NATIONAL RESEARCH UNIVERSITY
HIGHER SCHOOL OF ECONOMICS

Faculty of Computer Science
Bachelor's Programme "Data Science and Business Analytics"

Research Project Report on the Topic:

Data anlysis of single of single-cell sequencing: study of brain tumor using spatial transcriptomics data

Fulfilled by:

Student of the Group БПАД212
Orlova Anastasia Grigorevna

(signature) **02.06.2024** *(date)*

Assessed by the Project Supervisor:

Poptsova Maria Sergeevna

Head of the International Laboratory for Bioinformatics

Faculty of Computer Science, HSE University

(signature) *(date)*

Moscow 2024

Annotation

Russian

В данном проекте планируется изучить все тонкости работы новых технологии в биоинформатике: одноклеточное секвенирование(single-cell RNA sequencing, scRNA-seq) глиом с применением пакета Seurat.

Глиома — это общий термин для группы опухолей, которые возникают из глиальных клеток головного и спинного мозга. Глиомы могут различаться по типу глиальных клеток, из которых они происходят, и по степени злокачественности.

После освоения технологии одноклеточного секвенирования, в данной работе необходимо будет опробовать данный метод на реальных данных по опухоли мозга.
С помощью этого проекта и проделанного анализа, мы сможем сделать выводы о том, какие гены активны в каждой клетке, выявить различные клеточные подтипы и как клетки взаимодействуют друг с другом и с их микроокружением.

English

In this project, it is planned to study all the subtleties of the work of new technologies in bioinformatics: single-cell sequencing (single-cell RNA sequencing, scRNA-seq) of gliomas using the Seurat package.

Glioma is a general term for a group of tumors that arise from glial cells in the brain and spinal cord. Gliomas can differ in the type of glial cells from which they originate and in the degree of malignancy.

After mastering the technology of single-cell sequencing, in this work it will be necessary to test this method on real data on a brain tumor.
With the help of this project and the analysis done, we will be able to draw conclusions about which genes are active in each cell, identify different cell subtypes and how cells interact with each other and with their microenvironment.

Key terms

- Single-cell sequencing - ScRNA-seq
- Seurat
- glioma
- clustering
- tumor microenvironment
- Ligand-receptor iteractions

Introduction

Relevance

Single-cell sequencing is the latest approach in the study of brain tumors, which is of great importance for understanding the mechanisms of disease development and progression. This method opens up new prospects for the development of targeted therapeutic strategies and improved diagnosis.

The study of brain tumors at the cellular level makes a significant contribution to translational medicine and molecular oncology, providing new data for: the development of personalized treatment approaches that take into account the genetic heterogeneity of the tumor and its interaction with the microenvironment, as well as to improve diagnostic tools to more accurately determine the stages of tumor development and predict the response to treatment.

Thus, single-cell sequencing is a key tool in the fight against brain tumors, opening up new horizons for health science and providing hope for more effective treatment of patients.

Goals and objectives

In this project, we plan to study the single-cell sequencing method in bioinformatics. Our goal is to apply the technology to the analysis of brain tumor data. We will focus on single-cell sequencing of a single tumor sample to identify the active genes in each cell and identify different cellular subtypes in brain tumor samples. This project will help us draw conclusions about cellular heterogeneity and interactions in the selected brain tumor and data on it.

Data description and preprocessing

In this work, we selected data in the form of a single sample of high-grade glioma. Gene Expression Omnibus (GEO) was used as a resource. High expression of the IDH gene was found in this sample.

Using the R package Seurat, we read the data and created a SeuratPackage from it. Then we carried out quality control (QC). This stage is necessary to ensure the reliability and accuracy of the analysis of single-cell RNA sequence (scRNA-seq) data. scRNA-seq data may contain noise and artifacts due to technical limitations and biological variability. QC helps to filter out low-quality cells and get rid of data that can distort the results of the analysis. The main QC parameters include the number of unique genes (nFeature_RNA), the total number of reads (nCount_RNA) and the percentage of mitochondrial genes (percent.mt), which may indicate damaged or dying cells. Before QC, many highly variable genes are isolated, including noise. After QC, only truly variable genes remain, which improves the accuracy of the analysis.

Results

After QC, the data becomes cleaner and more homogeneous, which increases the accuracy of the analysis of single-cell RNA sequences.

Clusterization

Next, we clustered the processed SeuratObject. We conducted clustering of single-cell RNA sequencing data following a specific sequential process that included several key steps.

At first, the data was normalized to eliminate technical artifacts and bring them to a comparable appearance. This is important to ensure the accuracy of subsequent analysis, as normalization helps to balance gene expression between cells.

Then we selected highly variable genes that show the greatest variability in expression. This helps to focus on biologically significant signals and eliminate less informative genes. In this case, 7,500 of the most variable genes were selected.

After that, we visualized the selected highly variable genes to see their distribution and make sure that the choice was correct. The top 10 most variable genes were marked for additional analysis.The next step was to scale the data. Scaling aligns the values of gene expression, which makes them more comparable with each other and prepares data for principal component analysis (PCA).

Next, we performed a principal component analysis (PCA) to reduce the dimensionality of the data. PCA helps to identify the main sources of variation in the data, which facilitates subsequent clustering.

We visualized the results of the PCA, including graphs of the loads of the main components, a heat map for the first main component and an Elbow raft. The Elbow raft was used to determine the optimal number of components to be considered in the analysis.

After that, we built a graph of the nearest neighbors using the K-Nearest Neighbors (KNN) algorithm. Based on this graph, we performed clustering using the Louvain algorithm to determine cell clusters. Clusters were obtained at different resolution levels (different levels of detail), in order to choose the most appropriate clustering level, a resolution of 0.03 was selected.

