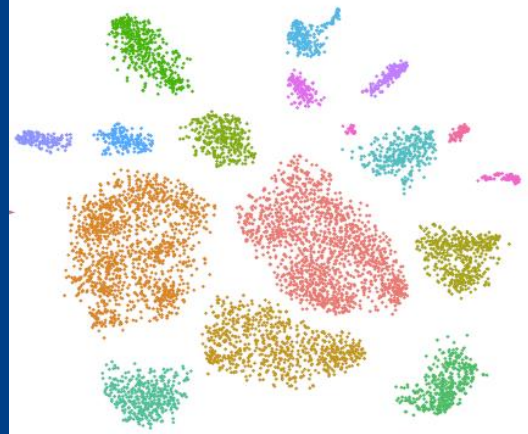
浙江大学
ZHEJIANG UNIVERSITY

# 单细胞其他组学数据分析

——空间转录组、表观组



褚琴洁　qinjiechu@zju.edu.cn

2023年11月6日

# 课件PPT

**http://ibi.zju.edu.cn/bioinplant/courses/scomics/**

所有课程示例数据和代码

链接：https://pan.baidu.com/s/1rmtjWqArlODqeRmn5_pEYA

提取码：scoo

# 上节回顾

- Seurat使用巩固
- 数据整合（批次效应矫正）
- 差异基因富集分析
- 拟时序分析
- 细胞通讯分析
- ……





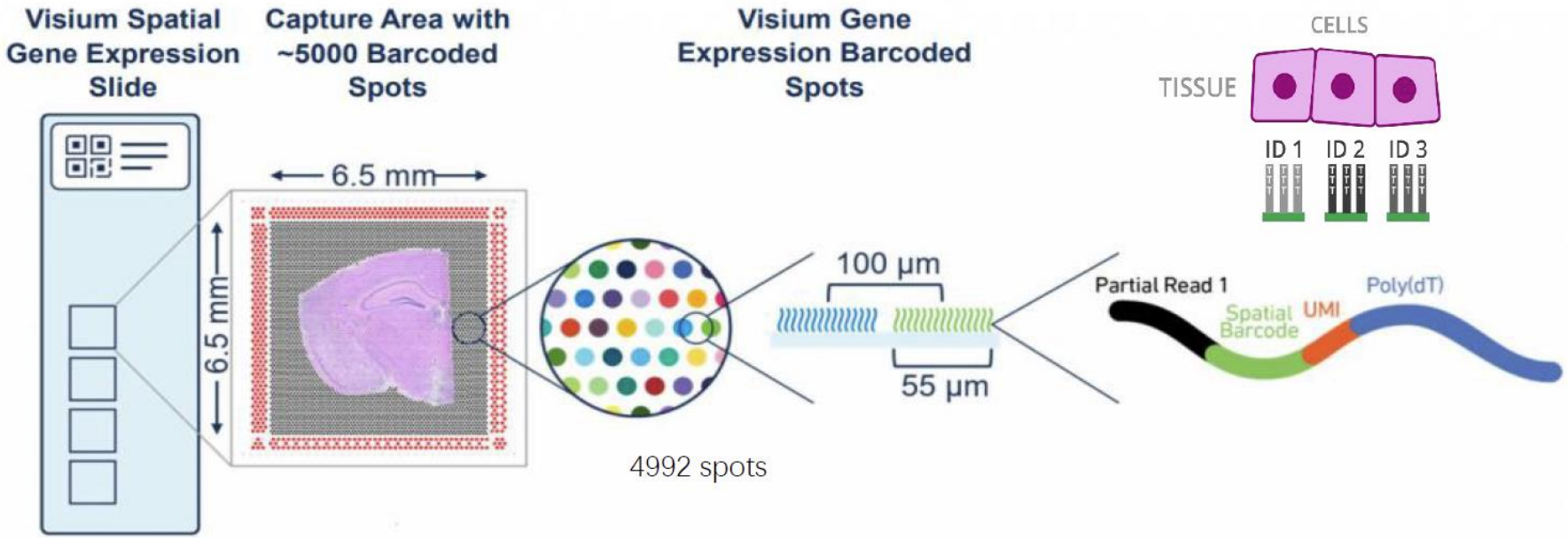






# 上节回顾：10X Visium

**Visium：**在组织原位检测全转录组基因表达的一种技术，使得我们在检测基因表达水平的同时，获得基因在组织内部空间表达的位置信息。Visium平台由Spatial transcriptomics技术发展而来。




- 将新鲜的冷冻组织切片成像，观察组织结构，将切片置于含有RNA捕获探针载玻片上。
- 对组织切片进行固定和透化，使RNA释放，结合相应的捕获探针。
- 以捕获的RNA为模板合成cDNA，制备测序文库，将制备好的文库上机测序并进行数据可视化。

# 问题与分析内容

- 空间转录组数据分析存在的问题

## 问题

- 如何将空间数据与表达数据关联在一起?
- 有了空间转录组数据，如何与单细胞转录组数据联用?
- 做了多层切片如何展示真实的三维空间的转录本信息?

## 分析内容

- 标准化 Normalization
- 降维和聚类 Dimensional reduction and clustering
- 检测空间可变特征 Detecting spatially-variable features
- 交互式可视化 Interactive visualization
- 与单细胞RNA-seq数据的整合 Integration with single-cell RNA-seq data
- 使用多个切片 Working with multiple slices

# Seurat 分析 10X Visium 数据

（1）数据来源：小鼠脑切片，有两个连续的前部切片和两个（匹配的）连续的后部切片

Seurat 中存储空间数据的格式：

- Gene-spot表达矩阵（spot可能包含多个细胞）
- 数据采集时通过HE染色获得的组织切片图像
- 将原始高分辨率图像与用于可视化的低分辨率图像关联的缩放因子

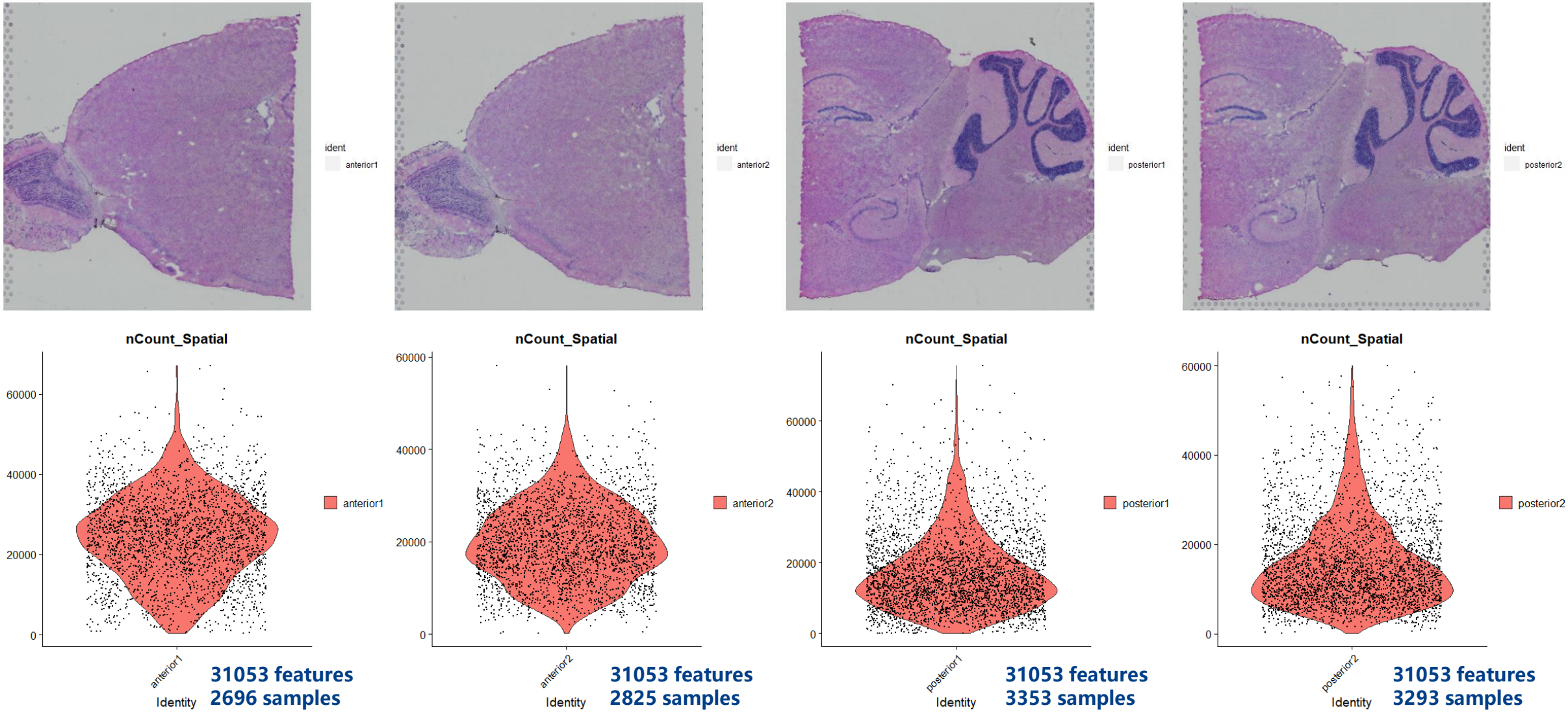
```
> brain
An object of class Seurat
31053 features across 2696 samples within 1 assay
Active assay: Spatial (31053 features, 0 variable features)
 1 image present: anterior1
```

```
> brain@assays$Spatial[1:4,1:4]
4 x 4 sparse Matrix of class "dgCMatrix"
       AAACAAGTATCTCCCA-1 AAACACCAATAACTGC-1 AAACAGAGCGACTCCT-1 AAACAGCTTTCAGAAG-1
Xkr4                    .                  .                  .                  .
Gm1992                  .                  .                  .                  .
Gm37381                 .                  .                  .                  .
Rp1                     .                  .                  .                  .
```

| | | |
|---|---|---|
| brain | S4 [31053 x 2696] (Seurat::Seurat | S4 object of class Seurat |
| assays | list [1] | List of length 1 |
| Spatial | S4 [31053 x 2696] (Seurat::Assay) | S4 object of class Assay |
| counts | S4 [31053 x 2696] (Matrix::dgCM | S4 object of class dgCMatrix |
| data | S4 [31053 x 2696] (Matrix::dgCM | S4 object of class dgCMatrix |
| scale.data | double [0 x 0] | |
| key | character [1] | 'spatial_' |
| assay.orig | NULL | Pairlist of length 0 |
| var.features | logical [0] | |
| meta.features | list [31053 x 0] (S3: data.frame) | A data.frame with 31053 rows and 0 columns |
| misc | NULL | Pairlist of length 0 |
| meta.data | list [2696 x 5] (S3: data.frame) | A data.frame with 2696 rows and 5 columns |
| active.assay | character [1] | 'Spatial' |
| active.ident | factor | Factor with 1 level: "anterior1" |
| graphs | list [0] | List of length 0 |
| neighbors | list [0] | List of length 0 |
| reductions | list [0] | List of length 0 |
| images | list [1] | List of length 1 |
| anterior1 | S4 [599 x 600] (Seurat::VisiumV1) | S4 object of class VisiumV1 |
| image | double [599 x 600 x 3] | 0.722 0.722 0.722 0.718 0.718 0.718 0.722 0.718 0.722 0.722 0.718 0.718 0.722 0. ... |
| scale.factors | list [4] (S3: scalefactors) | List of length 4 |
| spot | double [1] | 0.172117 |
| fiducial | double [1] | 144.5412 |
| hires | double [1] | 0.172117 |
| lowres | double [1] | 0.05163511 |
| coordinates | list [2696 x 5] (S3: data.frame) | A data.frame with 2696 rows and 5 columns |
| spot.radius | double [1] | 0.012439 |
| assay | character [1] | 'Spatial' |
| key | character [1] | 'anterior1_' |
| project.name | character [1] | 'anterior1' |
| misc | list [0] | List of length 0 |
| version | list [1] (S3: package_version, num | List of length 1 |
| commands | list [0] | List of length 0 |
| tools | list [0] | List of length 0 |

# Seurat 分析 10X Visium 数据

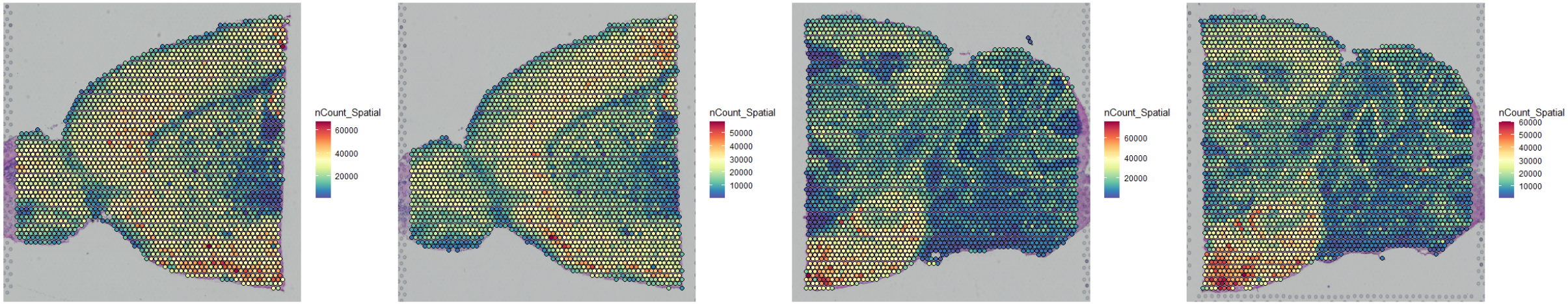
（1）数据来源：小鼠脑切片，有两个连续的前部切片和两个（匹配的）连续的后部切片



# Seurat 分析 10X Visium 数据

（2）预处理：标准化消除测序深度的影响（与处理scRNA-seq数据类似）

- 组织的细胞密度在空间上存在差异的话，会导致spot之间巨大的异质性（不同spot的分子计数的差异不仅是技术上的，还取决于组织解剖学）
- 标准化方法（LogNormalize函数）强制每个spot具有相同的文库大小，这可能会带来误差
- 建议使用sctransform 构建正则化的负二项基因表达模型



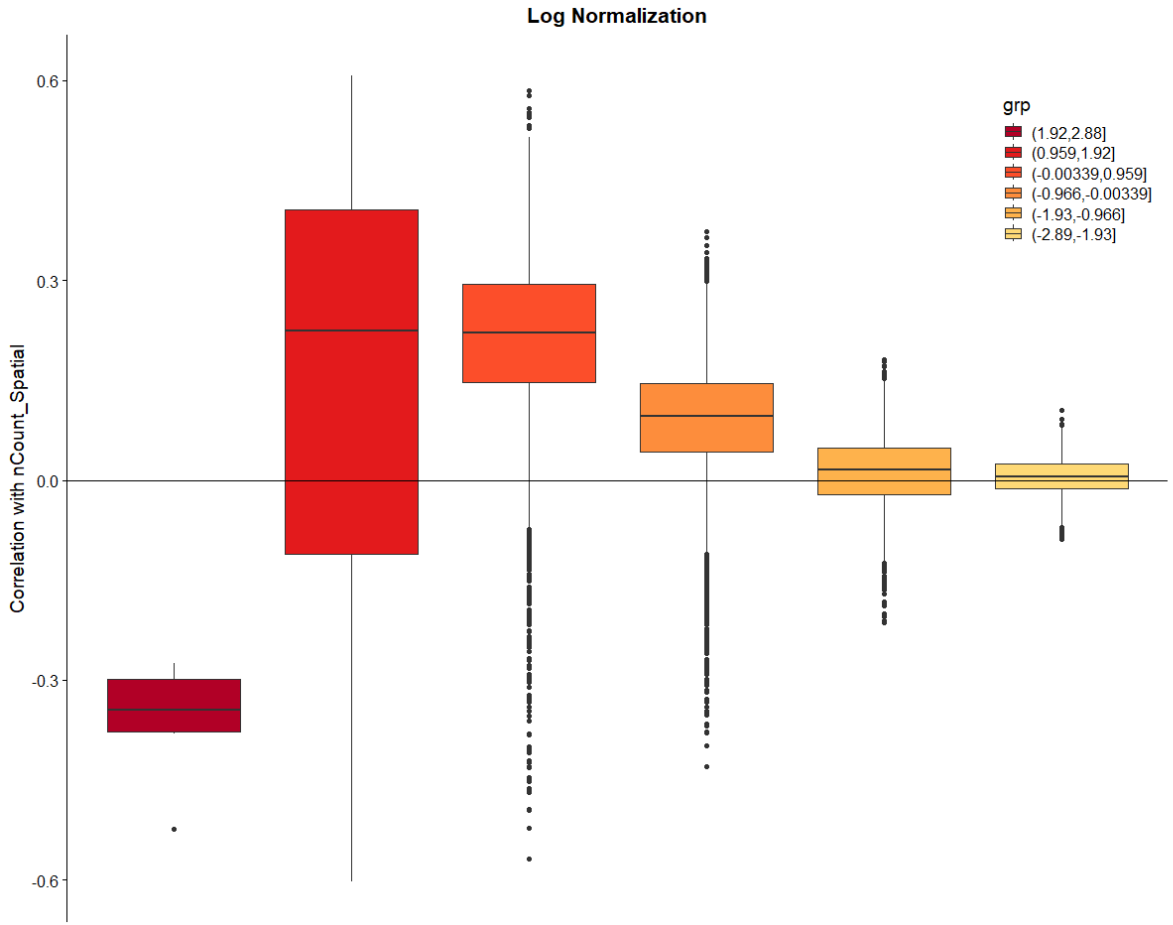

```
brain <- SCTransform(brain, assay = "Spatial", verbose = FALSE)
```

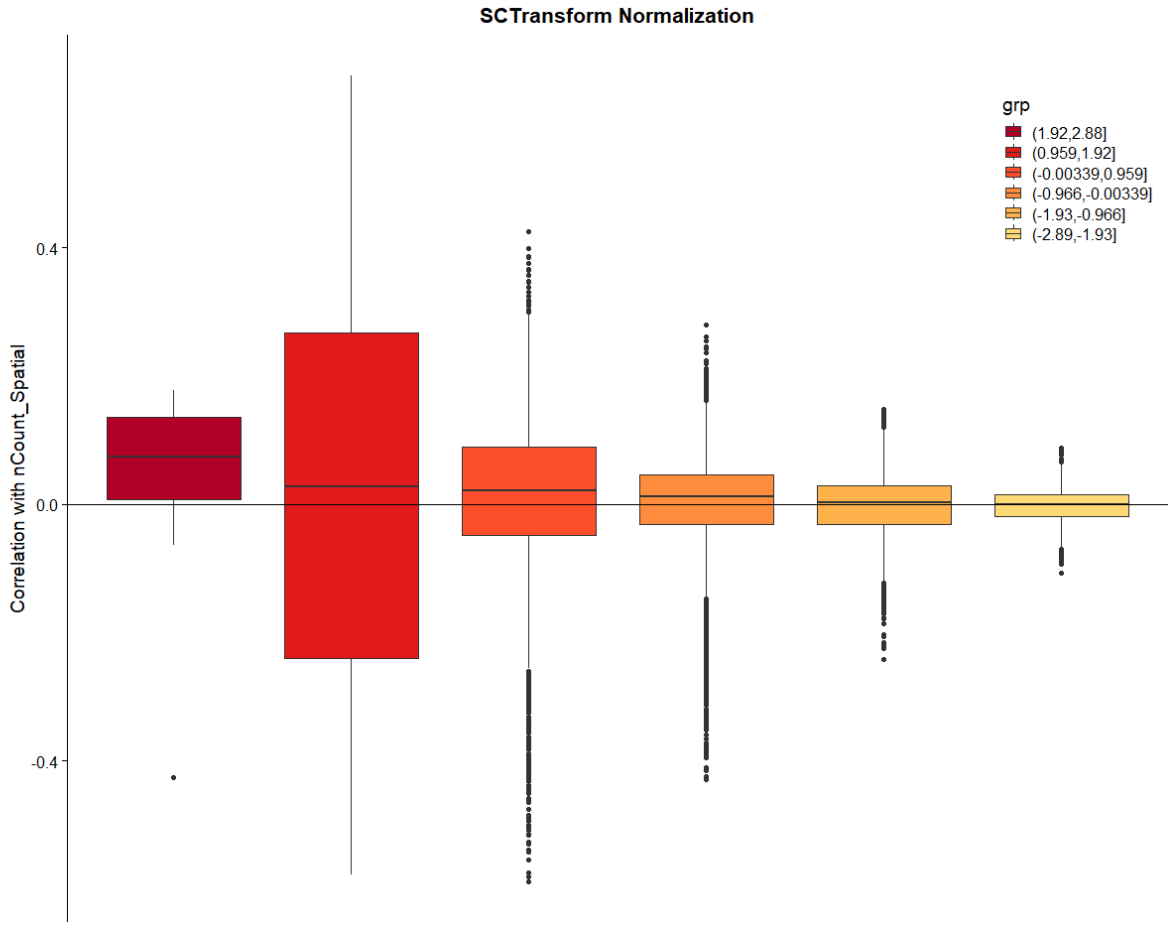
# 注意这里的默认参数：variable.features.n = 3000（默认选择的是3000个高变基因）

# Seurat 分析 10X Visium 数据

（2）预处理：标准化消除测序深度的影响（与处理scRNA-seq数据类似）

- SCTransform 与 LogNormalize 两种标准化方法之间的差异：计算每个基因与 nCount_Spatial 数量的相关性，根据基因的平均表达水平将其分组，发现LogNormalize（左）的前三组基因表达与 nCount_Spatial 相关性较强（即没有充分标准化基因表达值）

