Single-cell Sequencing Data Analysis Reveals Heterogeneity of Glioblastoma Multiforme

Yihan Chen[1 #, *], Jiayin Li[2 #], Wenna Chen[3 #]

[1] Xi'an Jiaotong-Liverpool University, Suzhou, China

[2] Dalian University of Technology, Dalian, China

[3] Wuhan University, Wuhan, China

* Corresponding Author Email: Yihan.Chen22@student.xjtlu.edu.cn

[#]These authors contribute equally to this work and share first authorship.

Abstract. Glioblastoma multiforme is a highly malignant brain tumor. The complex cellular heterogeneity and classification of cell groups have been key factors affecting tumor progression and treatment response. This paper analyzed GBM sequencing data through single-cell RNA sequencing. Firstly, flitering genes and cells according to some specific thresholds. After nomalizing the gene expression matrix, some high-variance genes were selected, and then this paper applied principle component analysis to reduce the dimensions of genes. To identify the cell types, this paper implemented louvain clustering to get 16 clusters, followed by cell annotation. Meanwhile, marker genes were used to find functional pathways and here some conclusions were made about brain cancer research.

Keywords: scRNA seq; PCA; Louvain Clustering; Enrichment Analysis; Glioblastoma Multiforme.

1. Introduction

Glioblastoma multiforme (GBM) is a highly heterogeneous brain tumor and is also a common primary intracranial tumor and one of the most malignant gliomas [1]. Due to invasive cells outside the tumor core are usually not affected by local therapy, GBM grows diffusely and is difficult to cure. The main challenge of treatment failure is its heterogeneity. At present, there is still a lack of research on the single-cell level of GBM and the interaction between neurons, microglia, immune cells, and vascular cells in the tumor microenvironment [2].

Single cell sequencing (scRNA-seq) can accurately determine the gene structure and expression status of a single cell by sequencing the RNA or DNA of a single cell [3]. To some extent, its solves the problem that tissue sequencing (Bulk cell sequencing) cannot reflect the real situation of each cell.

In this paper, scRNA-seq was used for dimension reduction and clustering. GBM cells were clustered by PCA dimensionality reduction clustering, and the cells of each population were annotated by CellMarker database. Finally, through functional enrichment analysis of differentially expressed genes (DEGs), biological pathways related to cancer and immune system regulation will be discovered. These key pathways provide a basis for finding the target for the treatment of GBM and will make a great contribution to the therapy of cancer in the future.

2. Methods

2.1. Data Pre-processing

In this paper, we selected and visualized the following three commonly used QC metrics for screening cells: the number of unique genes detected in each cell, the total number of molecules detected within the cell, and the percentage of reads mapping to the mitochondrial genome [4].

Next, the filtered data was used scaling normalization, dividing all counts in each cell by a cell-specific scale factor, often called the "size factor". The characteristic expression measure for each cell

was normalized by the total expression, multiplied by the scale factor, and the result was logarithmically transformed.

We performed a line transformation ('scaling') on all genes, centering the expression mean of each gene to 0, while expressing the variance as 1 for dominating the variance due to their own relatively large values.

2.2. PCA

We used principal component analysis (PCA) to reduce the dimension of BGM cells. PCA performs orthogonal transformation to convert linear related variables into a series of principal components. It is widely used in 2D or 3D data visualization after clustering of scRNA-seq data. When the number of genes increases, the dimensionality reduction is achieved by extracting the first few principal components that can explain most of the variations.

Here, we performed dimensionality reduction on cells via Seurat in the order of PCA scores. In addition, an "elbow map" was used to rank the principal components according to the percentage of variance explained by each function.

2.3. Louvain Clustering

In this paper, the Louvain algorithm is used for clustering in single cell analysis. Louvain method is an algorithm for detecting communities in large networks. Seurat uses the Louvain algorithm to find cell types on the SNN map of cells [5].

Modularity is a approach to estimate the division of a community network. It can be simply understood as the weight of the edges within the community minus the total of the weights of all the edges to the community nodes. Its definition is as shown below.

$$S = \frac{1}{2n}\sum_{i,j}\left[X_{ij} - \frac{c_i c_j}{2n}\right]\delta(m_i, m_j) \tag{1}$$

X is adjacency matrix, X_{ij} representative the weight of the edge between node i and j. $c_i = \sum_j X_{ij}$ is the total of the weights of all connected edges of node i and node j. $n = \frac{1}{2}\sum_{ij} Xij$ is the sum of the weights of all edges. m_i stands for the community of node i, $\delta(c_i, c_j)$ function demonstrates whether node i and node j are in the identical community.

