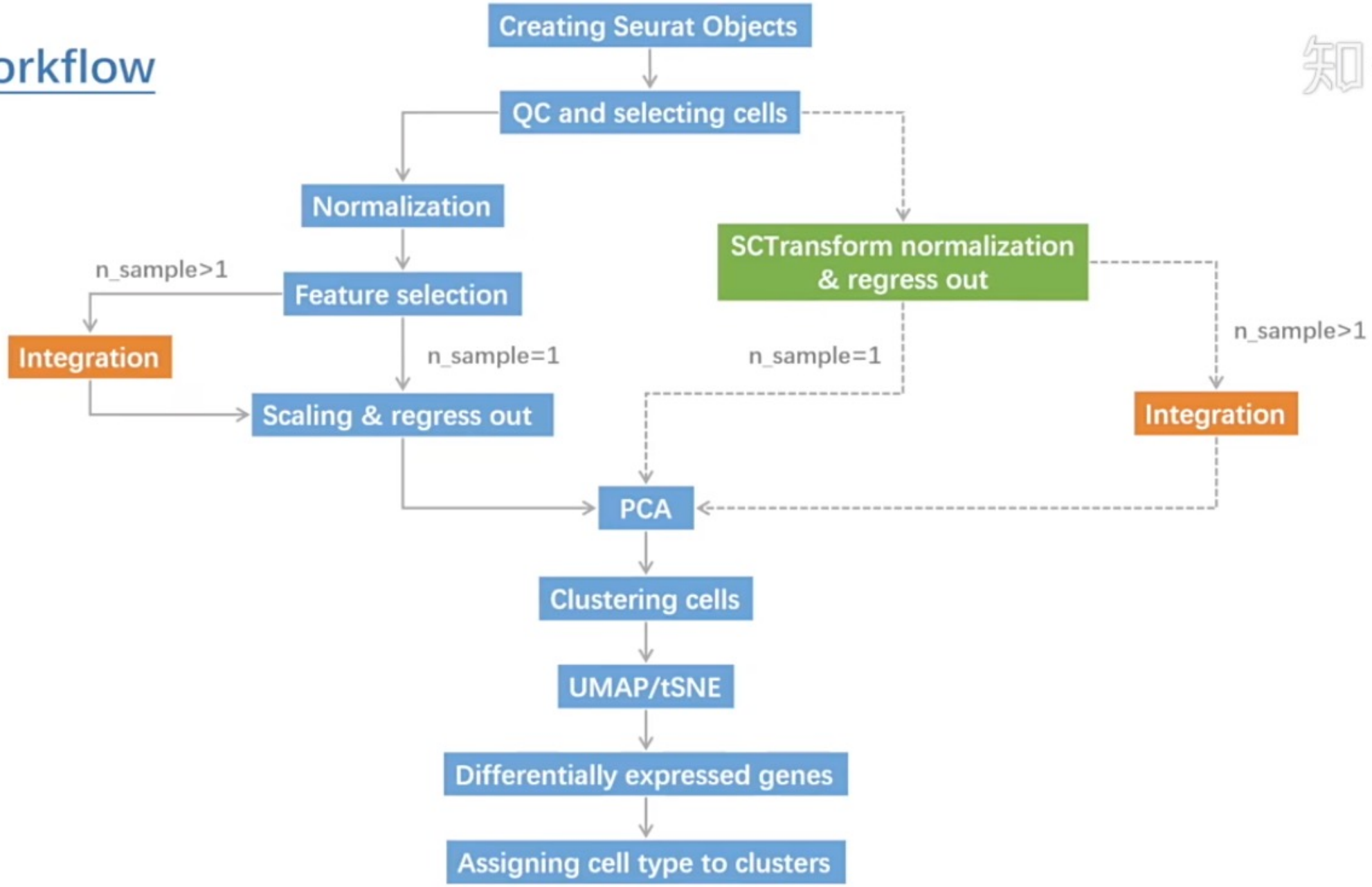
10X单细胞测序数据分析流程培训-2

生信培训

2022年4月

Official release of Seurat 4.0

Workflow


Creating Seurat Objects
QC and selecting cells
Normalization
SCTransform normalization & regress out
n_sample>1
Feature selection
Integration
n_sample=1
n_sample=1
n_sample>1
Scaling & regress out
Integration
PCA
Clustering cells
UMAP/tSNE
Differentially expressed genes
Assigning cell type to clusters

Seurat 安装

- https://satijalab.org/seurat/articles/install.html

- 最新版本安装

```
# Enter commands in R (or R studio, if installed)
install.packages('Seurat')
library(Seurat)
```

- Version 3

```
remotes::install_version("Seurat", version = "3.X.X")
```

一、创建 Seurat 对象

- 示例数据集：https://satijalab.org/seurat/articles/pbmc3k_tutorial.html 。

```
library(dplyr)
library(Seurat)
library(patchwork)

# Load the PBMC dataset
pbmc.data <- Read10X(data.dir = "../data/pbmc3k/filtered_gene_bc_matrices/hg19/")
# Initialize the Seurat object with the raw (non-normalized data).
pbmc <- CreateSeuratObject(counts = pbmc.data, project = "pbmc3k", min.cells = 3, min.features = 200)
pbmc
```

```
## An object of class Seurat
## 13714 features across 2700 samples within 1 assay
## Active assay: RNA (13714 features, 0 variable features)
```

- counts, #未标准化的数据，如原始计数或TPMs
- project, #设置Seurat对象的项目名称
- min.cells #包含至少在这些细胞检测到的features。
- min.features #包含至少检测到这些features的细胞