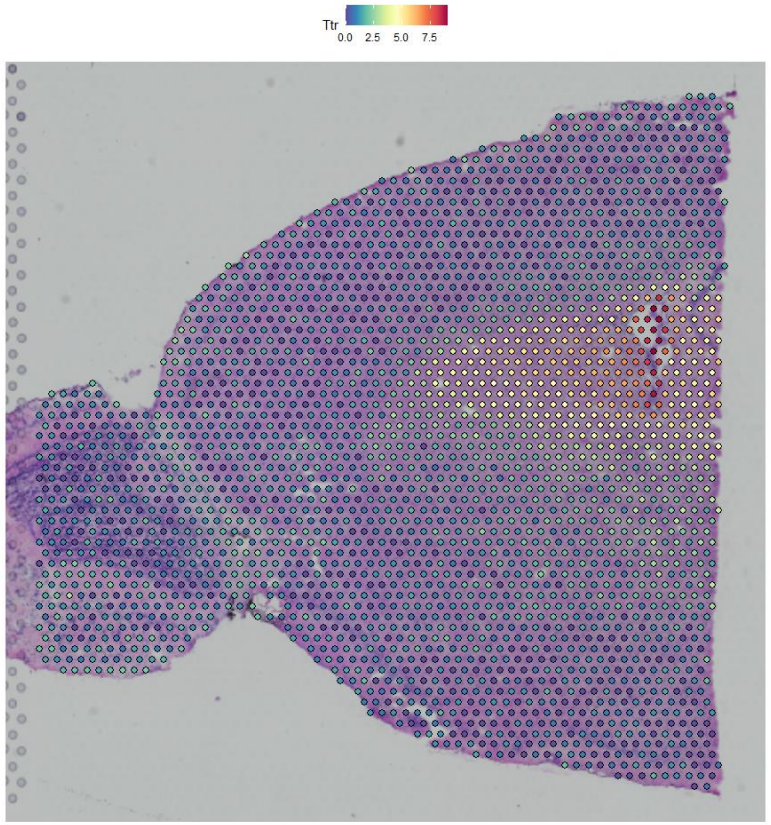


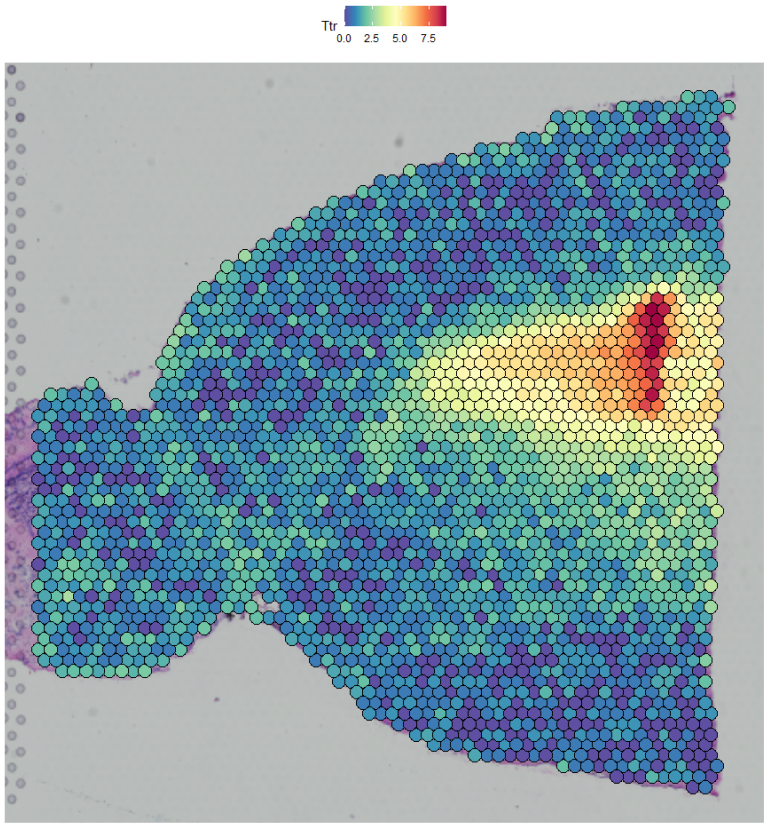
# Seurat 分析 10X Visium 数据

（2）预处理：基因表达的可视化（对于小鼠大脑数据，Hpca是海马体的标记基因，Ttr是脉络丛的标记基因）

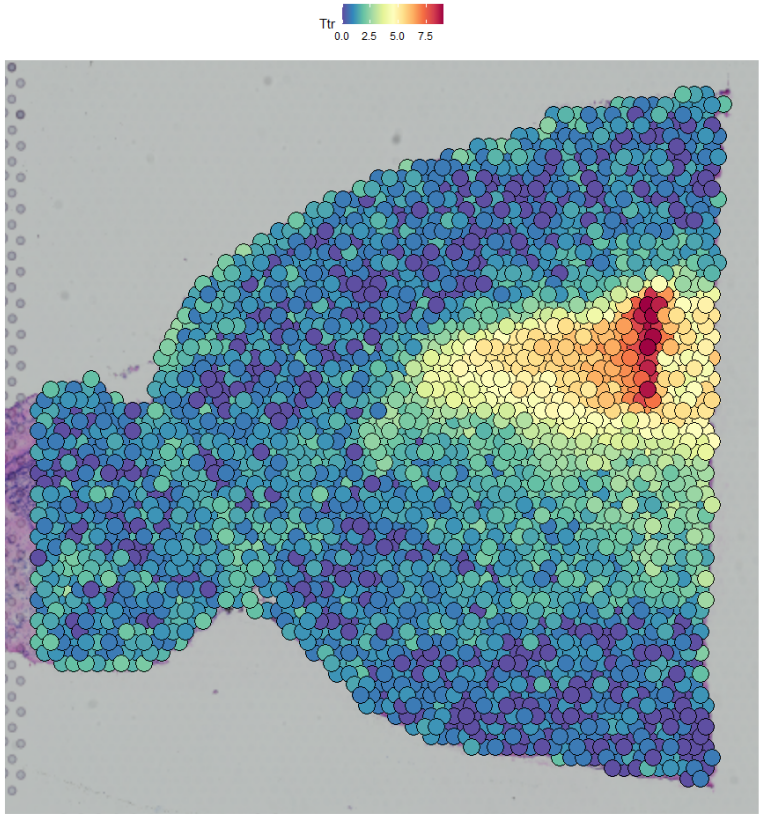
- pt.size.factor：缩放绘图斑点的大小




pt.size.factor = 1




pt.size.factor = 2

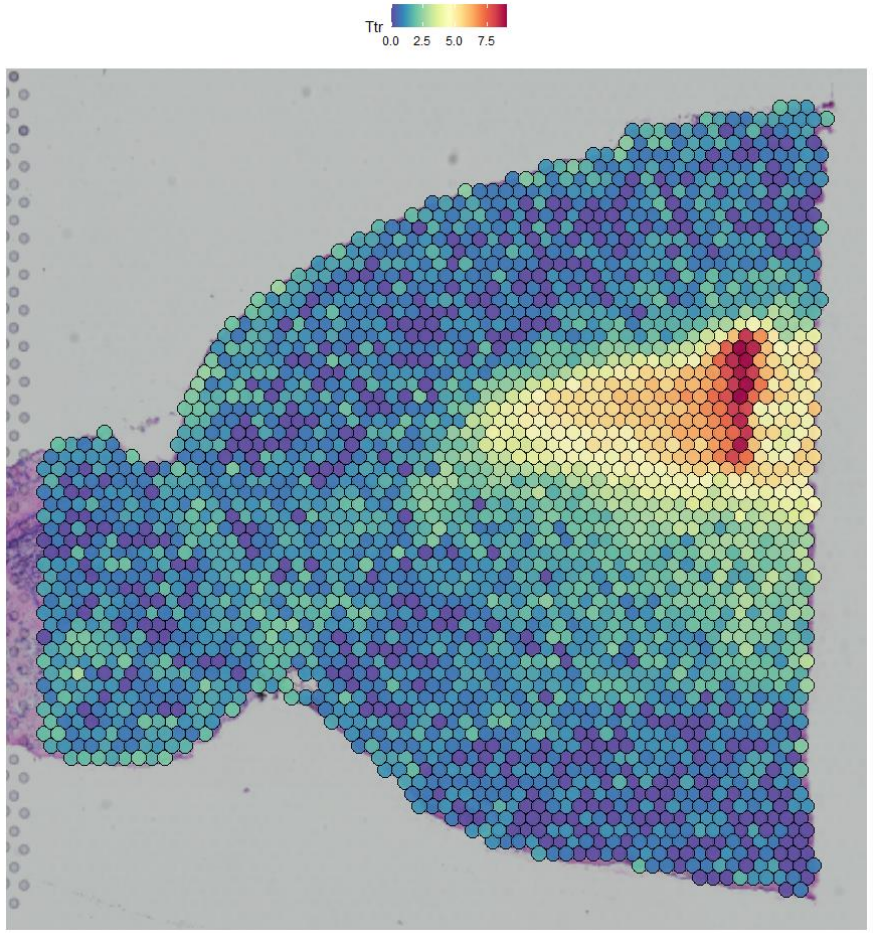



pt.size.factor = 2.5

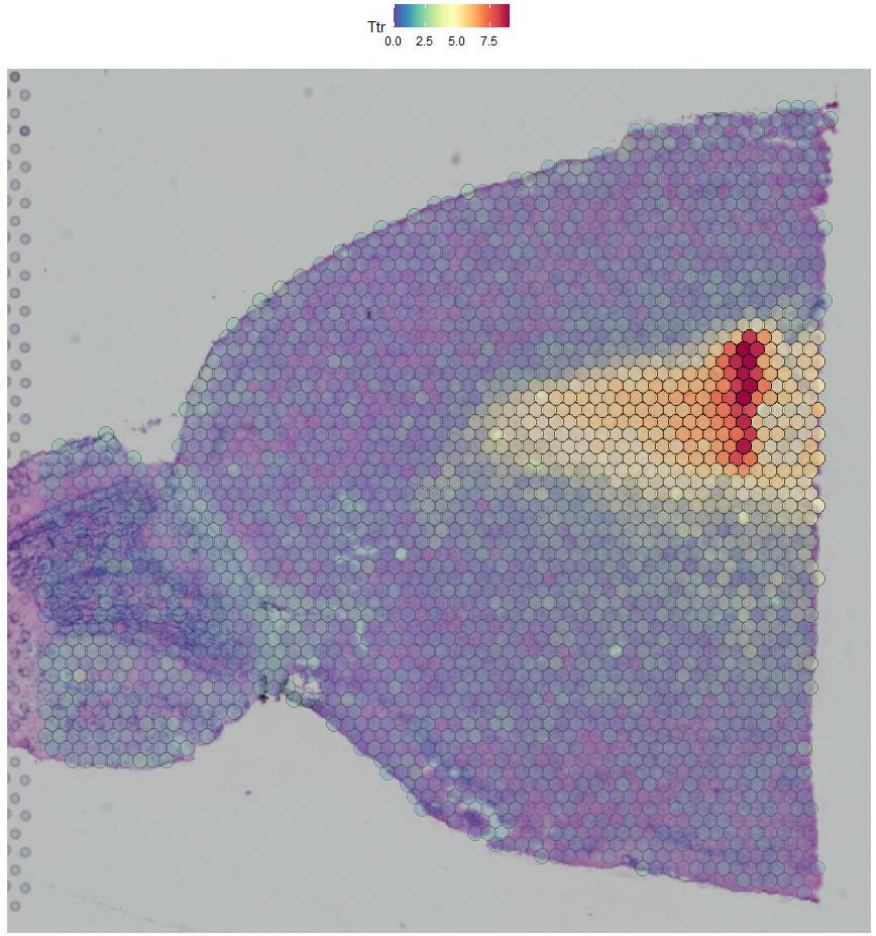
# Seurat 分析 10X Visium 数据

（2）预处理：基因表达的可视化（对于小鼠大脑数据，Hpca是海马体的标记基因，Ttr是脉络丛的标记基因）

- alpha：设置最小和最大的透明度，可以将表达较低的点设置为更透明




alpha = 0.9

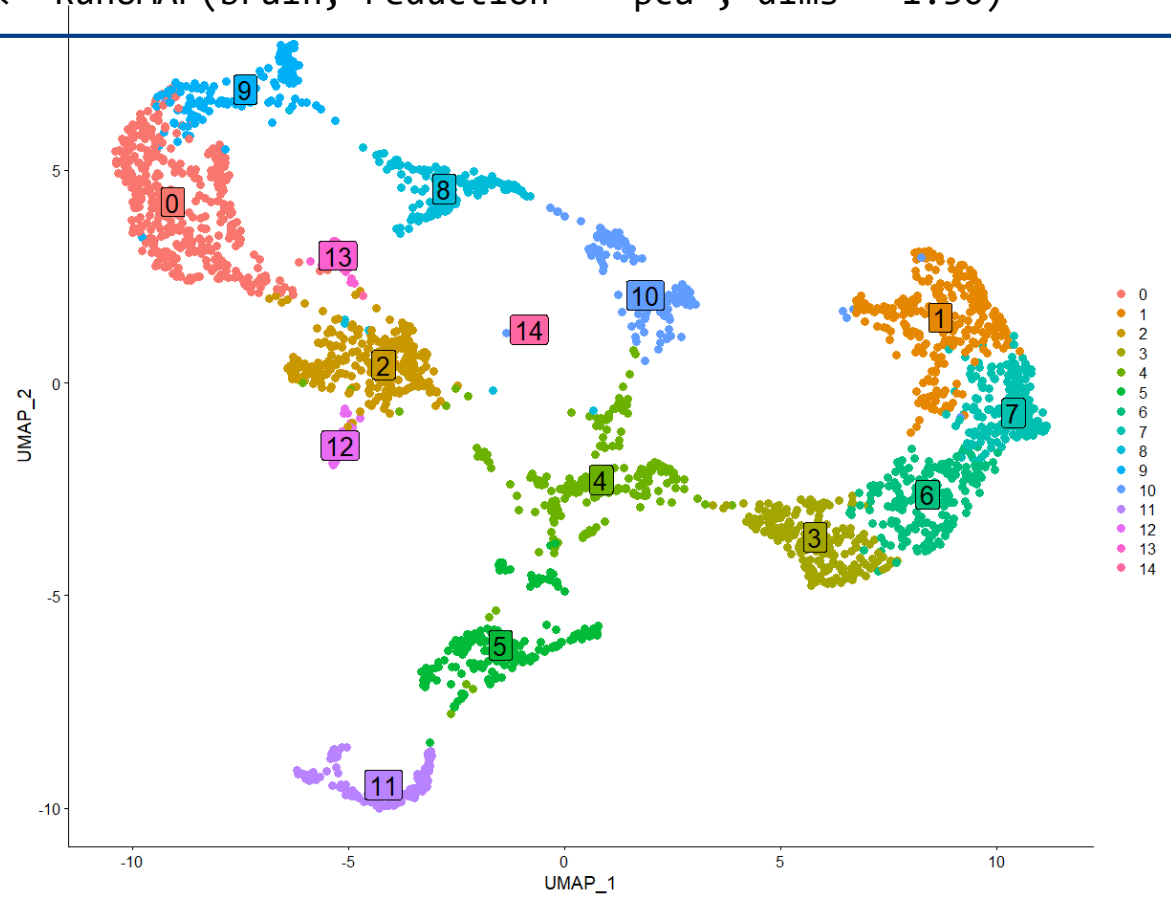


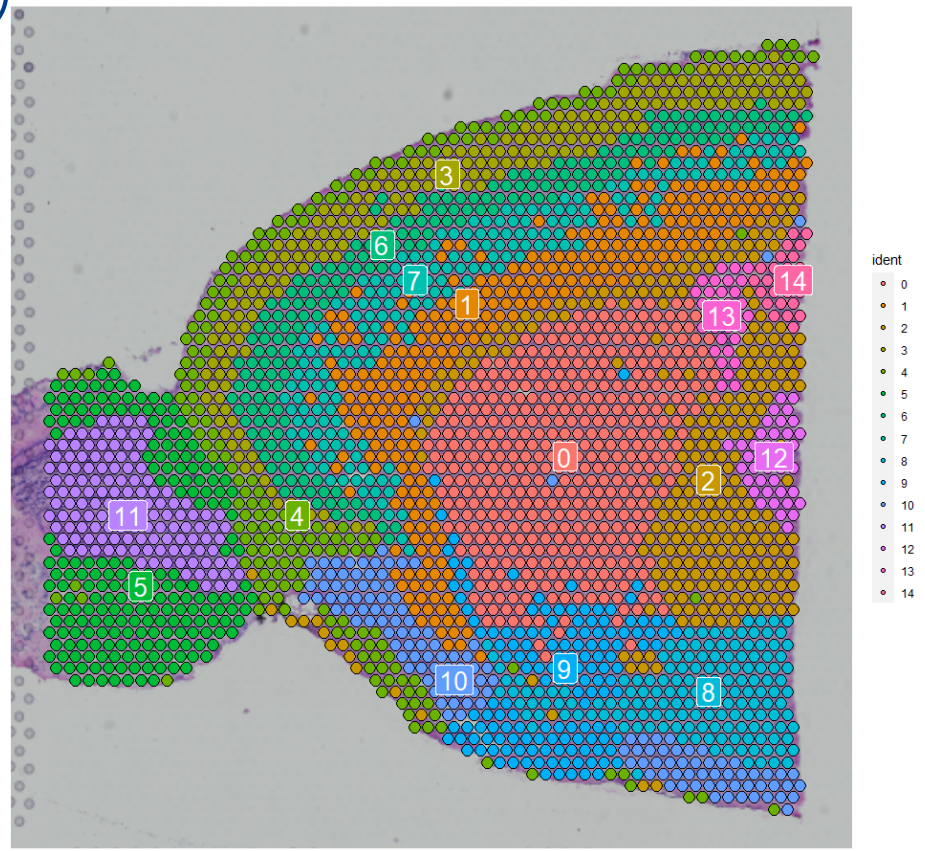

alpha = c(0.1, 1)

# Seurat 分析 10X Visium 数据

（3）降维、聚类、可视化：降维聚类与处理常规scRNA-seq数据类似

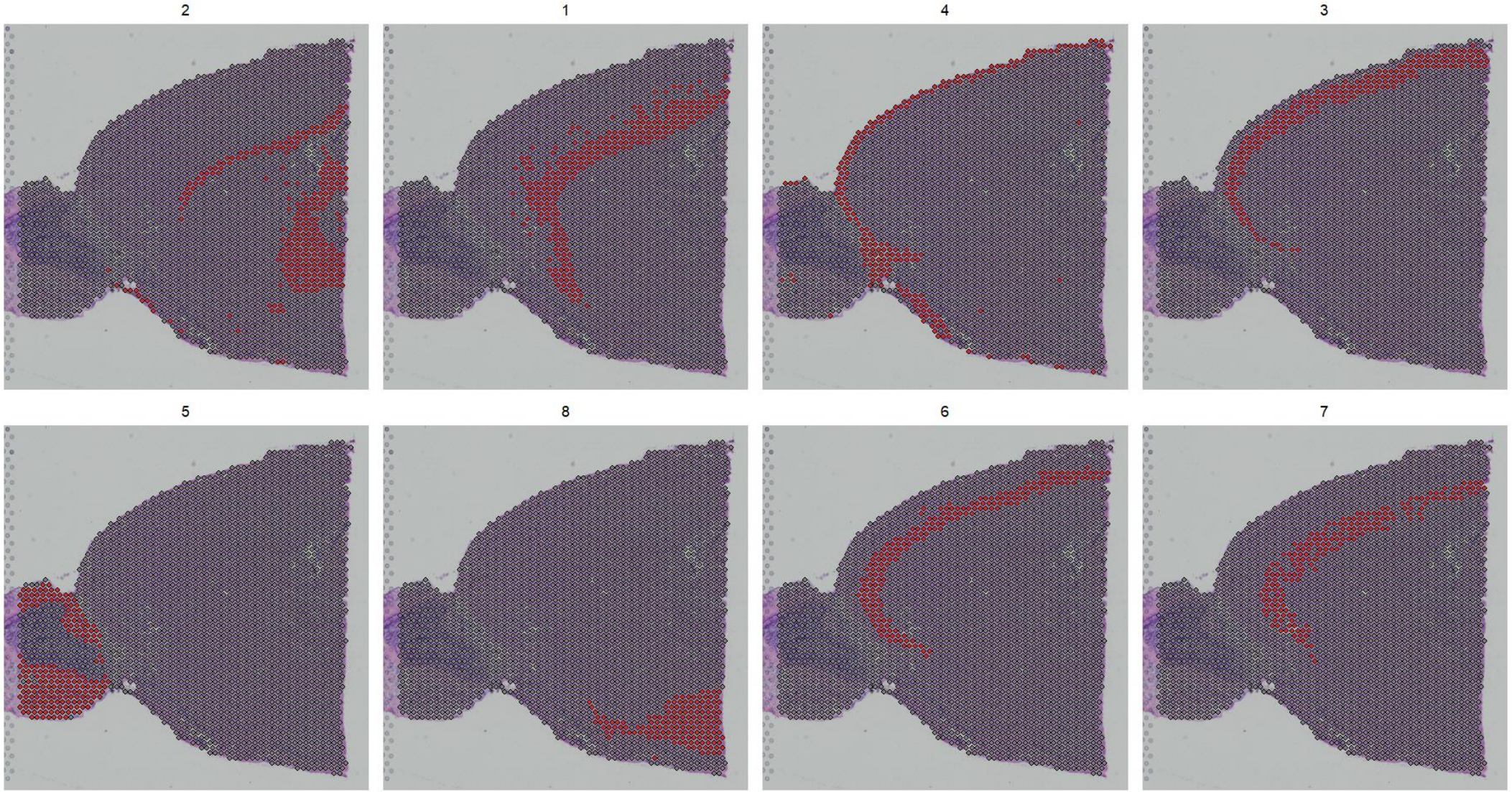
```
brain <- RunPCA(brain, assay = "SCT", verbose = FALSE)
brain <- FindNeighbors(brain, reduction = "pca", dims = 1:30)
brain <- FindClusters(brain, verbose = FALSE)
brain <- RunUMAP(brain, reduction = "pca", dims = 1:30)
```