This algorithm aims to maximize modularity S and it could support the continuous construction of a community construction with interior integration and exterior partially link via strategic optimization.

Module gain is the metric to estimate the result of its iteration, which is an optimization procedure.

$$\Delta S = \left[\frac{\sum_{\text{in}} + c_{i,in}}{2n} - \left(\frac{\sum_{\text{tot}} + c_i}{2n}\right)^2\right] - \left[\frac{\sum_{\text{in}}}{2n} - \left(\frac{\sum_{\text{tot}}}{2n}\right)^2 - \frac{c_i}{2n}\right] = \frac{1}{2m}\left(c_{i,in} - \frac{\sum_{\text{tot}} c_i}{n}\right) \tag{2}$$

$C_{i,in}$ is the total of the weights of node i such as cluster m, $\sum_{tot}$ is the total weight of cluster m, c_i is the total weight of the incoming node i.

Initialization is performed first. Each node is considered as a separate community with the same number of nodes.

Next is the first stage of iteration, that is, node transfer between communities. Each node i is assigned to the community where the adjacent nodes are in proper sequence, and the change in modularity before and after the distribution is calculated. Then repeat this step and continue to evaluate the node transfer between communities until the communities of all nodes are no longer changed, indicating the end of the transfer, which indicates that the local maximization of this iteration has been achieved.

The second phase is reconstructing the graph. In this round of the first phase, a new node i has been appended to community S, which is missing from the old community. Thus, to reconstruct the whole map it should be done. The graph is rebuilt, and all nodes of the community are reconstructed into a new community. The weight of the edge between the nodes in the community is renewed to

new node, and the weight of the edge between the communities is renewed too. Then repeat the previous step and continue to start the next round of two stages until the modularity of the whole graph remains unchanged.

2.4. Cell type annotation

Manual annotation includes knowing about genes and gene functions specific to each cell group to better identity cell annotations. The Cell Marker Database is aimed at providing a detailed and precise resource for cell markers of diverse cell types in human and mouse tissues. Based on cell marker, clusters were labeled by us according to the marker genes they obviously express.

2.5. Functional enrichment analysis of DEGs

To analyze the function of these differentially expressed genes (DEGs), using a biological analysis tool called Metascape online database (http://metascape.org/) provides a comprehensive set of functional annotation information of genes. Terms with a p-value < 0.01, a minimum count of 3, and an enrichment factor > 1.5. Kappa scores4 are used as the similarity metric when performing hierarchical clustering on the enriched terms, and sub-trees with a similarity of > 0.3 are considered a cluster. The most statistically significant term within a cluster is chosen to represent the cluster.

3. Results

3.1. Quality Control

We found that the percentage of reads that map to the mitochondrial genome in a cell was 0, so we did not screen further for this QC metric, therefore we used unique feature counts to screen the cells with unique feature counts greater than 7500 or less than 500 are eliminated. The figure 1 shows the relationship between the two QC indicators the figure 2 shows the visualization of the QC indicators after the screening.

