**糖尿病视网膜病变单细胞数据挖掘**

# 数据和方法

## 1.1、数据简介

## 从Gene Expression Omnibus (<https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi>)搜索并中下载相应的单细胞测序数据。

## 1.2、质量控制QC

用R包DropletUtils[1]检测各细胞的表达情况，并过滤掉没有任何细胞表达的barcoded。再根据各细胞UMI（unique molecular identifiers）数量进一步过滤。然后利用scater[2]包统计细胞中基因的表达情况，并根据线粒体基因（< 10%）和核糖体基因（>5%）表达的比例，对细胞进行过滤。最后对各基因进行统计。

## 1.3、数据预处理

利用Seurat[3]包中的NormalizeData对过滤后各样本的表达矩阵进行标准化。

## 1.4、主成分分析（PCA）

利用Seurat包中的FindVariableFeatures筛选出细胞间差异最明显的前2000个基因。在下游分析中关注这些基因有助于突出单细胞数据集中的生物信号。然后利用Seurat包中的ScaleData对表达数据进行线性缩放。最后用Seurat包中的RunPCA进行线性降维分析（PCA）。

## 1.5、细胞聚类及注释

挑选取标准差较大的主成分（PC），用Seurat包中的FindNeighbors和FindClusters进行细胞聚类分析。然后用Seurat包中的RunUMAP进行非线性降维分析（UMAP）。

## 1.6、确定marker基因

利用Seurat包中的FindMarkers计算各聚类与其他所有细胞的差异表达基因（log2FC大于等于0.1，细胞群最小表达比例0.25，p值小于等于0.05），从而获得marker基因（logFC前500个）。并根据已有marker基因[4]对细胞进行标注并做聚类展示图。

## 1.7、差异表达分析

利用Seurat包中的FindMarkers进行差异表达分析（细胞群最小表达比例0.25）。

## 1.8、基因富集分析

基于Gene Ontology[5]数据库，以及KEGG PATHWAY DATABASE[6]生物化学通路的数据库，对候选基因进行了功能富集分析。利用统计学算法（Fisher’s exact test）来找出一组基因和哪些具体的功能条目联系最大，分析结果中每个条目都对应一个统计值P-value来表示显著性。

## 1.9、细胞间互作网络分析

这里利用CellChat[7] v1.1，根据各类细胞对应的受体配体基因表达值，推断细胞与细胞间的通信。然后获取细胞间的受体配体对网络关系。

## 1.10、转录因子调控分析（SCENIC）

这里利用SCENIC[8]，根据相应的数据，推断其中对应各转录因子的调控活性。

# 分析结果

## 2.1、建立糖尿病视网膜病变的单细胞图谱

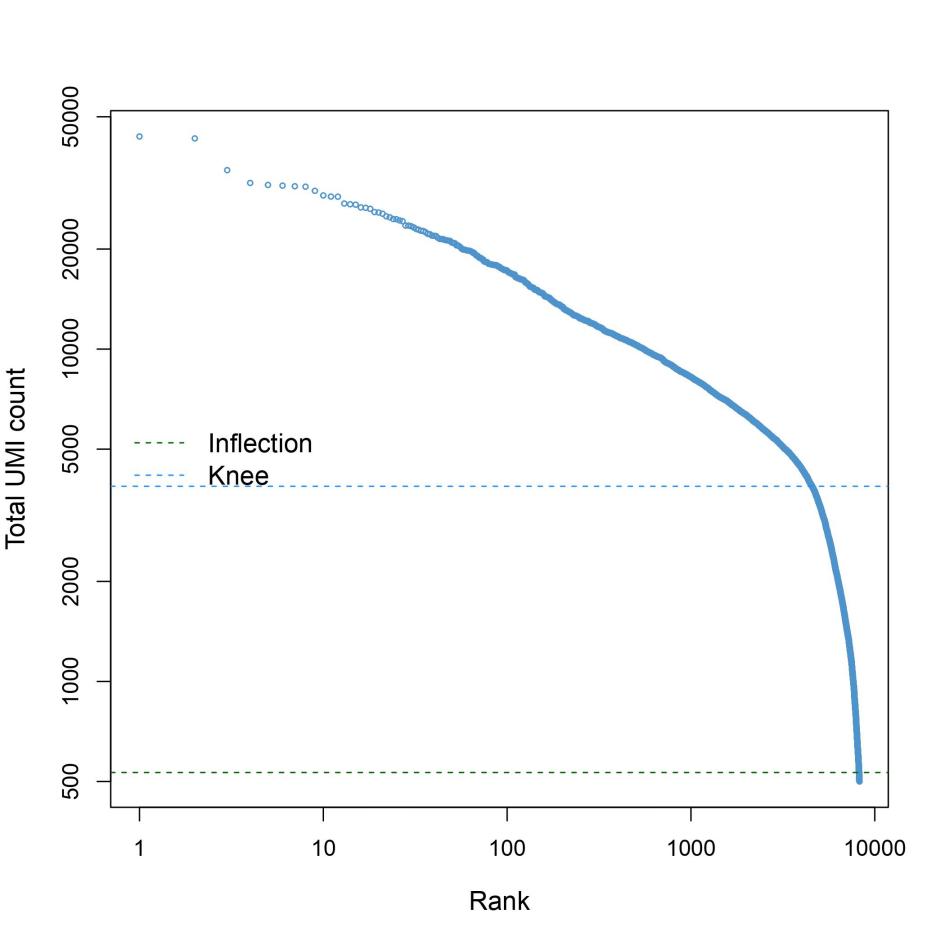
数据来源：GEO：GSE178121

### 2.1.1、质量控制QC

首先用R包DropletUtils中的barcodeRanks统计各barcode（对应每个细胞）的表达情况。再用emptyDrops过滤掉没有任何基因表达的barcode（该barcode中可能不包含任何细胞）。

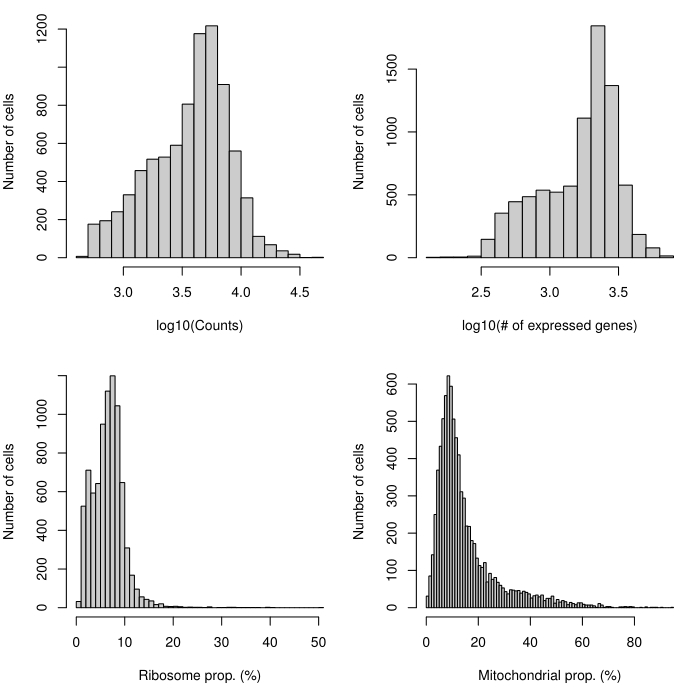
结果文件夹：QC

注：这里以GSM5380581为例展示结果



细胞表达情况图

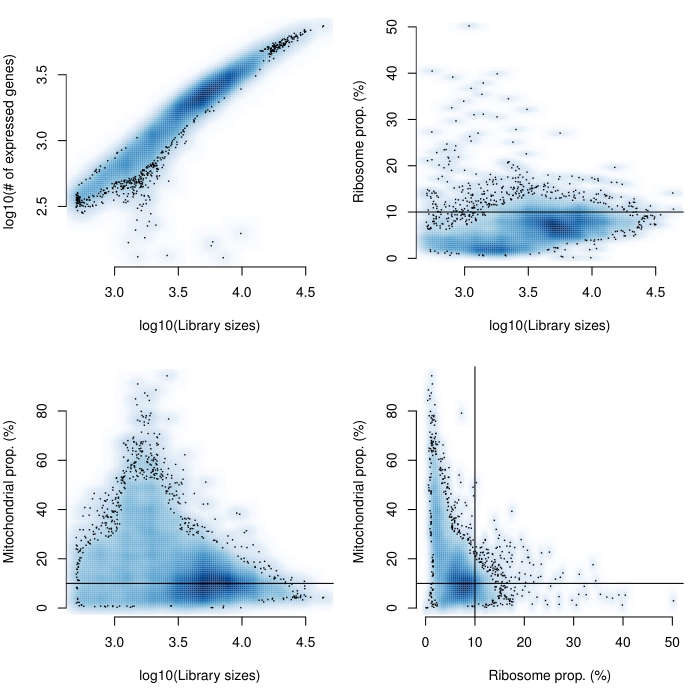
接下来用scater包中的perCellQCMetrics对表达数据进行统计。



基因表达统计

### 2.1.2、细胞过滤

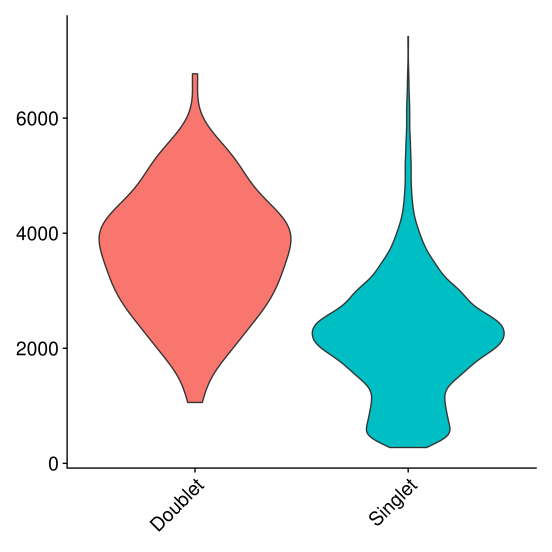
首先过滤掉表达基因（UMI）数小于200的细胞。然后统计线粒体及核糖体基因在每个细胞中表达的比例，并保留线粒体表达比例小于10%且核糖体基因表达比例大于5%的细胞。结果在QC文件夹中。