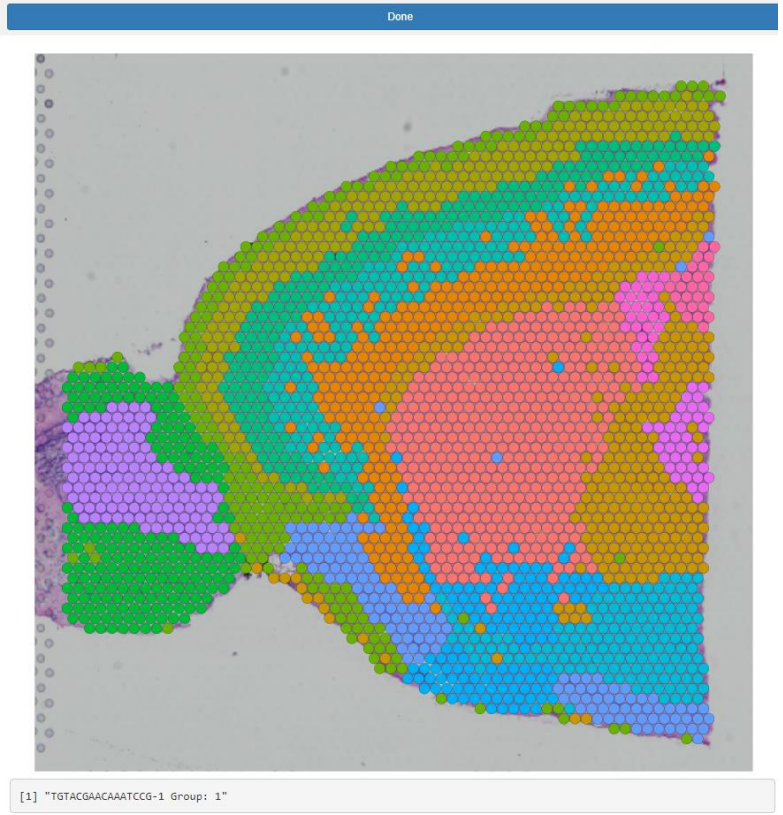
# Seurat 分析 10X Visium 数据

（3）降维、聚类、可视化：可视化感兴趣的簇（清晰地区分单个簇的空间位置）

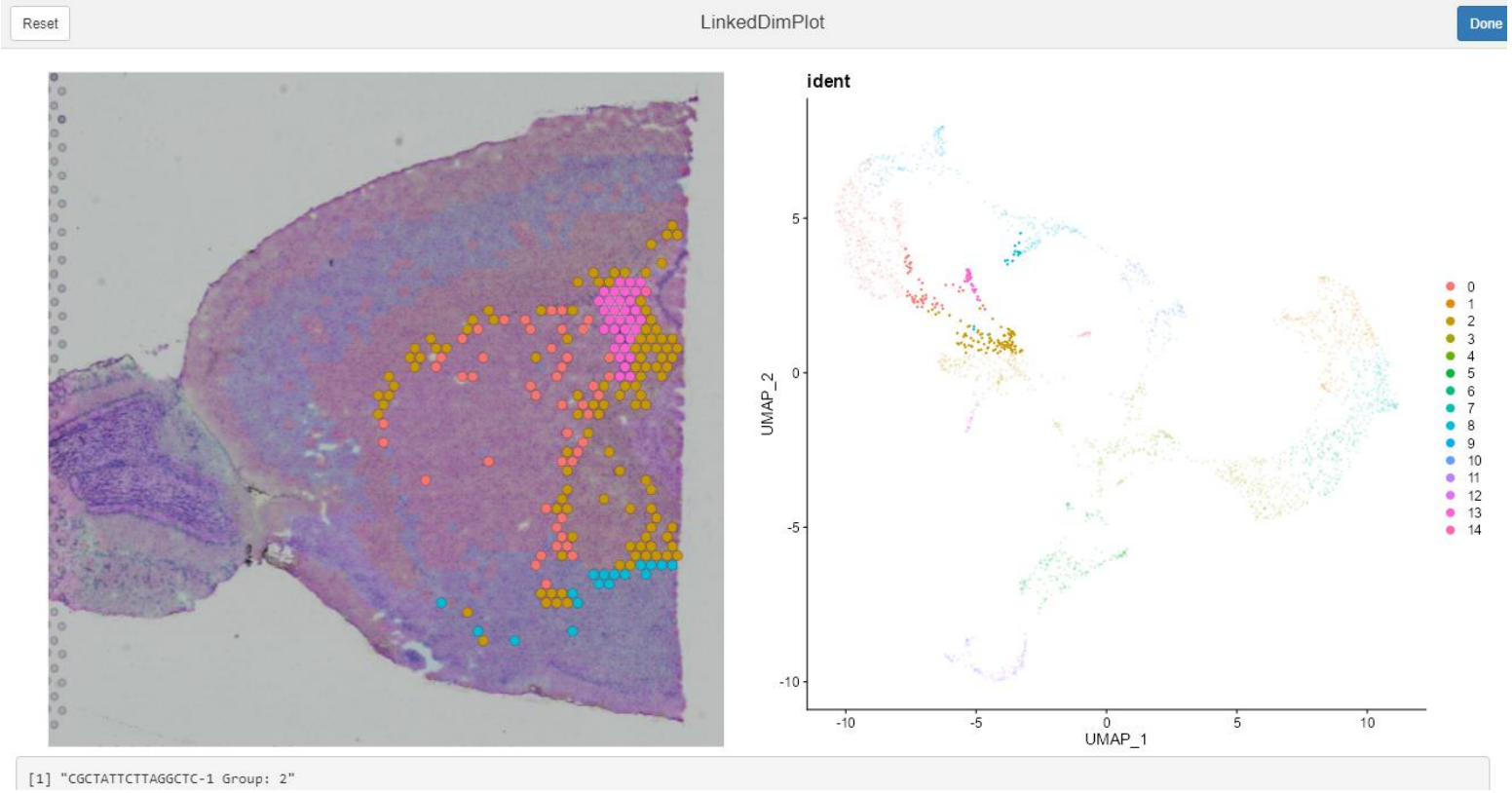


# Seurat 分析 10X Visium 数据

（3）降维、聚类、可视化：交互式绘图




- 鼠标移动会显示相应的spot ID和簇



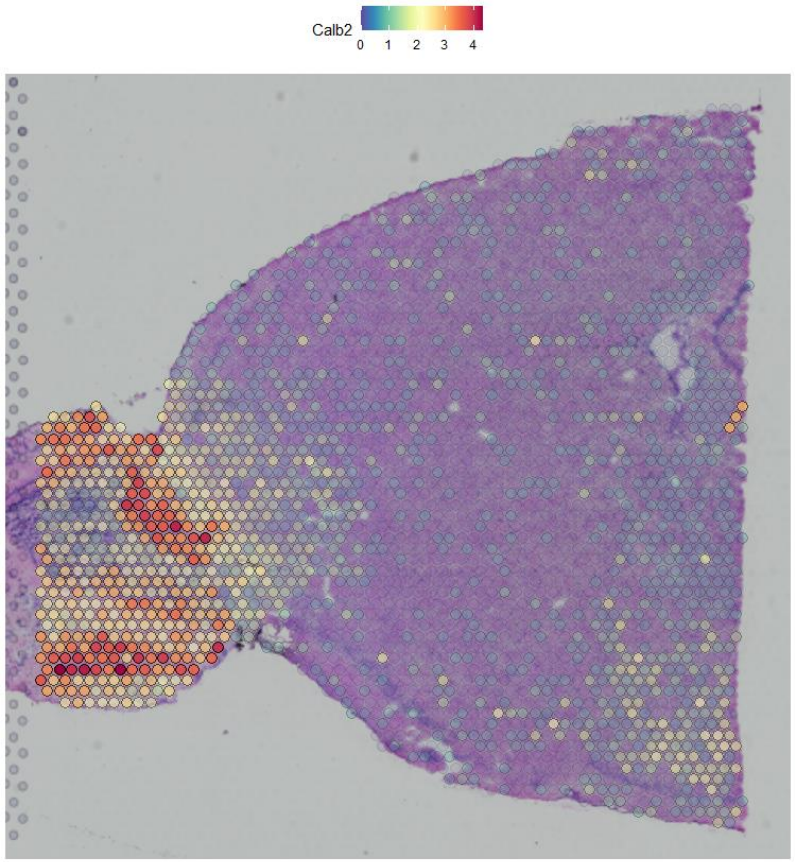

- 根据UMAP图选择spot在空间图上展示

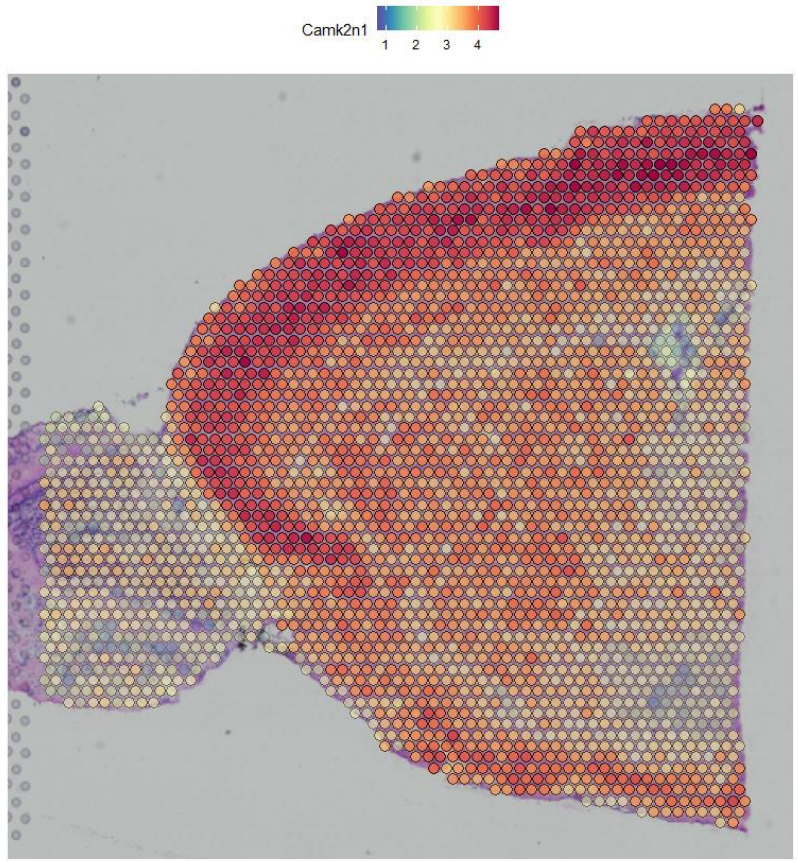
# Seurat 分析 10X Visium 数据

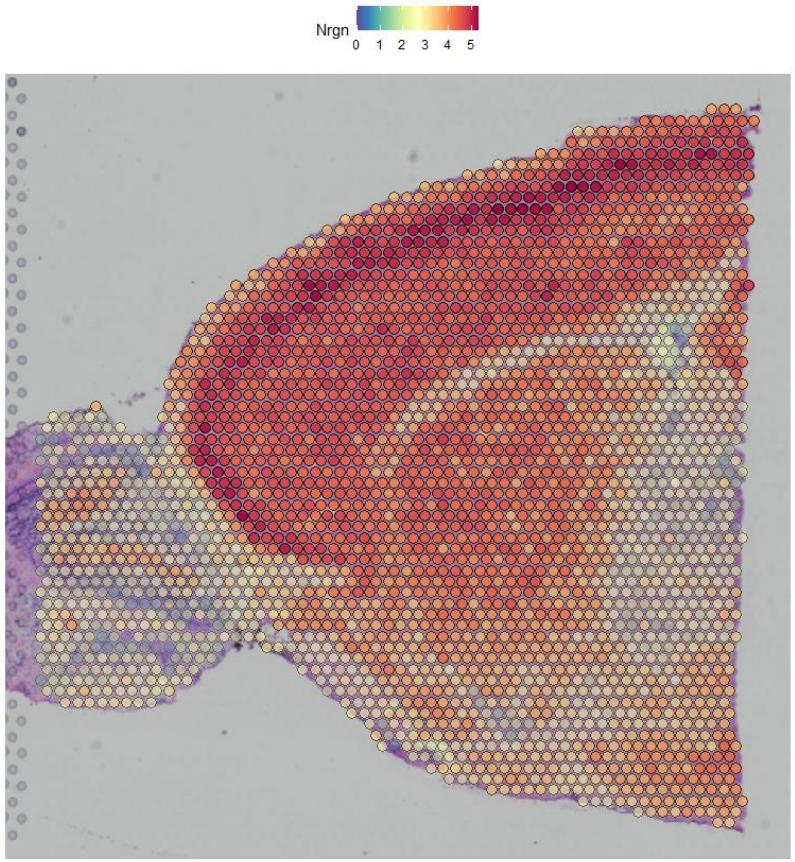
（4）空间高变基因的鉴定：两种方法 FindMarkers 函数和 FindSpatiallyVariableFeatures 函数

- FindMarkers 函数根据定义好的cluster或者celltype进行计算（和scRNA-seq数据分析类似）
- 例如：cluster5和6之间的差异基因

| | p_val | avg_log2FC | pct.1 | pct.2 | p_val_adj |
|---|---|---|---|---|---|
| Calb2 | 6.427214e-69 | 3.336874 | 1.000 | 0.537 | 1.135560e-64 |
| Camk2n1 | 1.519204e-68 | -2.450388 | 1.000 | 1.000 | 2.684130e-64 |
| Nrgn | 1.573095e-68 | -3.229826 | 0.971 | 1.000 | 2.779344e-64 |



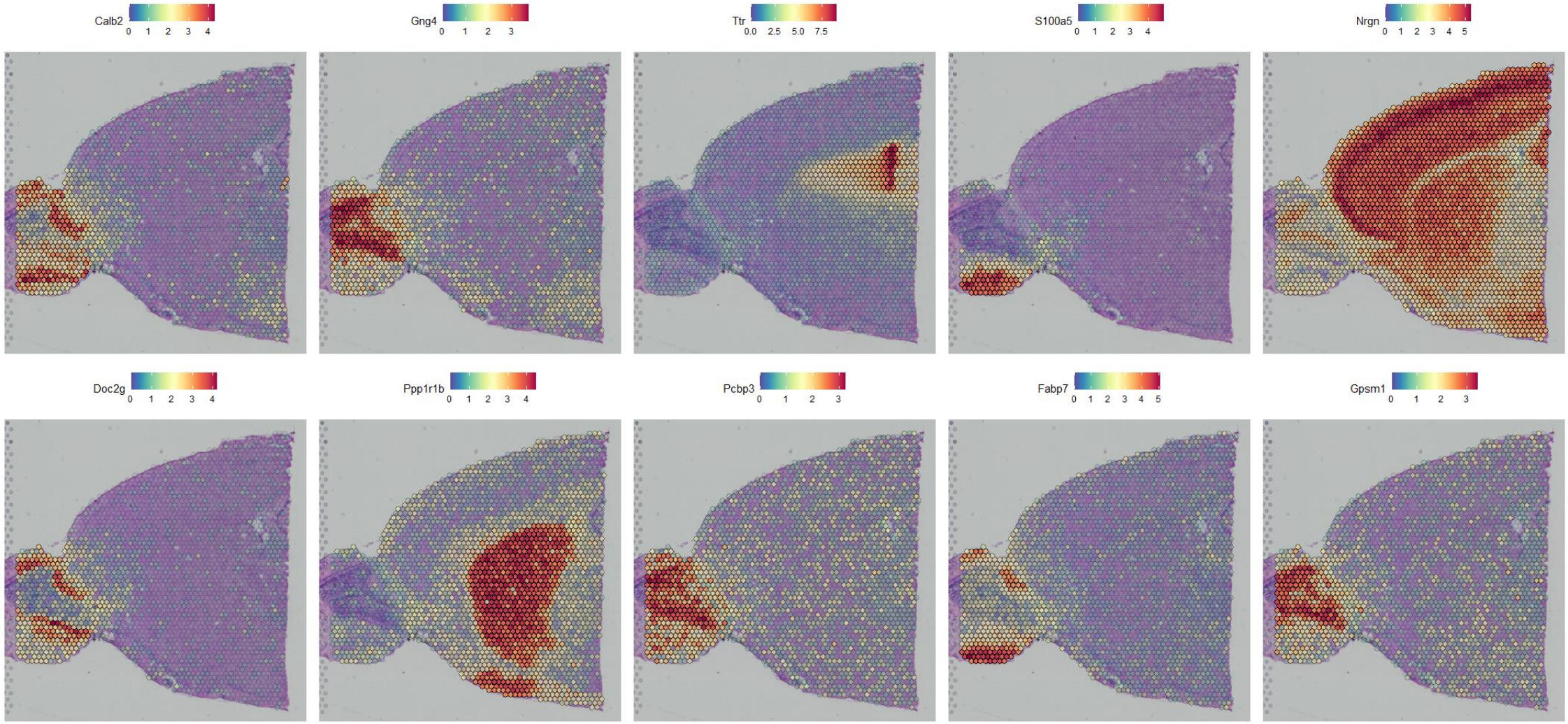



# Seurat 分析 10X Visium 数据

（4）空间高变基因的鉴定：两种方法 FindMarkers 函数和 FindSpatiallyVariableFeatures 函数

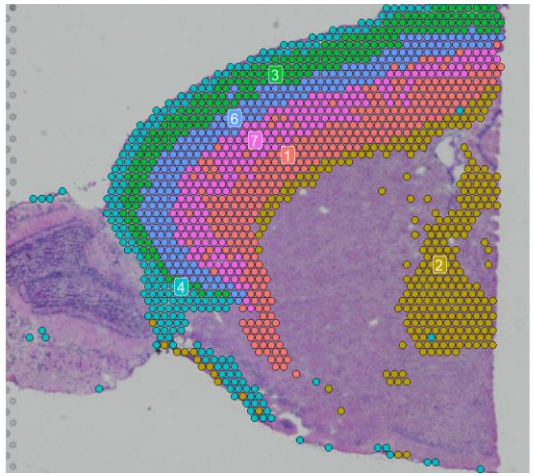
- FindSpatiallyVariableFeatures 函数：在没有cluster 或者 celltype等预先注释的情况下，返回在某些切片区域高表达的基因

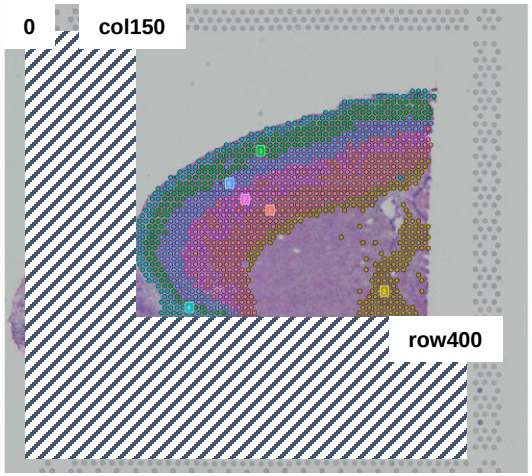


# Seurat 分析 10X Visium 数据

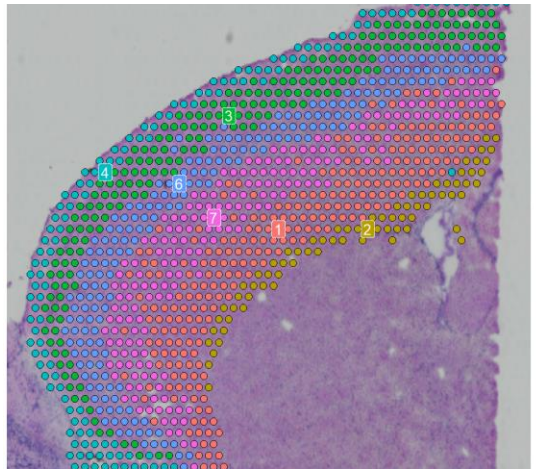
（5）空间转录组数据根据组织切片取子集：先选择cluster的子集，然后可以根据空间位置进行进一步细分

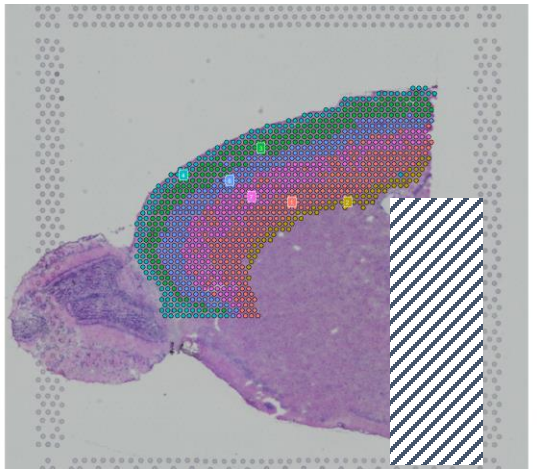
```
cortex <- subset(brain, idents = c(1, 2, 3, 4, 6, 7))
```



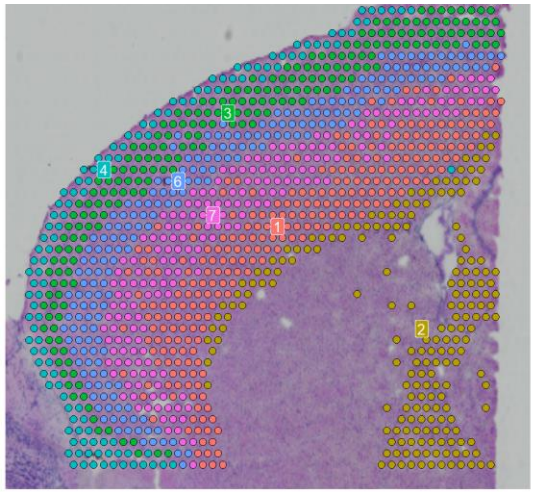


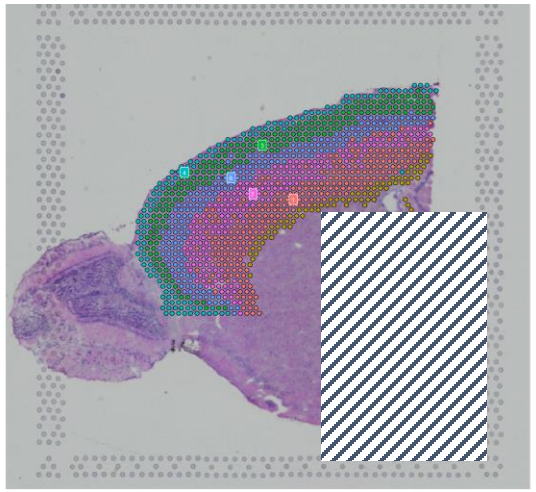

```
cortex <- subset(cortex, anterior1_imagerow > 275 &
anterior1_imagecol > 370, invert = TRUE)
```



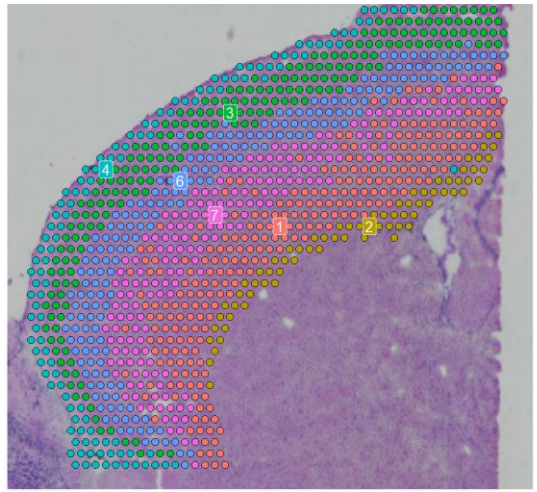


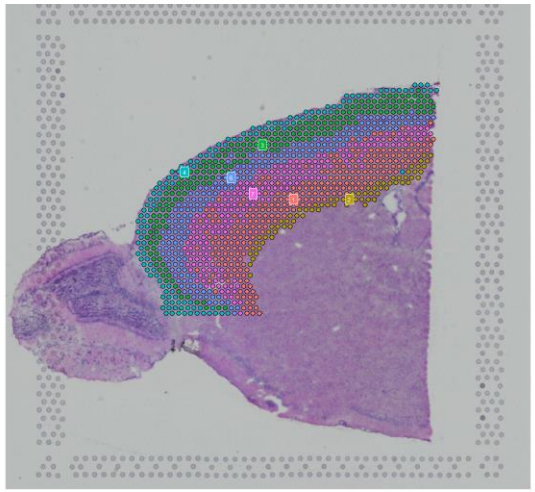
```
cortex <- subset(cortex, anterior1_imagerow > 400 |
anterior1_imagecol < 150, invert = TRUE)
```





```
cortex <- subset(cortex, anterior1_imagerow > 250 &
anterior1_imagecol > 440, invert = TRUE)
```

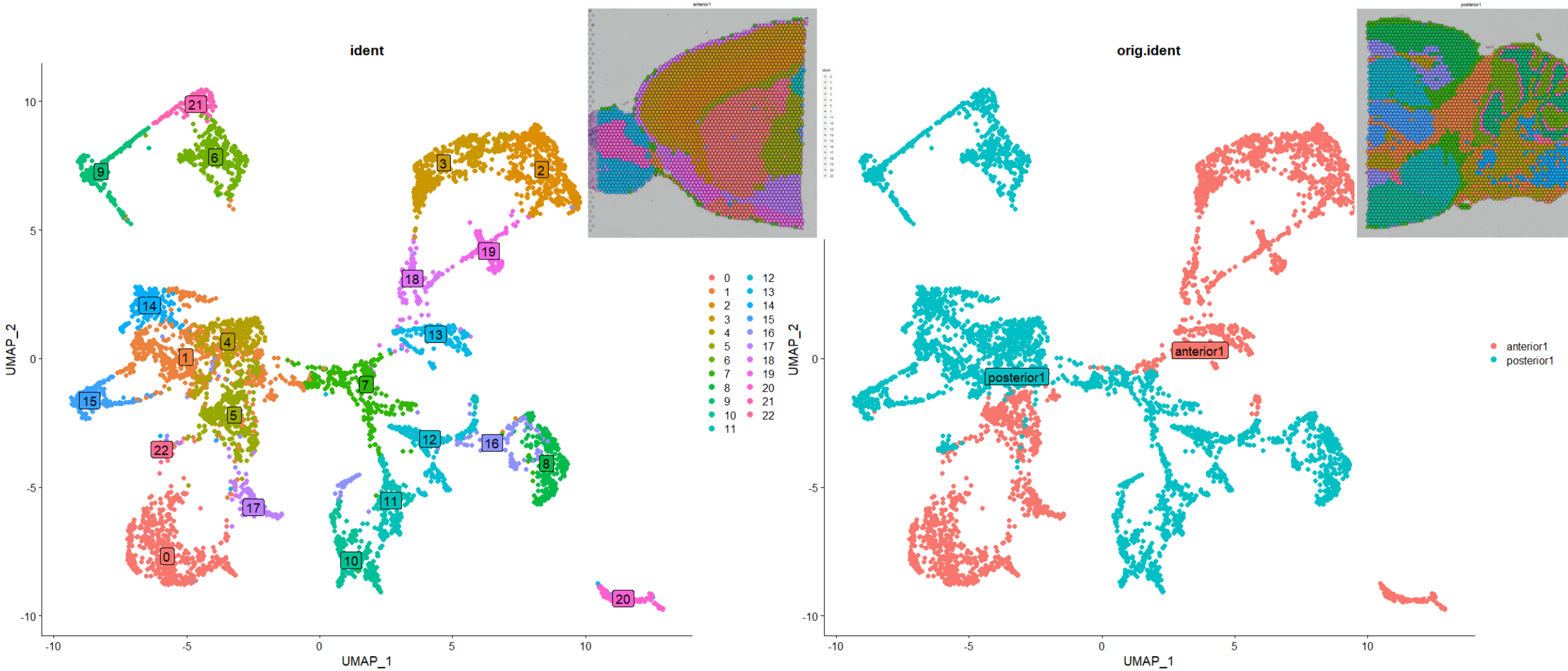




# Seurat 分析 10X Visium 数据

## （6）多张芯片数据整合：小鼠大脑的多张切片

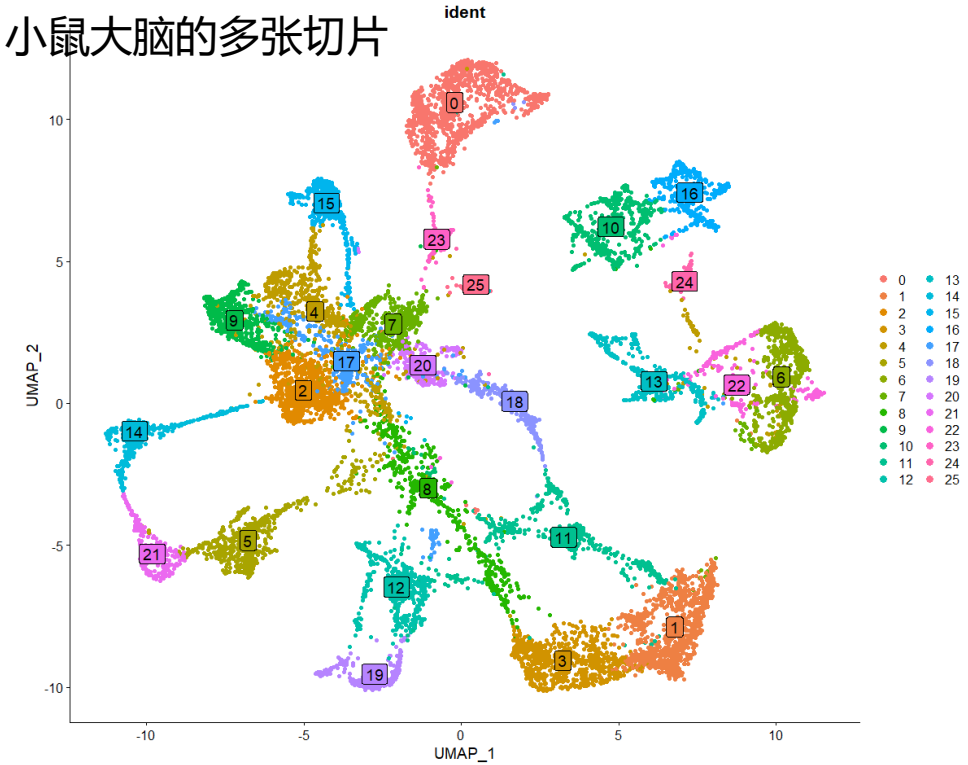
- 整合前后脑两张切片：分别SCT标准化之后merge整合，之后降维、聚类、可视化