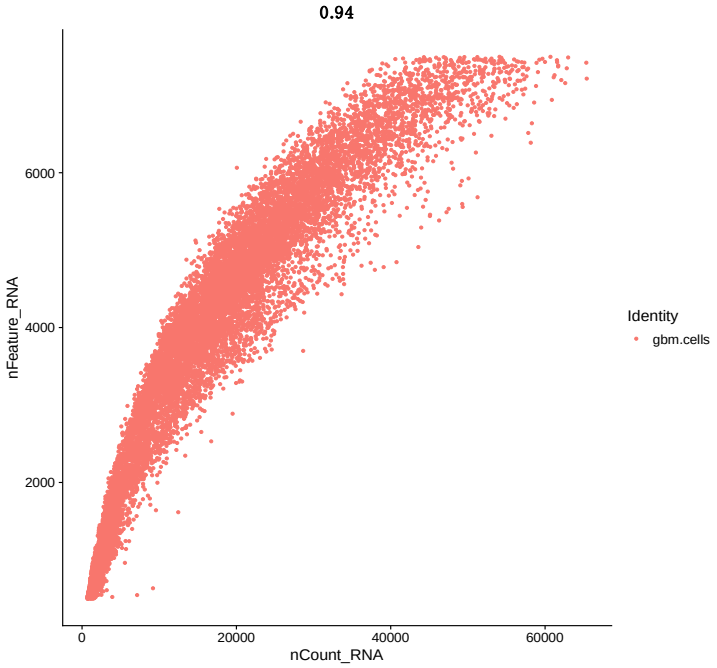

Fig. 1 Visualization of QC Metrics

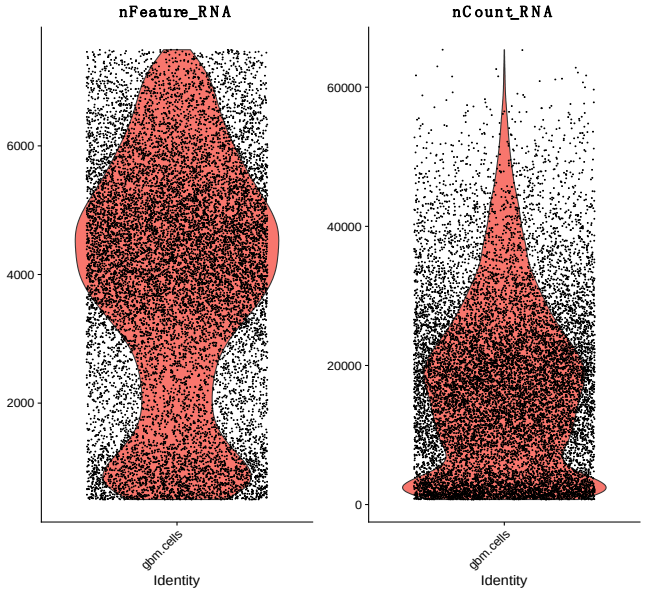

Fig. 2 Visualization of QC Metrics after Screening Feature Selection

The 2,000 genes with the largest cell-to-cell expression differences (largest variance) were selected. (Figure 3) We also marked the names of the top ten genes for use in subsequent analysis.

It can be observed that there are 17878 Non-variable Count screened out, and the top ten highly variable genes in the Variale Count are Cldn5, Ptgds, Ctla2a, Crym, Lyz2, Spp1, Mgp, Ly6c1, Cxcl9, and Apod.

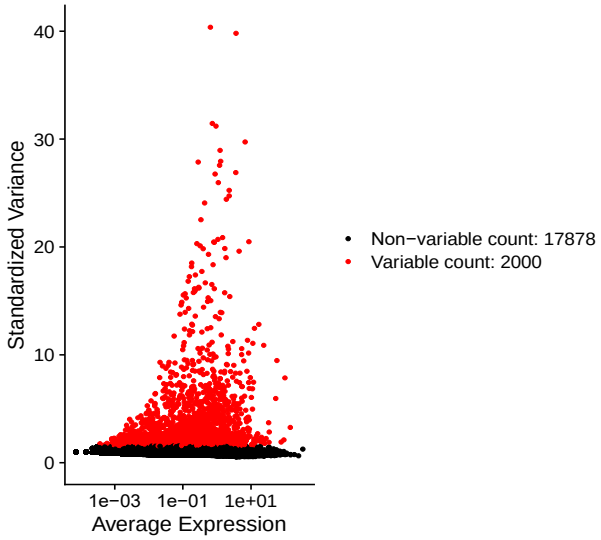

Fig. 3 Visualization of QC Metrics after Screening

3.2. Dimensionality reduction

Principal component analysis is the basis of clustering. Here, PCA dimensionality reduction were performed on the scaled data and principal components to biological covariates are mapped to find the basis of cell clustering. Seurat package accomplished by R was applied to identify major cell types [7]. Highly variable genes were generated and used to perform PCA Significant principal components were determined using JackStraw analysis and visualization of heatmaps focusing on PC 1 to 8 (Figure 4).

For each principal component, we selected the top 20 genes and ranked them according to PCA scores. The first two principal component with feature genes was showed in Figure 5.