二、标准预处理流程

1. QC和细胞筛选

- 常用的质控指标：
  - *每个细胞在检测到的特异基因数*
    - ✓ 低质量细胞或空液滴通常只能检测到非常少的基因
    - ✓ 两个或多个细胞被同时捕获通常会有很高的基因数
  - *每个细胞检测到的分子总数（与基因密切相关）*
  - *每个细胞的线粒体基因比例*
    - ✓ 低质量/濒死细胞常表现出广泛的线粒体污染
    - ✓ 使用PercentageFeatureSet()函数计算线粒体QC指标
    - ✓ 使用所有以MT-开头的基因作为一组线粒体基因

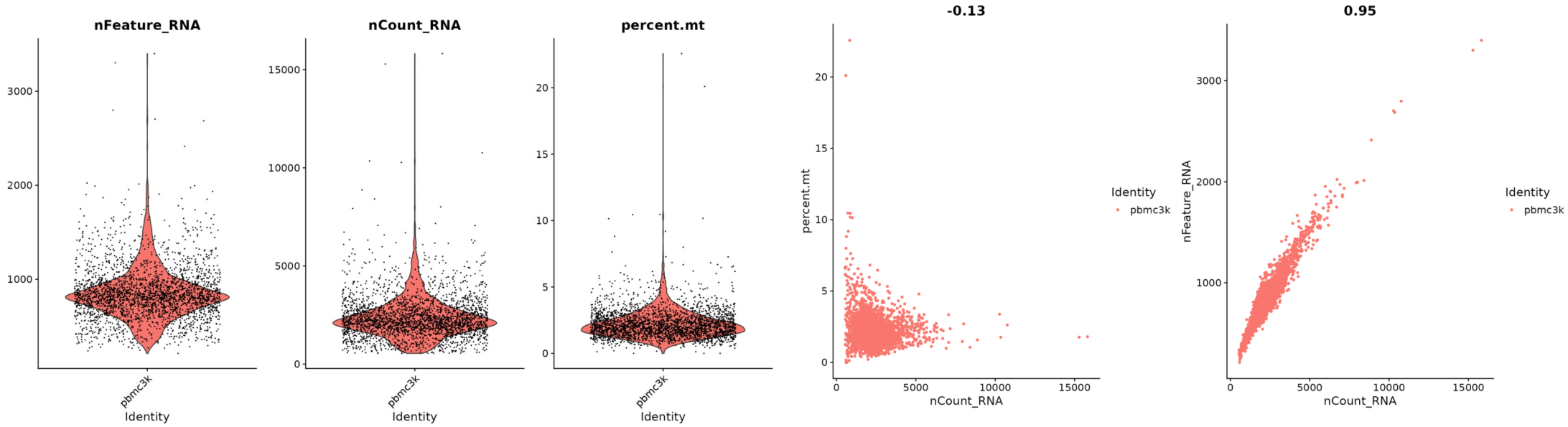
```
pbmc[["percent.mt"]] <- PercentageFeatureSet(pbmc, pattern = "^MT-")
```

```r
# Show QC metrics for the first 5 cells
head(pbmc@meta.data, 5)
```

```
##                  orig.ident nCount_RNA nFeature_RNA percent.mt
## AAACATACAACCAC-1      pbmc3k       2419          779  3.0177759
## AAACATTGAGCTAC-1      pbmc3k       4903         1352  3.7935958
## AAACATTGATCAGC-1      pbmc3k       3147         1129  0.8897363
## AAACCGTGCTTCCG-1      pbmc3k       2639          960  1.7430845
## AAACCGTGTATGCG-1      pbmc3k        980          521  1.2244898
```

- nFeature_RNA代表每个细胞测到的基因数目。
- nCount_RNA代表每个细胞测到所有基因的表达量之和。
- percent.mt代表测到的线粒体基因的比例。

```r
VlnPlot(pbmc, features = c("nFeature_RNA", "nCount_RNA", "percent.mt"), ncol = 3)
plot1 <- FeatureScatter(pbmc, feature1 = "nCount_RNA", feature2 = "percent.mt")
plot2 <- FeatureScatter(pbmc, feature1 = "nCount_RNA", feature2 = "nFeature_RNA")
plot1 + plot2
```

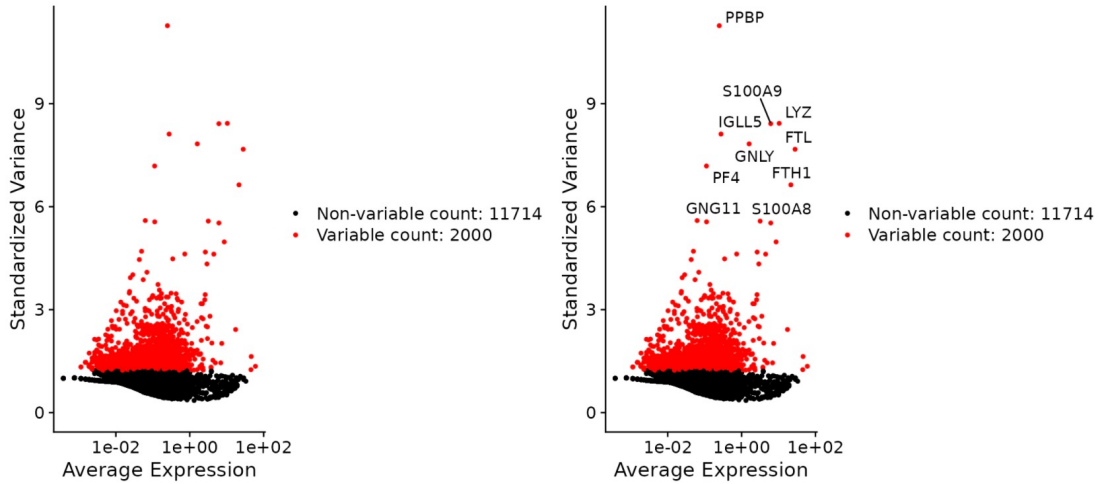


2.归一化数据

Normlization：LogNormalize, CLR, RC

```
pbmc <- NormalizeData(pbmc, normalization.method = "LogNormalize", scale.factor = 10000)
pbmc <- NormalizeData(pbmc)
```