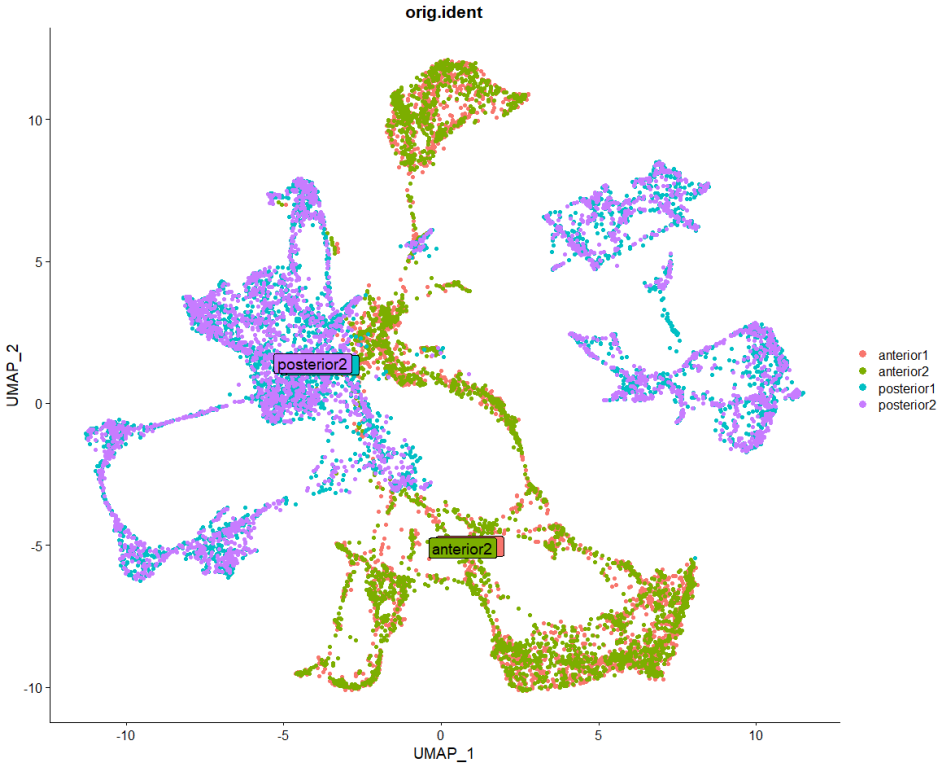


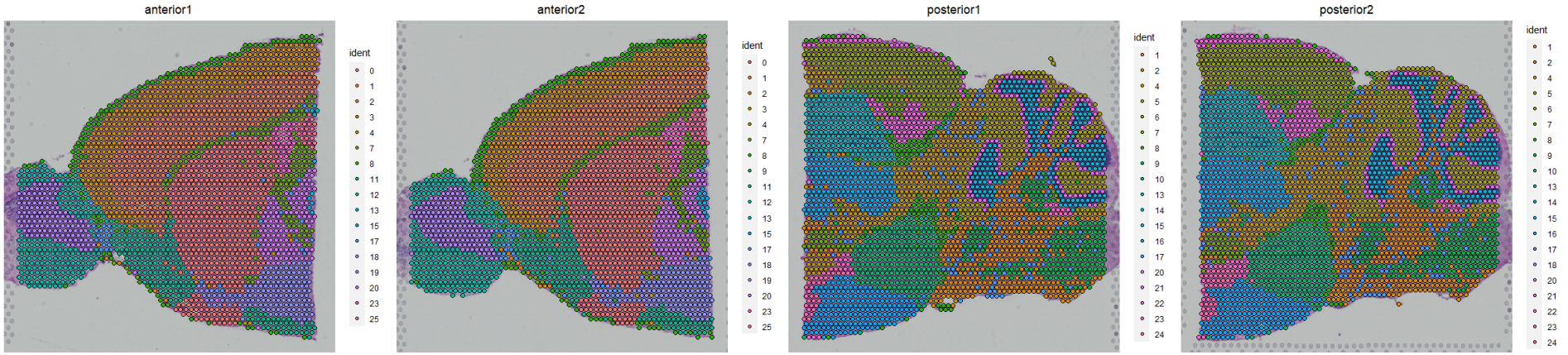
# Seurat 分析 10X Visium 数据

## （6）多张芯片数据整合：小鼠大脑的多张切片

- 整合4张切片小鼠脑切片：
- 连续切片的重复性较好

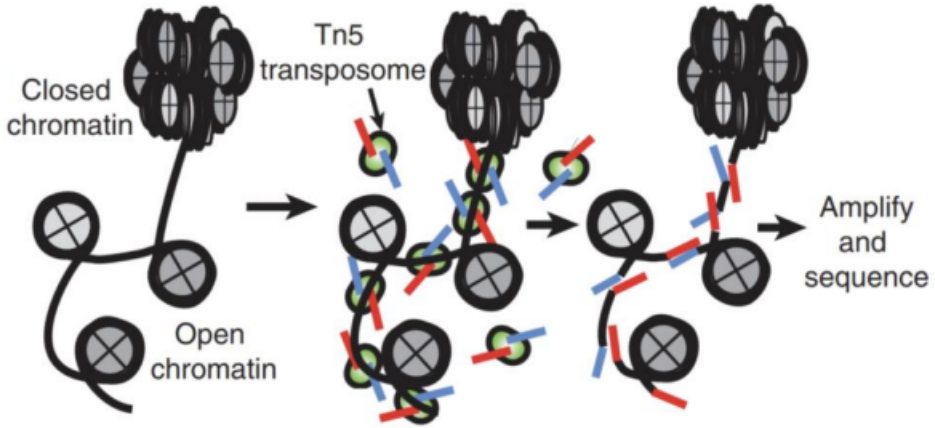


# 上节回顾：什么是scATAC-seq

单细胞染色质可及性测序**(scATAC-Seq)**：鉴定每个细胞的开放染色质区域，即染色体上可以被转录因子、核酸酶等结合的染色质区域。

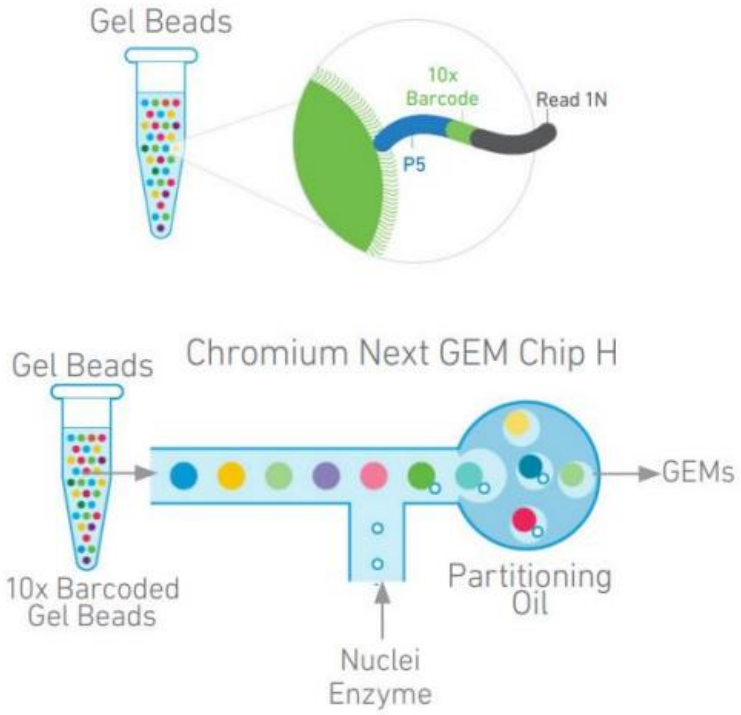


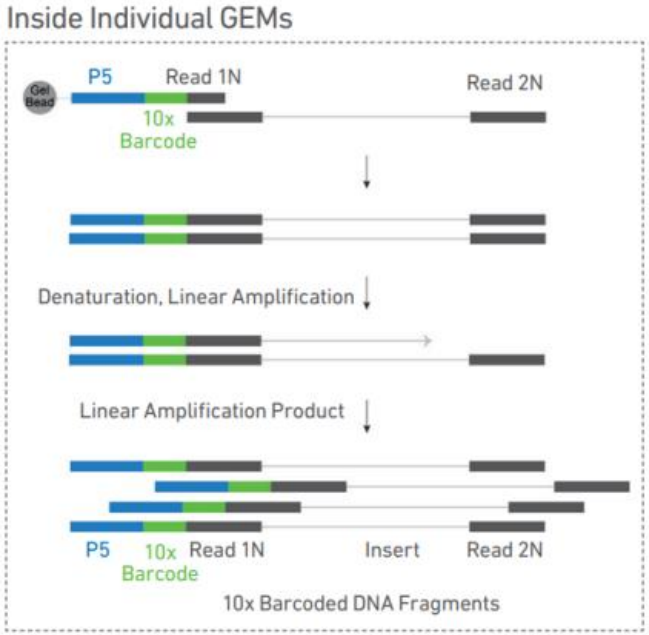

ATAC-Seq

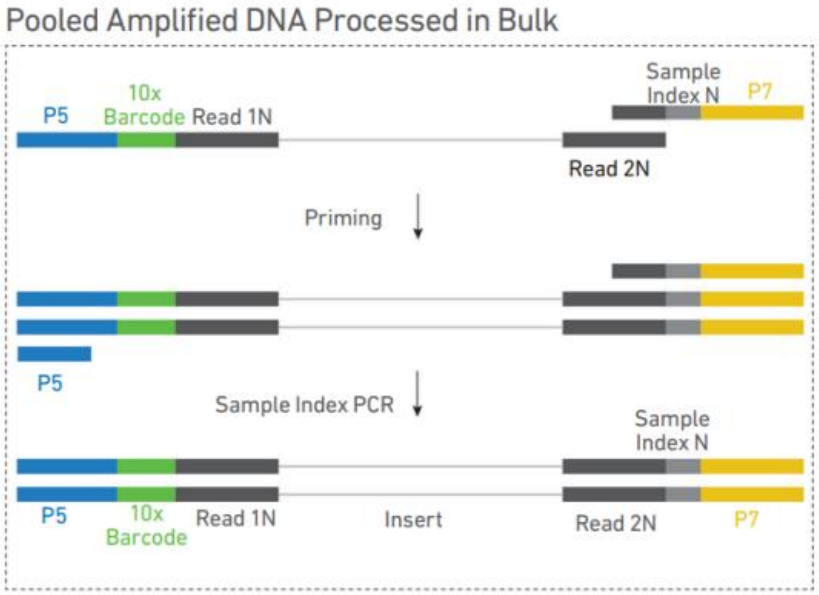
- DNA转座酶：识别并切断开放的染色质区域
- 地将携带已知DNA序列标签的转座复合物（即带着上图红色蓝色测序标签的转座酶Tn5）加入到细胞核中，插入到开放的染色质区域，再利用已知序列的标签进行PCR后测序

# 上节回顾：什么是scATAC-seq

单细胞染色质可及性测序：基于10X Genomics




- ATAC胶珠上的序列中不用UMI，因为基因组只有一对序列，无需像RNA一样定量
- 序列末端用接头引物Read 1N代替PolyT
- scATAC-seq的DNA片段没有PolyA尾，取而代之的是Tn5酶转座剪切时插入的adaptors片段，可以与胶珠上的Read 1N序列互补

# Cell Ranger ATAC

Cell Ranger ATAC是处理Chromium平台产生的scATAC-seq数据的软件，包括以下四个部分：

## cellranger-atac mkfastq

将Illumina测序仪生成的原始BCL文件转换为FASTQ文件，对bcl2fastq函数进行包装整合，并且适用于10X Genomics平台文库

## cellranger-atac count

进行ATAC数据分析，具体包括：

- Read filtering and alignment
- Barcode counting
- Identification of transposase cut sites
- Detection of accessible chromatin peaks
- Cell calling
- Count matrix generation for peaks and transcription factors
- Dimensionality reduction
- Cell clustering
- Cluster differential accessibility

## cellranger-atac aggr

整合一次实验的多个样本（重复），具体包括：

- Normalization of input runs to same median fragments per cell (sensitivity)
- Detection of accessible chromatin peaks
- Count matrix generation for peaks and transcription factors for the aggregate data
- Dimensionality reduction
- Cell clustering
- Cluster differential accessibility
- Chemistry batch correction

## cellranger-atac mkfastq

对count或者aggr生成的文件修改参数进行二次分析，包括：

- Cell calling
- Dimensionality reduction, Cell clustering
- Cluster differential accessibility

# 从原始数据到矩阵（cellranger-atac）

（1）软件下载和基因组建库

- cellranger-atac下载地址： https://support.10xgenomics.com/single-cell-atac/software/downloads/latest（解压后即可直接使用）
- 基因组索引文件：①官网直接下载；②根据基因组FASTA文件和GTF文件构建基因组索引（cellranger-atac mkref）

```
cellranger-atac-2.1.0
├── bin
│   ├── atac
│   ├── cellranger-atac
│   ├── _cellranger_atac_internal
│   └── tenkit
├── builtwith.json
├── cellranger-atac -> bin/cellranger-atac
├── external
│   ├── anaconda
│   ├── arc_testrun_files
│   ├── cellranger_atac_tiny_fastq
│   └── martian
├── lib
│   ├── bin
│   ├── plsa
│   └── python
├── LICENSE
├── mro
│   ├── atac
│   ├── rna
│   └── tenkit
├── sourceme.bash
└── sourceme.csh
```

- cellranger-atac解压后文件

```
cellranger-atac cellranger-atac-2.1.0
Process 10x Genomics Chromium Single Cell ATAC data

USAGE:
    cellranger-atac <SUBCOMMAND>

OPTIONS:
    -h, --help       Print help information
    -V, --version    Print version information

SUBCOMMANDS:
    count        Count reads from a single Single Cell ATAC
                 library
    mkfastq      Run bcl2fastq on Single Cell ATAC sequencing
                 data
    mkref        Create a cellranger-atac-compatible reference
                 package
    aggr         Aggregate data from multiple `cellranger-atac
                 count` runs
    reanalyze    Re-run secondary analysis (dimensionality
                 reduction, clustering, etc) on a completed
                 `cellranger-atac count` or `cellranger-atac
                 aggr` run
    testrun      Run a tiny cellranger-atac count pipeline to
                 verify software integrity
    upload       Upload analysis logs to 10x Genomics support
    sitecheck    Collect linux system configuration information
    help         Print this message or the help of the given
                 subcommand(s)
```

- cellranger-atac使用帮助

```
refdata-cellranger-arc-GRCh38-2020-A-2.0.0
├── fasta
│   ├── genome.fa
│   ├── genome.fa.amb
│   ├── genome.fa.ann
│   ├── genome.fa.bwt
│   ├── genome.fa.fai
│   ├── genome.fa.pac
│   └── genome.fa.sa
├── genes
│   └── genes.gtf.gz
├── reference.json
├── regions
│   ├── motifs.pfm
│   ├── transcripts.bed
│   └── tss.bed
└── star
    ├── chrLength.txt
    ├── chrNameLength.txt
    ├── chrName.txt
    ├── chrStart.txt
    ├── exonGeTrInfo.tab
    ├── exonInfo.tab
    ├── geneInfo.tab
    ├── Genome
    ├── genomeParameters.txt
    ├── SA
    ├── SAindex
    ├── sjdbInfo.txt
    ├── sjdbList.fromGTF.out.tab
    ├── sjdbList.out.tab
    └── transcriptInfo.tab
```

- cellranger-atac基因组索引文件

# 从原始数据到矩阵（cellranger-atac）

## （2）cellranger-atac mkfastq 转换 bcl 格式文件为 fastq 格式文件

- 输入两个文件：bcl 文件和 csv 文件

```
cellranger-atac-tiny-bcl-1.0.0
├── Config
│   ├── HiSeqControlSoftware.Options.cfg
│   ├── hjn3kbcx2_Jun-21-18 12-28-52_Effective.cfg
│   ├── RTAStart.bat
│   └── Variability_HiSeq_C.bin
├── Data
│   └── Intensities
│       ├── BaseCalls
│       ├── config.xml
│       ├── L001
│       ├── Offsets
│       └── RTAConfiguration.xml
├── InterOp
│   ├── ControlMetricsOut.bin
│   ├── CorrectedIntMetricsOut.bin
│   ├── ErrorMetricsOut.bin
│   ├── ExtractionMetricsOut.bin
│   ├── ImageMetricsOut.bin
│   ├── IndexMetricsOut.bin
│   ├── QMetricsOut.bin
│   └── TileMetricsOut.bin
├── RTAComplete.txt
├── RunInfo.xml
└── runParameters.xml
```

- BCL文件是由Illumina测序仪生成的测序文件，包含测序的信息。包括簇数和碱基信息。

- CSV文件共包含三列，分别为Lane, Sample, Index

```
Lane,Sample,Index
1,test_sample,SI-P01-H10
```

| 列 | 说明 |
| --- | --- |
| Lane | 芯片流动槽上的通道。可以是单个通道或多个（如2-4），星号 “*”表示所有通道。 |
| Sample | 样本的名称。该名称将是所有生成的FASTQ的前缀，并将对应于所有下游分析过程中的--sample参数。样本名称必须符合Illumina bcl2fastq的命名要求。只允许使用字母、数字、下划线和连字符；不允许使用其他符号，包括点（“.”）。 |
| Index | 用于10X平台文库构建的样本Index，如SI-NA-A12。 |

# 从原始数据到矩阵（cellranger-atac）

（2）cellranger-atac mkfastq 转换 bcl 格式文件为 fastq 格式文件

- `cellranger-atac mkfastq --id=tiny-bcl --run=/public/home/chuqj/bioinfor_test/scATAC/cellranger-atac-tiny-bcl-1.0.0 --csv=cellranger-atac-tiny-bcl-simple-1.0.0.csv` 运行结果：

```
/public/home/chuqj/bioinfor_test/scATAC/tiny-bcl/outs/
├── fastq_path
│   ├── HJN3KBCX2
│   │   └── test_sample
│   │       ├── test_sample_S1_L001_I1_001.fastq.gz
│   │       ├── test_sample_S1_L001_R1_001.fastq.gz
│   │       ├── test_sample_S1_L001_R2_001.fastq.gz
│   │       └── test_sample_S1_L001_R3_001.fastq.gz
│   ├── Reports
│   │   └── html
│   │       ├── HJN3KBCX2
│   │       ├── index.html
│   │       ├── Report.css
│   │       └── tree.html
│   ├── Stats
│   │   ├── AdapterTrimming.txt
│   │   ├── ConversionStats.xml
│   │   ├── DemultiplexingStats.xml
│   │   ├── DemuxSummaryF1L1.txt
│   │   ├── FastqSummaryF1L1.txt
│   │   └── Stats.json
│   ├── Undetermined_S0_L001_I1_001.fastq.gz
│   ├── Undetermined_S0_L001_R1_001.fastq.gz
│   ├── Undetermined_S0_L001_R2_001.fastq.gz
│   └── Undetermined_S0_L001_R3_001.fastq.gz
├── input_samplesheet.csv
└── interop_path
    └── IndexMetricsOut.bin
```

**I1** — Sample index

```
@D000684:1170:HJN3KBCX2:1:1101:0:53 1:N:0:GACCGCCA
GACCGCCA
+
GGGGGIII
```

**R1** — Read 1

```
@D000684:1170:HJN3KBCX2:1:1101:0:53 1:N:0:GACCGCCA
GAGTAGGGCTGAGACTGGGGTGGGGCCTTCTATGGCTGAGGGGAGTCAGG
+
GAGGGGGIIGIGGIGIIIIIIIIGIGIIIIIIIIIIIIGGGGIGIIIIII
```

**R2** — Barcode

```
@D000684:1170:HJN3KBCX2:1:1101:0:53 2:N:0:GACCGCCA
NTGCTCAAGCAGAAAG
+
#<GGAGGGIGIGGIIG
```

**R3** — Read 2

```
@D000684:1170:HJN3KBCX2:1:1101:0:53 3:N:0:GACCGCCA
GCCTCACCCCACTACTAGGCCTCCTCCTAGCAGCAGCAGGCAAATCAGC
+
GAGGGIIIIGGGIGIIIIIIGIIIIIIIIGGGIIIIIIIIIGGIIIIIIA
```

# 从原始数据到矩阵（cellranger-atac）

## （3）cellranger-atac count

- 输出文件格式描述

| File Name | Description |
|---|---|
| singlecell.csv | Per-barcode fragment counts & metrics 细胞表型信息 |
| possorted_bam.bam | Position sorted BAM file |
| possorted_bam.bam.bai | Position sorted BAM index |
| summary.json | Summary of all data metrics |
| web_summary.html | HTML file summarizing data & analysis |
| peaks.bed | Bed file of all called peak locations |
| raw_peak_bc_matrix.h5 | Raw peak barcode matrix in hdf5 format |
| raw_peak_bc_matrix | Raw peak barcode matrix in mex format 原始矩阵 |
| analysis | Directory of analysis files |
| filtered_peak_bc_matrix.h5 | Filtered peak barcode matrix in hdf5 format |
| filtered_peak_bc_matrix | Filtered peak barcode matrix 筛选后矩阵 |
| fragments.tsv.gz | Barcoded and aligned fragment file 所有片段文件 |
| fragments.tsv.gz.tbi | Fragment file index |
| filtered_tf_bc_matrix.h5 | Filtered tf barcode matrix in hdf5 format |
| filtered_tf_bc_matrix | Filtered tf barcode matrix in mex format |
| cloupe.cloupe | Loupe Browser input file |
| summary.csv | summary metrics in CSV form |
| peak_annotation.tsv | Peak-gene associations based on genome proximity |
| peak_motif_mapping.bed | Peak motif associations. Note that one peak could be associated with multiple transcription factor motifs. |

# Signac 进行 scATAC-seq 数据分析

Signac是Seurat的一个扩展功能R包，可以用来分析、解释和探索单细胞染色质数据集。

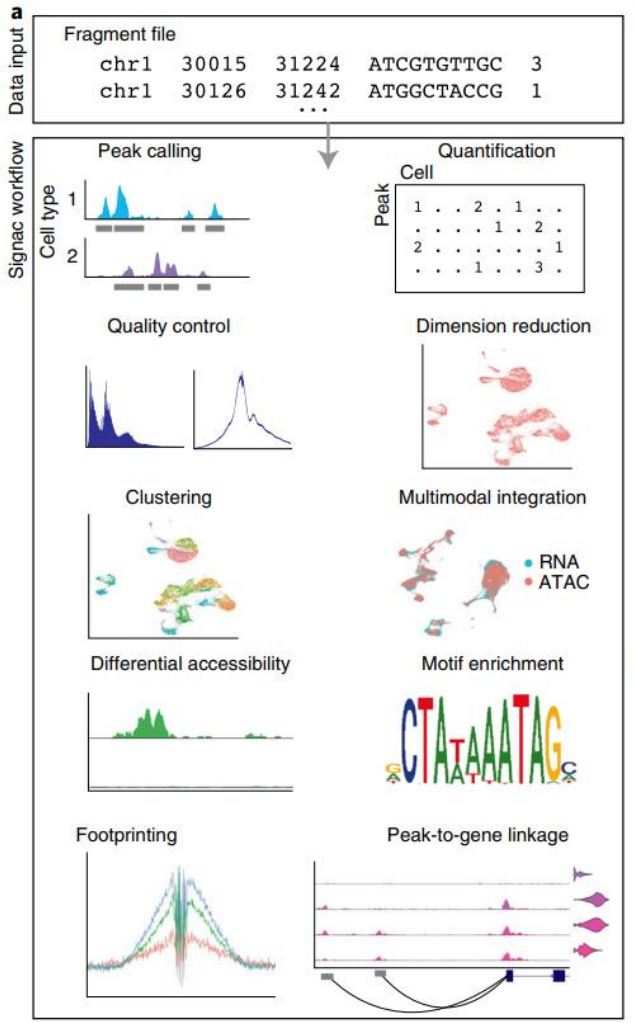
主要包括以下功能：

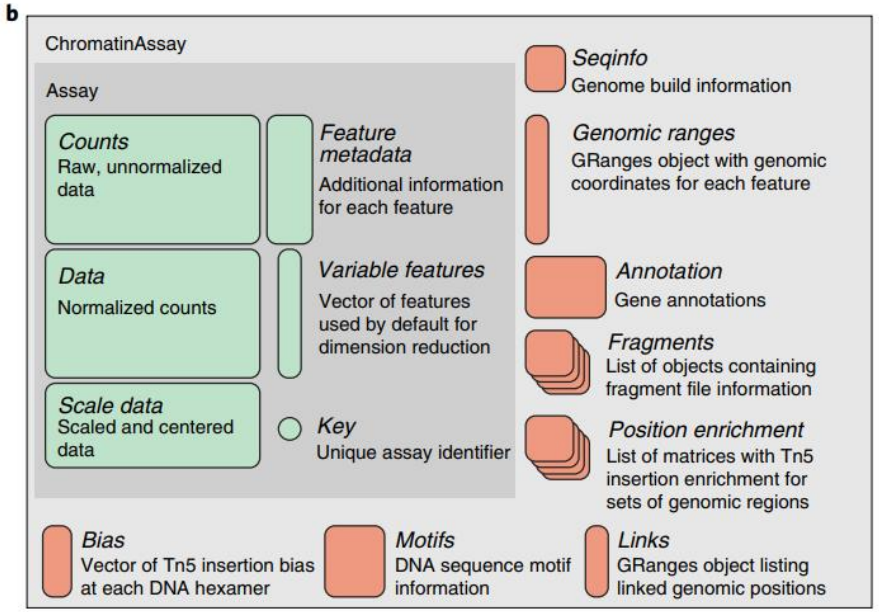
- 单细胞数据质控指标的计算
- 数据降维、可视化和细胞聚类
- 细胞类型特异性peak的鉴定
- 多样本scATAC-seq数据集的整合
- 与scRNA-seq数据集的整合
- Motif富集分析

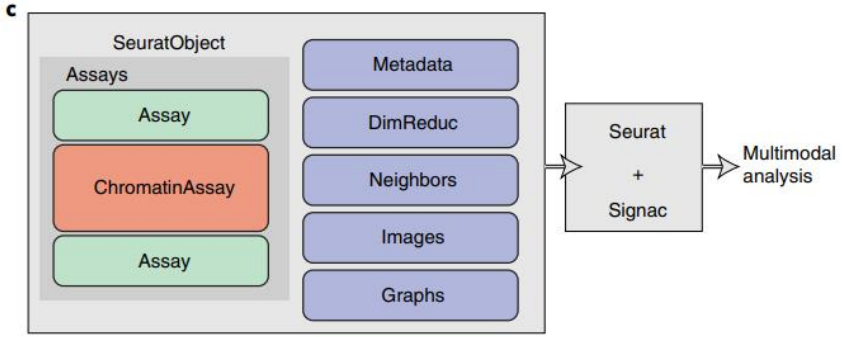
nature methods ARTICLES
https://doi.org/10.1038/s41592-021-01282-5
Check for updates

## Single-cell chromatin state analysis with Signac

Tim Stuart[1,2] ✉, Avi Srivastava[1,2], Shaista Madad[1,2], Caleb A. Lareau[3] and Rahul Satija[1,2] ✉

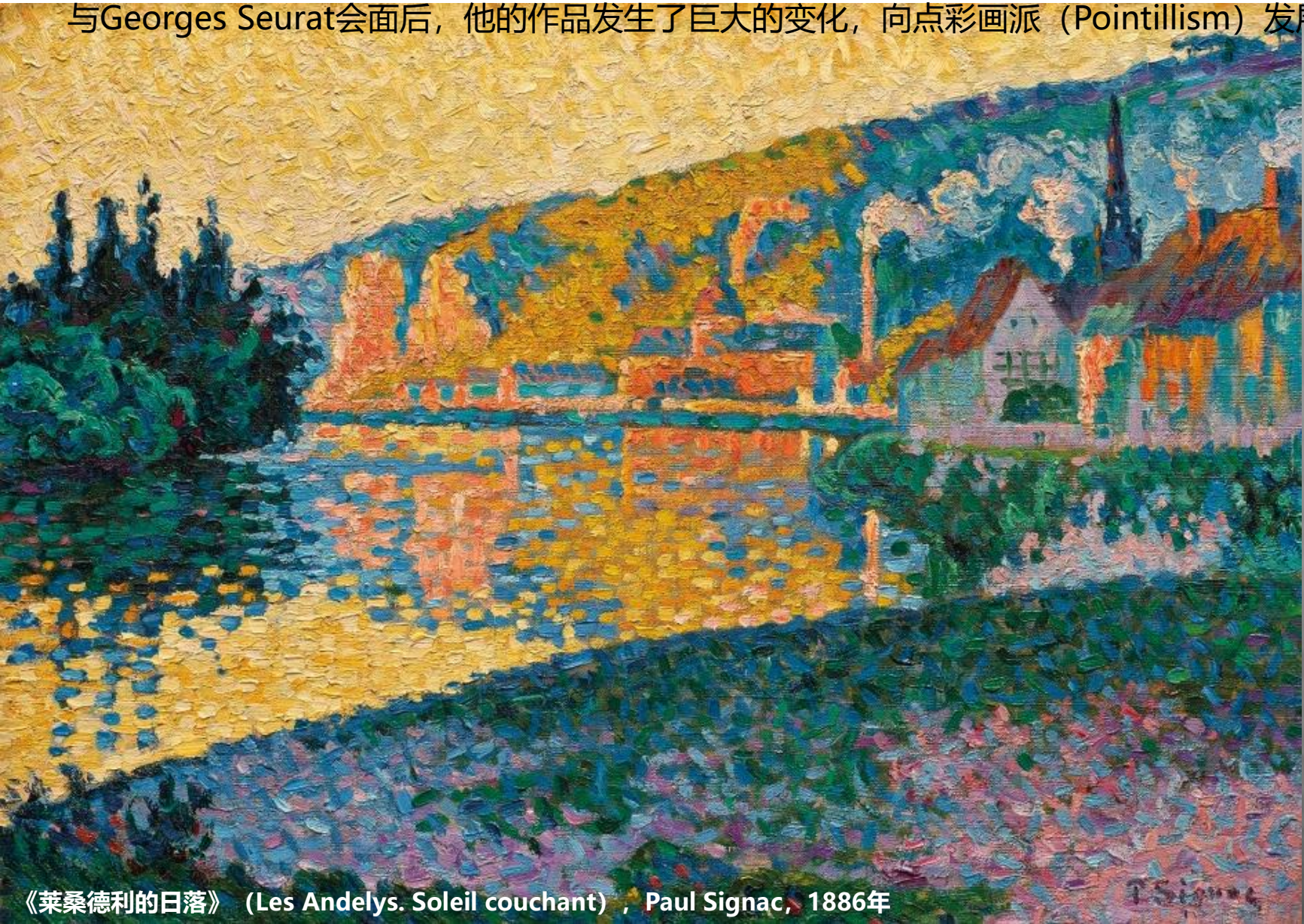


# Paul Signac 点彩派画家

Paul Signac（1863-1935）出生于巴黎，早年曾学习建筑学。1880年代初在莫奈影响下绘制的第一批印象派作品，在1884年与Georges Seurat会面后，他的作品发生了巨大的变化，向点彩画派（Pointillism）发展。

