: dropout-based feature selection for scRNASEq. Bioinformatics. 2018;35(16):2865–7.
37. Blondel VD, Guillaume J-L, Lambiotte R, Lefebvre E. Fast unfolding of communities in large networks. Journal of statistical mechanics: theory experiment. 2008;2008(10):P10008.
38. Butler A, Hoffman P, Smibert P, Papalexi E, Satija R. Integrating single-cell transcriptomic data across different conditions, technologies, and species. Nat Biotechnol. 2018;36(5):411–20.
39. Ritchie ME, Phipson B, Wu D, Hu Y, Law CW, Shi W, Smyth GK. Limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Res. 2015;43(7):e47.
40. Puram SV, Tirosh I, Parikh AS, Patel AP, Yizhak K, Gillespie S, Rodman C, Luo CL, Mroz EA, Emerick KS, et al. Single-cell transcriptomic analysis of primary and metastatic tumor ecosystems in head and neck cancer. Cell. 2017;171(7):1611–24.
41. McCallum A, Nigam K, Ungar L: Efficient Clustering of High-Dimensional Data Sets with Application to Reference Matching. In: *Proceeding of the Sixth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining: 03/23 2000*; 2000: 169–178.

42. Suomi T, Seyednasrollah F, Jaakkola MK, Faux T, Elo LL. ROTS: An R package for reproducibility-optimized statistical testing. PLoS Comp Biol. 2017;13(5):e1005562.
43. Zhou Y, Zhou B, Pache L, Chang M, Khodabakhshi AH, Tanaseichuk O, Benner C, Chanda SK. Metascape provides a biologist-oriented resource for the analysis of systems-level datasets. Nature communications. 2019;10(1):1523–3.
44. Van der Maaten L, Hinton G. Visualising data using t-SNE. Journal of Machine Learning Research. 2008;9(Nov):2579–605.
45. Tirosh I, Izar B, Prakadan SM, Wadsworth MH 2nd, Treacy D, Trombetta JJ, Rotem A, Rodman C, Lian C, Murphy G, et al. Dissecting the multicellular ecosystem of metastatic melanoma by single-cell RNA-sEq. Science. 2016;352(6282):189–96.
46. Neftel C, Laffy J, Filbin MG, Hara T, Shore ME, Rahme GJ, Richman AR, Silverbush D, Shaw ML, Hebert CM, et al. An integrative model of cellular states, plasticity, and genetics for glioblastoma. Cell. 2019;178(4):835–49.
47. Venteicher AS, Tirosh I, Hebert C, Yizhak K, Neftel C, Filbin MG, Hovestadt V, Escalante LE, Shaw ML, Rodman C, et al. Decoupling genetics, lineages, and microenvironment in IDH-mutant gliomas by single-cell RNA-sEq. Science. 2017;355(6332):eaai8478.
48. Kita D, Takino T, Nakada M, Takahashi T, Yamashita J, Sato H. Expression of dominant-negative form of Ets-1 suppresses fibronectin-stimulated cell adhesion and migration through down-regulation of integrin α5 expression in U251 glioma cell line. Cancer Res. 2001;61(21):7985.
49. Berezowska S, Schlegel J. Targeting ErbB receptors in high-grade glioma. Curr Pharm Des. 2011;17(23):2468–87.
50. Musah-Eroje A, Watson S. Adaptive changes of glioblastoma cells following exposure to hypoxic (1% oxygen) tumour microenvironment. Int J Mol Sci. 2019;20(9):2091.
51. Yuan Y, Yan Z, Miao J, Cai R, Zhang M, Wang Y, Wang L, Dang W, Wang D, Xiang D, et al. Autofluorescence of NADH is a new biomarker for sorting and characterizing cancer stem cells in human glioma. Stem Cell Res Ther. 2019;10(1):330–0.
52. Trinh AL, Chen H, Chen Y, Hu Y, Li Z, Siegel ER, Linskey ME, Wang PH, Digman MA, Zhou Y-H. Tracking functional tumor cell subpopulations of malignant glioma by phasor fluorescence lifetime imaging microscopy of NADH. Cancers (Basel). 2017;9(12):168.