At the final stage, we used the UMAP (Uniform Manifold Approximation and Projection) method to further reduce the dimensionality of the data and visualize clusters in 2D space. This allowed us to visually see the data structure and the relationships between clusters. (Fig.1)

As a result, we get four clusters, which we plan to annotate later.

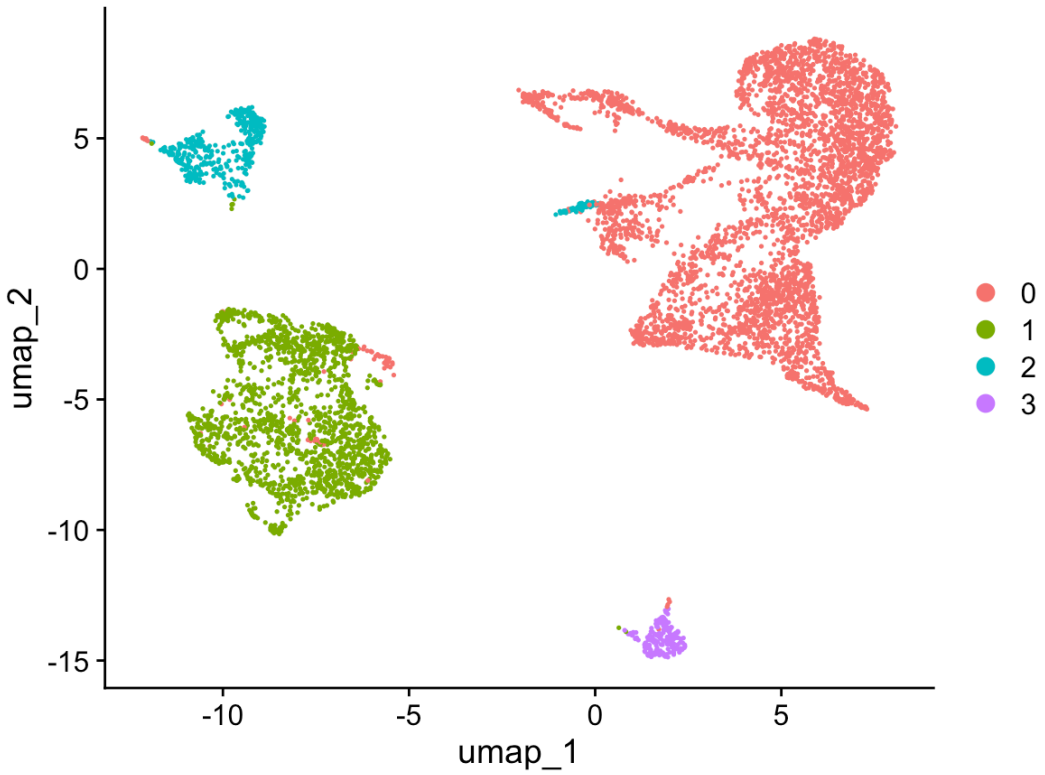
umap_2
umap_1
5
0
-5
-10
-15
-10
-5
0
5
0
1
2
3

Fig. 1a

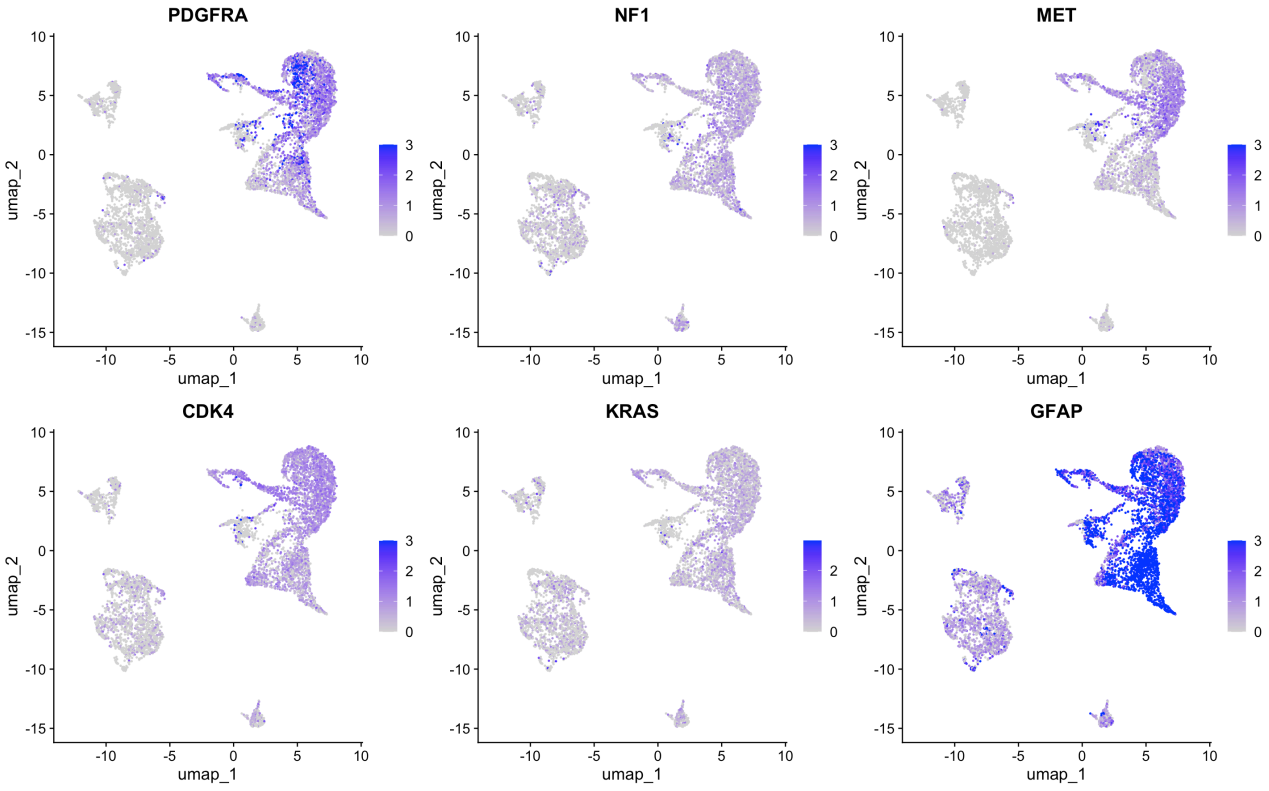
PDGFRA
NF1
MET
CDK4
KRAS
GFAP
umap_1
umap_2