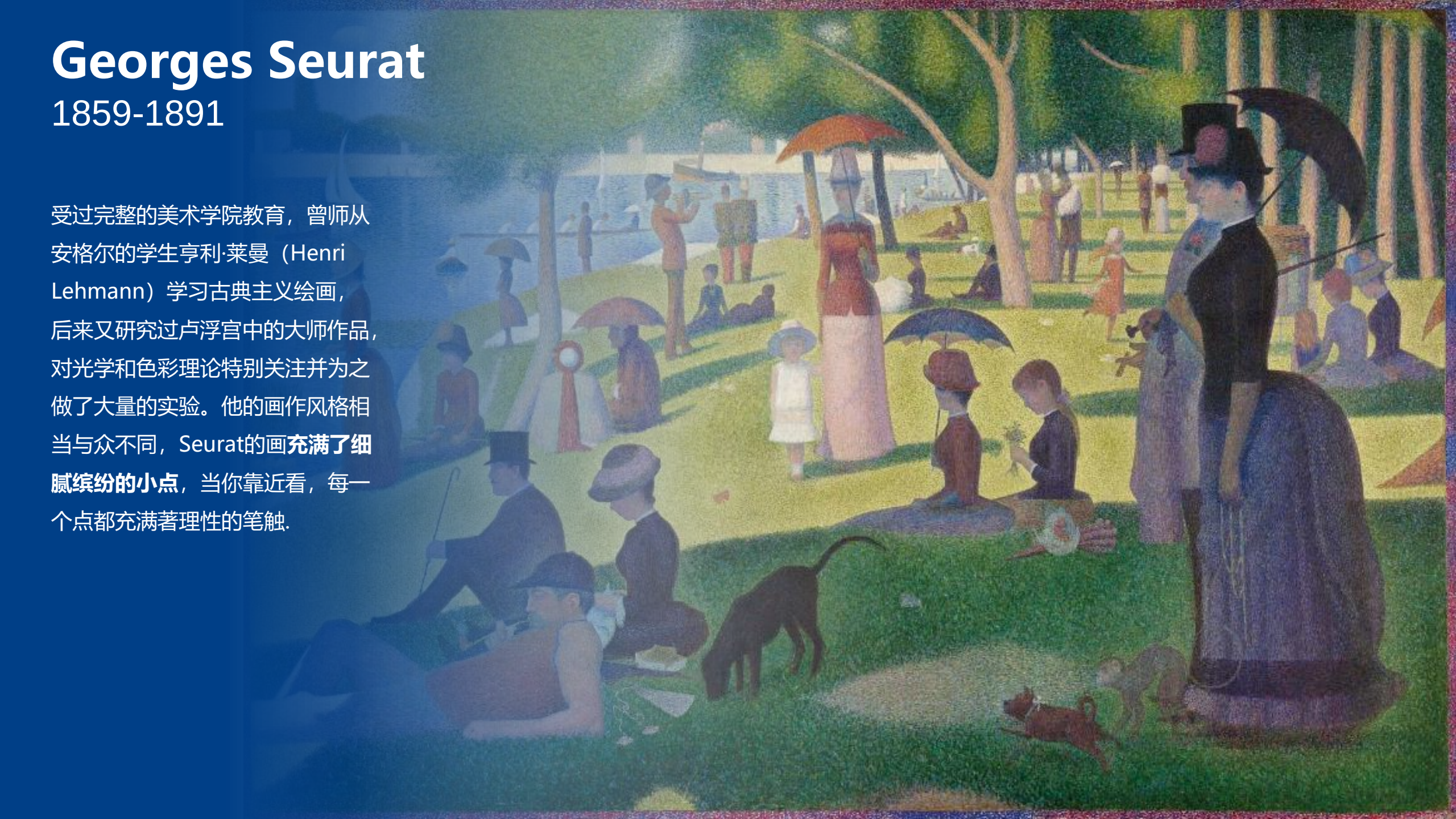


《莱桑德利的日落》（Les Andelys. Soleil couchant），Paul Signac，1886年

# Georges Seurat

1859-1891

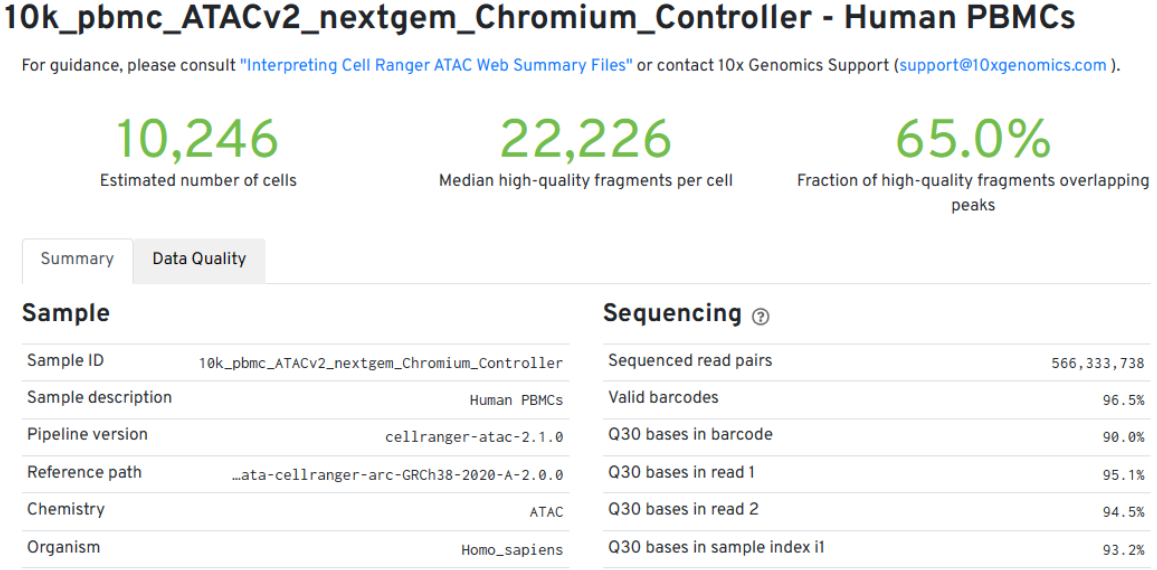
受过完整的美术学院教育，曾师从安格尔的学生亨利·莱曼（Henri Lehmann）学习古典主义绘画，后来又研究过卢浮宫中的大师作品，对光学和色彩理论特别关注并为之做了大量的实验。他的画作风格相当与众不同，Seurat的画**充满了细腻缤纷的小点**，当你靠近看，每一个点都充满著理性的笔触.



# Signac 进行 scATAC-seq 数据分析

## （1）数据加载

| Output Files | Format details | Size | md5sum |
|---|---|---|---|
| Clustering analysis | | 178 MB | e5486e686aabd9d2cb4beb2b060d5667 |
| Peak by cell matrix (filtered) | | 351 MB | a0870f3ef44ecfc8ce462c53faec5067 |
| Peak by cell matrix HDF5 (filtered) | | 170 MB | 647d691830e264c1d723ccfa75b44efc |
| Transcription Factor by cell matrix (filtered) | | 29.9 MB | 3469bffbf00a84aaaebf3932761ae920 |
| Transcription Factor by cell matrix HDF5 (filtered) | | 8.61 MB | b8b58ebef12d5e2881a8f06a2eaa716b |
| Fragments (TSV) | | 2.64 GB | 772c3330f03fd0ba3a6dbb6bda3f3cad |
| Fragments index (TBI) | | 1.15 MB | 6ced7825c817dda059600e70be27f466 |
| Peak annotation (CSV) | | 9.2 MB | d07d475587d2b6fa6dc5065e76ed9a58 |
| Peak-motif associations (BED) | | 32.6 MB | 09752d4d366690f78796d645ebd9fa0f |
| Peaks (BED) | | 3.95 MB | d81ab269baa773279133d319efcfc046 |
| Position-sorted alignments (BAM) | | 46.9 GB | 7f1af4f8192b6ee7b47effb25856916b |
| Position-sorted alignments index (BAI) | | 5.77 MB | 84518ef367f0f40e004d48883f818a58 |
| Peak by cell matrix (raw) | | 376 MB | 9cd3cc8941f0e125d22dfda0f4483943 |
| Peak by cell matrix HDF5 (raw) | | 183 MB | 93921721b7e45dd597c74426d1f1033f |
| Per Barcode metrics (CSV) | | 32.8 MB | 33b97f1c4fcffca84d24bcbdcf7d4dbf |
| Summary Metrics CSV | | 1.12 kB | e6c57059b4c58ff36d8fdced48b12fbb |
| Summary Metrics JSON | | 24.7 kB | c8d1f5d63722f2fabe23fd6a0164b8b1 |
| Summary HTML | | 3.19 MB | 9355506ebc60a169c3882695261bd57e |
| Loupe Browser file | | 1.44 GB | 66962ecaf25e70a9c9a0faa401d0c7ab |




## Peak-cell 矩阵

- 横坐标为每个peak区域，纵坐标为每个细胞。类似于单细胞的gene-cell表达矩阵
- 每个peak区域代表预测的每一个染色质开放区域
- 矩阵内的数值代表该细胞在这个peak位置的Tn5酶结合的个数

## Fragment文件

- 文件为一个列表，其中包含了所有单细胞中的所有的唯一片段
- 文件很大，其作用就是展示一个细胞中所有的片段，而不是仅包含在peak里面的片段

# Signac 进行 scATAC-seq 数据分析

（1）数据加载

```
pbmc <- CreateSeuratObject(
  counts = chrom_assay,
  assay = "peaks",
  meta.data = metadata
)
```

```
> pbmc
An object of class Seurat
165434 features across 10246 samples within 1 assay
Active assay: peaks (165434 features, 0 variable features)
```

```
> pbmc@assays[["peaks"]]@counts
165434 x 10246 sparse Matrix of class "dgCMatrix"
  [[ suppressing 60 column names 'AAACGAAAGAGAGGTA-1', 'AAACGAAAGCAGGAGG-1', 'AAACGAAAGGAAGAAC-1' ... ]]
  [[ suppressing 60 column names 'AAACGAAAGAGAGGTA-1', 'AAACGAAAGCAGGAGG-1', 'AAACGAAAGGAAGAAC-1' ... ]]

chr1-9772-10660     . . . . . .
chr1-180712-181178  . . . . . 2 . . . . . .
chr1-181200-181607  . . . . . 2 . . . . . .
chr1-191183-192084  . . . . . .
chr1-267576-268461  . . 2 . . . . . .
chr1-270850-271755  . . . . . .
chr1-273946-274792  . . . . . .
chr1-585753-586648  . . 2 . . . . . 2 . . . . . .
```

Peak-cell 矩阵

```
> head(pbmc@meta.data)
                    orig.ident nCount_peaks nFeature_peaks total duplicate chimeric unmapped lowmapq mitochondrial nonprimary passed_filters is__cell_barcode excluded_reason
AAACGAAAGAGAGGTA-1 SeuratProject        32618          12605 59781     33159        1      633    3228           221         17          22522                1               0
AAACGAAAGCAGGAGG-1 SeuratProject        13293           5392 19399      9003        1      231    1137             2          3           9022                1               0
AAACGAAAGGAAGAAC-1 SeuratProject        36155          13306 64452     31013        4      549    4538           182          4          28162                1               0
AAACGAAAGTCGACCC-1 SeuratProject        40155          14049 72316     38705        5      622    3831            11          6          29136                1               0
AAACGAACAAGCACTT-1 SeuratProject        18998           7115 32569     16889        7      343    2108            89          5          13128                1               0
AAACGAACAAGCGGTA-1 SeuratProject        39218          14514 81259     40822        3      800    5299            15         15          34305                1               0
                   TSS_fragments DNase_sensitive_region_fragments enhancer_region_fragments promoter_region_fragments on_target_fragments blacklist_region_fragments
AAACGAAAGAGAGGTA-1          9962                                0                           0                           0                9962                                0
AAACGAAAGCAGGAGG-1          5362                                0                           0                           0                5362                                0
AAACGAAAGGAAGAAC-1         13887                                0                           0                           0               13887                                0
AAACGAAAGTCGACCC-1         15159                                0                           0                           0               15159                                0
AAACGAACAAGCACTT-1          7542                                0                           0                           0                7542                                0
AAACGAACAAGCGGTA-1         15157                                0                           0                           0               15157                                0
```

细胞表型信息

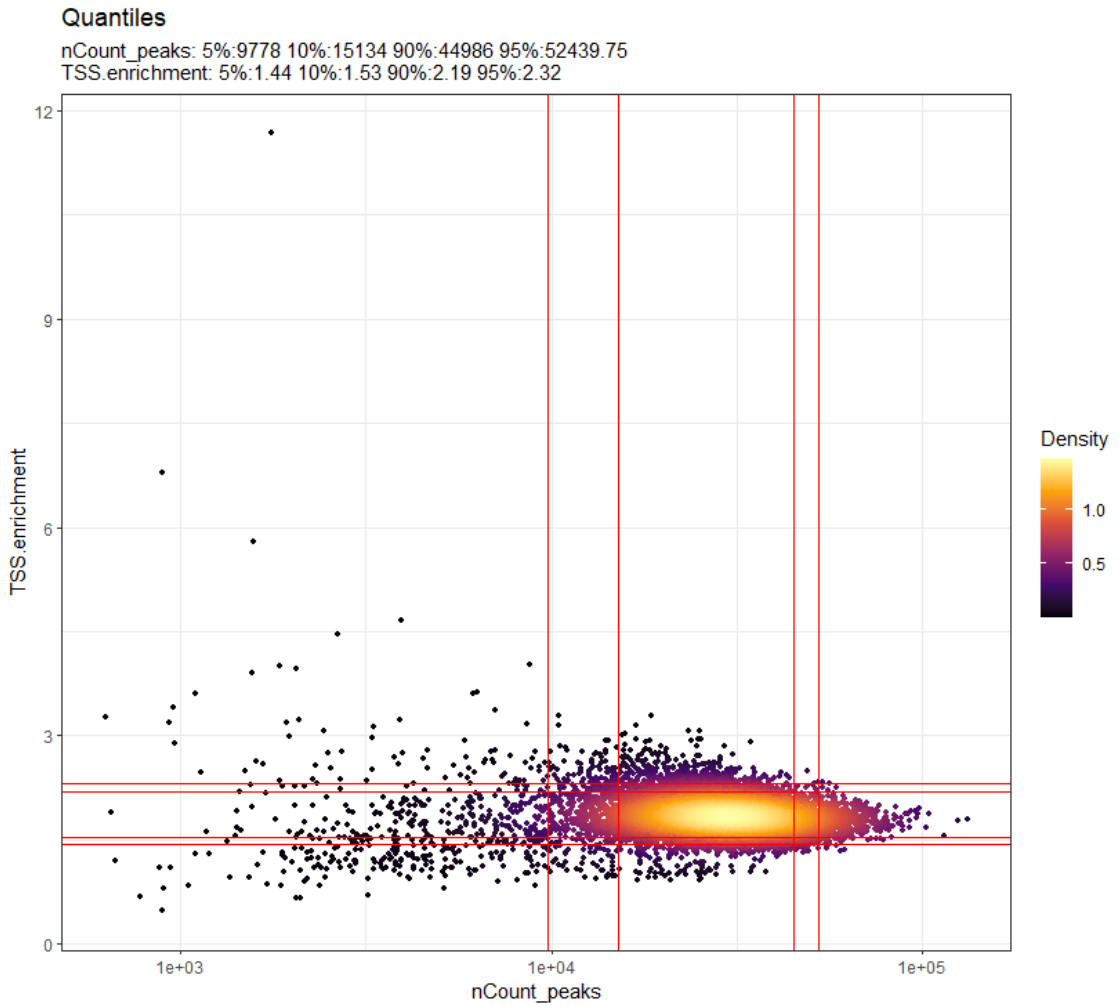
# Signac 进行 scATAC-seq数据分析

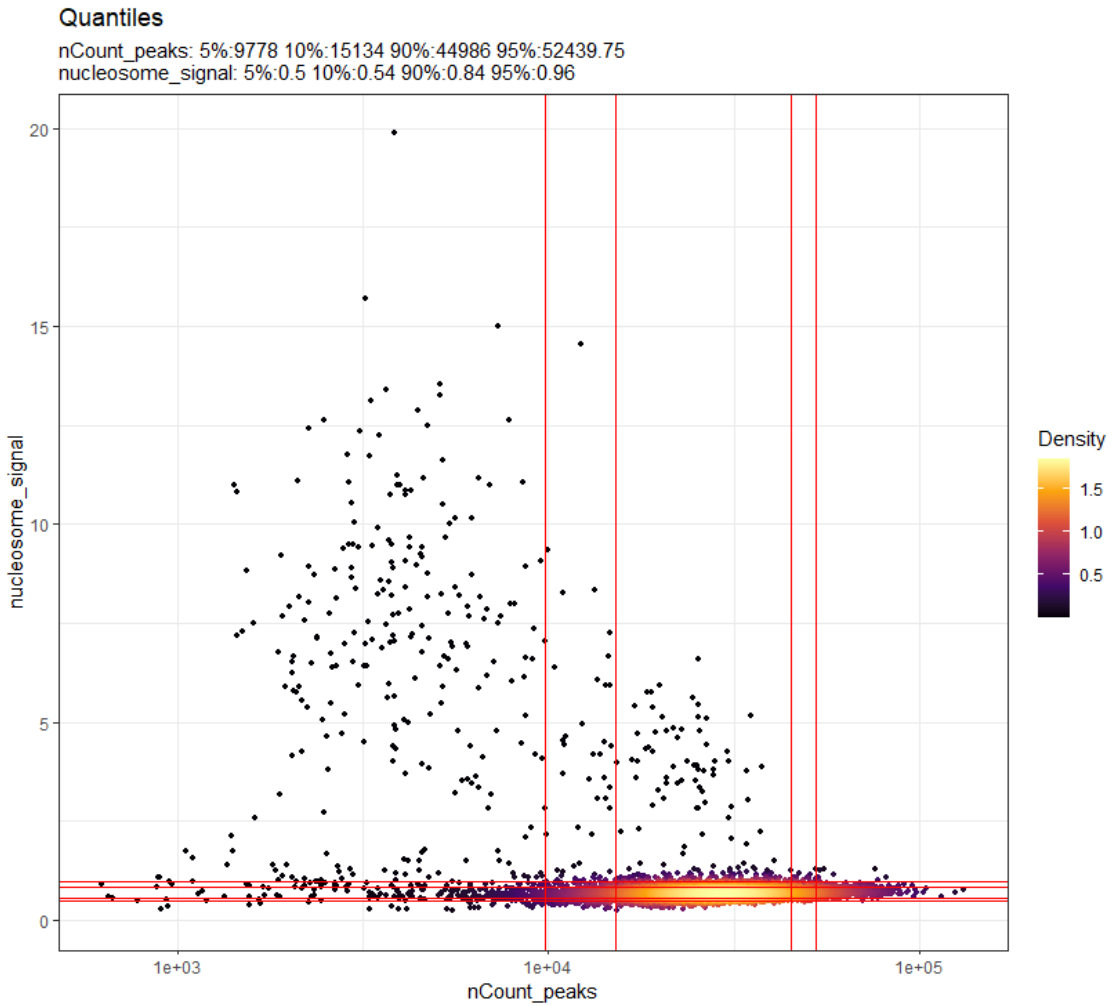
（2）质量控制：5个质量控制指标

| 名称 | 描述 |
| --- | --- |
| Nucleosome banding pattern<br>核小体绑定模式 | 绘制片段大小的直方图（由PE读序确定）来展示核小体在染色体上的绑定模式。<br>量化每个单细胞中核小体绑定与无核小体片段的近似比率（**越低越好**）。<br>NucleosomeSignal函数计算该指标，结果存储在metadata中，列名称为nucleosome_signal。 |
| Transcriptional start site (TSS) enrichment score<br>转录起始位置富集分值 | 根据覆盖TSS的片段与TSS两边片段的比值来定义。<br>通常将具有较低的TSS富集分值的数据质量较差（**越高越好**）。<br>TSSEnrichment函数计算该指标，结果存储在metadata中，列名称为TSS.enrichment |
| Total number of fragments in peaks<br>比对到峰上的片段总数 | 度量细胞测序深度/复杂性。排除读序少的细胞（测序深度低），读序高的细胞可能是双细胞或多细胞。 |
| Fraction of fragments in peaks<br>比对到峰上的片段比例 | 该比例较低的细胞（即<15-20%）通常代表低质量细胞或技术误差。 |
| Ratio reads in genomic blacklist regions<br>比对到基因组blacklist区域的读序比例 | ENCODE项目提供了一个blacklist区域列表，这些区域通常与人为信号相关。读序比对到这些区域的比例较高的细胞（与比对到peaks区域比例相比）通常为技术误差，应将其删除。Signac包中包含了人类（hg19和GRCh38），小鼠（mm10），果蝇（dm3）和秀丽隐杆线虫（ce10）的 blacklis区域。 |

# Signac 进行 scATAC-seq数据分析

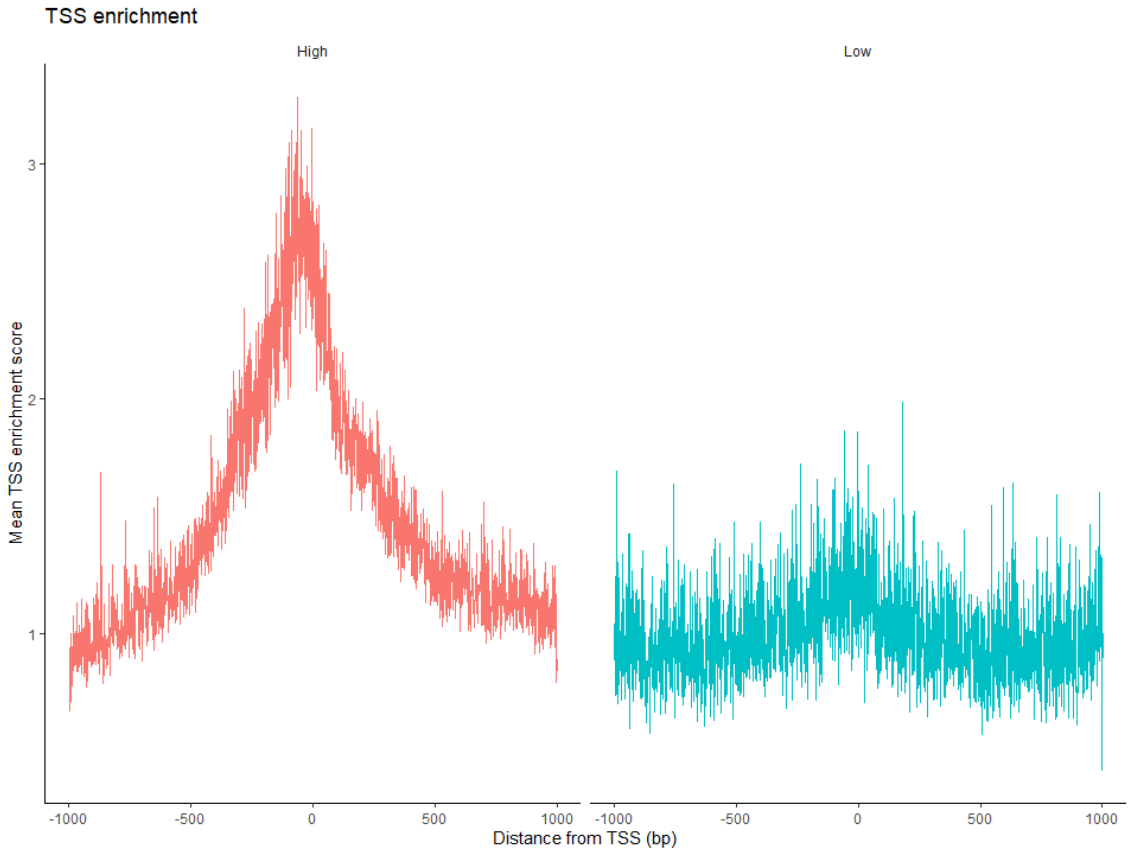
（2）质量控制：TSS.enrichment（越高越好）和nucleosome_signal （越低越好）