各种基因所占比例统计图

### 2.1.3、去除双细胞

最后用DoubletFinder包计算并去除双细胞。结果在QC文件夹中。

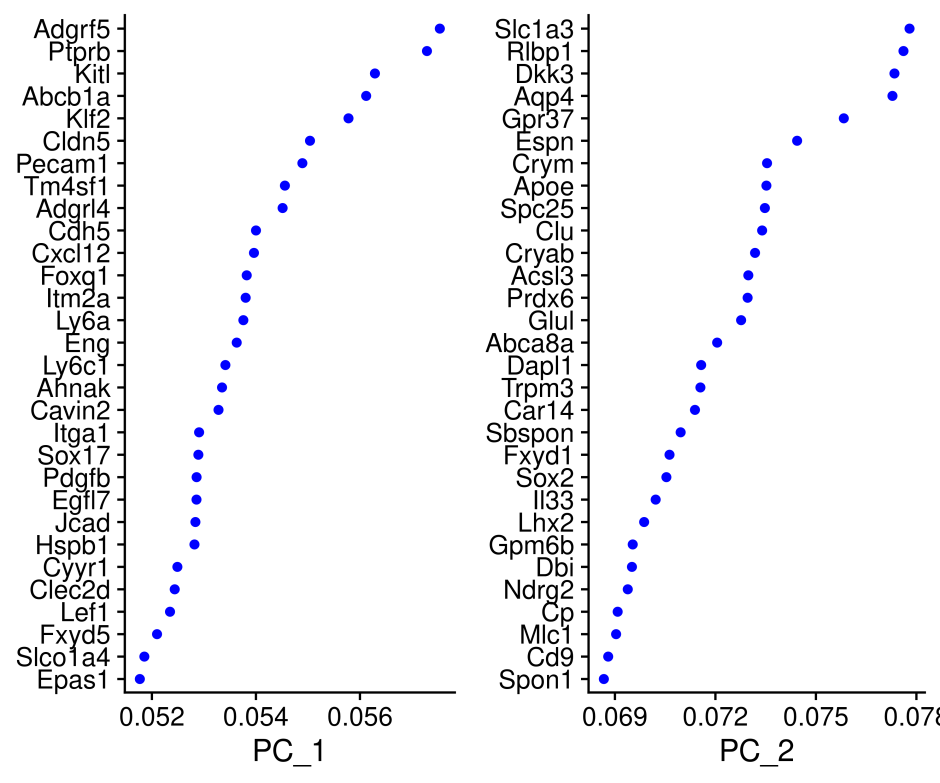


双细胞确定结果图

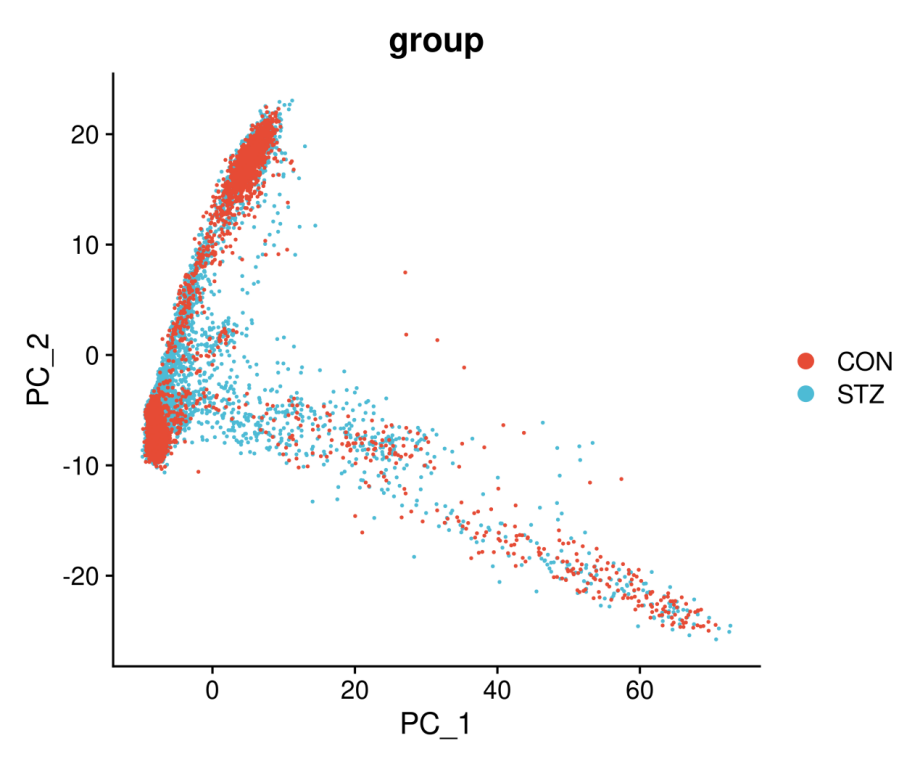
### 2.1.4、主成分分析（PCA）

用Seurat包中的ScaleData对数据进行线性缩放后，再用RunPCA对数据执行PCA进行降维分析。并筛选SD较大的PC进行后续分析。

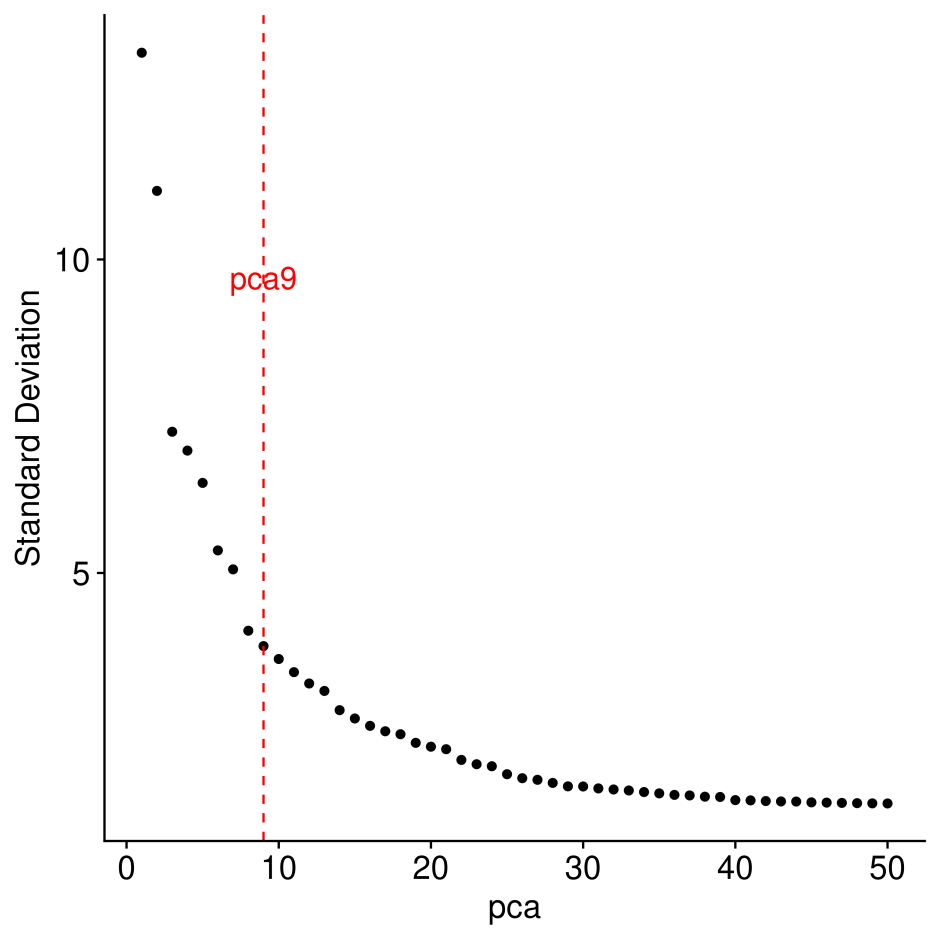
结果文件夹：cluster



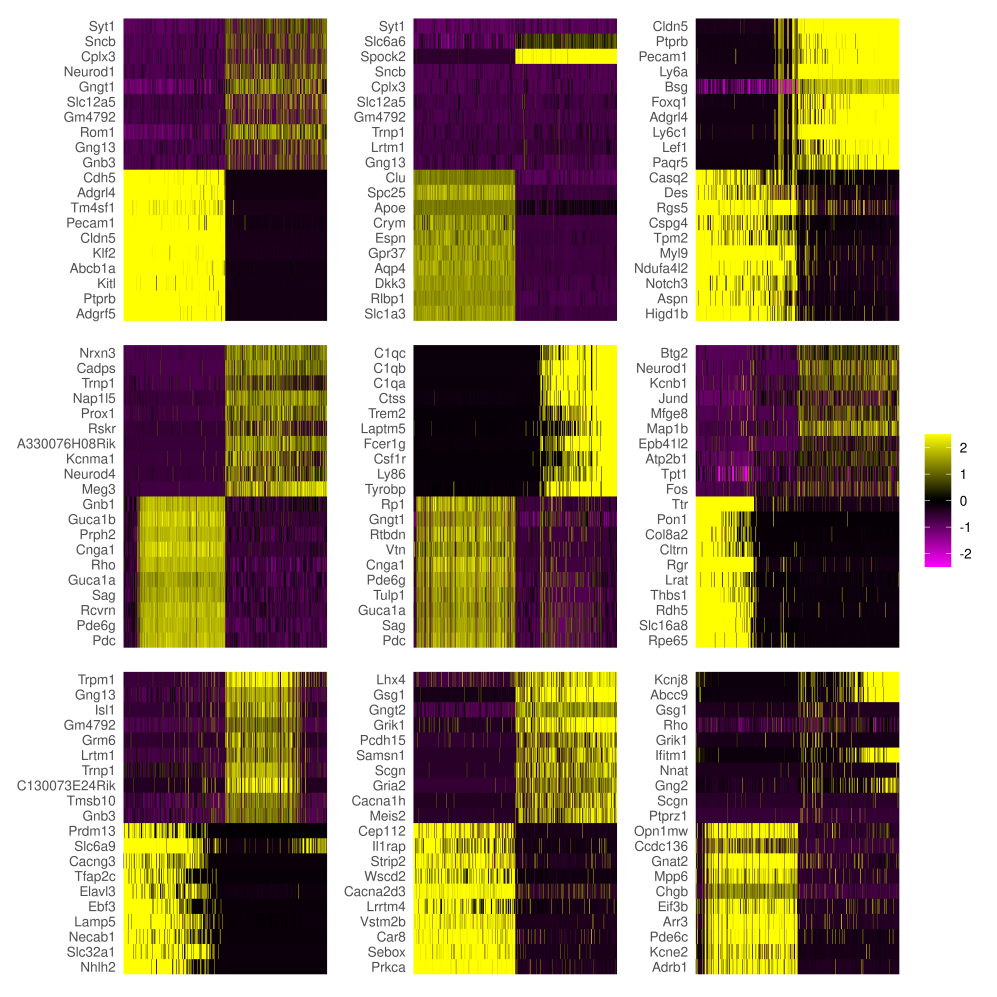