Fig. 1b

Defining cell types

After clustering the cells, it was necessary to determine what type of cells each cluster is. To do this, we combined two technologies: the differential expression method and manual detection by key markers.

Differential expression

First of all, we used differential expression.

We perform a search for marker genes for all clusters in our dataset. Only highly variable genes are used for this. We are looking for genes that are expressed higher in a certain cluster compared to the rest of the clusters. In this case, the marker gene should be expressed in at least 25% of the cells in the cluster, and the threshold for the logarithmic fold of expression change is set at 0.25.

After identifying the marker genes, we filter the results to leave only those genes that have a logarithmic fold of expression change greater than 1. This allows us to identify the genes with the most significant changes in expression. For each cluster, we select the top 15 such marker genes, which gives us the most informative genes for further analysis.

At the last stage, we create a heat map to visualize the expression of these marker genes in various clusters. The heat map allows you to visually see which genes are markers for each cluster, and how they are expressed in different groups of cells. To simplify perception, the legend on the heat map has been removed.

Using differential expression, we saw the most expressive genes in each cluster and, using this information, made an assumption to which cell type each cluster corresponds:Zero cluster: All expressed genes are different and belong to different cell types, making it impossible to clearly identify the cluster. Presumably, this cluster consists of cancer cells. (Fig.2)

First cluster: FCGR3A, AIF1 (also known as IBA1): These markers are typically associated with macrophages.

Second cluster: CD3D, CD3E, TRAC, TRBC2: These are markers for immune cells.

Third cluster: MAG, MOG, CLDN11: These are markers for oligodendrocytes.

Key markers

Then we decided to identify the null cluster. Since we assume that cluster zero is a tumor, and more specifically a high-grade glioma, I found well-known cancer markers on the Network of Cancer Genes(NCG) website and highlighted them on the cluster map. In addition to cancer markers, I highlighted markers for oligodendrocytes, t cells, and macrophages/macroglia to make sure that they express only in the initially declared clusters. I got the following results.