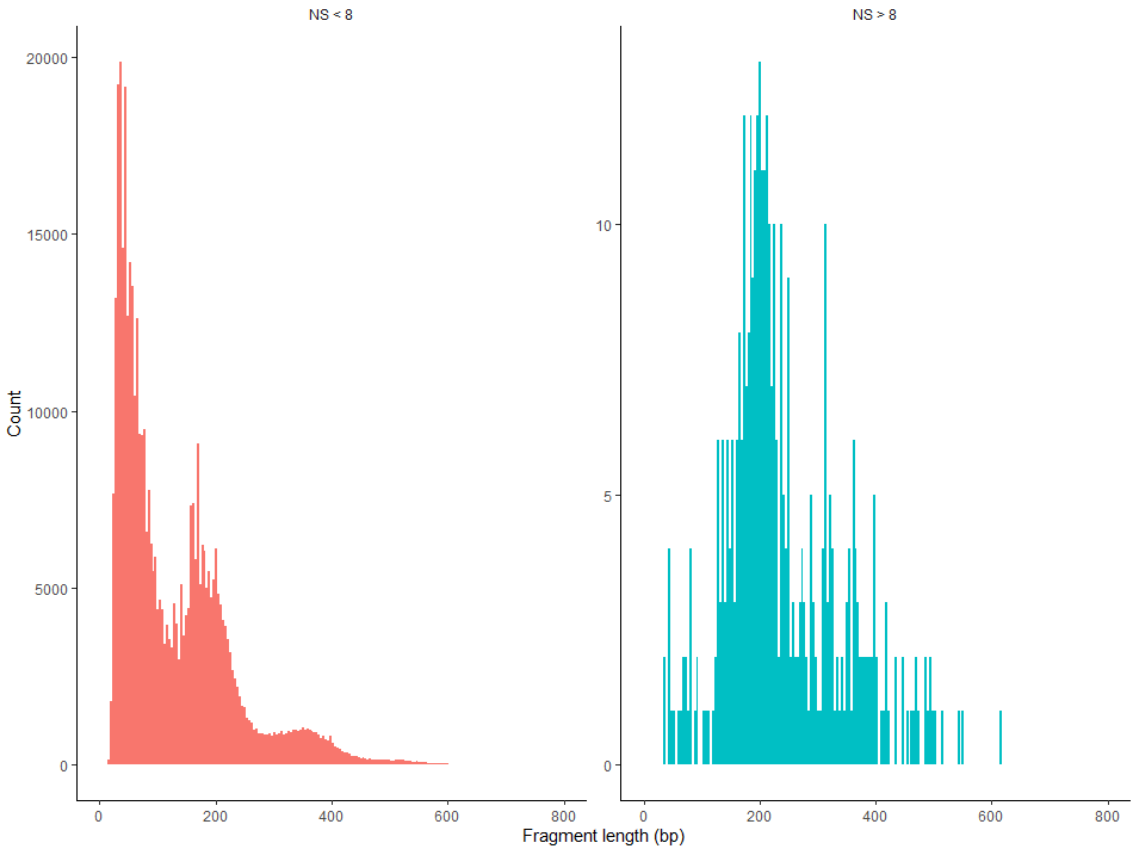



# Signac 进行 scATAC-seq数据分析

（2）质量控制：TSS.enrichment（越高越好）和nucleosome_signal （越低越好）




- 不同转录起始位置富集分值在TSS附近的分布不一样



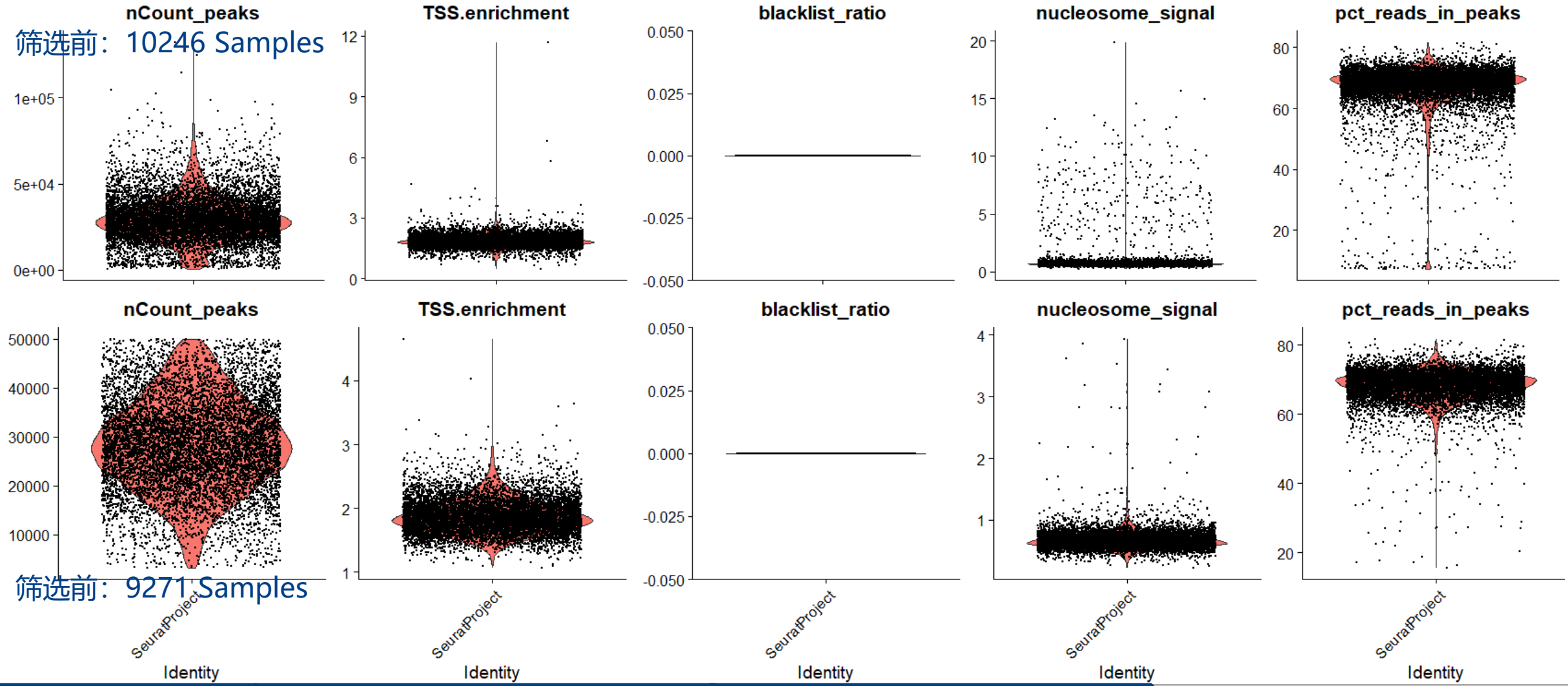

- 不同 nucleosome_signal 值有不同的片段长度分布模式

# Signac 进行 scATAC-seq数据分析

（2）质量控制：筛选低质量细胞前后比较

筛选标准：nCount_peaks > 3000 & nCount_peaks < 50000 & pct_reads_in_peaks > 15 & blacklist_ratio < 0.05 & nucleosome_signal < 4 & TSS.enrichment > 1

筛选前：10246 Samples




筛选前：9271 Samples

# Signac 进行 scATAC-seq数据分析

## （3）数据标准化、特征选择与线性降维

- 标准化（Normalization）：包括两方面，一是校正细胞测序深度的差异；而是在峰间进行归一化，为更罕见的峰提供更高的值。

```
> pbmc@assays[["peaks"]]@counts
165434 x 9271 sparse Matrix of class "dgCMatrix"
  [[ suppressing 60 column names 'AAACGAAAGAGAGGTA-1', 'AAACGAAAGCAGGAGG-1', 'AAACGAAAGGAAGAAC-1' ... ]]
  [[ suppressing 60 column names 'AAACGAAAGAGAGGTA-1', 'AAACGAAAGCAGGAGG-1', 'AAACGAAAGGAAGAAC-1' ... ]]

chr1-9772-10660     . . . . . .
chr1-180712-181178  . . . . . 2 . . . . . .
chr1-181200-181607  . . . . . 2 . . . . . .
chr1-191183-192084  . . . . . .
chr1-267576-268461  . . . . . 2 . . . . . .
chr1-270850-271755  . . . . . .
chr1-273946-274792  . . . . . .
chr1-585753-586648  . . . . . 2 . . . . . 2 . . . . . .
```

pbmc <- RunTFIDF(pbmc)

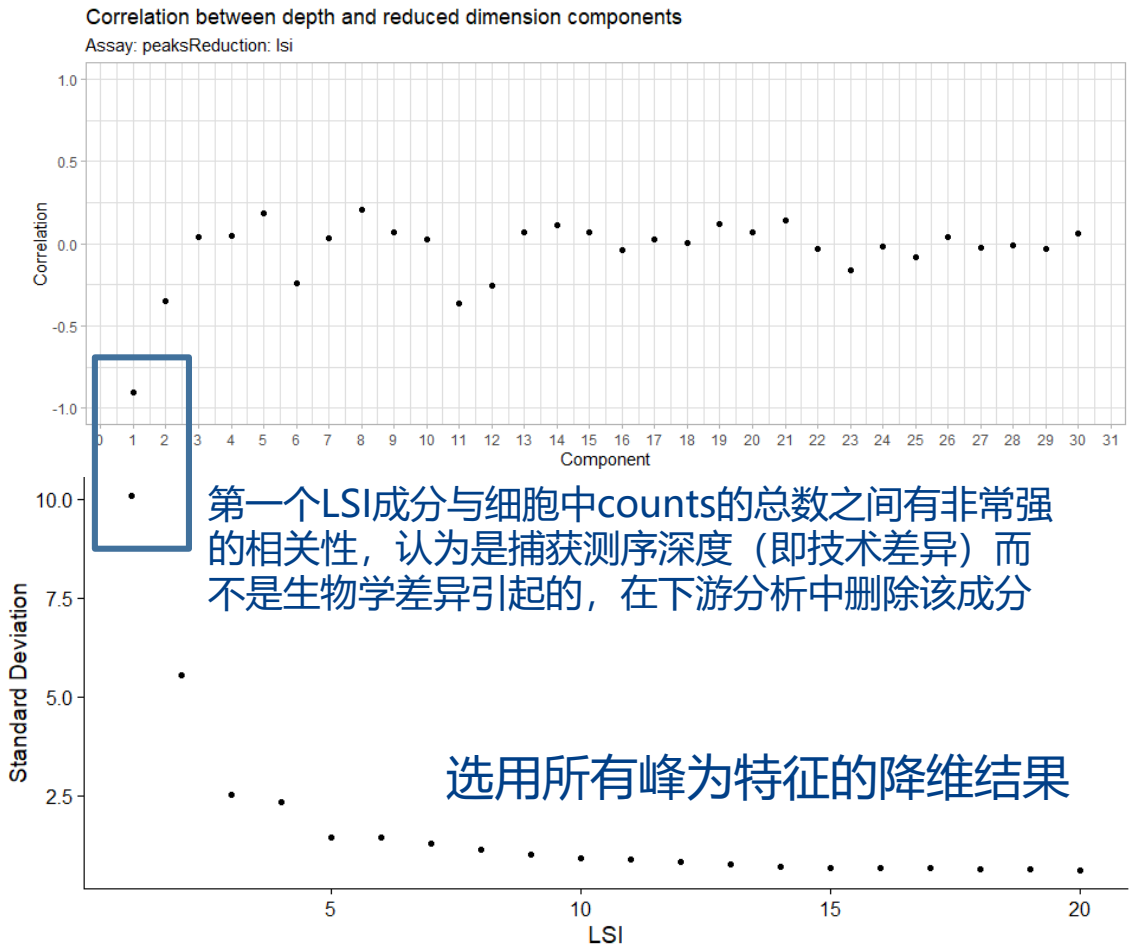
```
> pbmc@assays[["peaks"]]@data
165434 x 9271 sparse Matrix of class "dgCMatrix"
  [[ suppressing 60 column names 'AAACGAAAGAGAGGTA-1', 'AAACGAAAGCAGGAGG-1', 'AAACGAAAGGAAGAAC-1' ... ]]
  [[ suppressing 60 column names 'AAACGAAAGAGAGGTA-1', 'AAACGAAAGCAGGAGG-1', 'AAACGAAAGGAAGAAC-1' ... ]]

chr1-9772-10660     . . . . . .
chr1-180712-181178  . . . . . 3.10281 . . . . . .
chr1-181200-181607  . . . . . 3.740337 . . . . . .
chr1-191183-192084  . . . . . .
chr1-267576-268461  . . . . . 3.258034 . . . . . .
chr1-270850-271755  . . . . . .
chr1-273946-274792  . . . . . .
chr1-585753-586648  . . . . . 3.107148 . . . . . 3.291585 . . . . . .
```

# Signac 进行 scATAC-seq数据分析

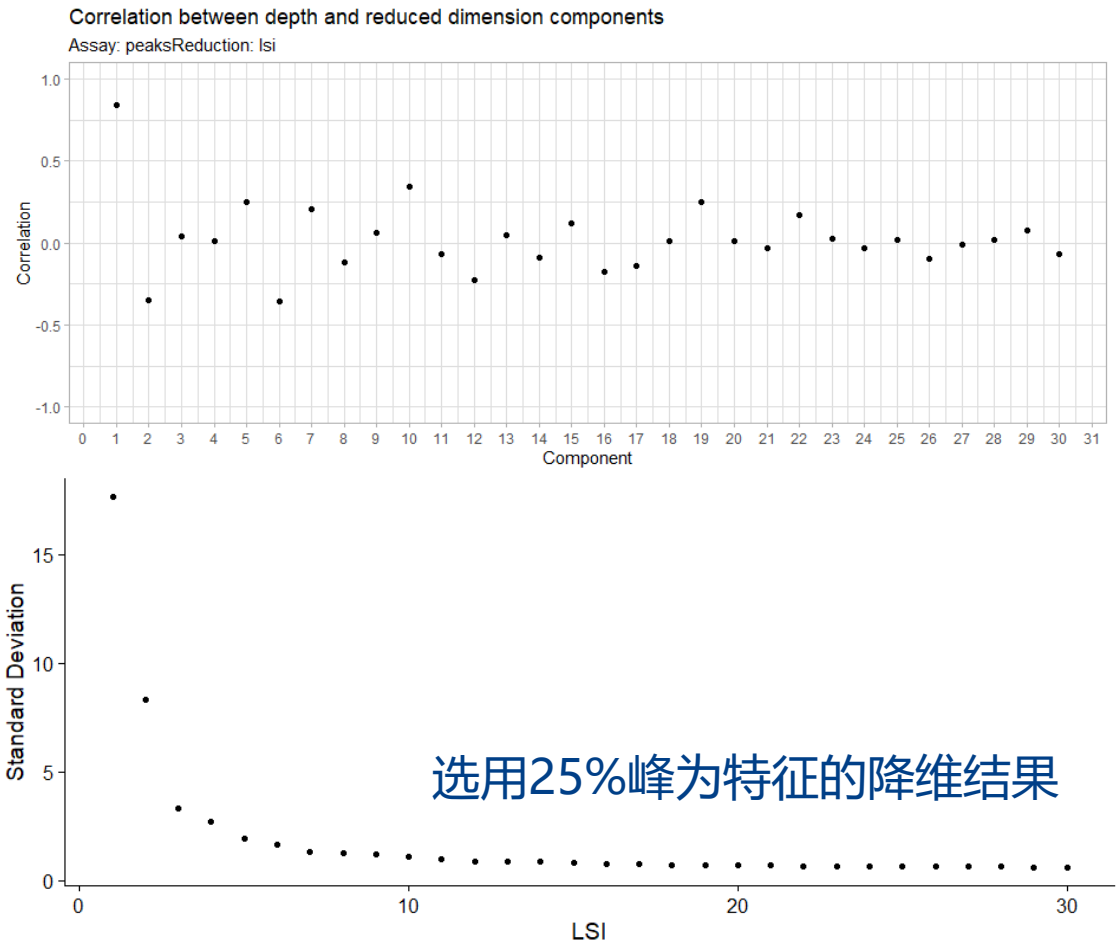
## （3）数据标准化、特征选择与线性降维

- 特征选择（Feature selection）：与scRNA-seq数据分析类似，n个特征峰用于后续的数据降维
- 降维（Dimensional reduction）：与scRNA-seq数据分析中的PCA类似，这里使用奇异值分解（SVD）返回对象的低维数据




第一个LSI成分与细胞中counts的总数之间有非常强的相关性，认为是捕获测序深度（即技术差异）而不是生物学差异引起的，在下游分析中删除该成分

选用所有峰为特征的降维结果

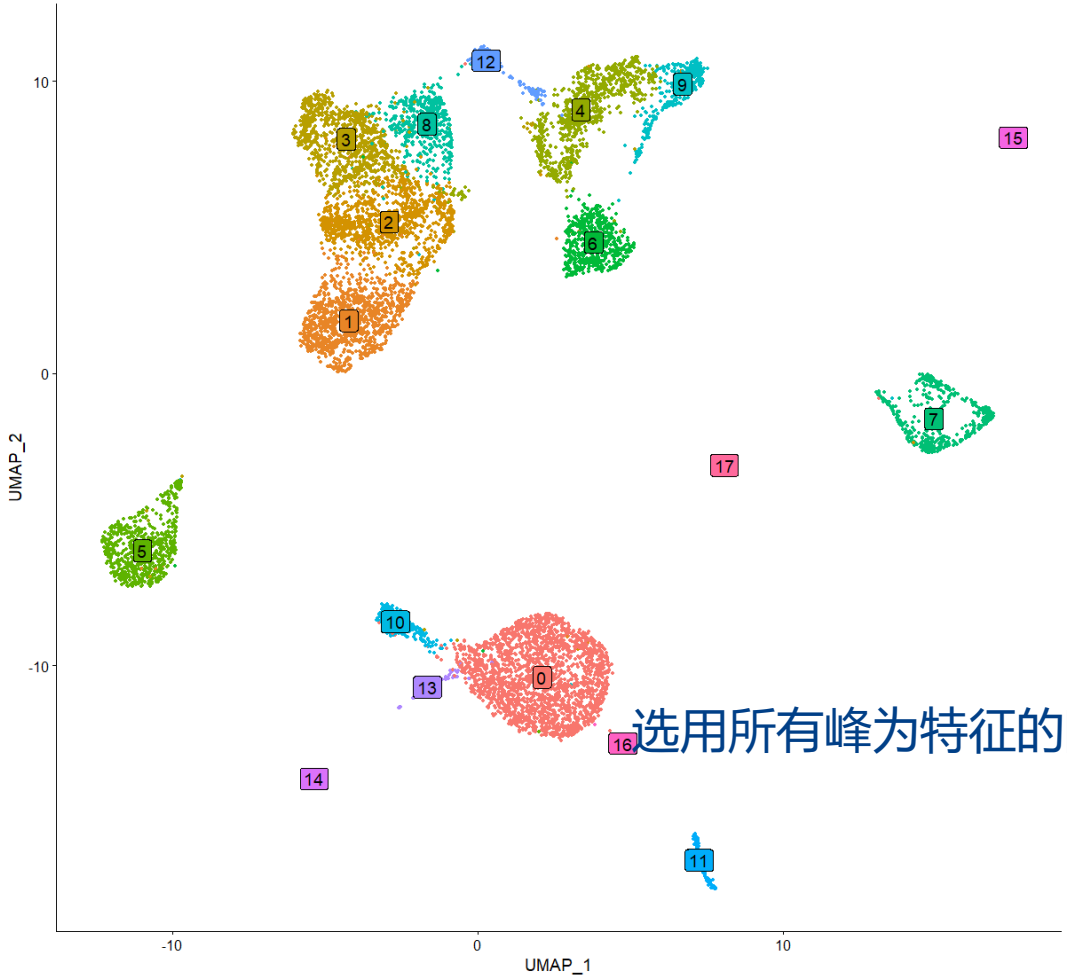



选用25%峰为特征的降维结果

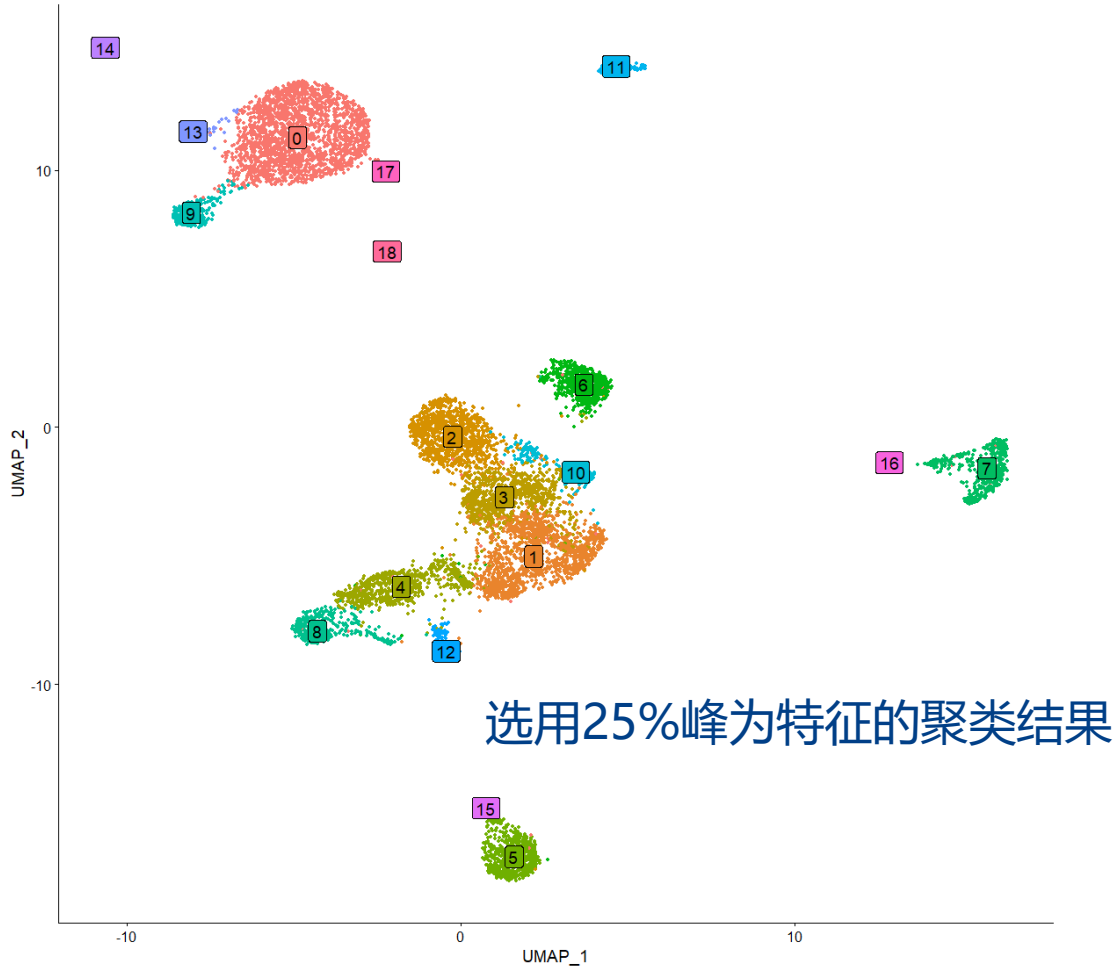
# Signac 进行 scATAC-seq数据分析

## （4）非线性降维与细胞聚类

- 与scRNA-seq数据分析类似，使用 RunUMAP，FindNeighbors 和 FindClusters 函数




选用所有峰为特征的降维结果



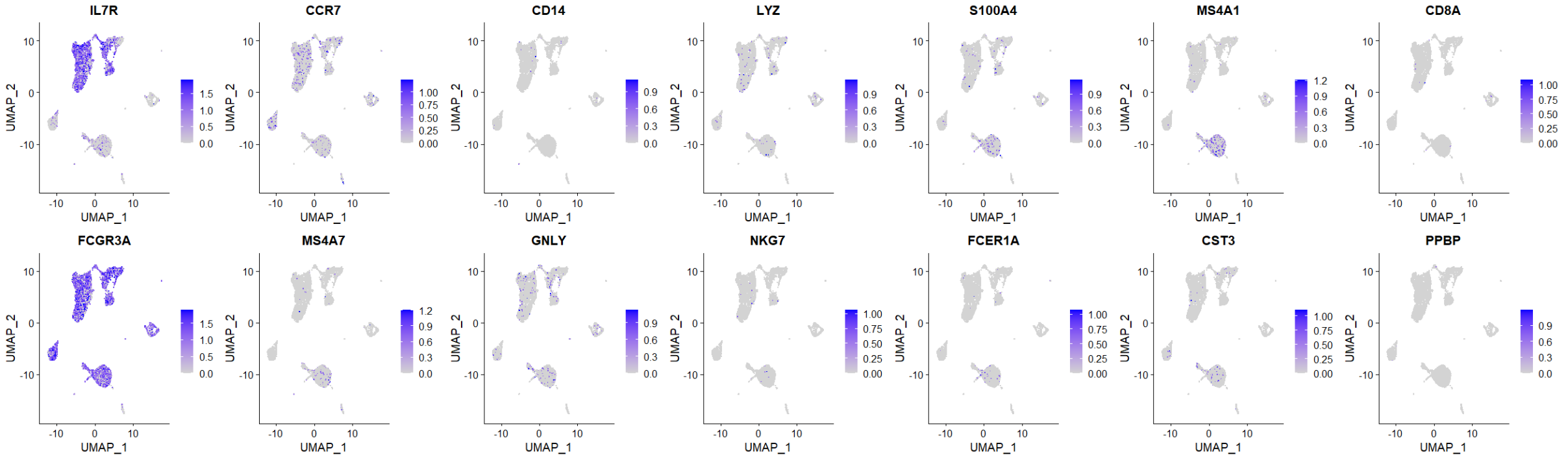

选用25%峰为特征的聚类结果

# Signac 进行 scATAC-seq数据分析

## （5）创建基因活性表达矩阵（gene activity matrix）

- 通过评估与每个基因相关的染色质可及性来量化基因组中每个基因的表达活性（将基因的区域扩展到TSS上游2kb区域）
- 基因活性矩阵比scRNA-seq数据的噪音大，因为假定了启动子区域的可及性与基因表达的相关性