Figures

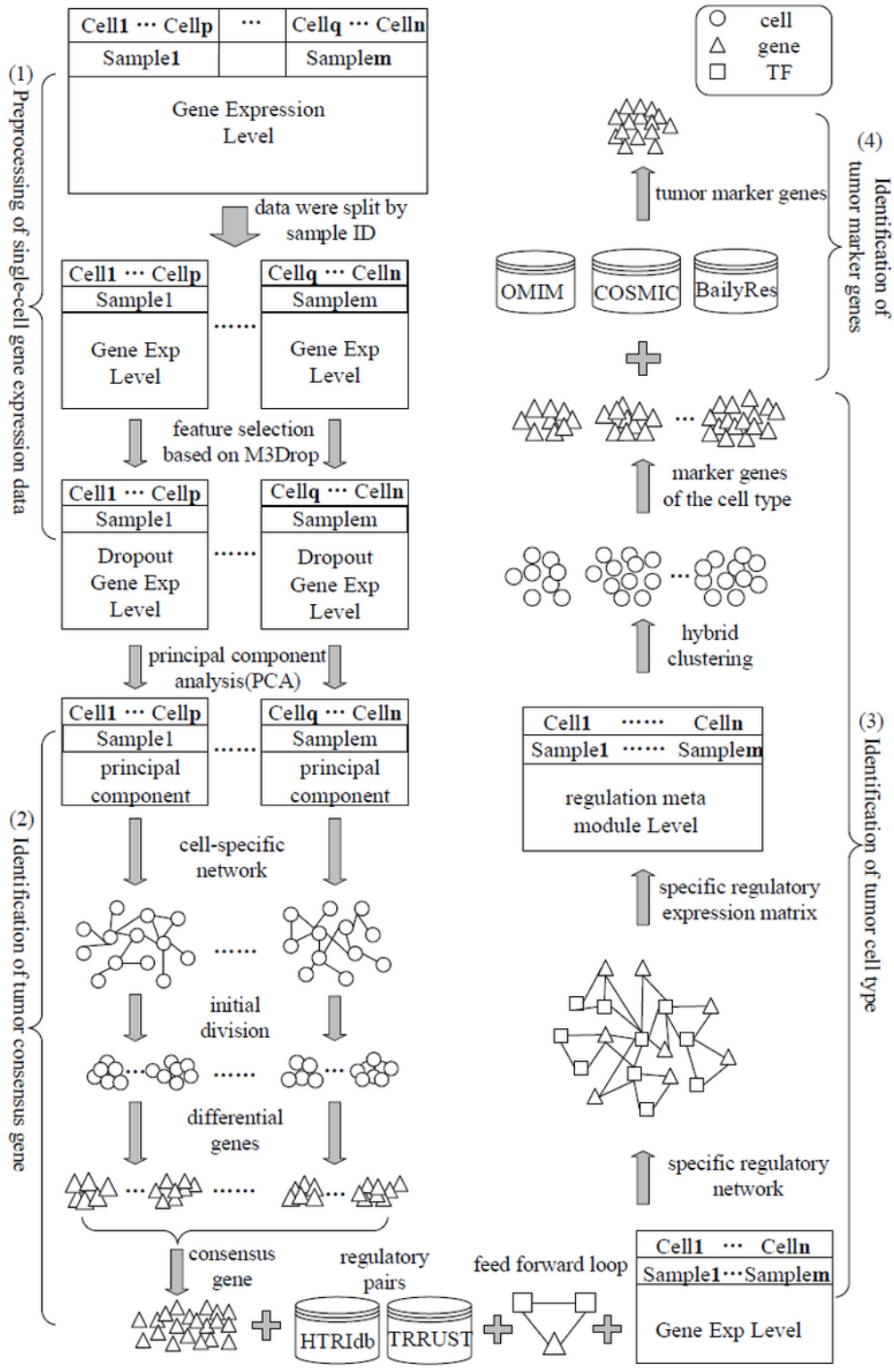

Figure 1

Overview of the computational framework.