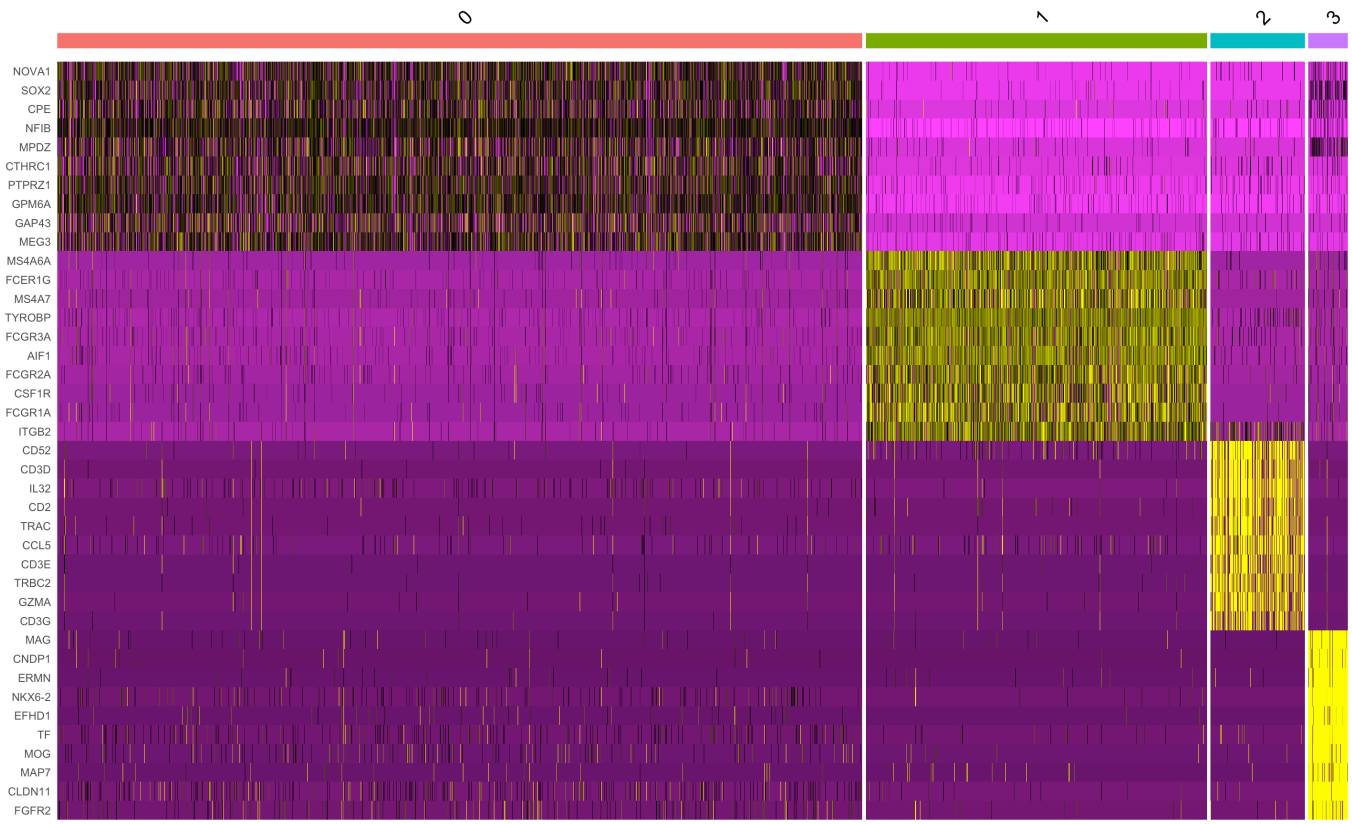

Fig. 2

We see that in the zero cluster such oncomercers as MET, PDGFRA, CDK4, NF1, KRAS are indeed overexpressed, which confirms our hypothesis that these are cancer cells (Fig. 1b). Moreover, it was noticed that genes such as GFAP (astrocyte marker) and OLIG1-2 (oligodendrocyte markers) are also vividly expressed in the zero cluster, which allows us to assume that this is not just a glioma, but classify it as an astrocytoma or oligodendrinoma.

I also noticed several interesting markers that are expressed simultaneously in cancer cells and immune cells, for example CCND1-2-3. Presumably, these markers can affect the tumor. These features will be discussed in more detail later in the paper.

So, we clustered the data, determined what type of cells each cluster has. Four types of cells were obtained: tumor cells, macrophages, immune cells and oligodendrocytes.

Researching each cluster individually

Next, I decided to take a closer look at large clusters in order to more clearly see the differentiation of cells within each cluster. I did the same with each cluster as with the common object earlier: I clustered, determined the most expressed genes in each subcluster using differential expression and tinting the most well-known markers manually, and then determined the cell type of each subcluster.

Macrophages

First of all, it was especially important to differentiate macrophages. Macrophages in the context of cancer play a dual role: M1 macrophages and M2 macrophages.

M1 macrophages have anti-tumor properties. They are activated by pro-inflammatory signals, destroy cancer cells and inhibit their growth. The presence of M1 macrophages in the tumor is usually associated with a good prognosis for the patient.

M2 macrophages, on the contrary, support tumor progression. They are activated by anti-inflammatory signals, promote the growth of new blood vessels, suppress the antitumor immune response and promote metastasis. A high number of M2 macrophages in the tumor is associated with a poor prognosis.Thus, M1 macrophages suppress cancer, and M2 macrophages support it. There are also other types of macrophages (M3/M0), but the most significant in the context of cancer was to identify M1-M2.

The macrophage cluster was divided into three subclusters, using markers they were identified as M1, M2, M3 macrophages, respectively. Among the most expressed genes, we were most interested in the MIF and PTEN genes, since they were the most expressive in both immune and cancer clusters (fig.3), and according to the results of other studies, it turned out that assumptions had already been made about the potential effect of these markers on cancer, which our study also confirms.

Immune cells

Then it was important to separate the immune cells - the main goal was to find T cells and B cells, T cells recognize and destroy cancer cells, whereas B cells can produce antibodies against tumor antigens, contributing to the immune response to cancer.

The immune cell subcluster also split into three subclusters, which we subsequently identified as T-cells, B-cells, and Plasma cells. After this analysis, we were interested in BTG1 and TNFAIP3 (fig.3) markers, since they were the most expressive in both immune and cancer clusters, and according to the results of other studies, it turned out that assumptions had already been made about the potential effect of these markers on cancer, which our study also confirms.