Top2主成分基因贡献度



PCA图



Pc筛选图（top9 pc）

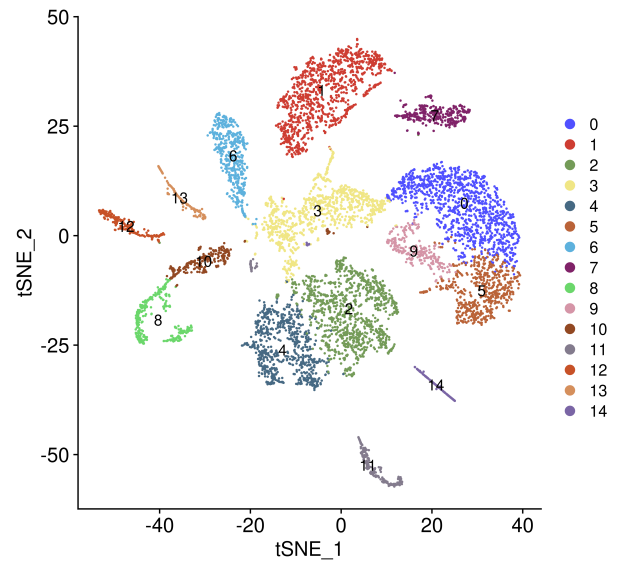
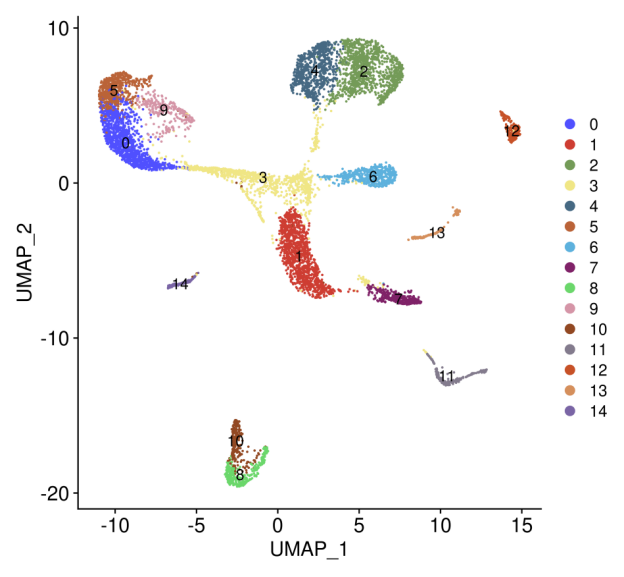


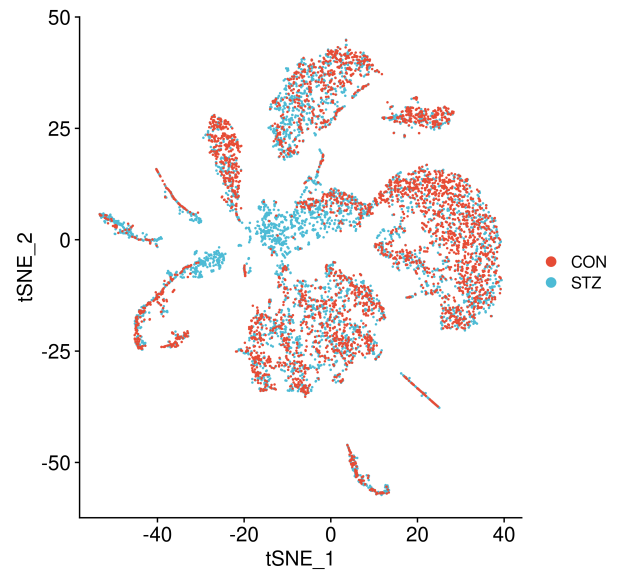
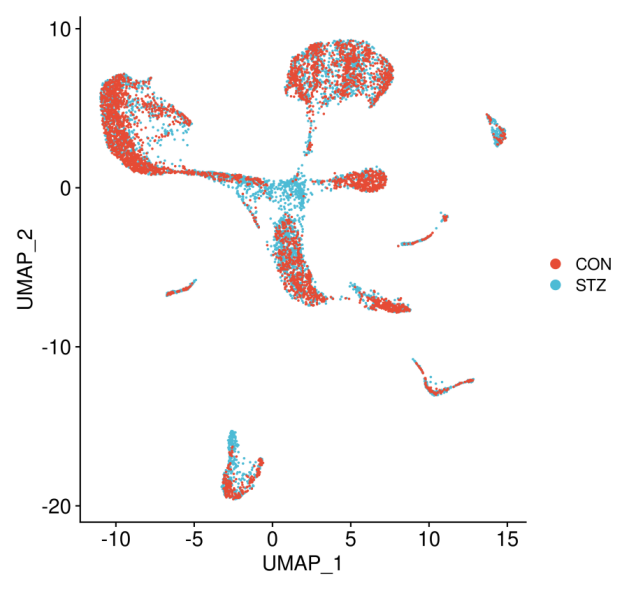
以及Top9主成分热图。

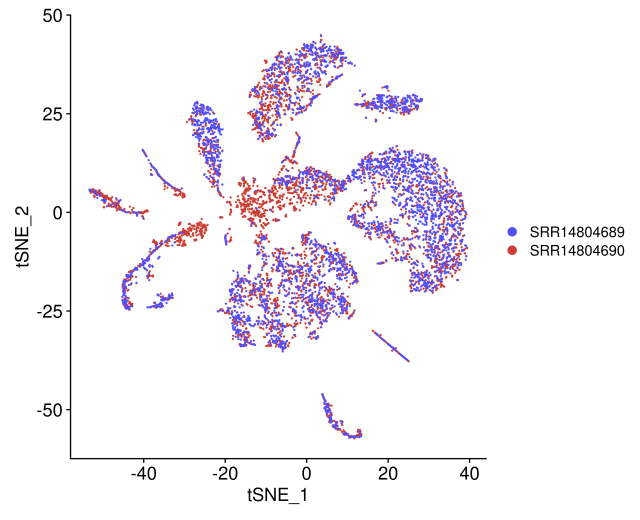
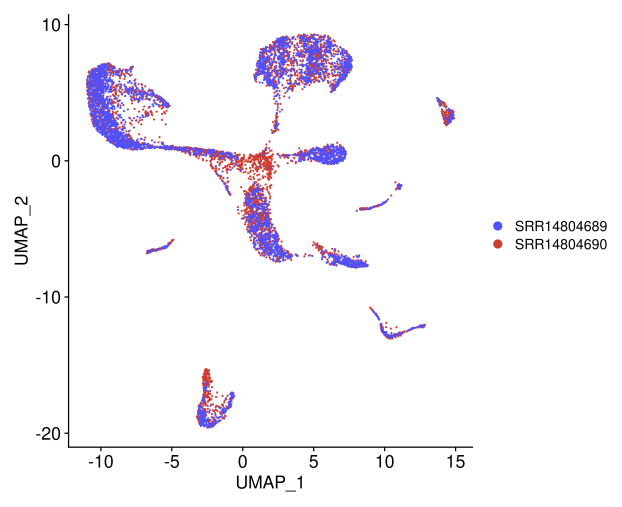
### 2.1.5、细胞聚类分群