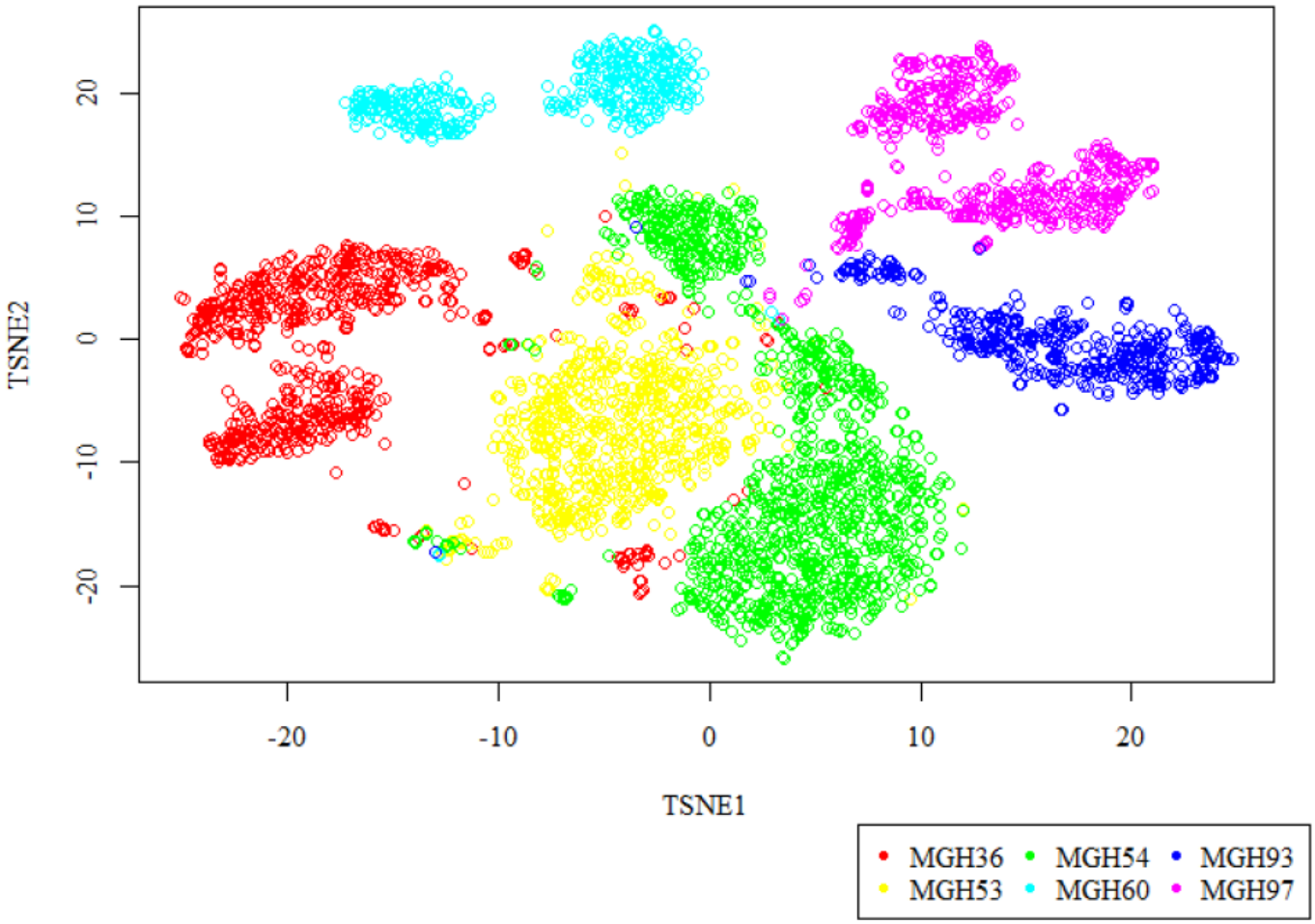

Figure 2

TSNE analysis of the entire single-cell gene expression data.