3.归一化数据




```
pbmc <- FindVariableFeatures(pbmc, selection.method = "vst", nfeatures = 2000)

# 识别前十的高变基因
top10 <- head(VariableFeatures(pbmc), 10)

plot1 <- VariableFeaturePlot(pbmc)
plot2 <- LabelPoints(plot = plot1, points = top10, repel = TRUE)
plot1 + plot2
```

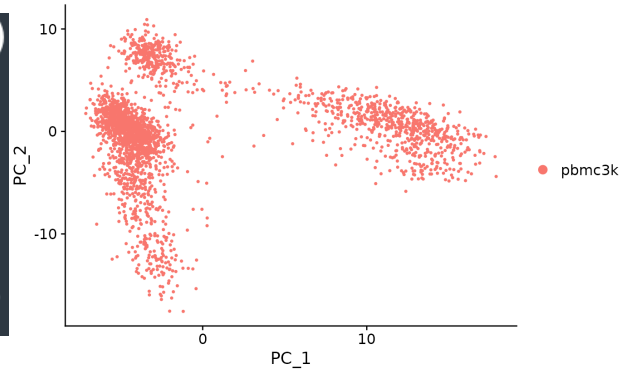
4.缩放数据

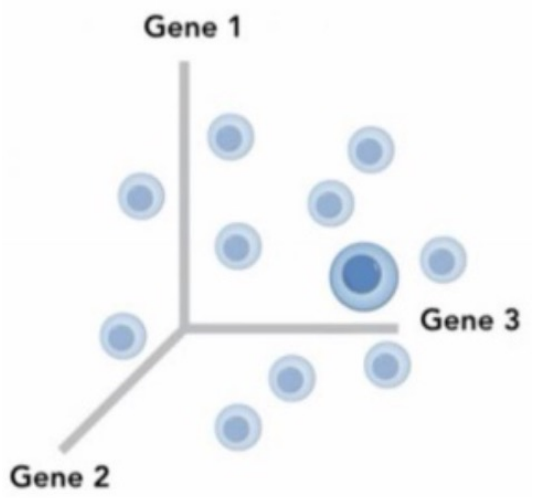
- 转换每个基因的表达值，使每个细胞的平均表达值为0
- 转换每个基因的表达值，使细胞间方差为1

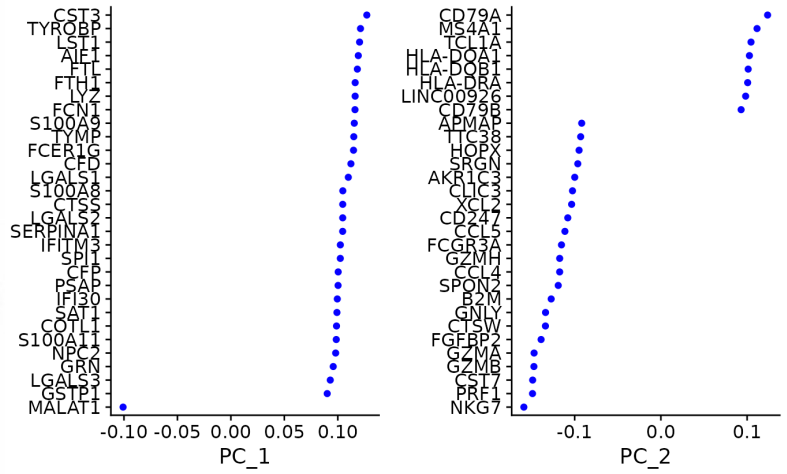
```
all.genes <- rownames(pbmc)
pbmc <- ScaleData(pbmc, features = all.genes)
```