这里利用UMAP和tSNE法，根据上述PCA中筛选出的PC对细胞进行聚类并标注展示。然后计算各聚类中的marker基因。

结果文件夹：cluster





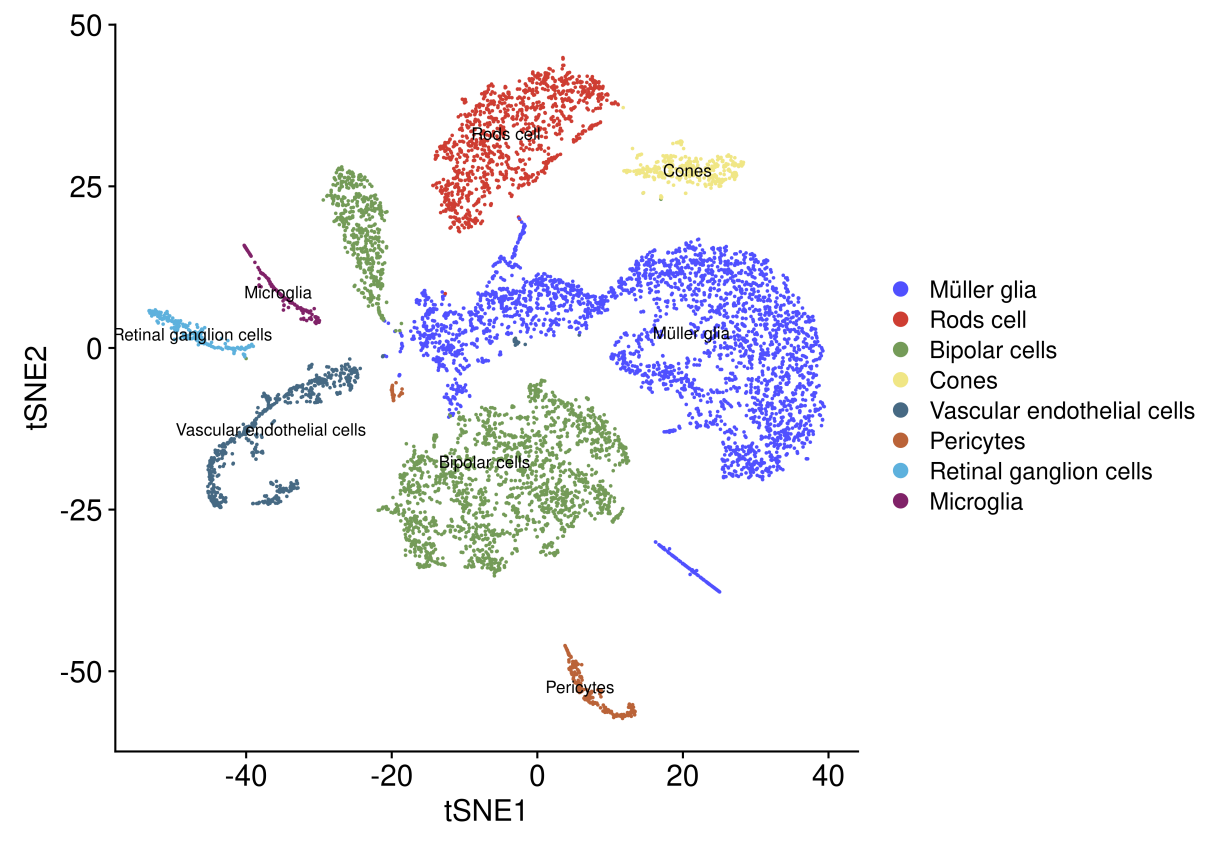
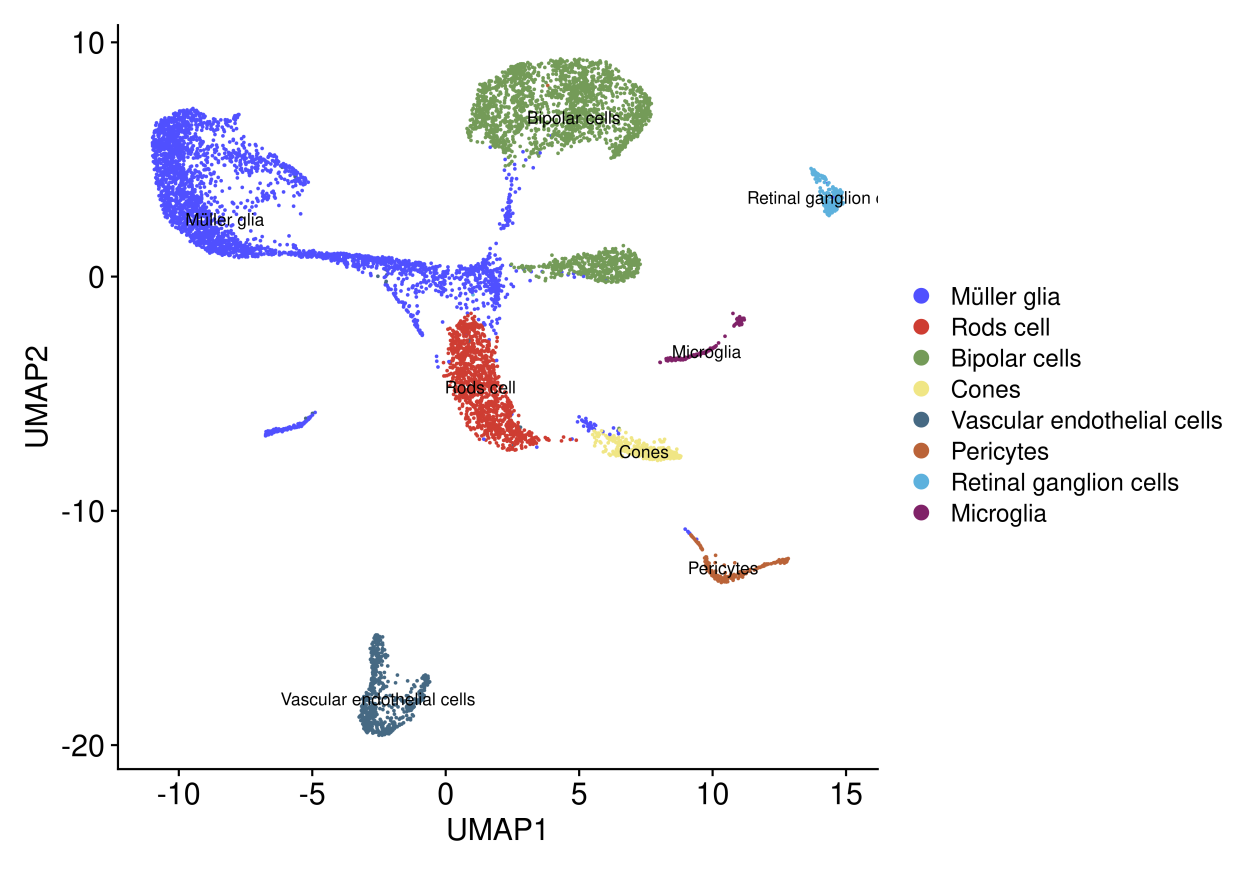