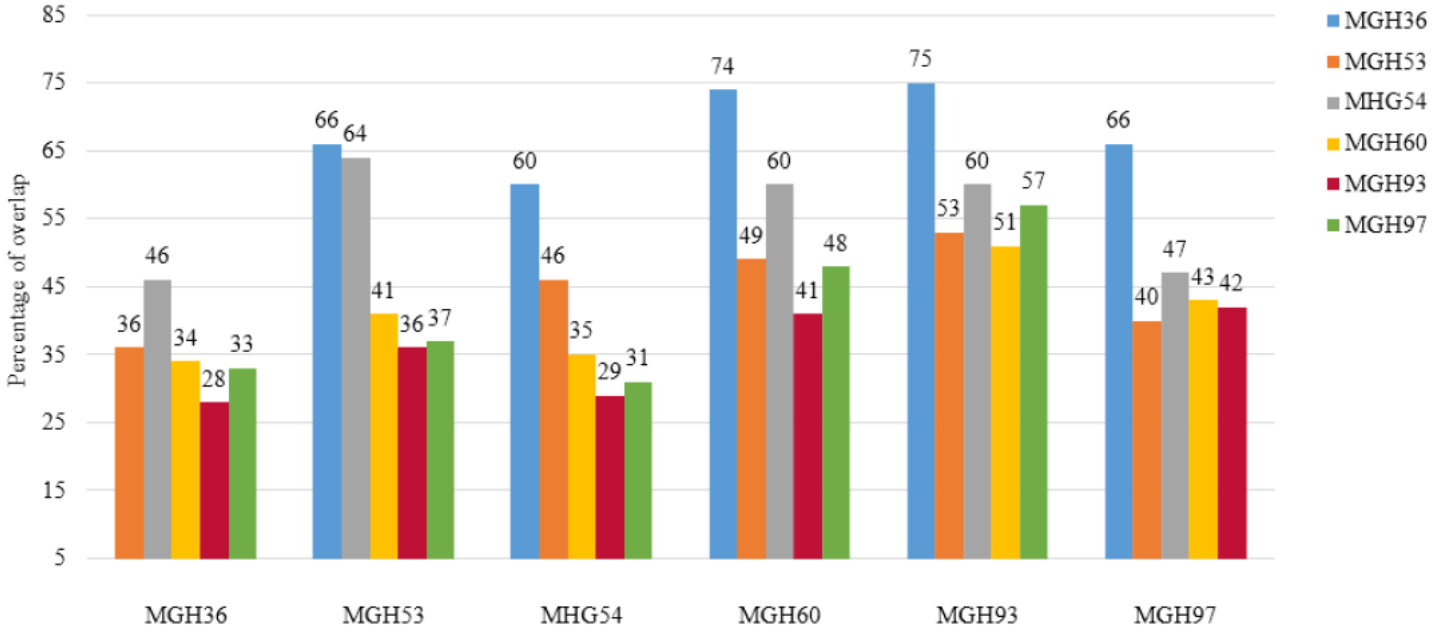

Figure 3

TSNE analysis of the entire single-cell gene expression data.