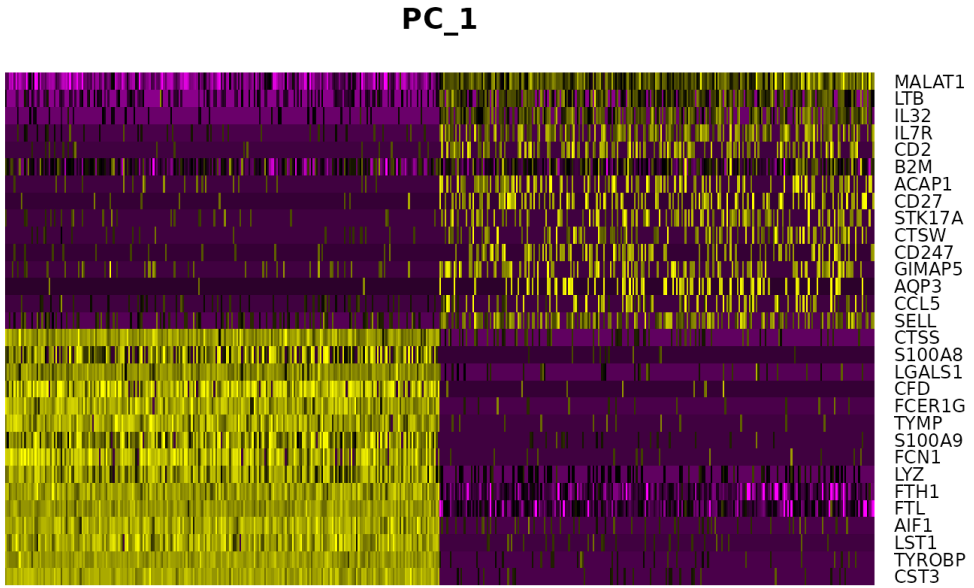
5.线性维度约化 PCA

```
pbmc <- RunPCA(pbmc, features = VariableFeatures(object = pbmc)
print(pbmc[["pca"]], dims = 1:5, nfeatures = 5)
VizDimLoadings(pbmc, dims = 1:2, reduction = "pca")    #图1
DimPlot(pbmc, reduction = "pca") #图2
DimHeatmap(pbmc, dims = 1, cells = 500, balanced = TRUE)#图3
DimHeatmap(pbmc, dims = 1:15, cells = 500, balanced = TRUE)#图4
```



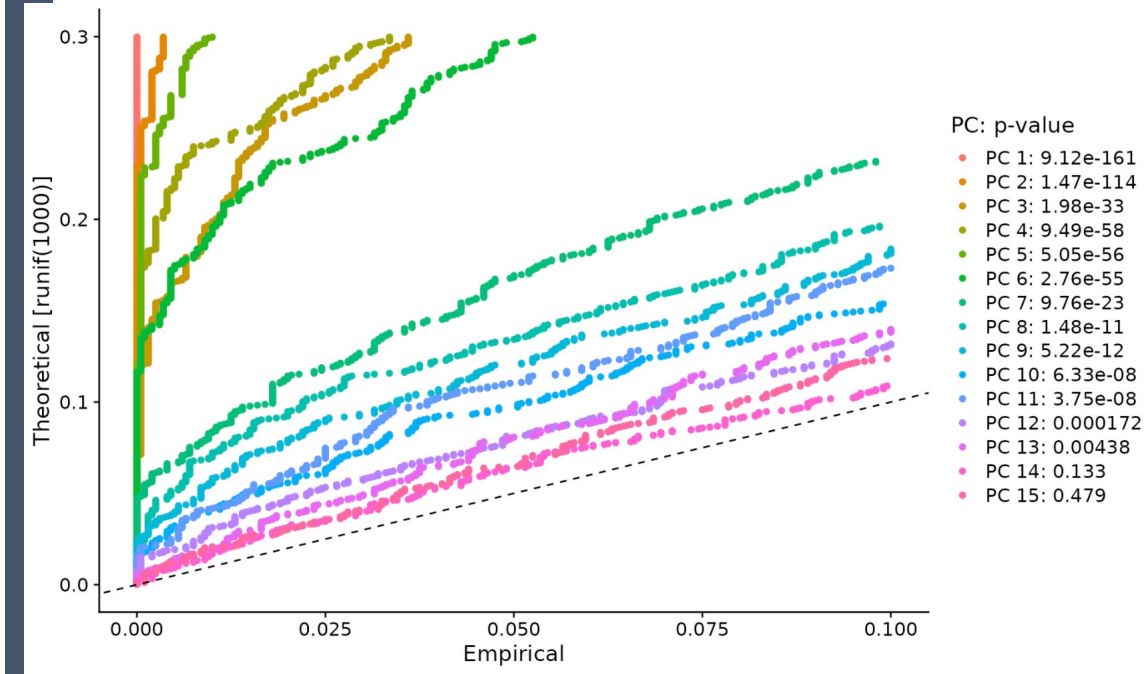


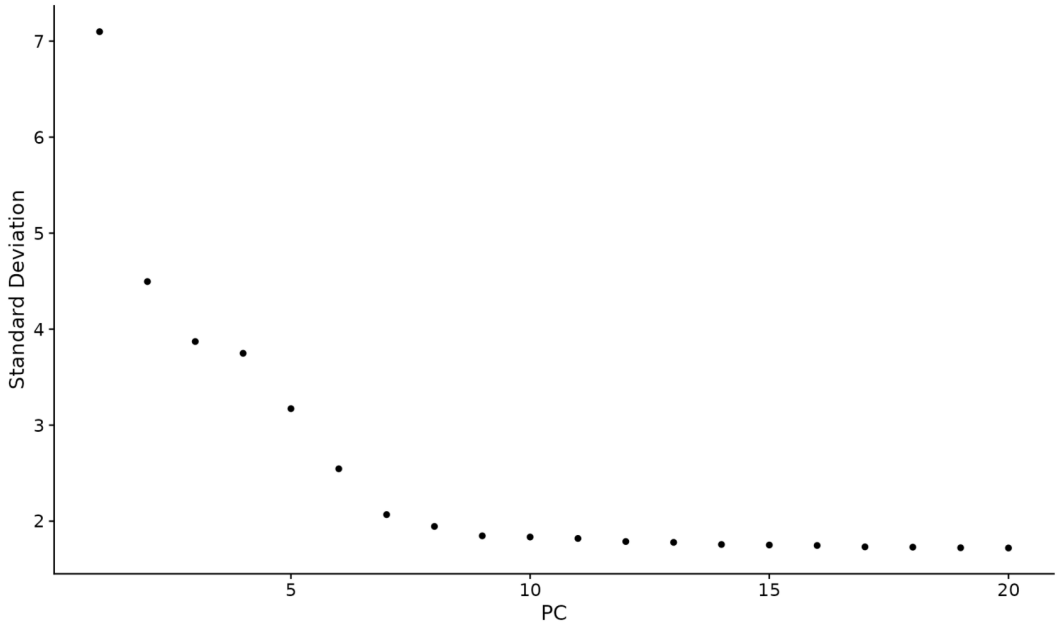




5.确定数据集的维度

```
pbmc <- JackStraw(pbmc, num.replicate = 100)
pbmc <- ScoreJackStraw(pbmc, dims = 1:20)
JackStrawPlot(pbmc, dims = 1:15)
ElbowPlot(pbmc)
```