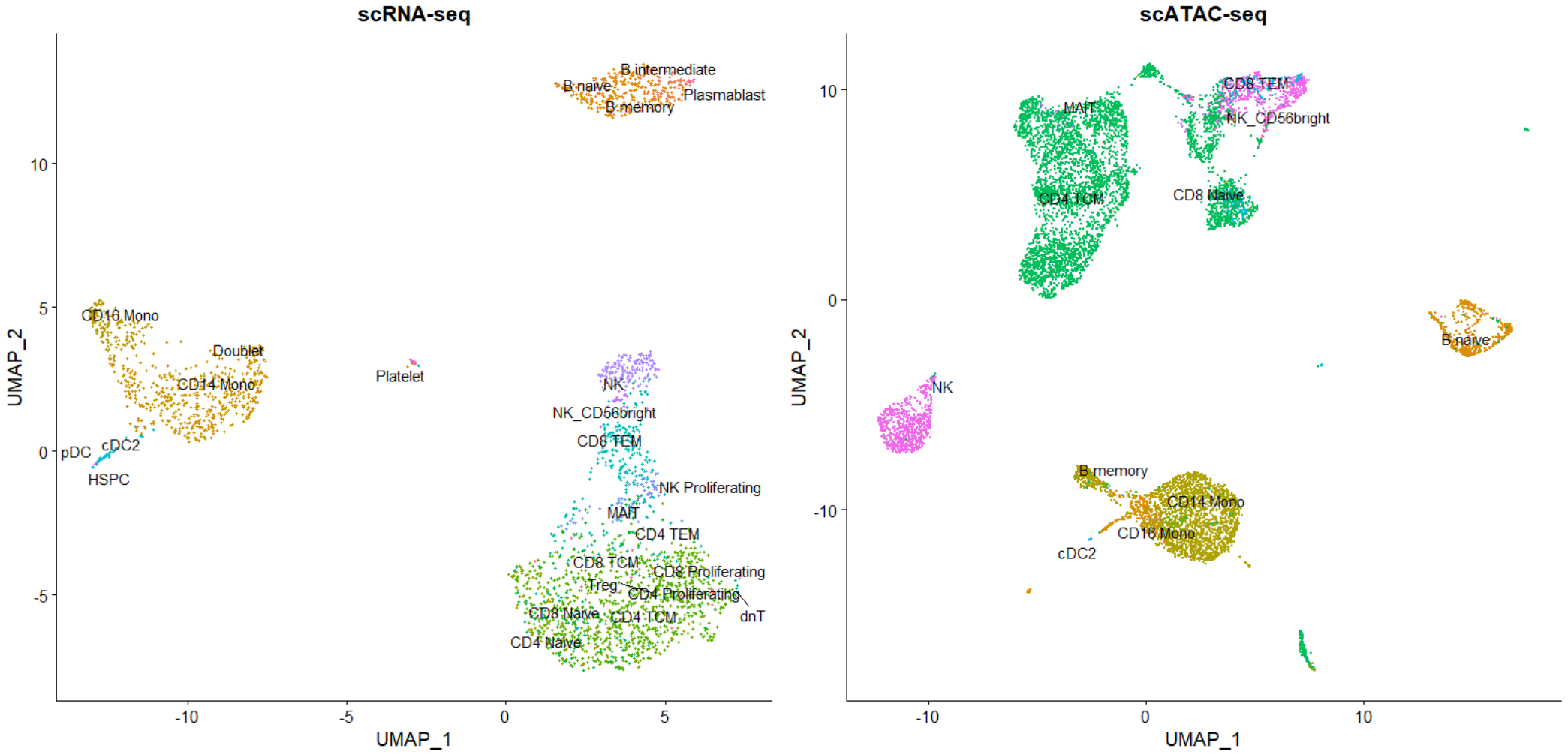
| Markers | Cell Type |
|---|---|
| IL7R, CCR7 | Naive CD4+ T |
| CD14, LYZ | CD14+ Mono |
| IL7R, S100A4 | Memory CD4+ |
| MS4A1 | B |
| CD8A | CD8+ T |
| FCGR3A, MS4A7 | FCGR3A+ Mono |
| GNLY, NKG7 | NK |
| FCER1A, CST3 | DC |
| PPBP | Platelet |



# Signac 进行 scATAC-seq数据分析

## （6）与scRNA-seq数据整合辅助scATAC-seq数据的注释

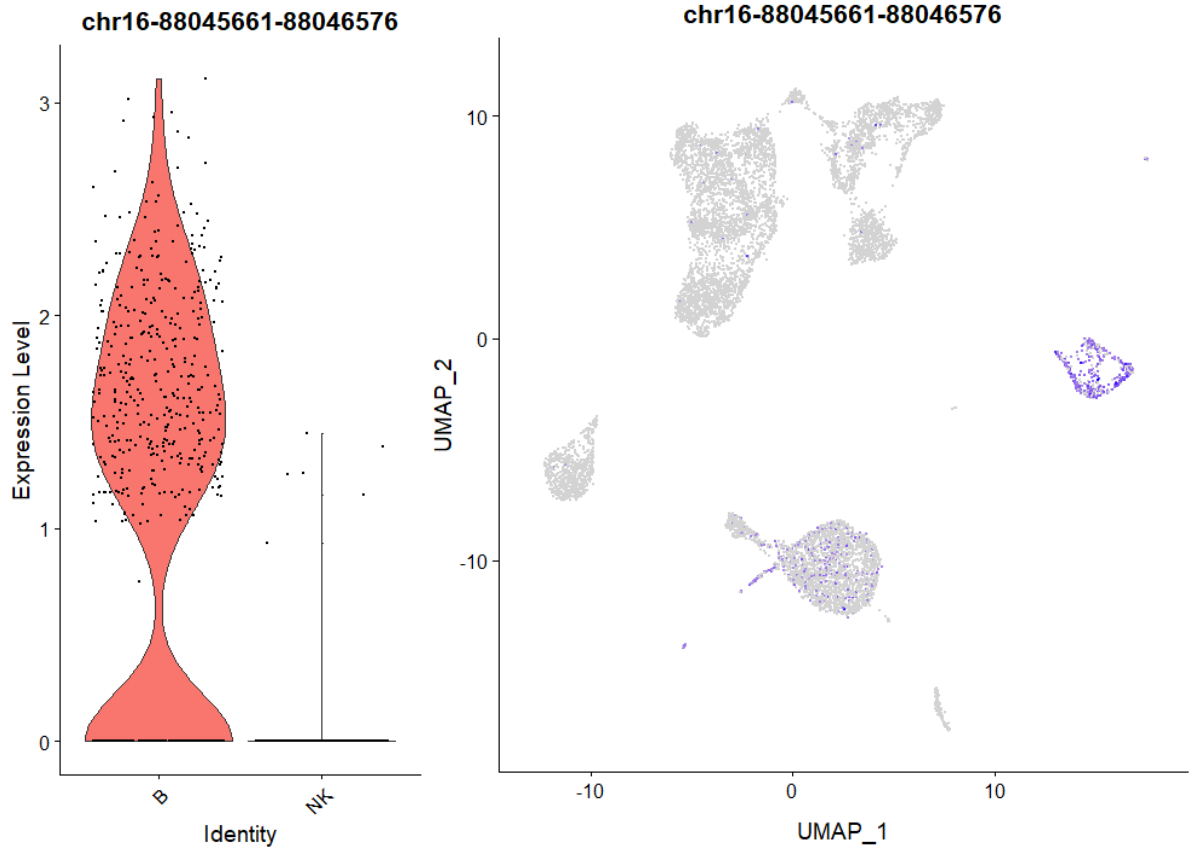
- FindTransferAnchors 函数寻找RNA和ATAC数据之间的锚点，TransferData 函数进行标签转移



# Signac 进行 scATAC-seq数据分析

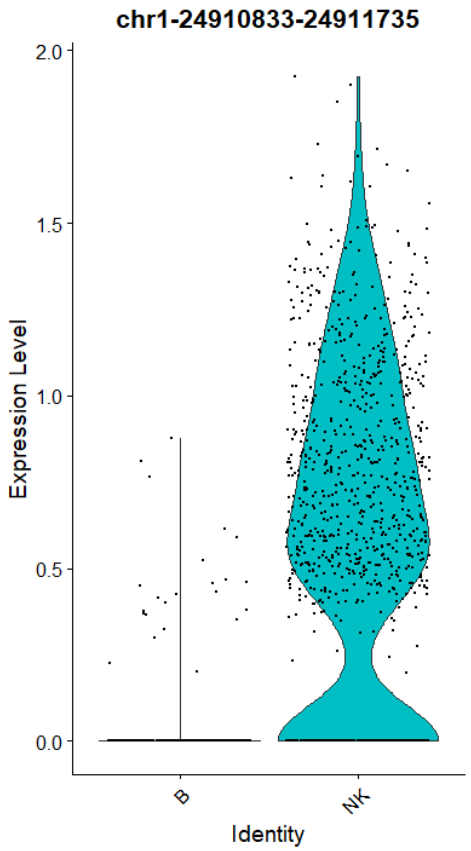
## （7）不同细胞簇之间染色质可及性差异峰的鉴定

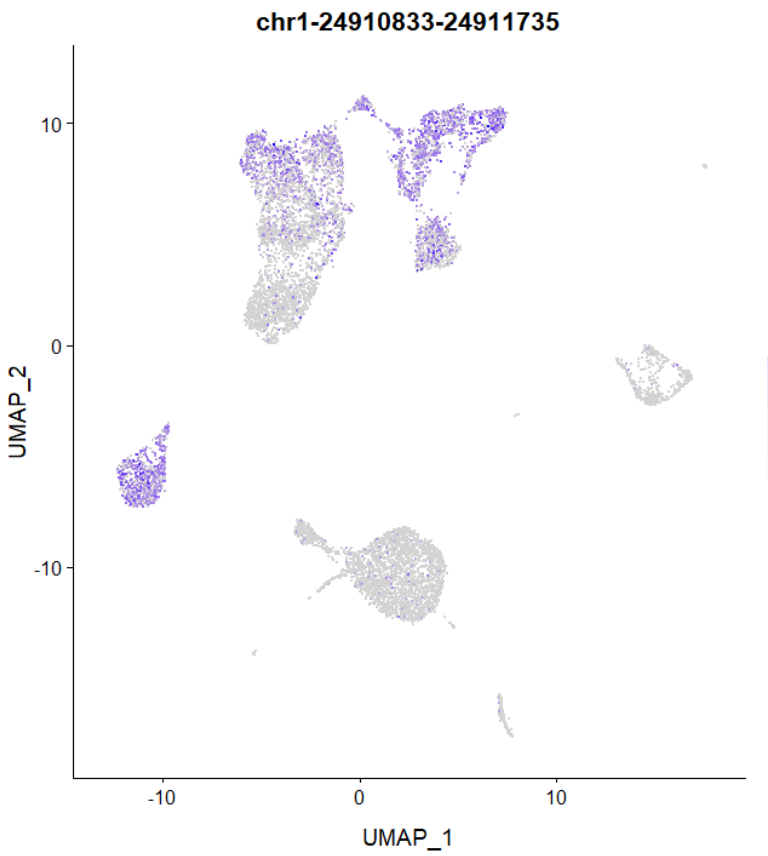
- 使用逻辑回归模型（LR）进行差异可及性（DA）分析







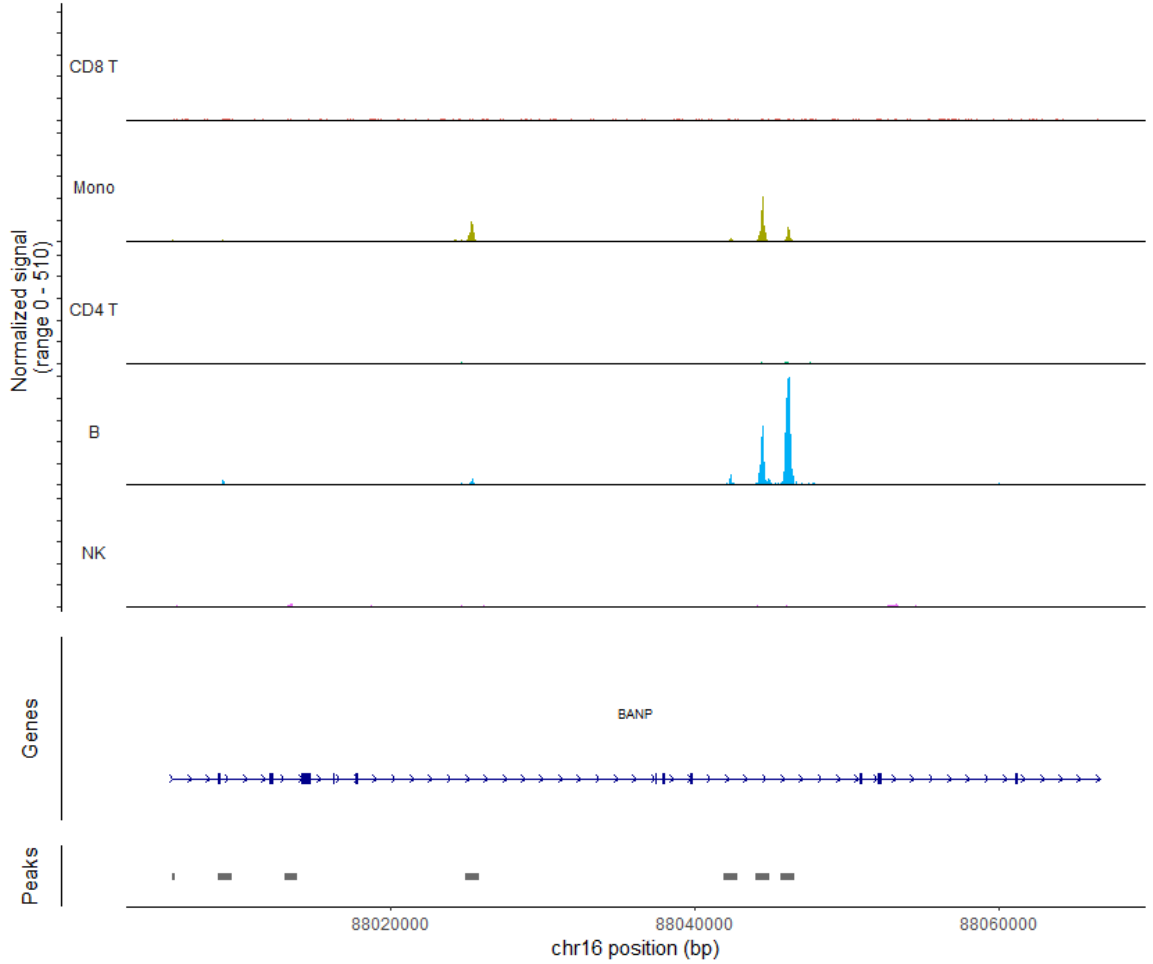


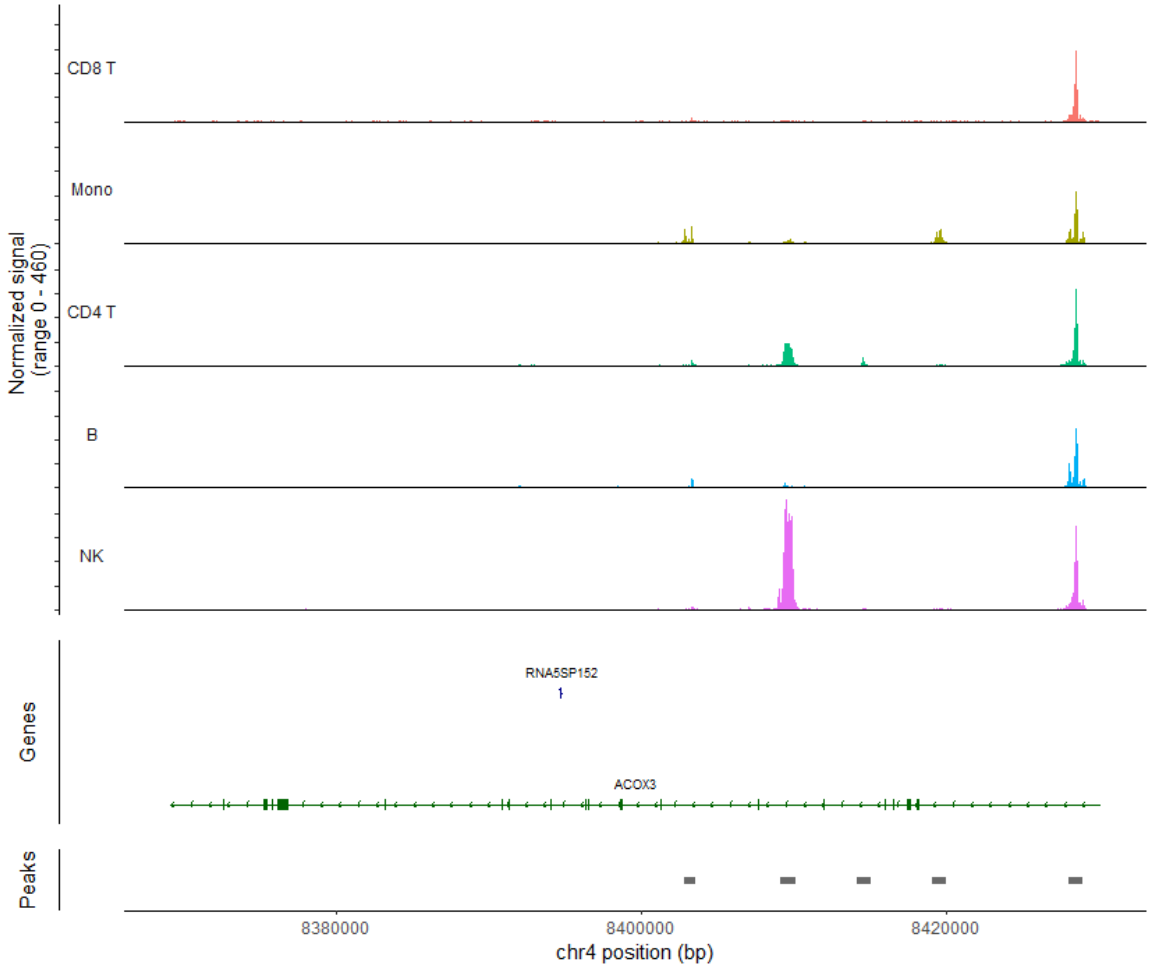



# Signac 进行 scATAC-seq数据分析

## （7）不同细胞簇之间染色质可及性差异峰的鉴定

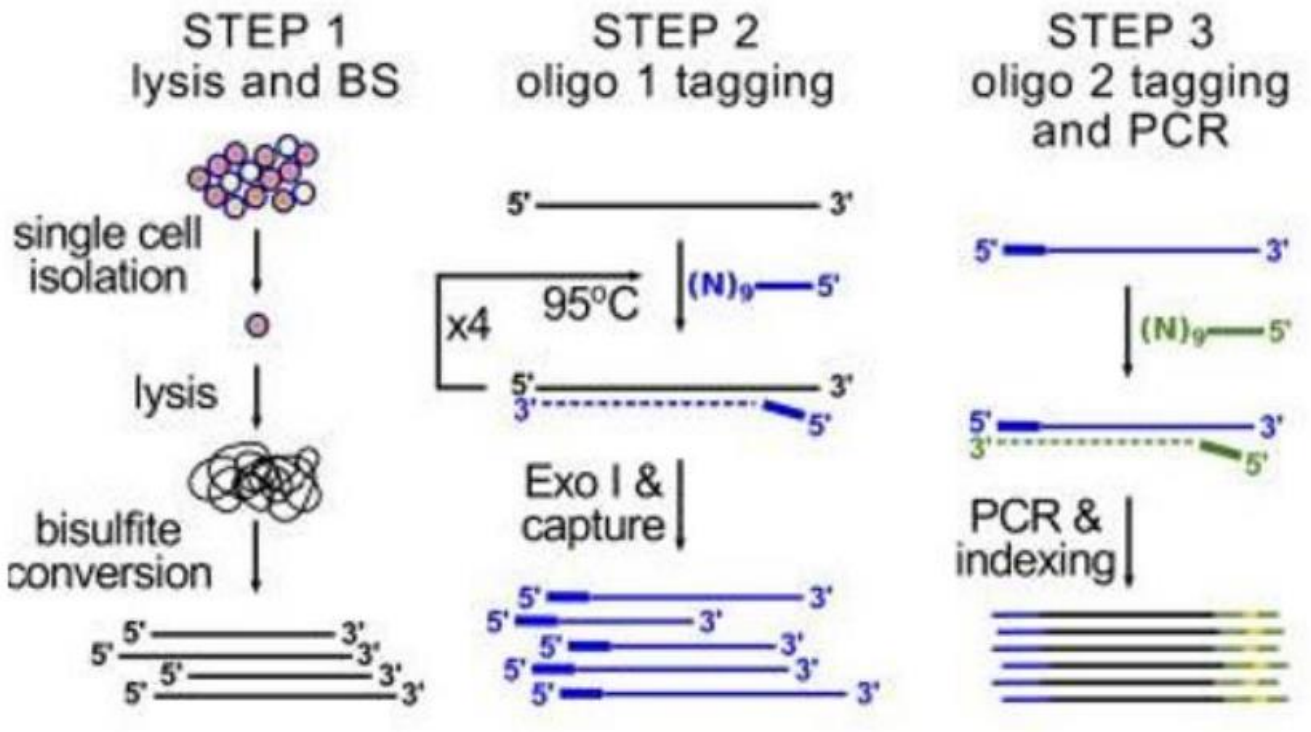
- 使用ClosestFeature函数提供的基因注释信息，可找到与每个峰最接近的基因



# 上节回顾：什么是单细胞DNA甲基化

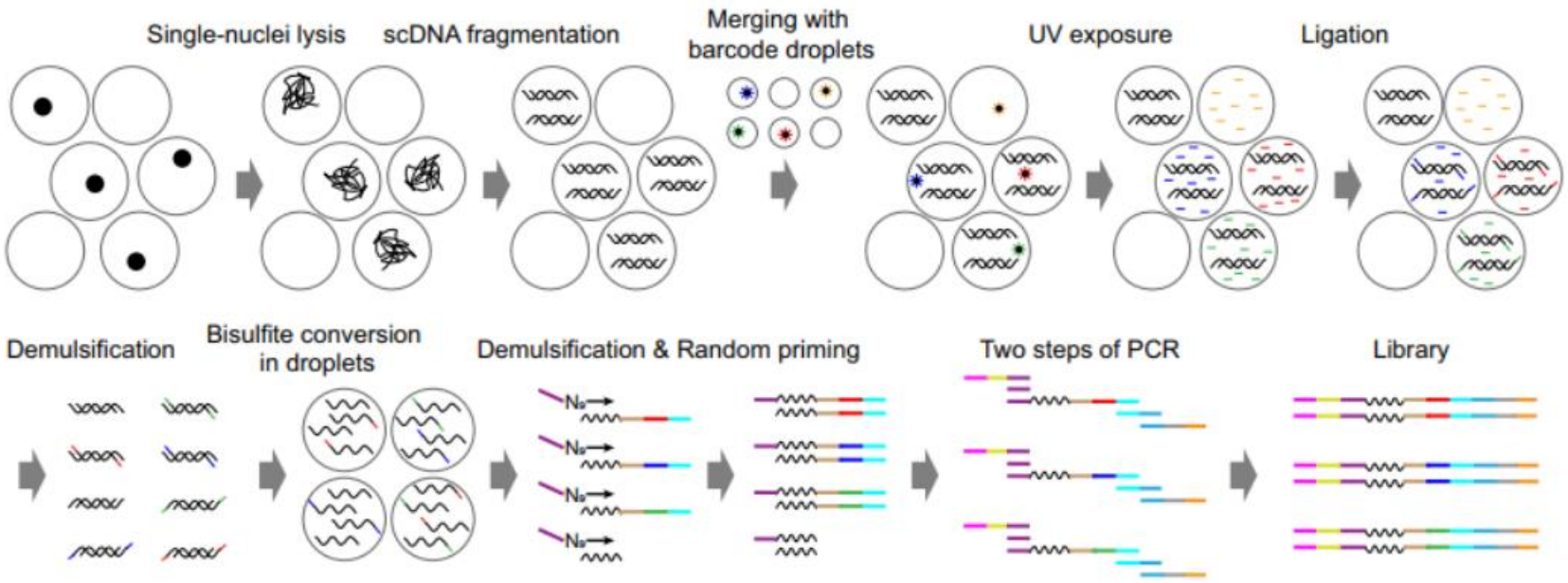
**单细胞DNA甲基化测序**：可以揭示不同细胞之间的表观遗传差异和表观遗传调控机制。DNA甲基化是一种重要的表观遗传修饰方式，它可以影响基因的表达和功能，从而对细胞的生长、分化、发育和疾病等方面产生影响。




使用亚硫酸氢盐处理DNA样本，将未甲基化的胞嘧啶（C）转化为脱氧尿嘧啶（T），而甲基化的胞嘧啶不被转化。然后对DNA进行测序，通过比对测序数据的C和T，可以推断出单个细胞中DNA的甲基化状态。

# 上节回顾：什么是单细胞DNA甲基化

基于液滴的高通量**单细胞亚硫酸氢盐测序平台（Drop-BS）**，利用液滴微流体技术实现了超高通量，并可在2天内制备多达10000个单细胞的亚硫酸氢盐测序文库。