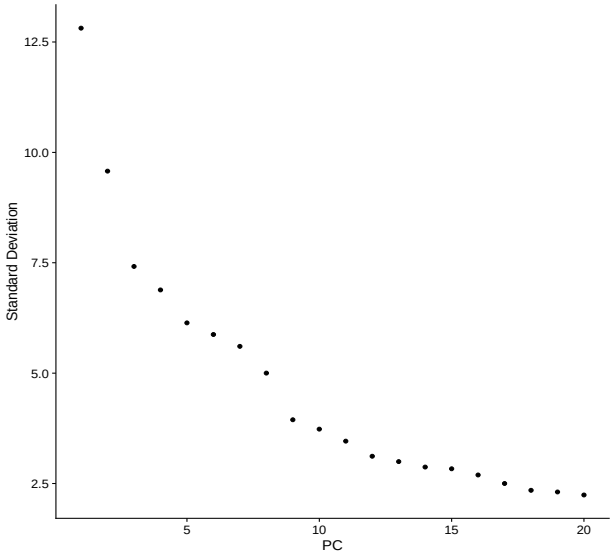

Fig. 4 Scatter Plot Showing the Standard Deviation of the Top PC

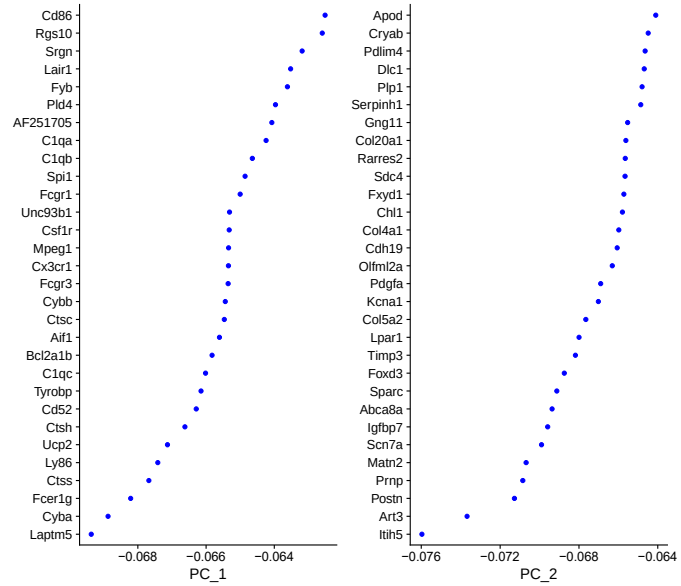

Fig. 5 Scatter Plot Showing the Scores of the Top Two PCs

3.3. Clustering

Cell populations are formed based on the similarity of gene expression so they could be clustered based on the similarity of gene expression to infer cell types. In this paper, Louvain algorithm was used for clustering. Each cell was connected to its K nearest neighbor cell, and the distance between cells was based on the Euclidean distance calculated by PC dimensionality reduction expression space. The proximity of the two subgroups represents the correlation between them. By clustering with the R package Seurat, we obtained 16 levels so 16 groups of cells were clustered.

3.4. Marker Genes Detection

After cell clustering, we could see the differential genes of the cells in different groups. To interpret our clustering results, the genes that drive separation between clusters were identified. These marker genes allow us to assign biological meaning to each cluster based on their functional annotation.

Firstly, we extracted this type of Seurat object, and used the FingMarkers function to set grouping and comparison. A volcanic map was used to show the differential genes (Figure 6). FC was fold change, which represented the ratio of the mean value of expression between two groups. After taking

the logarithm with the base of 2, it was log2fc. In this paper, we chose the absolute value of log2fc more than 1 as the screening standard of differential genes [8].

Then we took the corrected p-value (FDR, false discovery rate) and used negative log10 to get the coordinate value. The larger the value, the more significant the difference was. Because differential expression analysis of transcriptome sequencing is an independent statistical hypothesis test for many gene expression values, there will be false positive problems. Therefore, in the process of differential expression analysis, correction was used to adjust the significance p-value obtained from the original hypothesis test, and finally FDR<0.01 was used as the key index for differential expression gene screening [9].