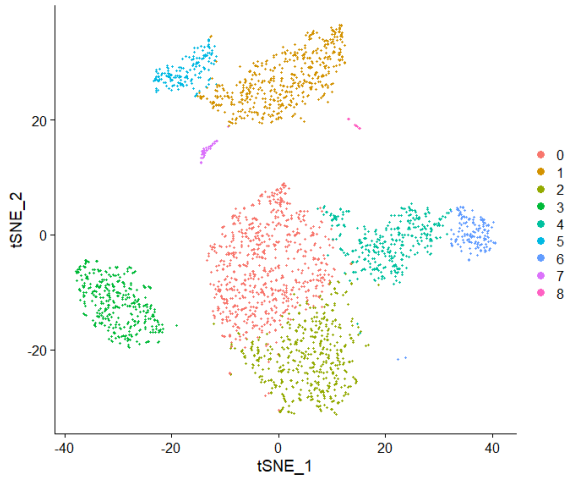


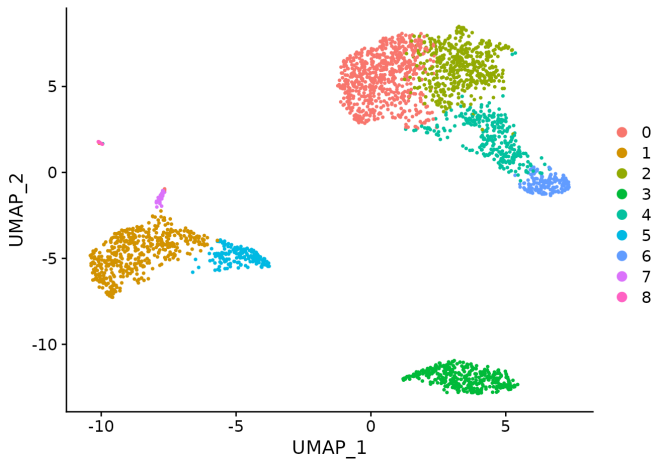
6.细胞聚类及非线性降维可视化（UMAP/TSNE）

```
pbmc <- FindNeighbors(pbmc, dims = 1:10)
pbmc <- FindClusters(pbmc, resolution = 0.5)
```

```
pbmc <- RunUMAP(pbmc, dims = 1:10)
DimPlot(pbmc, reduction = "umap")

# 使用TSNE聚类
pbmc <- RunTSNE(pbmc, dims = 1:10)
DimPlot(pbmc, reduction = "tsne")
# 显示在聚类标签
DimPlot(pbmc, reduction = "tsne", label = TRUE)
```

```
#保存rds，用于后续分析
saveRDS(pbmc, file = "../output/pbmc_tutorial.rds")
```





7.发现差异表达特征（cluster bioers）

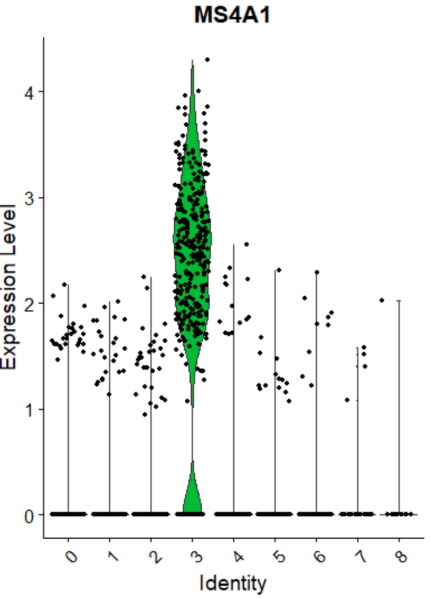
```r
# 发现聚类一的所有biomarkers
cluster1.markers <- FindMarkers(pbmc, ident.1 = 1, min.pct = 0.25)
head(cluster1.markers, n = 5)

# 查找将聚类5与聚类0和3区分的所有标记
cluster5.markers <- FindMarkers(pbmc, ident.1 = 5, ident.2 = c(0, 3), min.pct = 0.25)
head(cluster5.markers, n = 5)

# 与所有其他细胞相比，找到每个簇的标记，仅报告阳性细胞
pbmc.markers <- FindAllMarkers(pbmc, only.pos = TRUE, min.pct = 0.25, logfc.threshold = 0.25)
pbmc.markers %>% group_by(cluster) %>% top_n(n = 2, wt = avg_logFC)
cluster1.markers <- FindMarkers(pbmc, ident.1 = 0, logfc.threshold = 0.25, test.use = "roc", only.pos = TRUE)
```