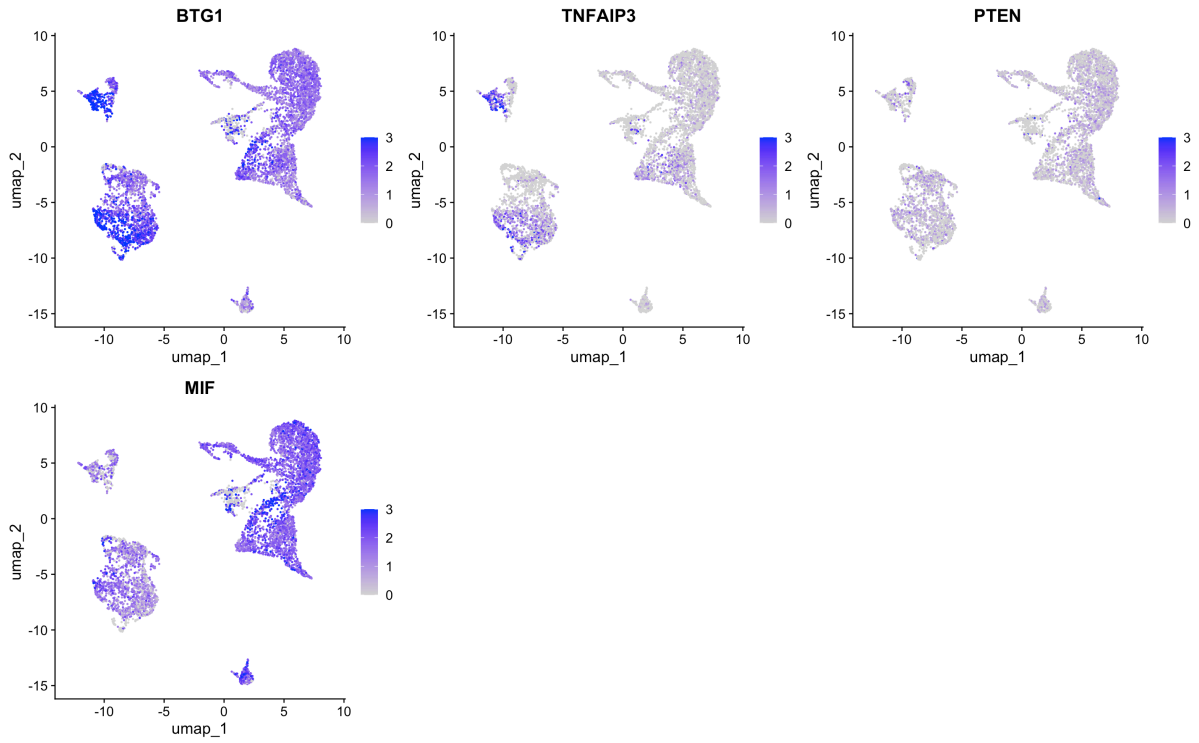

Fig. 3

Cancer cells

Then I found it interesting to cluster cancer cells to see the differentiation of cell functions within the tumor. This cluster is divided into 6 subclusters. Then we gave a name to each subcluster corresponding to the description of the functions of this cluster.

Cluster 0: This cluster is characterized by high expression of genes related to oligodendrocytes (OLIG1, OLIG2), neuronal differentiation (SOX4, SOX8), and neuronal activity (GRIA2, BCAN). These cells retain some properties of oligodendrocytes, participating in intercellular communication and axon support. We named this cluster «OPC» - precursors of oligodendrocytes.

Cluster 1: This cluster includes cells with astrocytic or glial properties, indicated by genes GFAP and VIM. Genes IGFBP7 and CLU are involved in cell protection and survival, while genes ID3 and S100A10 regulate cell differentiation and interactions. We named this cluster «Glial Support».

Cluster 2: This cluster consists of actively proliferating cells with high expression of genes related to cell proliferation and division (TOP2A, TK1, BIRC5). Genes CENPK and CENPM are involved in mitosis, while genes CA12 and OASL are associated with metabolic processes and cell protection. We named this cluster «Active Proliferation».

Cluster 3: This cluster is characterized by cells adapted to stress conditions and inflammatory responses. High expression of genes related to iron metabolism (FTH1, FTL), inflammation (IL32, NDRG1), and cell protection (SQSTM1, OSGIN1). We named this cluster «High Survival and Inflammation».

Cluster 4: This cluster includes cells with high expression of genes regulating cell adhesion and migration (RAB13, LPP), lipid metabolism (APOE), and immune interactions (CD86, ADAR). These cells are actively involved in cell migration and intercellular interactions. We named this cluster «Migration and Adhesion».

Cluster 5: This cluster consists of cells with high expression of genes related to immune response and inflammation (HAVCR2, CLEC7A). Genes FCGR1B and LILRB4 are involved in phagocytosis and antigen presentation, while genes CYBB and RNASE6 regulate cellular stress and metabolism. We named this cluster «Immune Inflammatory Response».
Then we combined all the subclusters and got a final map that includes all the subclusters (fig 4).

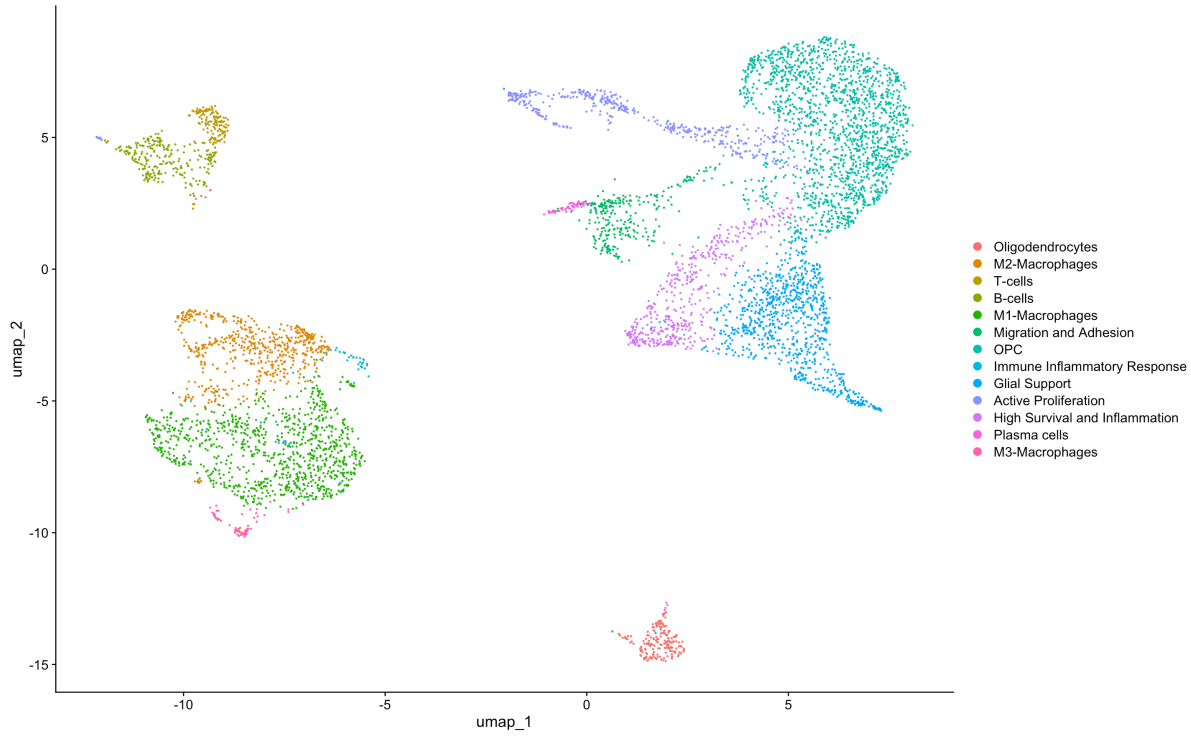

Fig. 4

Cell-cell communications analysis

The purpose of the analysis is to investigate intercellular communication to understand the interactions between different cell types. This is important for studying the mechanisms of diseases, immune responses and cellular development.