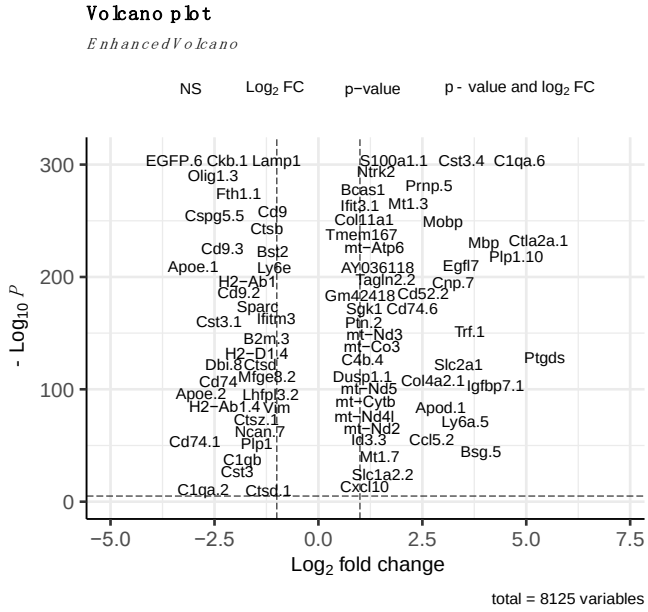

Fig. 6 Volcano plot showing differential expression of feature genes

Through volcanic map of the top gene of cell subpopulations, we observed the positive genes (log2FC > 1, Log10P < 0.01) and negative genes (log2FC < -1, Log10P < 0.01). Therefore, we could further confirm the ideal clustering results, and made sure marker genes were used as subpopulations by comparison. Red dots represented differentially expressed genes while blue dots stood for genes that were not differentially expressed. We could see that C1qa.6, Ctla2a.1, Plp1.10, Mbp e.g., were obvious differential expression positive genes, and these differential genes would provide the basis for the next cell annotation.

3.5. Cellular heterogeneity within glioblastoma

The heterogeneity of tumors leads to cancer progression and treatment failure. After strict quality control and normalization, we analyzed a total of 14302 cells from two samples. All high-quality single cells were then clustered and annotated into 16 different cell types, including natural G3 cells from Schwalie et al., oligodendrocyte precursor cells(OPCs), epithelial cells, neuron-glial antigen (NG2) cells, neural progenitor cells, astrocytes, neurons, macrophages, smooth muscle cells, type I spiral ganglion neurons(SGNs), late activates neural stem cells, microglial cells, oligodendrocytes (OLs), natural killer cells (NK), mural cells, and endothelial cells(Figure 7). Among these, G3 cells were characterized by the specific expression of Cd44 and Meox2[9]. The protein encoded by CD44 is a cell surface receptor that plays a part in cell-cell interaction, cell adhesion and migration, helping them sense and respond to changes in the tissue microenvironment. [10].It participates in extracellular matrix components such as hyaluronic acid (HA), collagen, growth factors, cytokines or proteases and serves as a platform for signal transduction by assembling, via its cytoplasmic domain, protein complexes containing receptor kinases and membrane proteases [11]. And Meox2 is a Protein Coding gene. Gene Ontology (GO) annotations related to this gene include DNA-binding transcription factor

activity and RNA polymerase II cis-regulatory region sequence-specific DNA binding. It may have a regulatory role when quiescent vascular smooth muscle cells reenter the cell cycle (By similarity). Also, it acts as a negative regulator of angiogenesis [12] and activates expression of CDKN1A and CDKN2A in endothelial cells, acting as a regulator of vascular cell proliferation [13]. While it activates CDKN1A in a DNA-dependent manner, it activates CDKN2A in a DNA-independent manner [14]. Oligodendrocyte precursor cells can not only differentiate into oligodendrocytes but also form synaptic connections with neurons, which interact with each other and participate in the composition of neural network structure. The functions of oligodendrocytes include the formation of myelin sheath of axons in the central nervous system: the secretion of neurotrophic factors to promote the survival and function of neurons and glia; the expression of inhibitory proteins to prevent the overgrowth of nerve fibers. They are closely related to central nervous system diseases such as multiple sclerosis, traumatic spinal cord injury and oligodendroglioma, and participate in the pathological process of these diseases.

Distributed throughout the grey and white matter, NG2 cells (also known as polydendrocytes) are a population of CNS cells, which can produce oligodendrocytes, protoplasmic astrocytes, and some neurons in the body. Besides, polydendrocytes receive synaptic input from neurons, which suggests that they are integrated into the neural network [15].Also, NG2 is phosphorylated at the threonine 2314 (thr2314) site and co-localized with β-1 integrin on the microprocesses on the surface of apical cells, which is beneficial to cell proliferation. On the other hand, THR2256 phosphorylation leads to the co-localization of NG2 and β 1 integrin on the lamellar foot of the cell front, which is related to enhanced cell movement. NG2+ cells play an important role in the proliferation of GBM. The NG2+ cells obtained by classification proliferate faster and are more invasive than NG2- cells in vivo. It is worth noting that fatal tumors also originate from NG2- cells. Neural progenitor cells construct complex and delicate cerebrum through the process of self-renewal, division, and differentiation. Astrocytes have many processes. They can extend and fill between the cell bodies and processes of nerve cells, support and separate nerve cells, and participate in the formation of the blood-brain barrier. Because astrocytes can produce and secrete some neurotransmitters and express some neurotransmitter receptors, they can respond to some neuroactive substances. In addition, astrocytes can biotransform exogenous compounds and help regulate the ion microenvironment around neurons. It is worth noting that GBM originates from astrocytes. Type I SGNs transmit complex acoustic information from the inner hair cells (IHCs) to the brain stem.