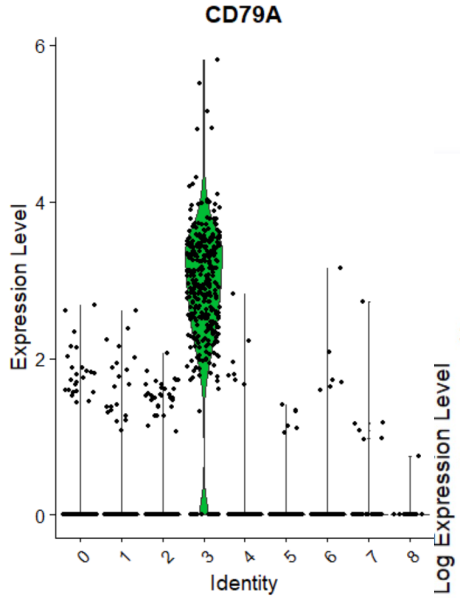
名称	解释
gene	基因名称
cluster	该基因对应的cluster
pct.1	在当前cluster细胞中检测到该基因表达的细胞比例
pct.2	在其它cluster细胞中检测到该基因表达的细胞比例
avg_logFC	两组间平均logFC，Seurat v4默认log2。正值表示特征在第一组中表达得更高
p_val	未调整P-value，数值越小越显著
p_val_adj	基于使用数据集中所有特征的bonferroni校正，校正后的p值

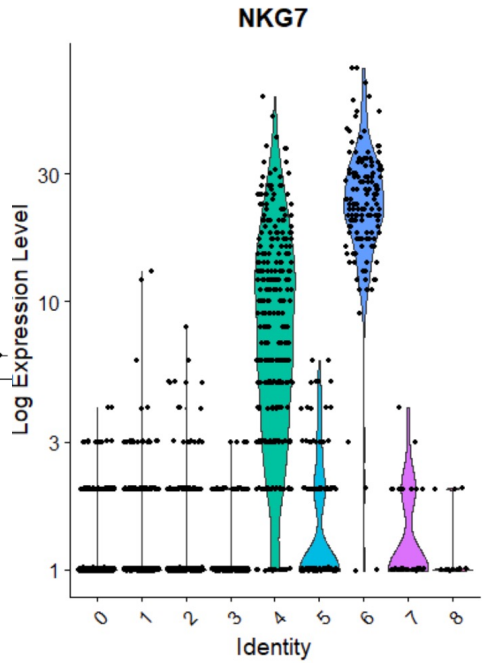
- **FindAllMarkers**：比较一个cluster与所有其他cluster之间的基因表达
- **FindMarkers**：比较两个特定cluster之间的基因表达

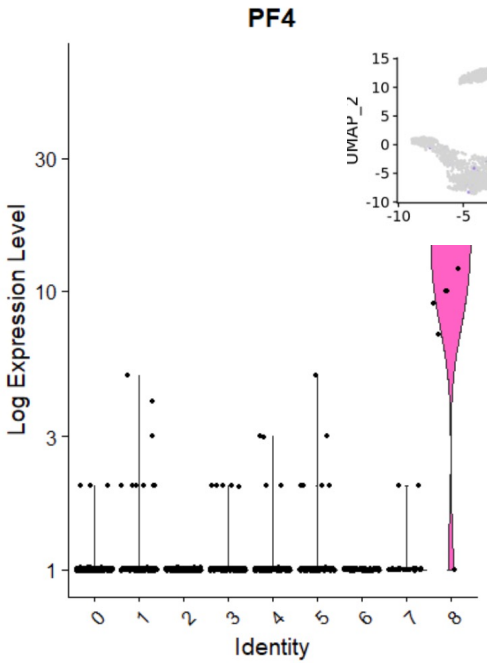
可视化，探索感兴趣的基因

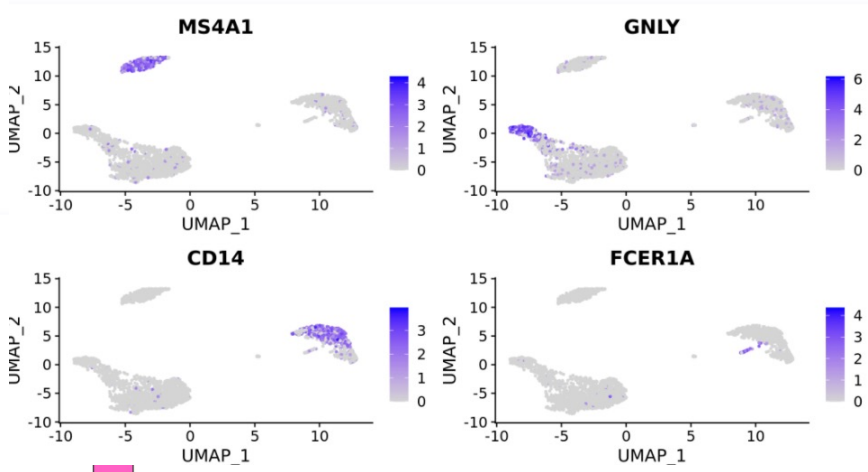
```r
# 绘图看看 -- 1
VlnPlot(pbmc, features = c("MS4A1", "CD79A"))
# 使用原始count绘制  -- 2
VlnPlot(pbmc, features = c("NKG7", "PF4"), slot = "counts", log = TRUE)
#  -- 3
FeaturePlot(pbmc, features = c("MS4A1", "GNLY", "CD3E", "CD14", "FCER1A", "FCGR3A", "LYZ", "PPBP", "CD8A"))
```