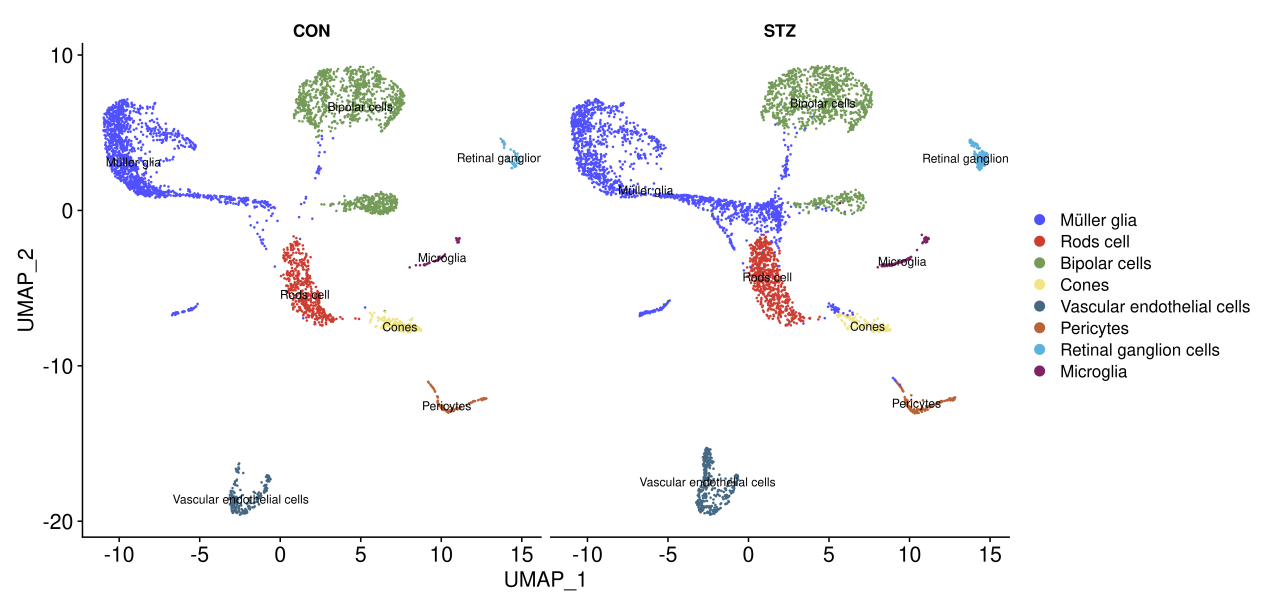
细胞聚类结果图

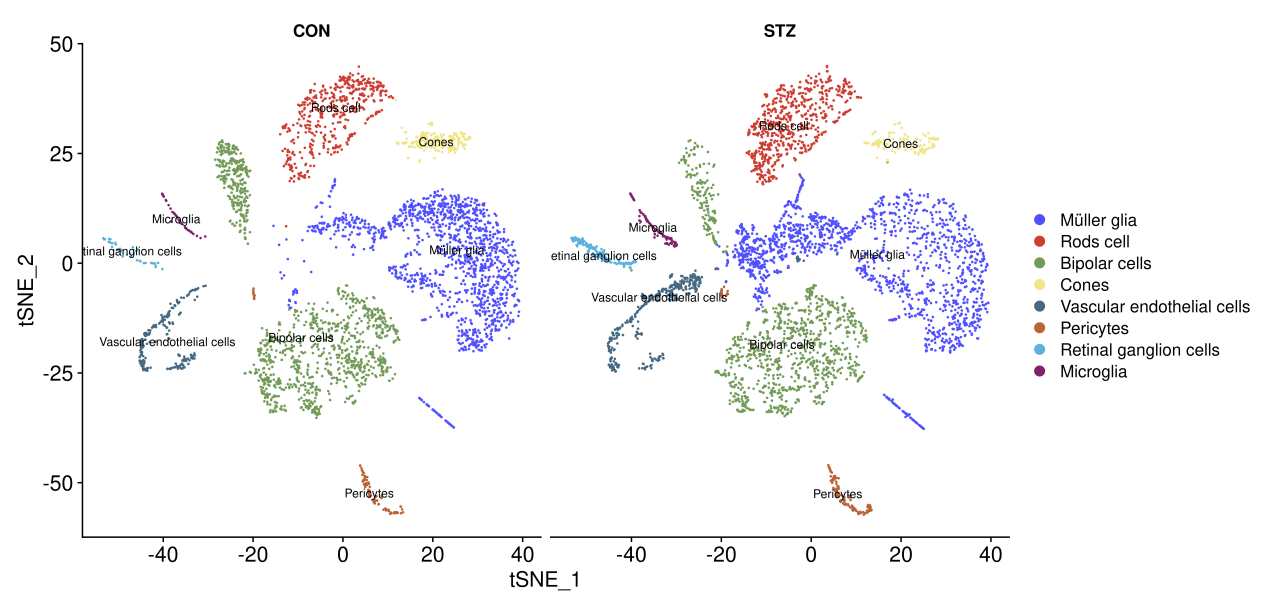
### 2.1.6、确定细胞类型

为了对聚类得到的细胞群进行身份鉴定，本研究结合细胞群经典marker及各细胞的标记基因及各细胞聚类的标记基因来确定各聚类的细胞类型。

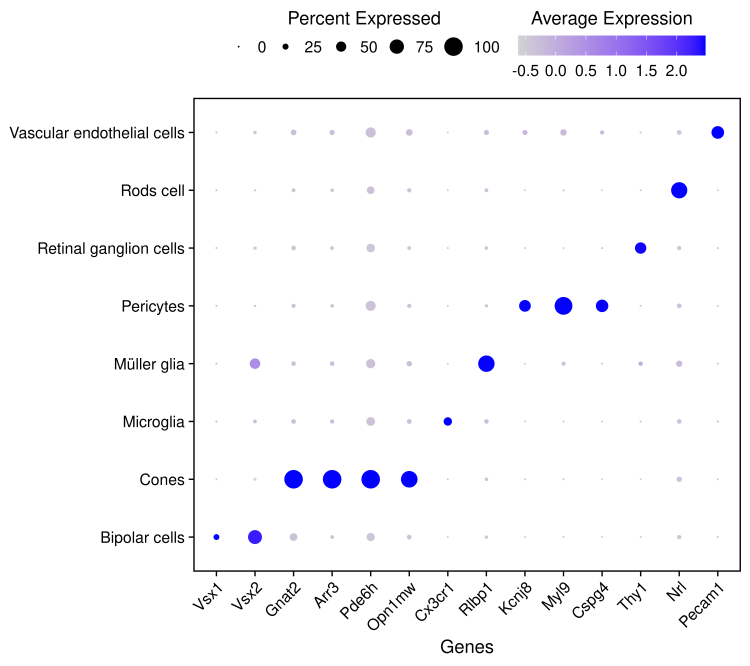
结果文件夹：cell\_type



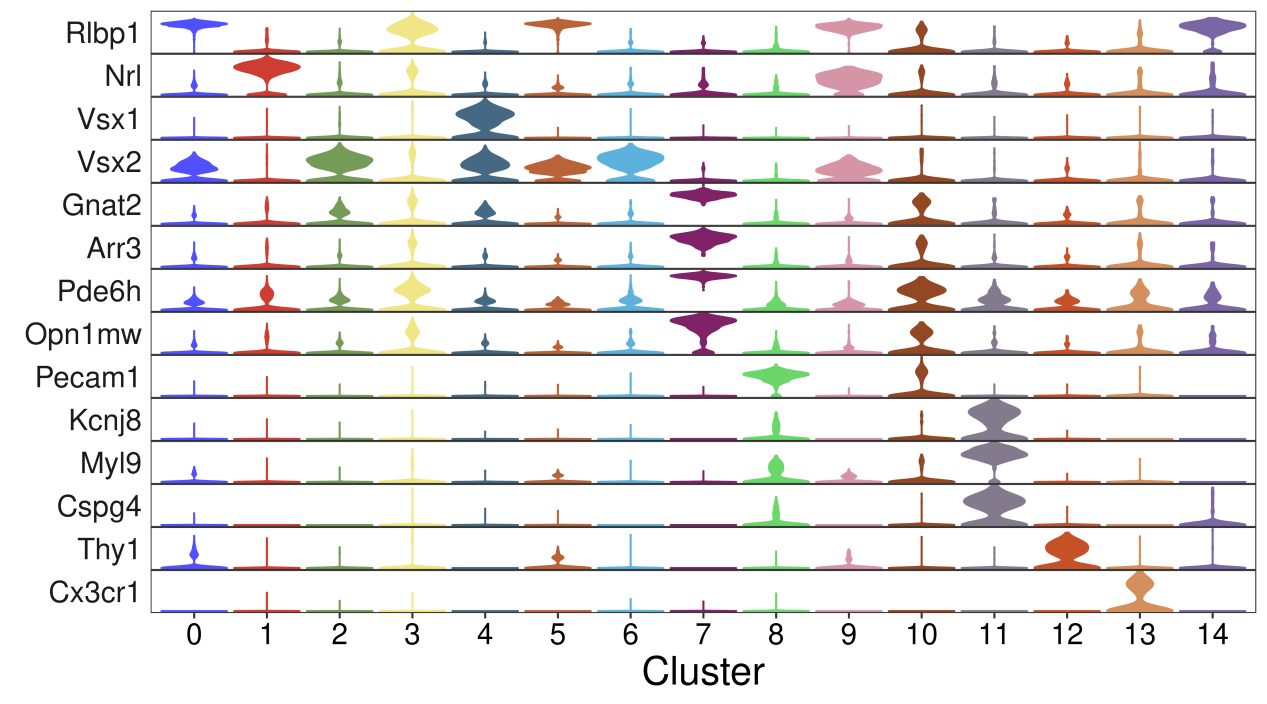


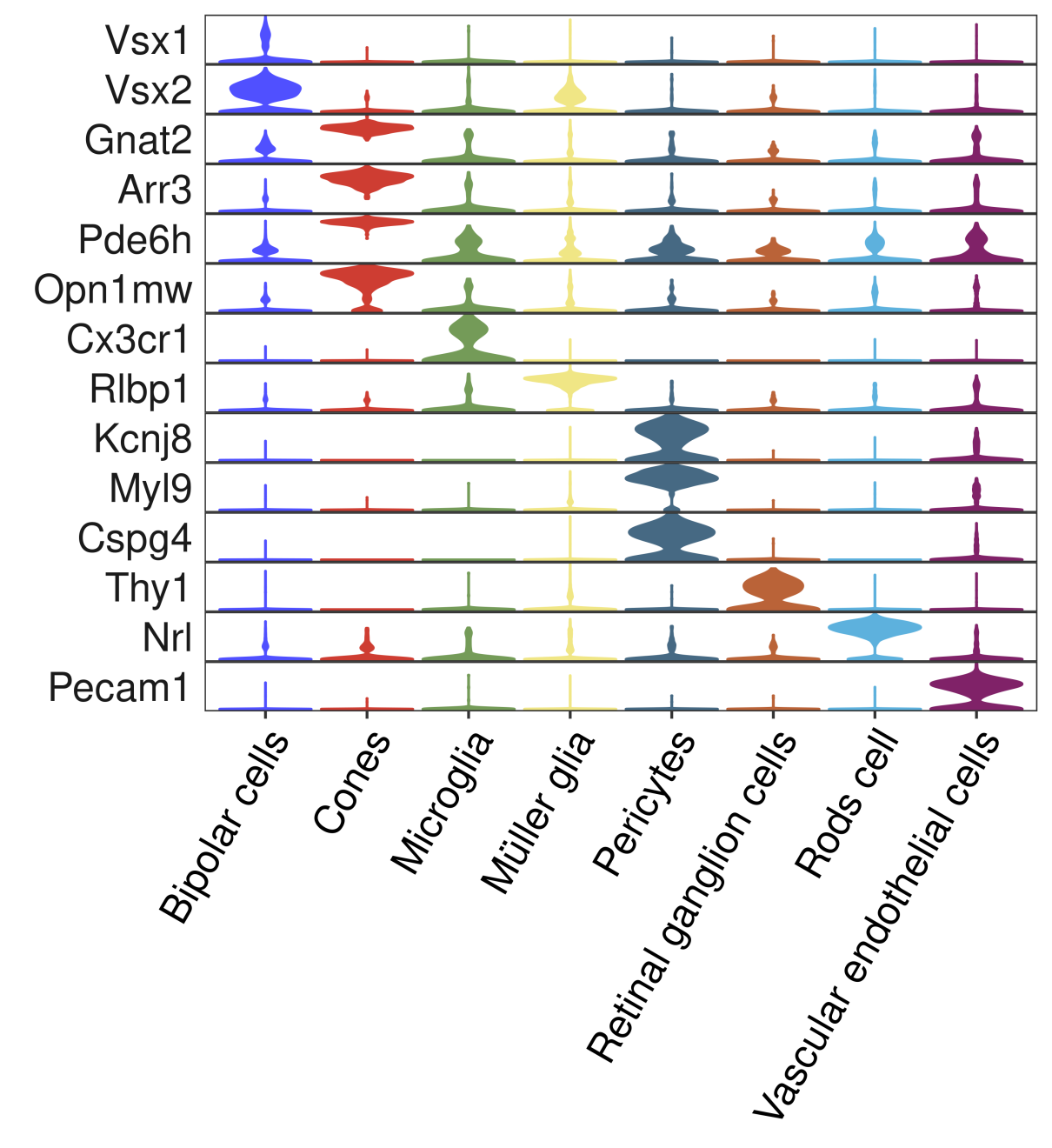


各聚类细胞类型



鉴定细胞所用marker基因在各类细胞中的表达情况图





鉴定细胞所用marker基因在各聚类和细胞中的表达情况小提琴图

**2.1.7、差异表达分析**

这里根据immune cells中各细胞的表达及分组情况，利用Seurat包中的FindMarkers进行差异表达分析，并筛选出差异表达基因（FC>1.5，p<0.05）。这里展示的结果是利用Seurat包中的FindMarkers对Müller glia细胞进行差异表达分析。

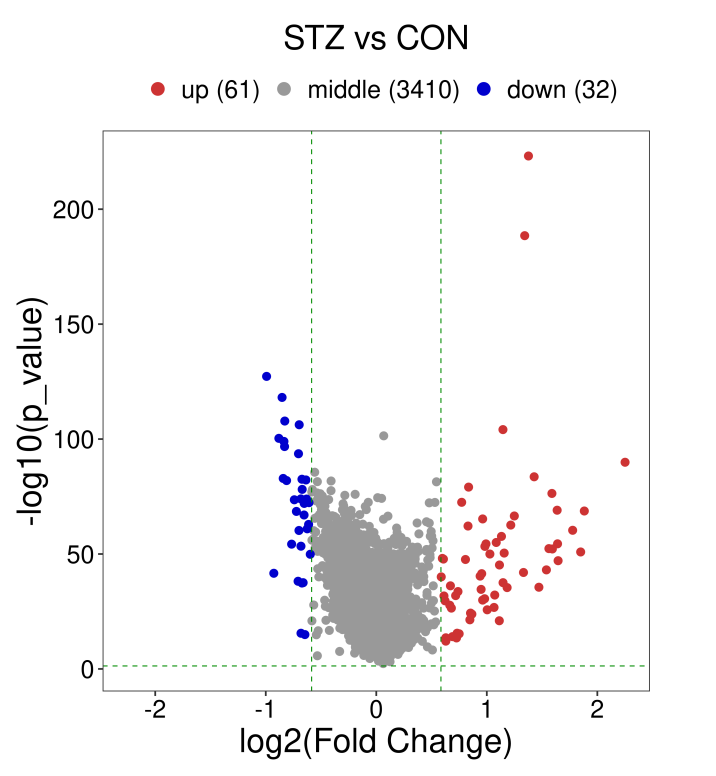
结果文件夹：diff/Müller\_glia/diff\_exp

差异基因统计总览表

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **数据信息** | | **差异阈值** | | **差异基因数** | | **样本数** | |
| **数据来源** | **平台** | **FC** | **q\_value** | **Up** | **Down** | **STZ** | **CON** |