Microglia are necessary for the proper development of the nervous system. At the early stage of brain development, about 20%-80% of neurons with long projection axons undergo apoptosis and are rapidly cleared. Extracellular signals such as hormones, neurotransmitters or secreted proteins can bind to specific receptors on neurons to promote neuronal survival or initiate programmed cell death, thereby regulating the number of neurons in the central nervous system. In addition, microglia cells also participate in the development of a range of neurodegenerative diseases, mediating the endogenous immune response to central nervous system injury and diseases. And they play a neurotoxic and neuroprotective role. NK are a group of important immune cells in the body, which are not only related to immune regulation, antiviral infection, and anti-tumor, but also involved in the origin of hypersensitivity and autoimmune diseases. Mural cells participate in regulating the growth and homeostasis of the embryonic and adult vasculature.

These results indicated that different cell subsets in GBM exhibited different states, reflecting different tumor biological properties, and provided more opinions about the molecular characteristics of GBM cells than before.

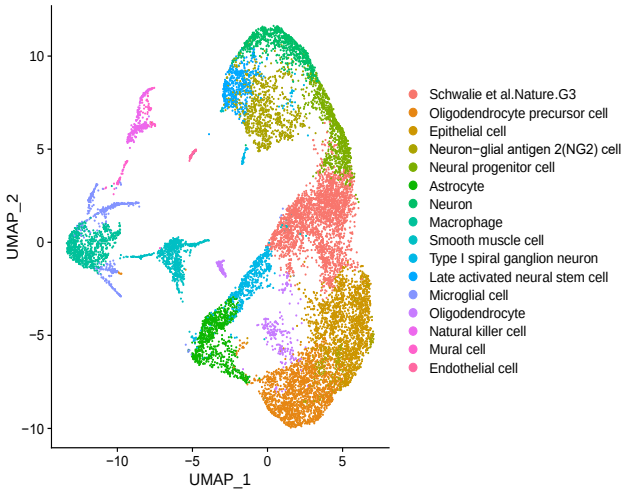

Fig. 7 UMAP diagram for visualization of clusters

3.6. Functional enrichment analysis of DEGs

To gain more insight into the biological functions of DEGs, we performed functional and pathway enrichment analyses using Metascape. The figure 8 showed the top 20 enriched GO terms. DEGs of biological process(BP) were involved in Innate Immune System, regulation of cell adhesion, cell activation, gliogenesis, positive regulation of cell migration, negative regulation of cellular component organization, hemostasis, actin cytoskeleton organization, regulation of cytokine production, response to wounding, regulation of cytoskeleton organization, negative regulation of immune system process, positive regulation of cell death, regulation of plasma membrane bounded cell projection organization, blood vessel morphogenesis, cell morphogenesis, regeneration, and response to metal ion. Metascape pathway analysis showed that DEGs were rich in L13a-mediated translational silencing of ceruloplasmin expression and signal transduction of RhoGTP enzyme. In addition, as shown in figure 9, the rich KEGG pathways included coronavirus disease - COVID-19, leukocyte transendothelial migration, osteoclast differentiation, pathways in cancer, yersinia infection, salmonella infection, and others. However, the interaction with COVID-19 is placed first. This implies that GBM cells have a rather close relationship with COVID-19, and new insights into coronaviruses can be obtained in the future by performing more in-depth downstream analyses to explore the interaction relationship and formation mechanism of the two.