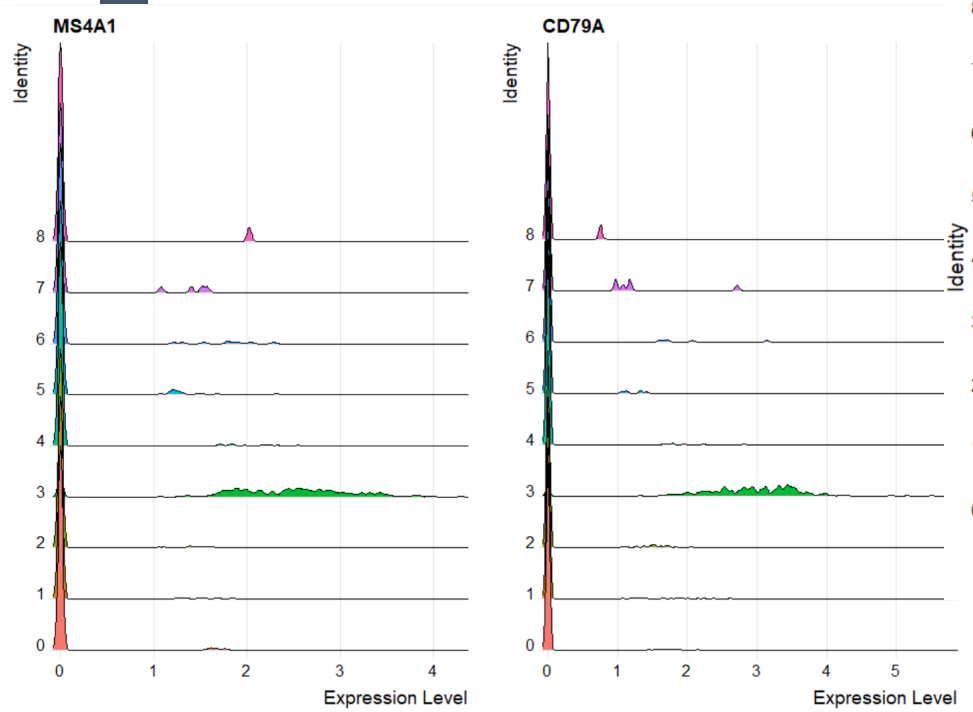


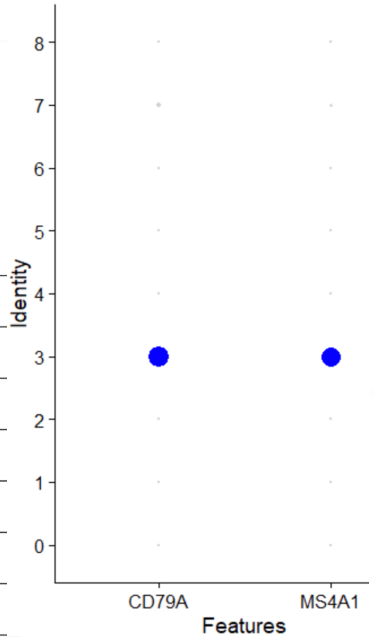


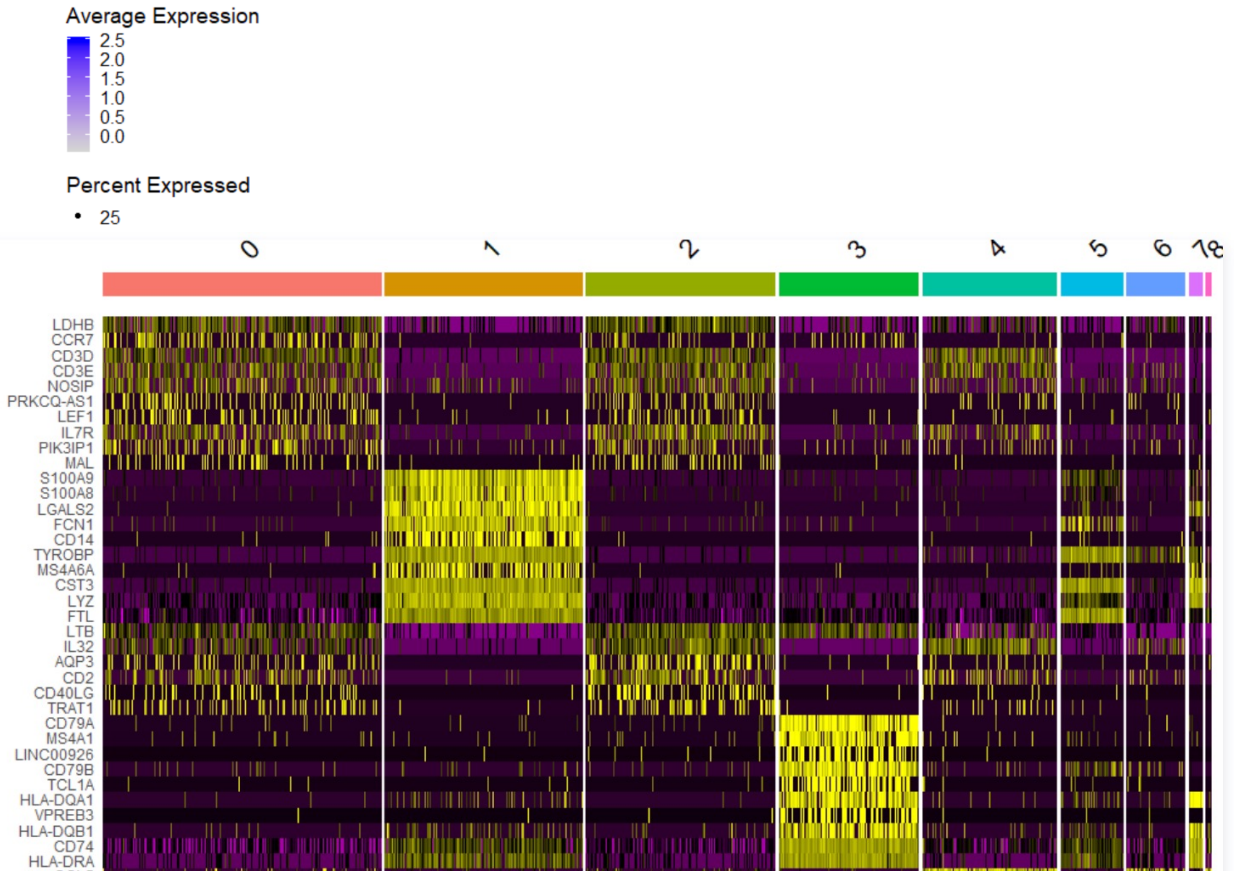




```r
#  -- 4
RidgePlot(pbmc, features = c("MS4A1", "CD79A"))
#  -- 5
DotPlot(pbmc, features = c("MS4A1", "CD79A"))
#  -- 6
top10 <- pbmc.ers %>% group_by(cluster) %>% top_n(n = 10, wt = avg_logFC)
DoHeatmap(pbmc, features = top10$gene) + NoLegend()
```