In the analysis process using the CellChat R package, we created an object to analyze cell type-based communications, connected a human ligand-receptor database, isolated relevant interactions, identified overexpressed genes and interactions, reduced data dimensionality with PCA, calculated and filtered interaction probabilities, and finally visualized the networks of interactions between different cell types.

In the graph below that shows a network of intercellular communications, you can see the interactions between different cell types. M1 macrophages interact with T cells, B cells, M2 macrophages, M3 macrophages and other cell types, with particularly strong and diverse connections to dendritic cells and oligodendrocytes. T cells interact with M1 macrophages, B cells, M2 macrophages and oligodendrocytes. B cells interact with T cells, M1 macrophages and M2 macrophages. M2 macrophages interact with M1 macrophages, T cells, B cells and oligodendrocytes. Dendritic cells interact with M1 macrophages, M2 macrophages and oligodendrocytes. Oligodendrocytes interact with M1 macrophages, T cells, B cells, M2 macrophages and dendritic cells. OPCs interact with glial cells involved in support and inflammatory response. Plasma cells interact with M1 macrophages and other immune cells. Glial support interacts with OPC, actively proliferating cells and cells involved in survival and inflammation (fig. 5)

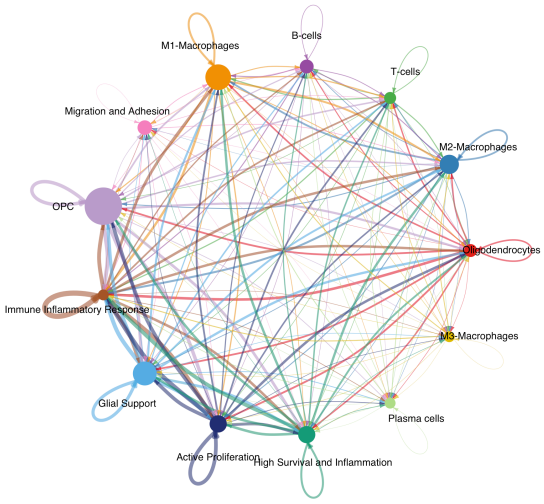

Fig. 5

Signaling networks

Then we built a bubble diagram of interactions between different cell types, where signal sources (stromal cells) interact with target cells (tumor cells). On the vertical axis, various signaling molecules or ligands (for example, VCAM 1, TNFSF, TGFBI, etc.) involved in intercellular interactions (fig. 6).