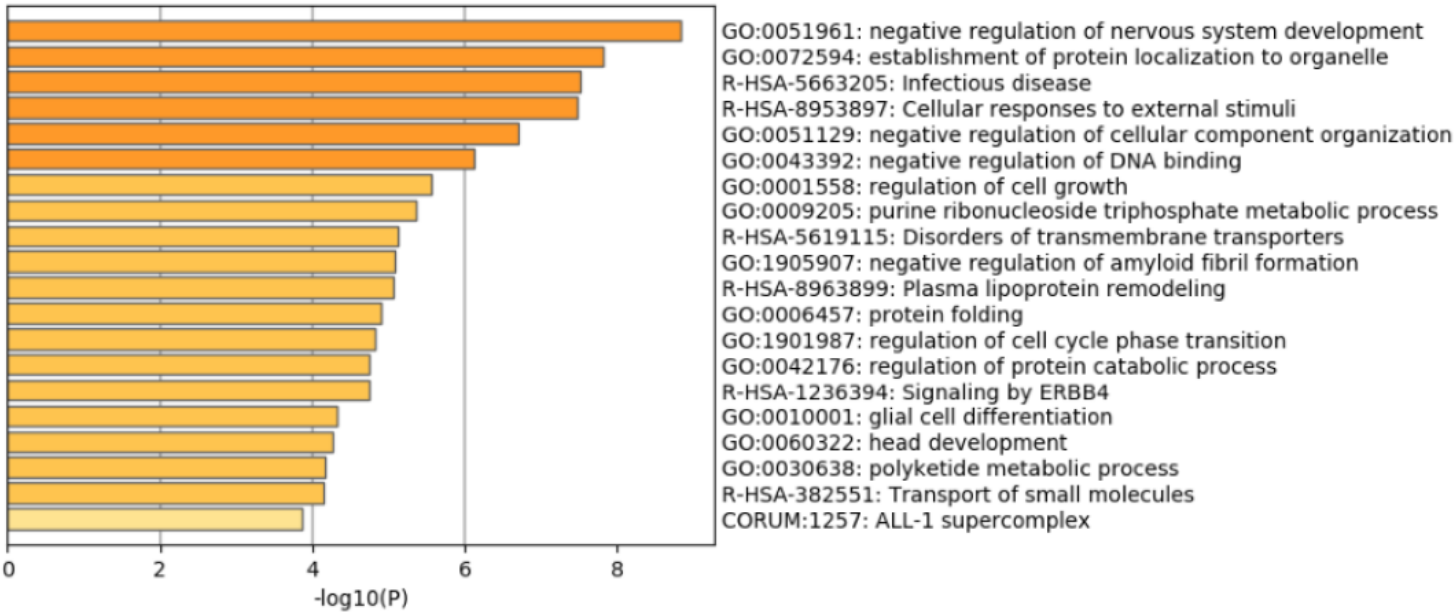

Figure 4

Enrichment analysis results of Metascape of cell type A.

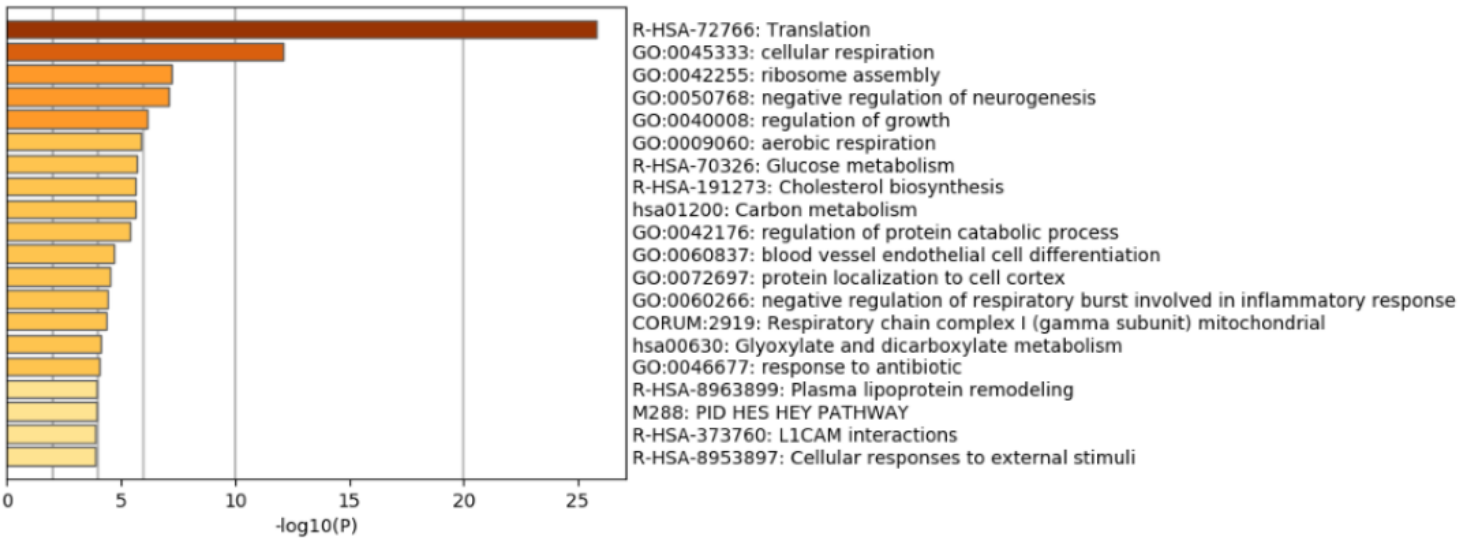

Figure 5

Enrichment analysis results of Metascape of cell type D.