| GSE178121 | 测序 | 1.5 | 0.05 | 61 | 32 | 1 | 1 |

STZ vs CON 部分差异表达结果表

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Gene\_Name** | **log2FC** | **FC** | **p\_value** | **q\_value** |
| Pcp2 | 2.250954738 | 4.759977447 | 1.23392E-90 | 2.29977E-86 |
| Pde6h | 1.88217492 | 3.686303665 | 1.98411E-69 | 3.69797E-65 |
| Opn1mw | 1.849189503 | 3.602977152 | 1.30103E-51 | 2.42487E-47 |
| Gng13 | 1.776723174 | 3.426470281 | 4.86119E-61 | 9.06029E-57 |
| Pcp4 | 1.644955623 | 3.127382372 | 8.23715E-48 | 1.53524E-43 |



差异表达火山图

**2.1.8、差异表达基因功能富集分析**

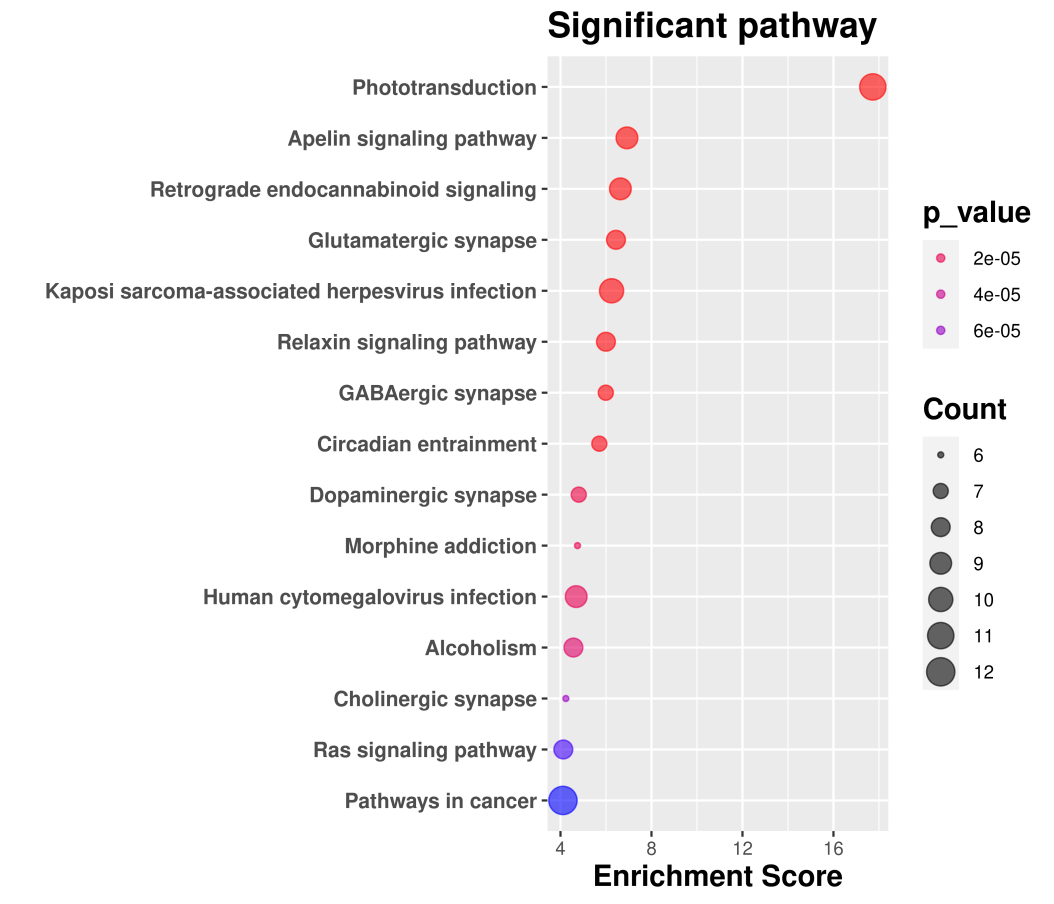
这里利用上述Müller glia差异表达基因进行GO、KEGG富集分析。这里只以Müller glia中KEGG富集结果进行展示，其他结果见相应文件。基因功能分3类：Biological Process (BP)、Molecular Function (MF)、 Cellular Component (CC)。