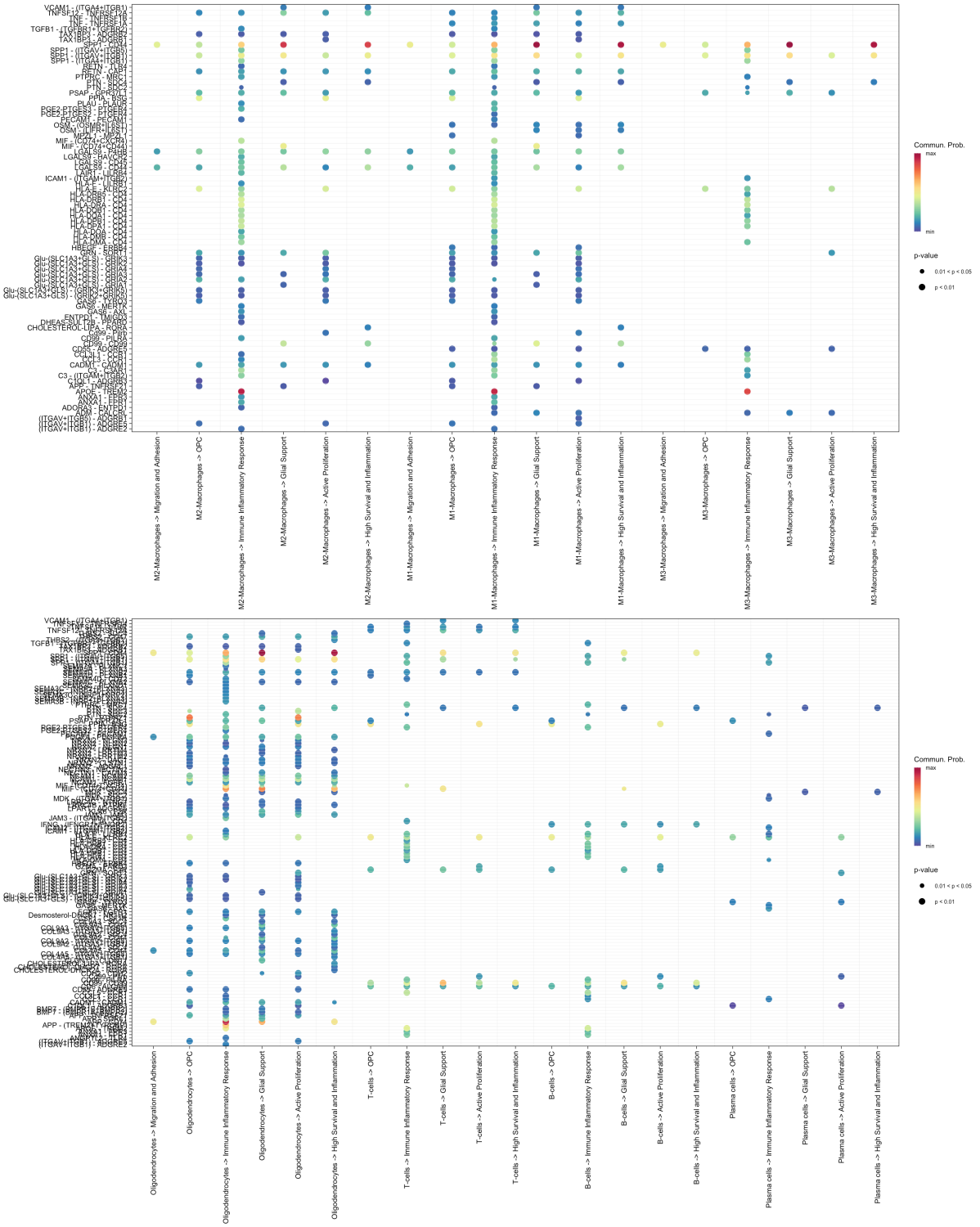

Fig. 6

From these bubble diagrams, we can see that the most active ligand-receptor pairs SPP1-CD44 and APOE-TREM2 - these pairs have the highest probability of communication. In addition, we can once again confirm the high communication between macrophages and cancer. In addition, a significant effect of the MIF - (CD74+CD44) pair on the "communication of cells" can be noted. Then we examined in more detail the effect of SPP1, MIF and APOE ligands on cellular communication, and in particular on glioma cells (fig. 7).

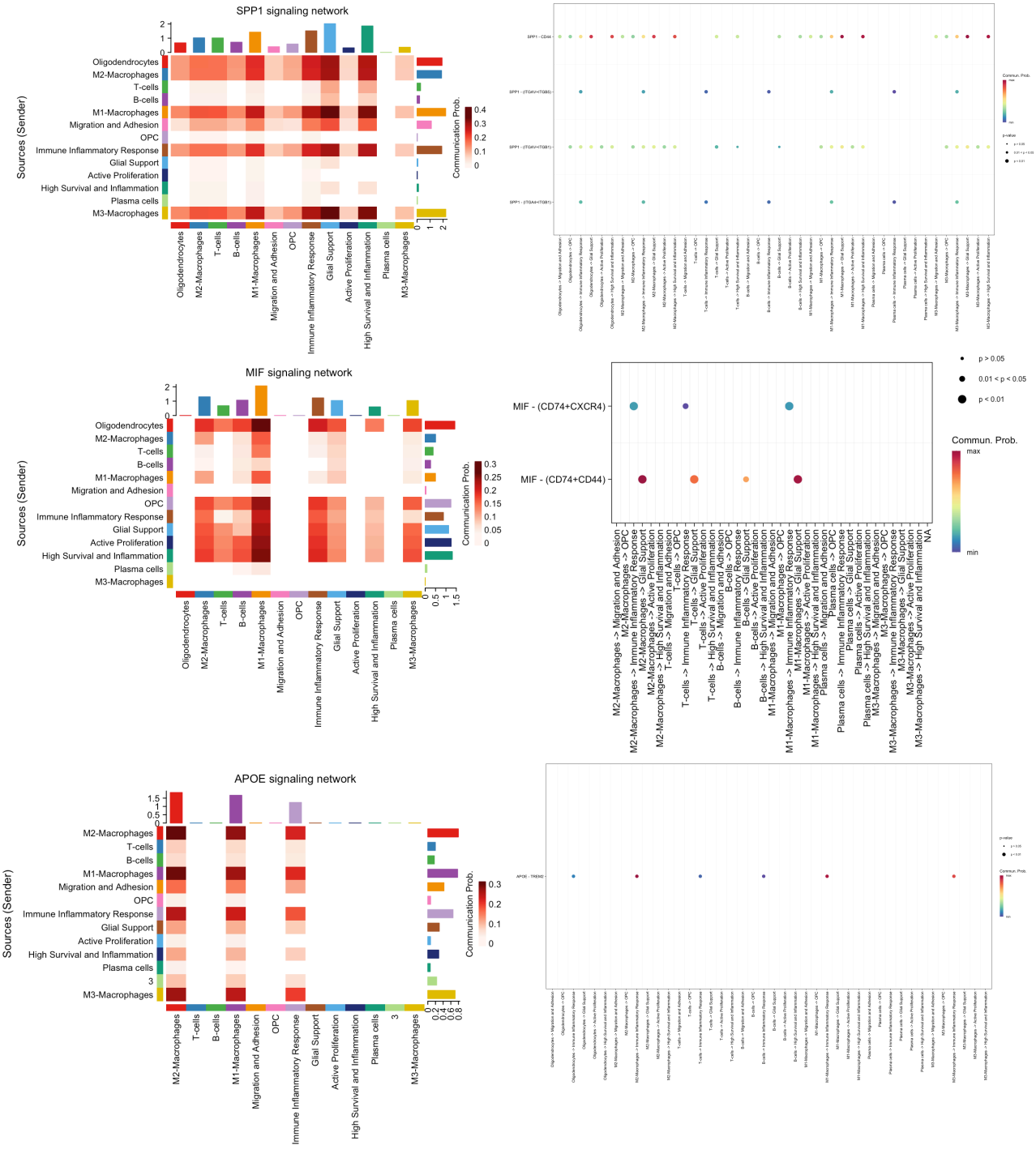

Fig. 7

Using bubble diagrams and graphs of the signaling networks of these three genes, it can be concluded that they clearly significantly contribute to the communication of cancer and stromal cells, so it can be assumed that they affect the growth and development of tumors. For example, the ligand-receptor interaction of MIF - (CD74+CD44) is particularly pronounced in the interaction of M1-M2 macrophages on "Glial support". The SPP1-CD44 pair similarly affects not only the effect of stromal cells, not only on "Glial support" cells, but also other cancer subtypes obtained. The graph of the APOE gene signaling network revealed communication only between macrophages and the «Immune Inflammatory Response».