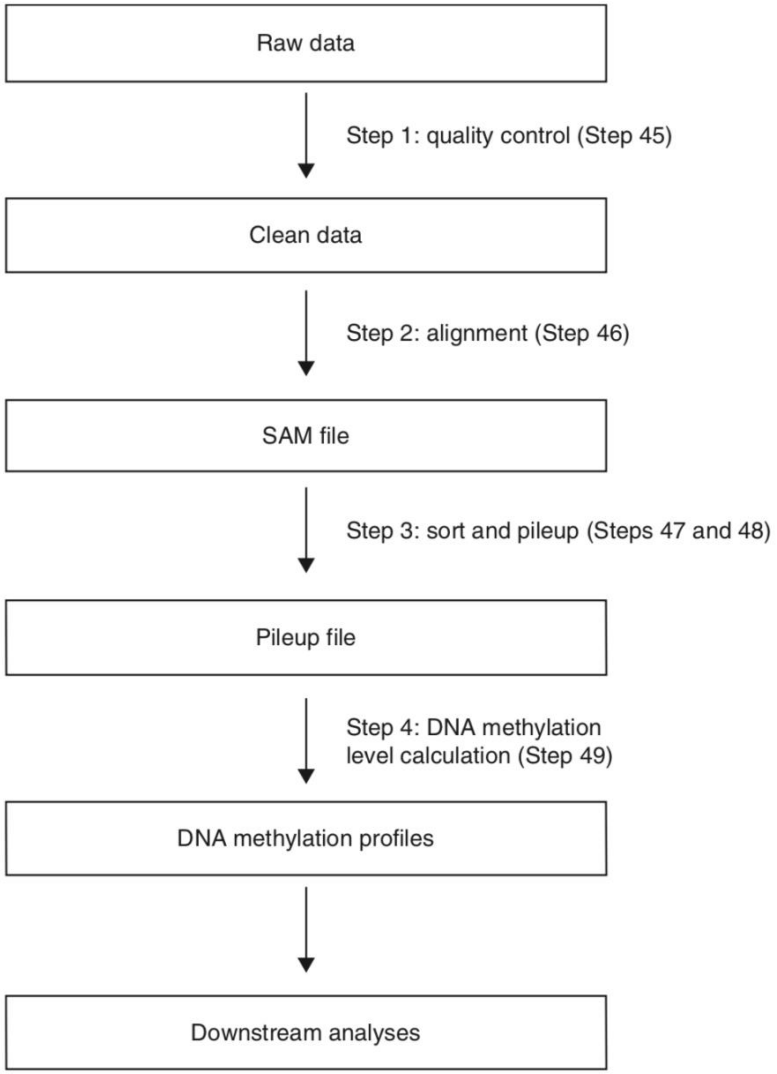



- 通过操纵液滴中的单个细胞和单个条形码，来自单细胞的DNA被独特的条形码进行标记。
- 条形码DNA在液滴中进行亚硫酸氢盐转换，转换后的DNA最终进行PCR生成测序文库。

Qiang Z. et al., *Nature Communications*, 2023

# 单细胞DNA甲基化数据分析

## 单细胞甲基化数据上游分析流程介绍




具体流程：

- 获得测序原始读序
- 质控去除低质量或接头污染的读序
- 利用bismark等软件将读序比对到参考基因组上，获得比对BAM文件，并且生成pileup文件
- 根据检测到的 C（甲基化读序）除以相同基因组位置上检测到的 C 和 T 的总数（总读序数）来计算每个胞嘧啶的 DNA 甲基化水平

nature protocols

Explore content ⌄ Journal information ⌄ Publish with us ⌄

nature › nature protocols › protocols › article

Published: 02 April 2015

**Profiling DNA methylome landscapes of mammalian cells with single-cell reduced-representation bisulfite sequencing**

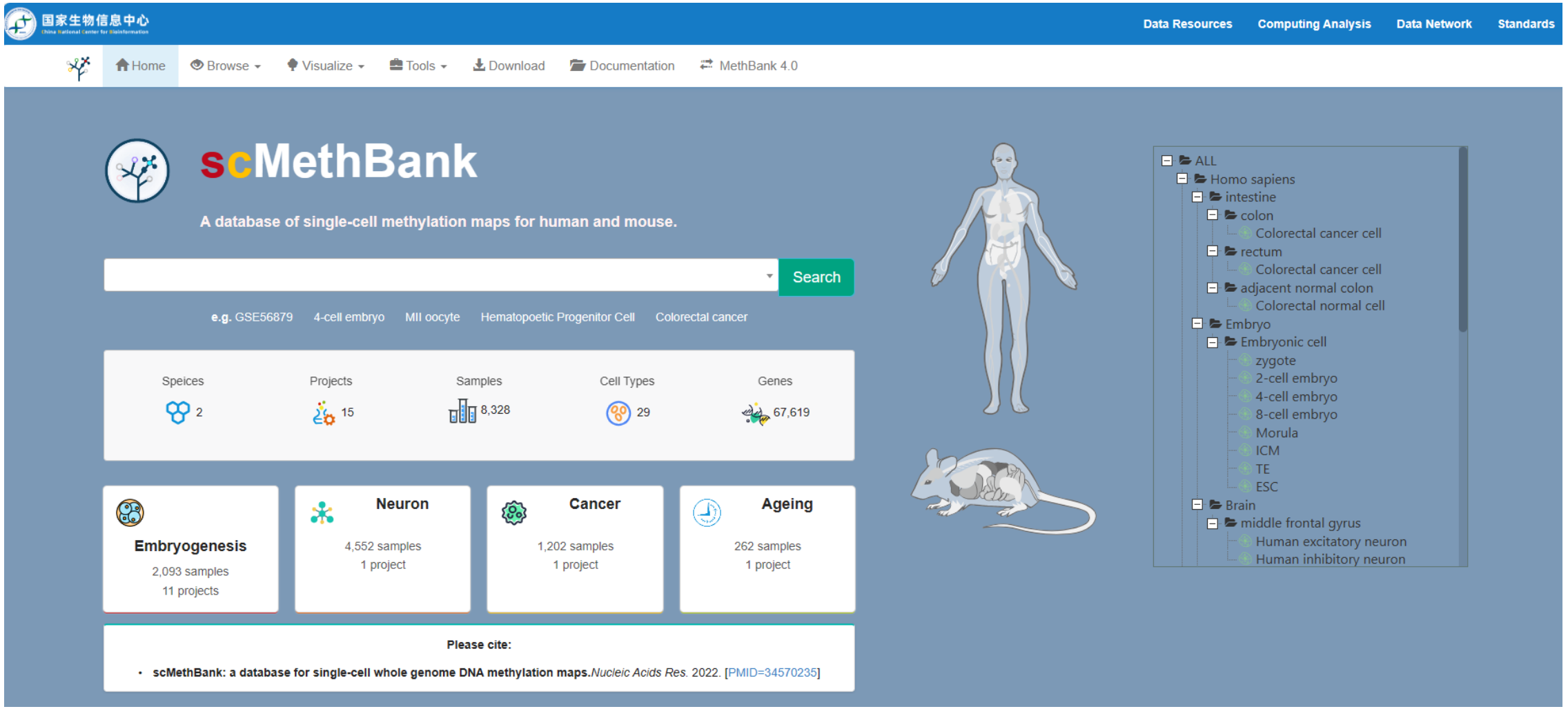
Hongshan Guo, Ping Zhu, Fan Guo, Xianlong Li, Xinglong Wu, Xiaoying Fan, Lu Wen ✉ & Fuchou Tang ✉

*Nature Protocols* **10**, 645–659 (2015) | Cite this article

**2407** Accesses | **86** Citations | **6** Altmetric | Metrics

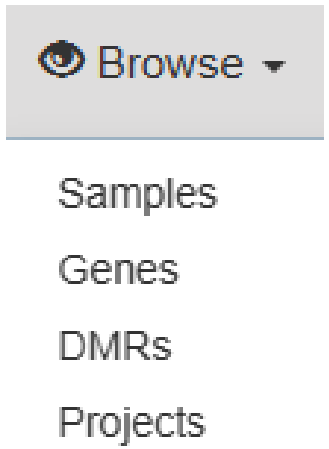
# 单细胞DNA甲基化数据库scMethBank

- scMethBank：**人类和小鼠**单细胞甲基化图谱数据库（来自两个物种的 15 个公共单细胞数据集的 8328 个样本，涉及29种细胞类型和两种疾病），提供浏览、搜索、可视化、下载功能和用户友好的在线工具



# scMethBank

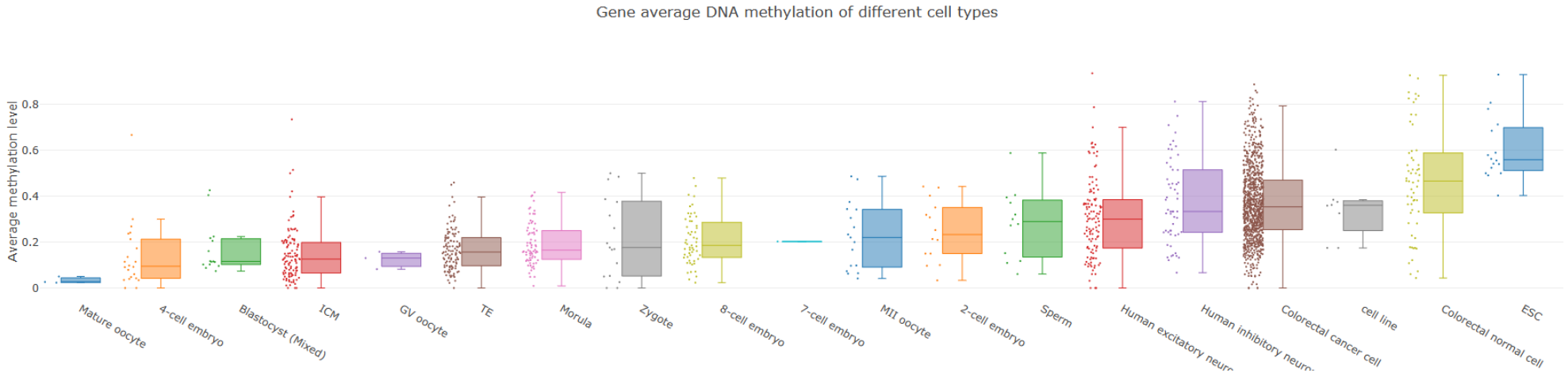
- 查看每个基因在不同样本中的甲基化水平





| Gene ID | Gene symbol | Chr | Start | End | Strand | Description | Sample Pattern |
|---|---|---|---|---|---|---|---|
| ENSG00000000003 | TSPAN6 | X | 100627108 | 100639991 | - | tetraspanin 6 | |
| ENSG00000000005 | TNMD | X | 100584936 | 100599885 | + | tenomodulin | |
| ENSG00000000419 | DPM1 | 20 | 50934867 | 50958555 | - | dolichyl-phosphate mannosyltransferase subunit 1, catalytic | |
| ENSG00000000457 | SCYL3 | 1 | 169849631 | 169894267 | - | SCY1 like pseudokinase 3 | |
| ENSG00000000460 | C1orf112 | 1 | 169662007 | 169854080 | + | chromosome 1 open reading frame 112 | |
| ENSG00000000938 | FGR | 1 | 27612064 | 27635185 | - | FGR proto-oncogene, Src family tyrosine kinase | |
| ENSG00000000971 | CFH | 1 | 196652043 | 196747504 | + | complement factor H | |
| ENSG00000001036 | FUCA2 | 6 | 143494812 | 143511720 | - | alpha-L-fucosidase 2 | |
| ENSG00000001084 | GCLC | 6 | 53497341 | 53616970 | - | glutamate-cysteine ligase catalytic subunit | |
| ENSG00000001167 | NFYA | 6 | 41072945 | 41099976 | + | nuclear transcription factor Y subunit alpha | |
| ENSG00000001460 | STPG1 | 1 | 24356999 | 24416934 | - | sperm tail PG-rich repeat containing 1 | |
| ENSG00000001461 | NIPAL3 | 1 | 24415802 | 24472976 | + | NIPA like domain containing 3 | |
| ENSG00000001497 | LAS1L | X | 65512582 | 65534775 | - | LAS1 like, ribosome biogenesis factor | |
| ENSG00000001561 | ENPP4 | 6 | 46129989 | 46146688 | + | ectonucleotide pyrophosphatase/phosphodiesterase 4 | |
| ENSG00000001617 | SEMA3F | 3 | 50155045 | 50189075 | + | semaphorin 3F | |

Gene Body



# scMethBank

- 查看DMR区域相关情况

## DMRs

A visual overview of the DMRs (Differentially Methylated Regions) between different two cell groups.

Select one dataset and cell groups, then click the submit button to generate a DMR overview of your selection.

**Selection of Datasets:**

GSE56879-- [single-cell genome-wide bisulfite sequencing f...

you can select one dataset in database

**Selection of Cell Groups:**

ESC VS MII oocyte

you can select any two cell groups

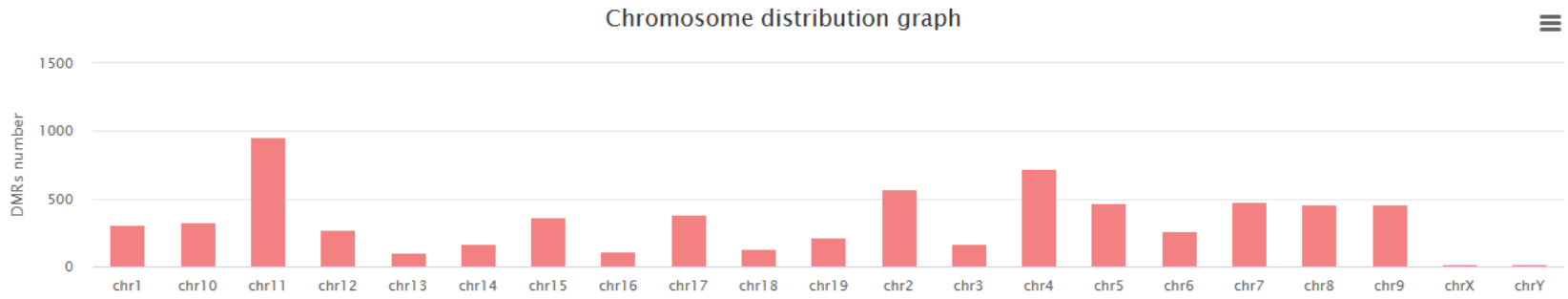
Submit | Reload

**Organism:** Mouse

**Project:** GSE56879

**Cell Groups:** ESC - MII oocyte

Chromosome Distribution | GO | KEGG | Genomic Location



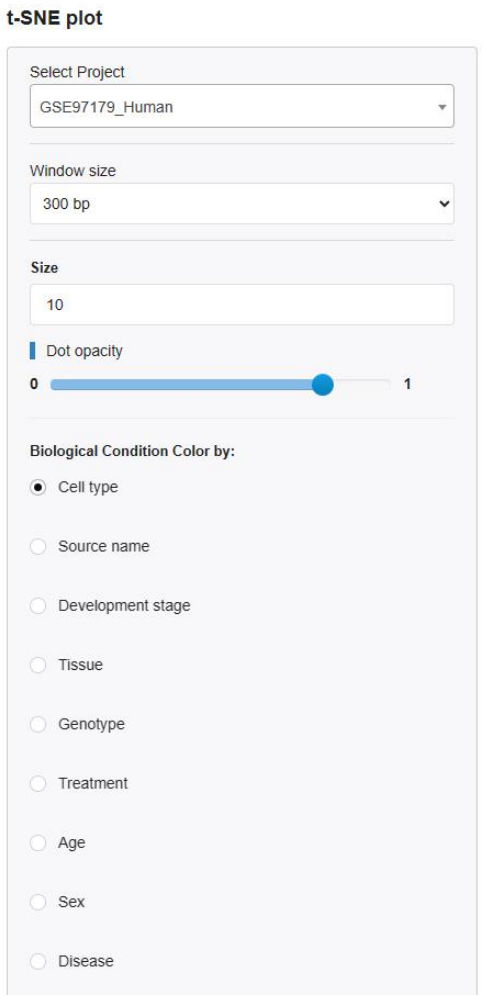

Download DMR Table

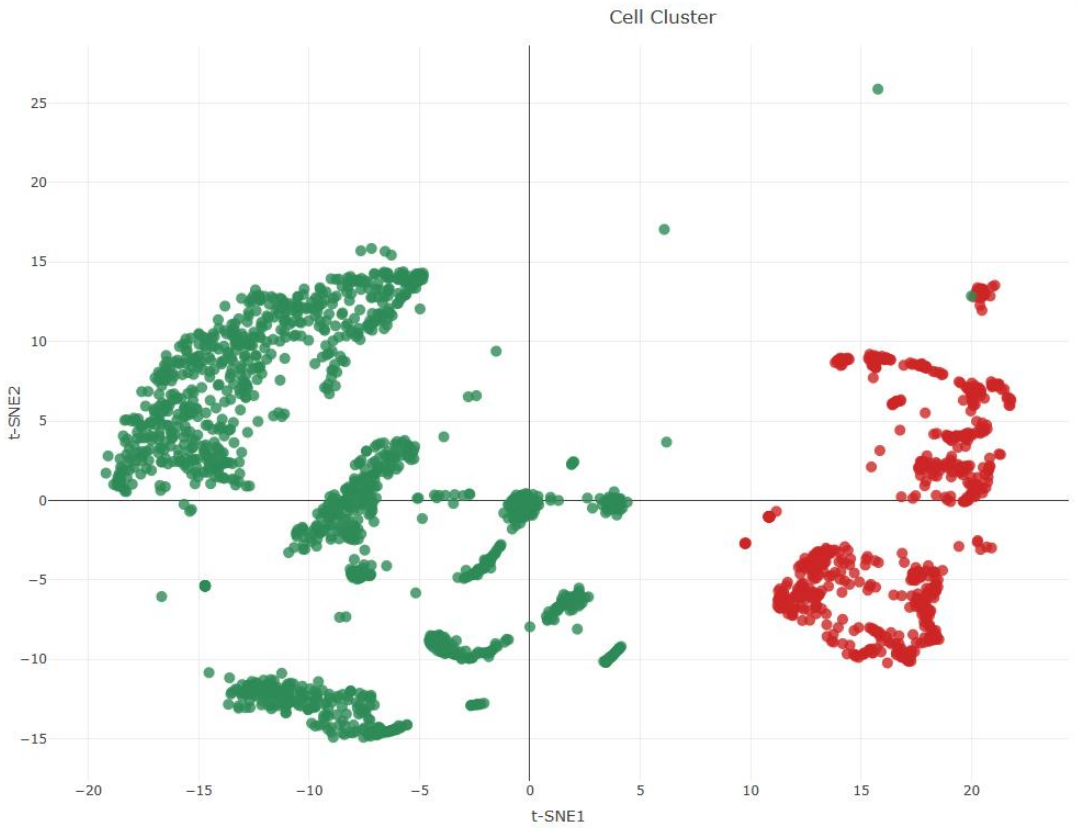
Search

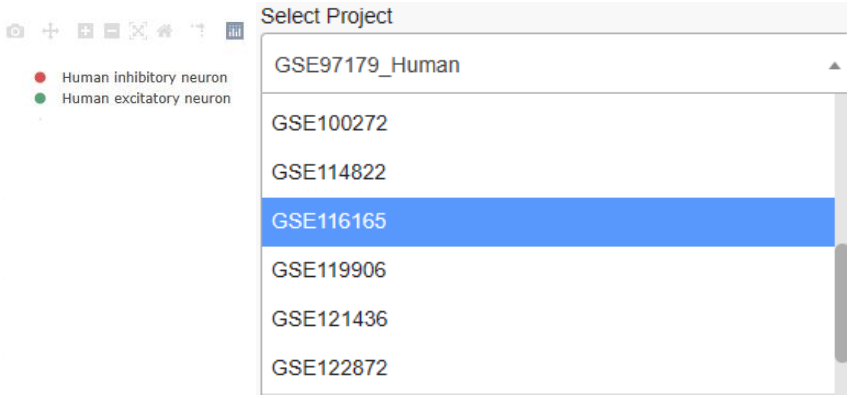
| Chr | Position | P Value | Differental Methylation Value | Annotation | Related Gene Ensemble ID | Related Gene Symbol | Methylatic |
|---|---|---|---|---|---|---|---|
| chr1 | 30087002-30087300 | 0.017 | 0.390 | Distal Intergenic | ENSMUSG00000099257 | Mir6342 | |

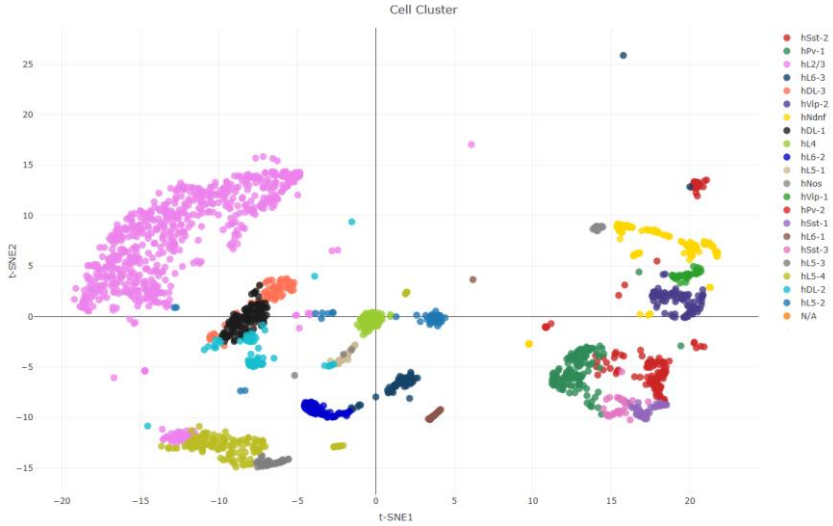
# scMethBank

- 基于甲基化水平的细胞聚类情况（不同的数据集分别聚类，没有整合；聚类图上无法查看每个细胞的甲基化水平）



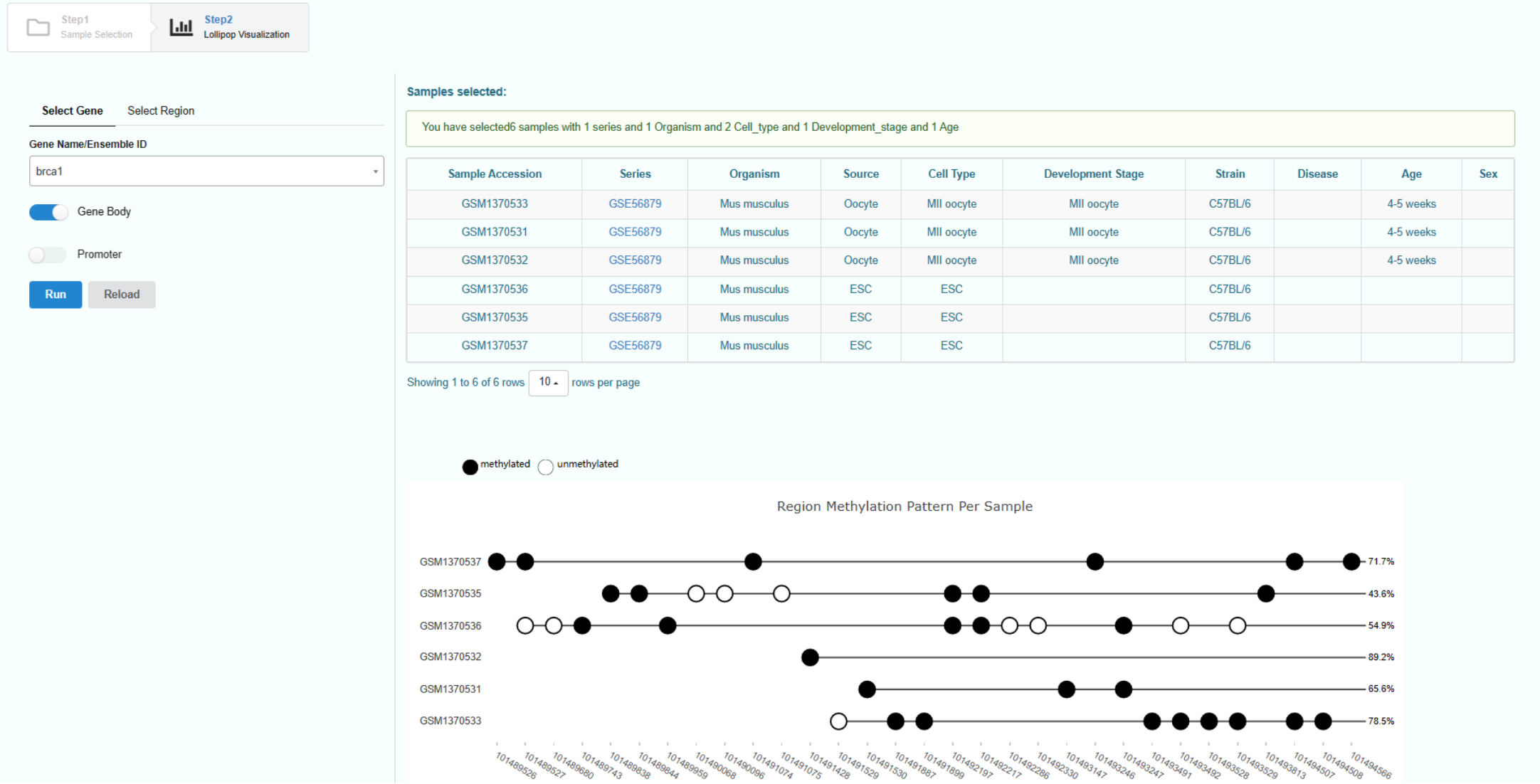



# scMethBank

- 选择不同的样本/细胞，查看特定基因在这些样本中的甲基化水平



# 小结

| | Cell1 | Cell2 | ... | CellN |
|---|---|---|---|---|
| *Gene1* | 3 | 2 | . | 13 |
| *Gene2* | 2 | 3 | . | 1 |
| *Gene3* | 1 | 14 | . | 18 |
| ... | . | . | . | . |
| ... | . | . | . | . |
| ... | . | . | . | . |
| *GeneM* | 25 | 0 | . | 0 |

单细胞转录组学

单细胞水平解析科学问题

空间转录组学

其他组学

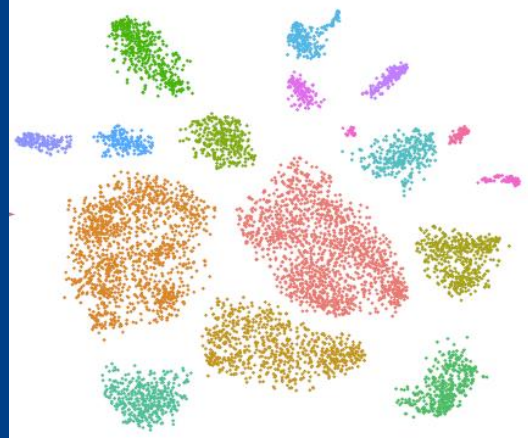
单细胞染色质可及性

单细胞DNA甲基化组




CellChat

浙江大学
ZHEJIANG UNIVERSITY



# 单细胞其他组学数据分析

——空间转录组、表观组

褚琴洁　qinjiechu@zju.edu.cn

2023年11月6日