结果文件夹： diff/Müller\_glia/functions

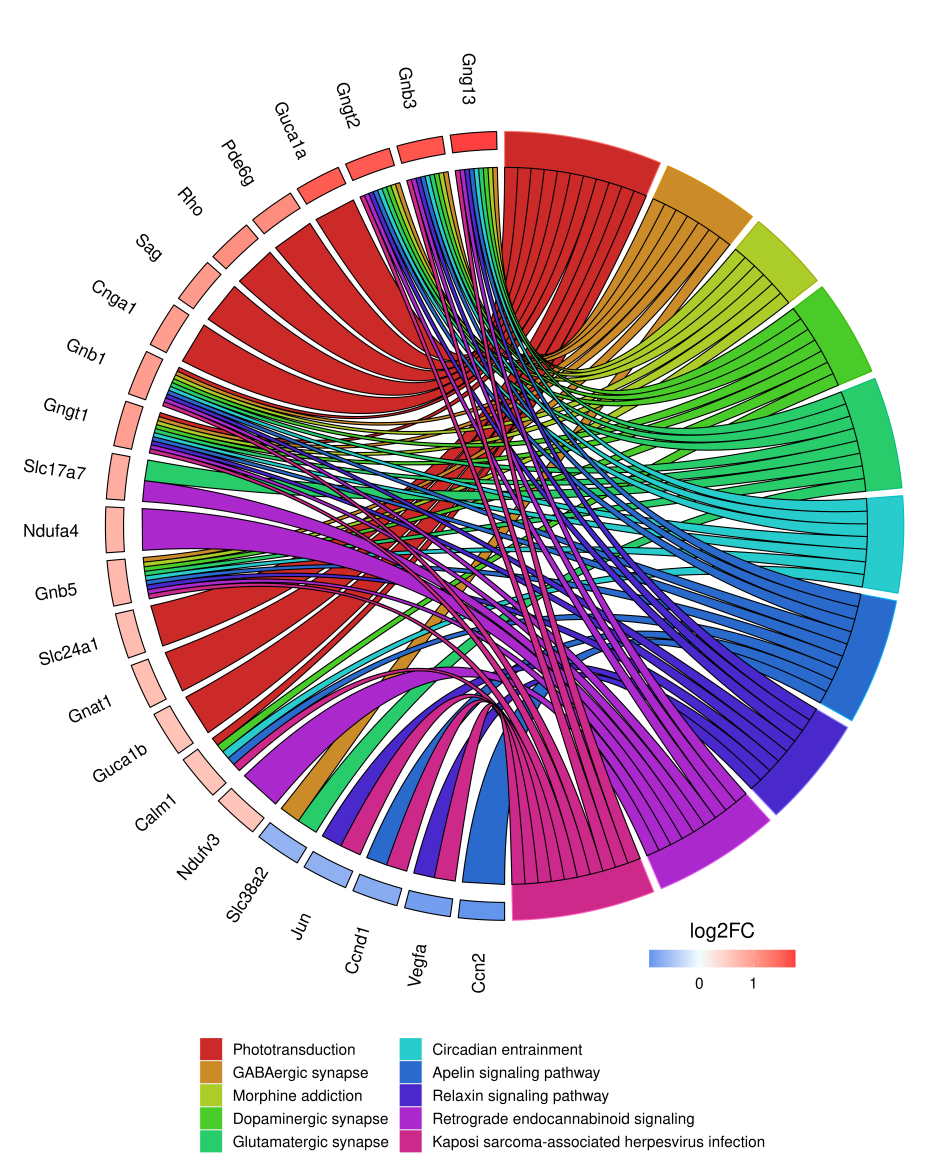
注：这里只以差异交集基因中KEGG图例富集结果进行展示，其他结果见相应文件。

KEGG富集部分结果表

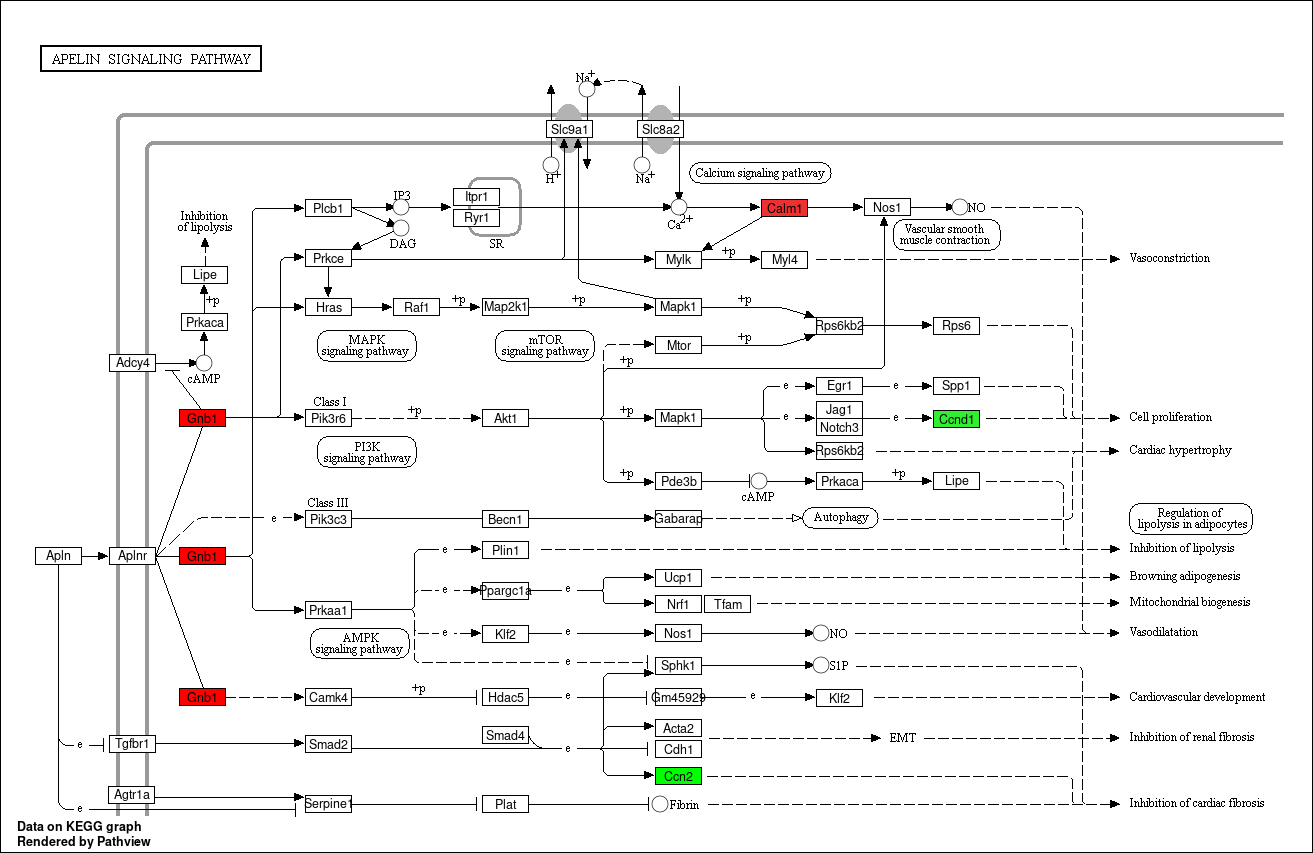
|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **ID** | **Term** | **p\_value** | **fdr** | **Enrichment\_Score** |
| mmu04744 | Phototransduction | 1.86E-18 | 2.87E-16 | 17.72932297 |
| mmu04371 | Apelin\_signaling\_pathway | 1.20E-07 | 9.23E-06 | 6.921200304 |
| mmu04723 | Retrograde\_endocannabinoid\_signaling | 2.33E-07 | 1.20E-05 | 6.632637994 |
| mmu04724 | Glutamatergic\_synapse | 3.64E-07 | 1.40E-05 | 6.439336212 |
| mmu05167 | Kaposi\_sarcoma-associated\_herpesvirus\_infection | 5.68E-07 | 1.75E-05 | 6.245861979 |
| mmu04926 | Relaxin\_signaling\_pathway | 1.01E-06 | 2.24E-05 | 5.997763125 |