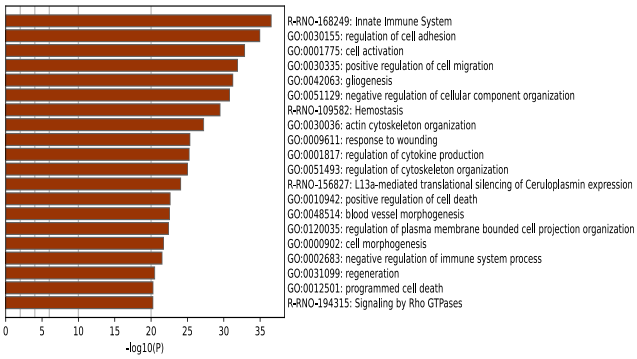

Fig. 8 Enrichment Analysis

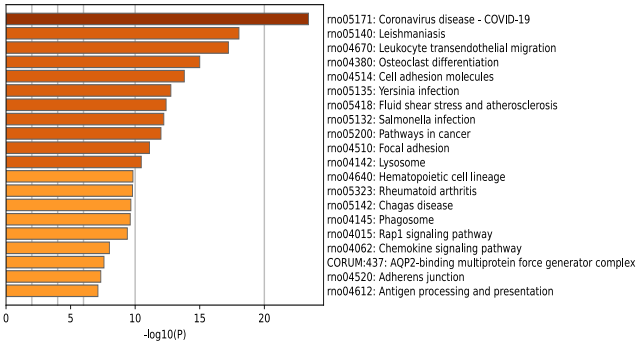

Fig. 9 KEGG pathway enrichment analysis of DEGs

To further gain the relationships between terms, a rich subset of terms was selected and presented as a network diagram (Figure 10 and Figure 11). The plot shows that there is a correlation between the regulation of cytokine production, the negative regulation of immune system processes, the regulation of cell adhesion, the activation of cells, the positive regulation of cell migration and the morphogenesis of blood vessels. And there is also a correlation between cell morphogenesis, regulation of plasma membrane bounded cell project, gliogenesis, negative regulation of cellular component organization, regulation of cytoskeleton organization, and actin cytoskeleton organization. Besides, there is a correlation between gliogenesis and programmed cell death. Most GBM pathways were closely related.

Fig. 10 Network of Enrichment analysis: (a) every color represents a cluster ID.

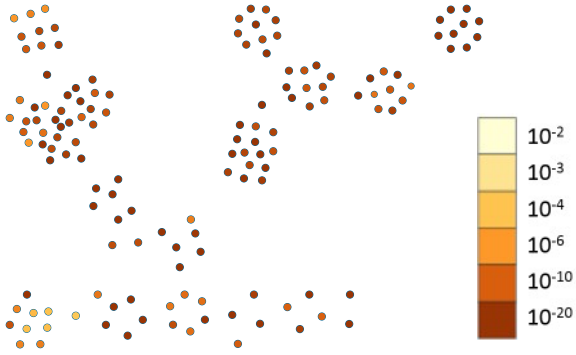

Fig. 11 Network of Enrichment analysis: (b) every color represents a p-value.

4. Discussion

Here, we used an integrative approach to confirm glioblastoma genetic heterogeneity, combining scRNA-seq statics form mice (14,302 cells in total). We controlled, filtered, and standardized the high-throughput single cell sequencing data of GBM, and then used PCA to reduce the dimension and cluster analysis. The cell samples were divided into 16 clusters. Through the analysis of differentially expressed genes by volcano map, the genes with significant differential expression were obtained. Next, cells of each group were annotated through CellMarker database. We found the cell groups of Oligodendrocyte precursor cell, Type I spiral ganglion neuron and Late activated neural stem cell, which may be related to the occurrence of cancer. Finally, the function of the cell population was obtained through go enrichment analysis. We observed that most of the differential genes were enriched in the development of immune nervous system and signal transduction and other functions. Single cell sequencing technology and subsequent data analysis are rapidly developing as tools for cancer research. Here we searched for genes and biological pathways related to the occurrence of cancer. However, it can be improved through subgroups finding related to prognosis and analysis of trajectory analysis of tumor cell differentiation, which will contribute to explaining the heterogeneity between tumors and finding new therapeutic targets for GBM.

5. Conclusions

In conclusion, Cldn5, Ptgds, Ctla2a, Crym, Lyz2, Spp1, Mgp, Ly6c1, Cxcl9, and Apod are involved in the process of GBM. From the results of this study, we have known that the negative regulation of immune system processes, the regulation of cell adhesion and the activation of cells were closely linked, and the regulation of cell adhesion was central. In addition, among cell morphogenesis, regulation of plasma membrane bounded cell project, gliogenesis, negative regulation of cellular component organization, regulation of cytoskeleton organization, and actin cytoskeleton organization, the regulation of the plasma membrane-bound cellular program was at the core. These results suggest that we should pay much attention to the regulation of cell adhesion, and the regulation of the plasma membrane-bound cellular program when considering ways to treat GBM. At the same, we should also pay attention to these two pathways' impacts on other pathways. Moreover, when considering the therapeutic targets of GBM, the cellular heterogeneity with GBM couldn't be ignored, which provides a lot of useful information. And it's worth noting that the coronavirus disease -- COVID-19 plays an important role in the origin and progress of GBM, indicating that GBM may be treated with the same methods used to treat COVID-19. while to confirm this possibility, we need further experimental studies in vitro and in vivo.