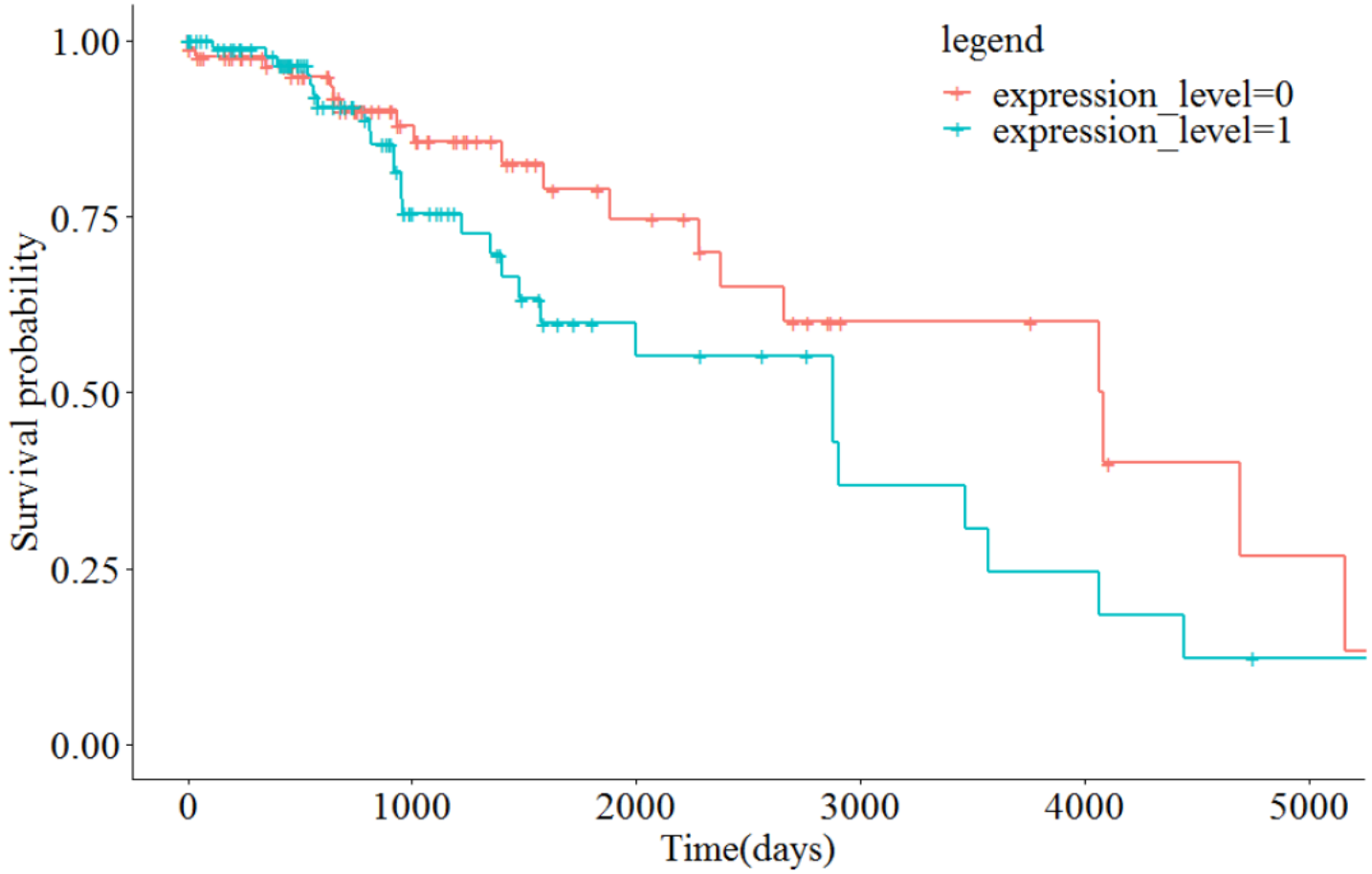

Figure 6

KM survival analysis results of ATP6V0B in OS survival data.