富集结果图气泡图



前10个条目对应基因圈图

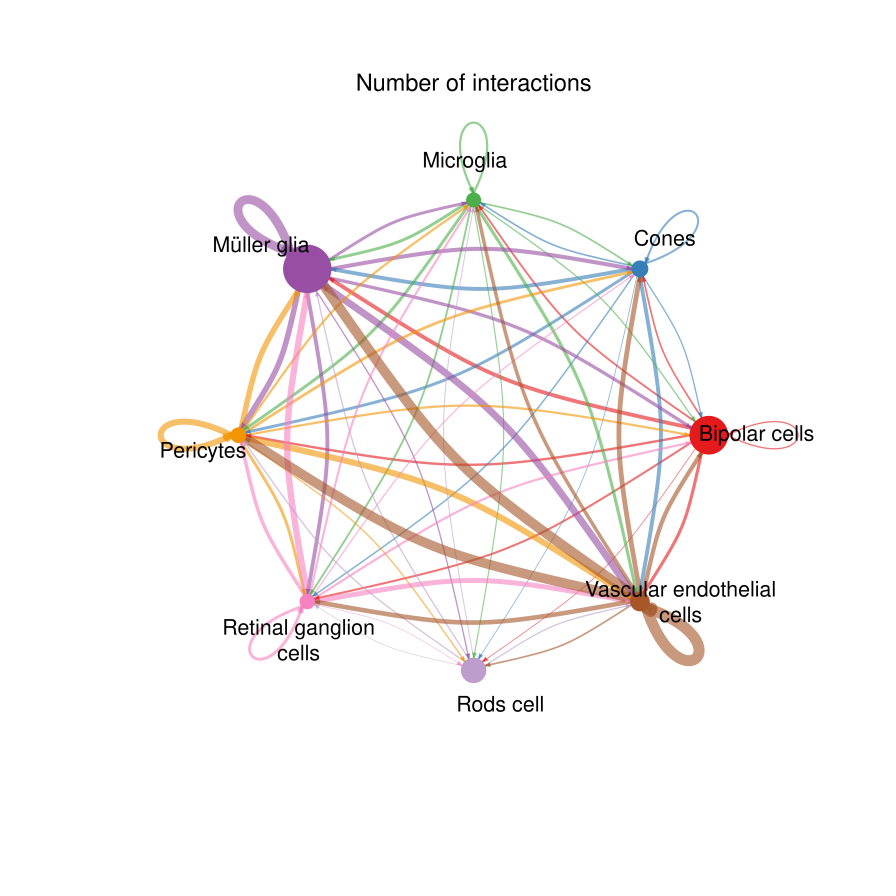


基因在KEGG通路中的位置图（在STZ vs CON中红色是上调基因，绿色下调）

### 2.1.9、细胞互作网络分析

为了探究疾病微环境中的细胞互作模式，本研究基于细胞互作分析探究了核心细胞亚群其他细胞群之间的crosstalk。这里用CellChat v1.1，根据各细胞的表达情况进行细胞与细胞间的通信分析。然后筛选出Müller glia相关的互作细胞以及相应的ligand-receptor关系对。

结果文件夹：cell-cell



细胞间互作网络图

注：连线宽度代表配体受体关系对数量（越宽关系对越多）

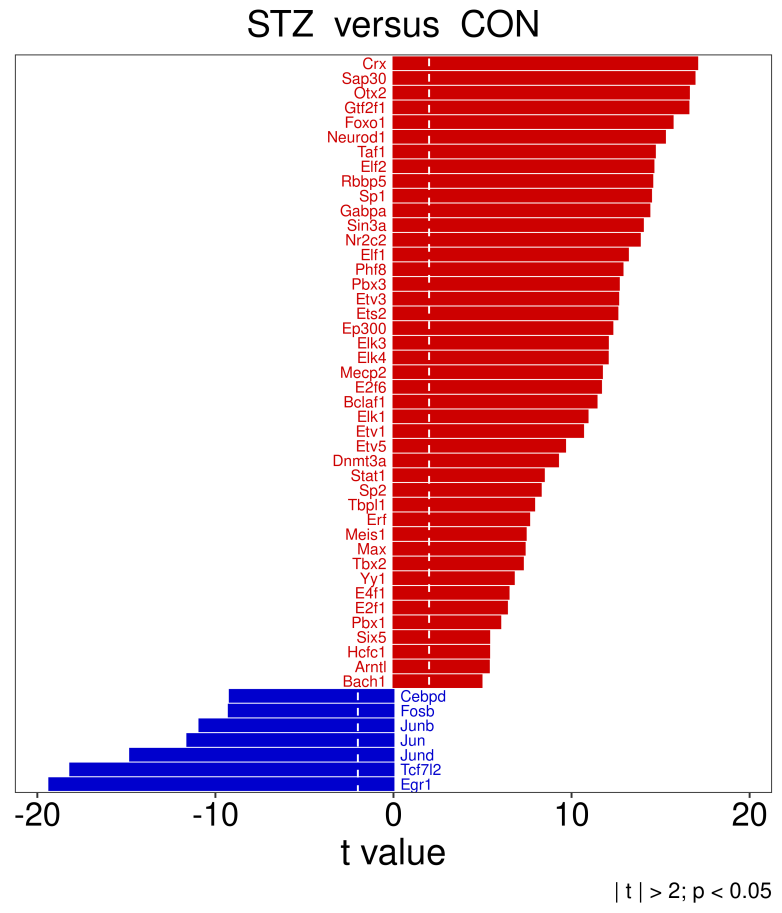
部分Müller glia相关的互作细胞以及相应的ligand-receptor关系对表

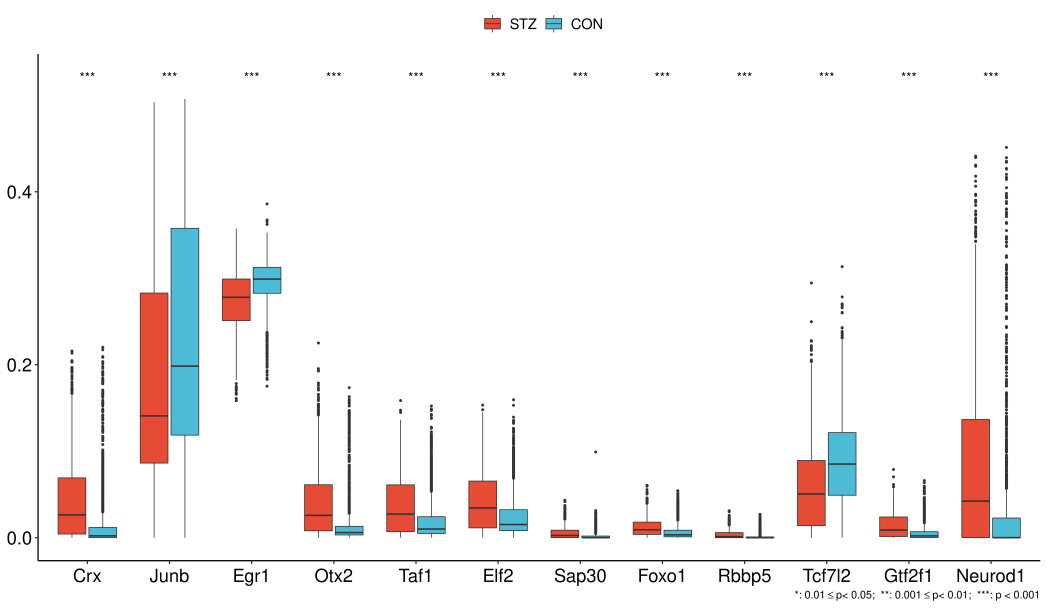
|  |  |  |  |
| --- | --- | --- | --- |
| **source** | **target** | **ligand** | **receptor** |
| Müller glia | Bipolar cells | Igf2 | Igf1r |
| Müller glia | Cones | Igf2 | Igf1r |
| Müller glia | Microglia | Igf2 | Igf1r |
| Müller glia | Müller glia | Igf2 | Igf1r |
| Müller glia | Pericytes | Igf2 | Igf1r |

### 2.1.10、转录因子调控分析

这里利用SCENIC包，根据相应的数据，推断Müller glia中对应各转录因子的调控活性。然后用t.test计算组间转录因子的调控活性差异情况（| t | > 2, p < 0.05）。

结果文件夹：scenic

Müller glia 细胞中组间scenic得分差异图（t绝对值top50）



Müller glia细胞DR与对照差异基因的转录因子差异结果图

**参考文献**

[1] LUN A T L, RIESENFELD S, ANDREWS T, et al. EmptyDrops: distinguishing cells from empty droplets in droplet-based single-cell RNA sequencing data [J]. Genome Biology, 2019, 20(1): 63.

[2] MCCARTHY D J, CAMPBELL K R, LUN A T L, et al. Scater: pre-processing, quality control, normalization and visualization of single-cell RNA-seq data in R [J]. Bioinformatics (Oxford, England), 2017, 33(8): 1179-86.

[3] HAO Y, HAO S, ANDERSEN-NISSEN E, et al. Integrated analysis of multimodal single-cell data [J]. bioRxiv, 2020, 2020.10.12.335331.

[4] PUTHUMANA J, THIESSEN-PHILBROOK H, XU L, et al. Biomarkers of inflammation and repair in kidney disease progression [J]. The Journal of Clinical Investigation, 2021, 131(3):

[5] ASHBURNER M, BALL C A, BLAKE J A, et al. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium [J]. Nature genetics, 2000, 25(1): 25-9.

[6] KANEHISA M, GOTO S. KEGG: kyoto encyclopedia of genes and genomes [J]. Nucleic acids research, 2000, 28(1): 27-30.

[7] JIN S, GUERRERO-JUAREZ C F, ZHANG L, et al. Inference and analysis of cell-cell communication using CellChat [J]. Nature communications, 2021, 12(1): 1088.

[8] AIBAR S, GONZÁLEZ-BLAS C B, MOERMAN T, et al. SCENIC: single-cell regulatory network inference and clustering [J]. Nature Methods, 2017, 14(11): 1083-6.