Survival analysis

Finally we investigated the clinical significance of increased expression of SP1, CD44, APOE and MIF, BTG1, TNFAIP3 in glioma. To do this, we used Kaplan-Meier curves.

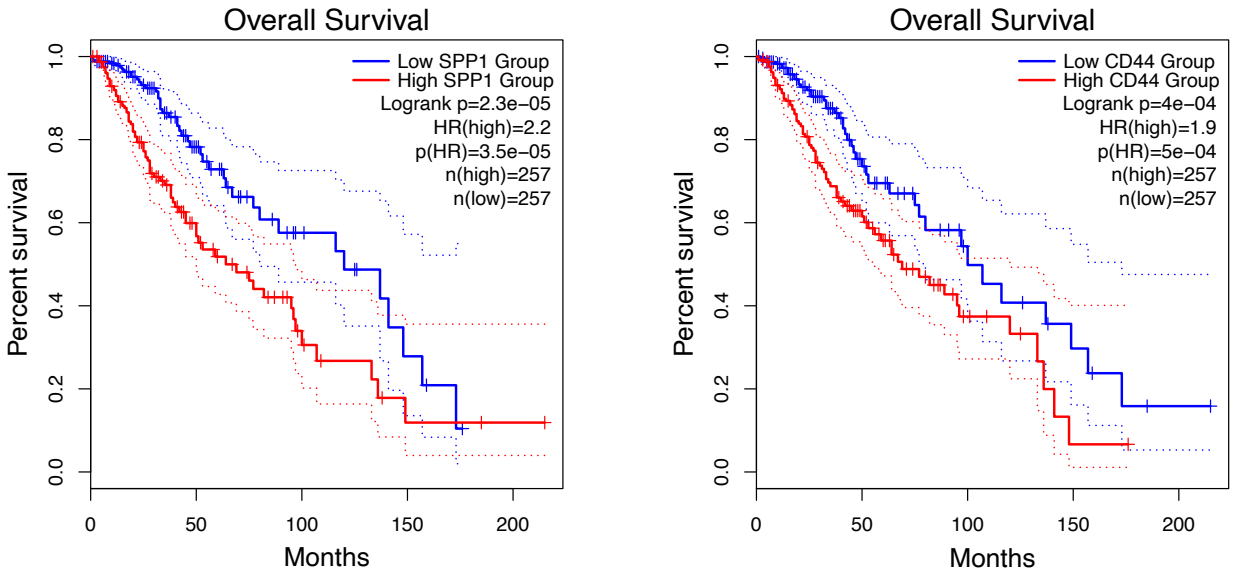

Fig. 8

We found that high SPP1, CD44 expression correlate with a shorter overall survival time of glioma patients (Fig. 8). Whereas high MIF, APOE expression correlate with a longer overall survival time of glioma patients. Such results regarding MIF and APOE are explicable with the protective immune functions of these genes. The TNFAIP3 and BTG1 graphs showed that there was no significant difference in survival between the group where these genes were highly expressed and low. As for SPP1, CD44, it can be assumed that these genes have a positive effect on the growth and development of gliomas, since life expectancy with high expression of these genes is significantly reduced.

Conclusion

Based on the results of this work, we studied and analyzed a sample of glioma and its microenvironment. The intercellular interactions within the tumor and the surrounding cells were studied. We have found markers that can potentially influence the development of glioma.

Along with identifying the tumor's intracellular heterogeneity, we also discovered a number of novel glioma symptoms. Initially, we discovered that this tumor is primarily composed of Tam cells, with a relatively low ratio of T cells to B cells. Our findings suggested that these macrophages probably had an anticancer function in gliomas. Furthermore, this tumor has five distinct subpopulations of glioma cells found in it. We discovered that the induction of tumor cells is significantly influenced by SPP1/CD44 signaling. Specifically, it has been demonstrated that SPP 1/ CD44 signaling in the perivascular niche enhances the stem cell-like characteristics and radiation resistance of glioma cells. Furthermore, it has been demonstrated that elevated SPP-1 expression causes glioma-associated macrophage infiltration and is linked to a poor prognosis in individuals with glioma. In a similar vein, elevated CD44 expression has also been linked to a worse prognosis for glioma patients and has been shown to contribute to the disease's advancement. Thus, our findings along with those of other research highlight the significance of tumor cells and macrophages mediating targeted interaction of SPP-1/CD44 in the treatment of gliomas.

List of references

https://satijalab.org/seurat/articles/get_started_v5_new

http://network-cancer-genes.org/

https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE185231

http://gepia2.cancer-pku.cn/#survival

https://www.genecards.org/

https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8602110/

https://pubmed.ncbi.nlm.nih.gov/21088135/

https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8858002/

https://www.genecards.org/cgi-bin/carddisp.pl?gene=BTG1

https://cancerci.biomedcentral.com/articles/10.1186/s12935-021-02089-2

https://www.nature.com/articles/s41417-022-00582-y