In summary, by this study, we have obtained a lot of useful evidence for further exploring the molecular mechanism of GBM, the selection of biomarkers, and the exploration of therapeutic targets. In the future, further experiments are needed to clarify the biological function.

References

[1] C. Neftel, J. Laffy, M. G. Filbin, T. Hara, and M. Suvà, "An Integrative Model of Cellular States, Plasticity, and Genetics for Glioblastoma," Cell, vol. 178, 2019.

[2] S. Darmanis, S. A. Sloan, D. Croote, M. Mignardi, et al., "Single-Cell RNA-Seq Analysis of Infiltrating Neoplastic Cells at the Migrating Front of Human Glioblastoma," Cell Reports, vol. 21, p. 1399, 2017.

[3] H. W. Lee, W. Chung, H. O. Lee, E. J. Da, A. Jo, J. E. Lim, et al., "Single-cell RNA sequencing reveals the tumor microenvironment and facilitates strategic choices to circumvent treatment failure in a chemorefractory bladder cancer patient," Genome Medicine, vol. 12, 2020.

[4] Ilicic T, Kim JK, Kolodziejczyk AA, Bagger FO, McCarthy DJ, Marioni JC, Teichmann SA. Classification of low quality cells from single-cell RNA-seq data. Genome Biol. 2016 Feb 17;17:29.

[5] V. D. Blondel, J. L. Guillaume, R. Lambiotte, and E. Lefebvre, "Fast unfolding of communities in large networks," 2008.

[6] Traag, Waltman, van, and Eck, "From Louvain to Leiden: guaranteeing well-connected communities," Scientific Reports, 2019.

[7] Y. Hao, S. Hao, E. Andersen-Nissen, W. M. Mauck, and R. Satija, "Integrated analysis of multimodal single-cell data," Cold Spring Harbor Laboratory, 2020.

[8] C. C. Qiu, Q. S. Su, S. Y. Zhu, and R. C. Liu, "Identification of Potential Biomarkers and Biological Pathways in Juvenile Dermatomyositis Based on miRNA-mRNA Network," BioMed Research International, vol. 2019, pp. 1-9, 2019.

[9] X. Lin, Y. Wu, S. Chen, X. Sun, Z. Ke, D. Chen, et al., "Identification of a five-mRNA signature as a novel potential prognostic biomarker in pediatric Wilms tumor," Molecular Genetics Genomic Medicine, vol. 8, 2020.

[10] Sánchez-Sendra B, Pérez-Debén S, González-Muñoz J F, et al. Prognostic value of IGF2 mRNA-binding protein 3 (IGF2BP3) intratumoral expression in melanoma patients at the time of diagnosis: comparative analysis of RT-qPCR versus immunohistochemistry[J]. Cancers, 2022, 14(9): 2319..

[11] Ma Y, Gu Y, Zhang X, et al. High Expression of MUC5AC, MUC5B, and Layilin Plays an Essential Role in Prediction in the Development of Plastic Bronchitis Caused by MPP[J]. Frontiers in Microbiology, 2022, 13..

[12] Sawada J, Hiraoka N, Qi R, et al. Molecular Signature of Tumor-Associated High Endothelial Venules That Can Predict Breast Cancer SurvivalMolecular Signature of Tumor-Associated HEVs [J]. Cancer Immunology Research, 2022: OF1-OF14

[13] Jiang J F, Zhou Z Y, Liu Y Z, et al. Role of Sp1 in atherosclerosis [J]. Molecular Biology Reports, 2022: 1-10.

[14] Brandt K J, Burger F, Baptista D, et al. Single-cell analysis uncovers osteoblast factor growth differentiation factor 10 as mediator of vascular smooth muscle cell phenotypic modulation associated with plaque rupture in human carotid artery disease [J]. International Journal of Molecular Sciences, 2022, 23(3): 1796.

[15] Pfeiffer F. Reciprocal Interactions between Oligodendrocyte Precursor Cells and the Neurovascular Unit in Health and Disease [J]. Cells, 2022, 11(12): 1954.

[16] Zhou Y, Zhou B, Pache L, Chang M, et al. Metascape provides a biologist-oriented resource for the analysis of systems-level datasets. Nat Commun. 2019 Apr 3; 10(1): 1523.