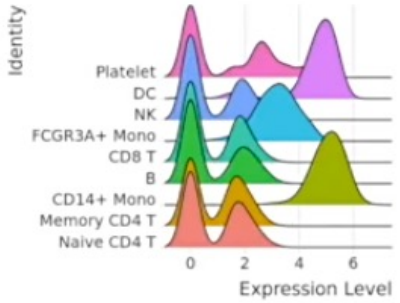


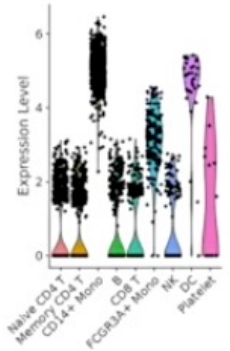
可视化功能

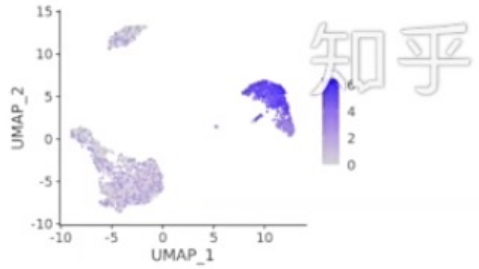
Visualization

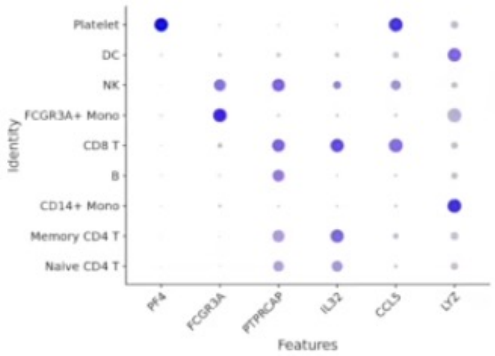
- RidgePlot
- VlnPlot
- FeaturePlot
- DotPlot
- DoHeatmap
- DimPlot
- FeatureScatter

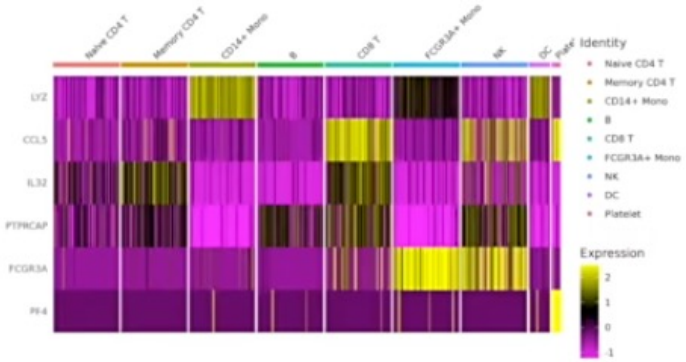
- ggplot2 theme
- patchwork
- Interactive plots

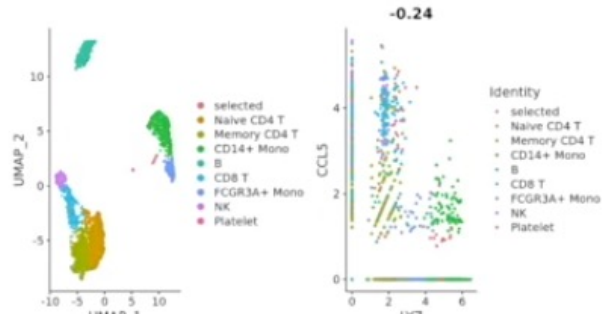


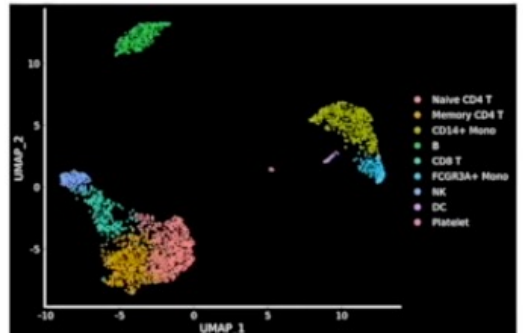










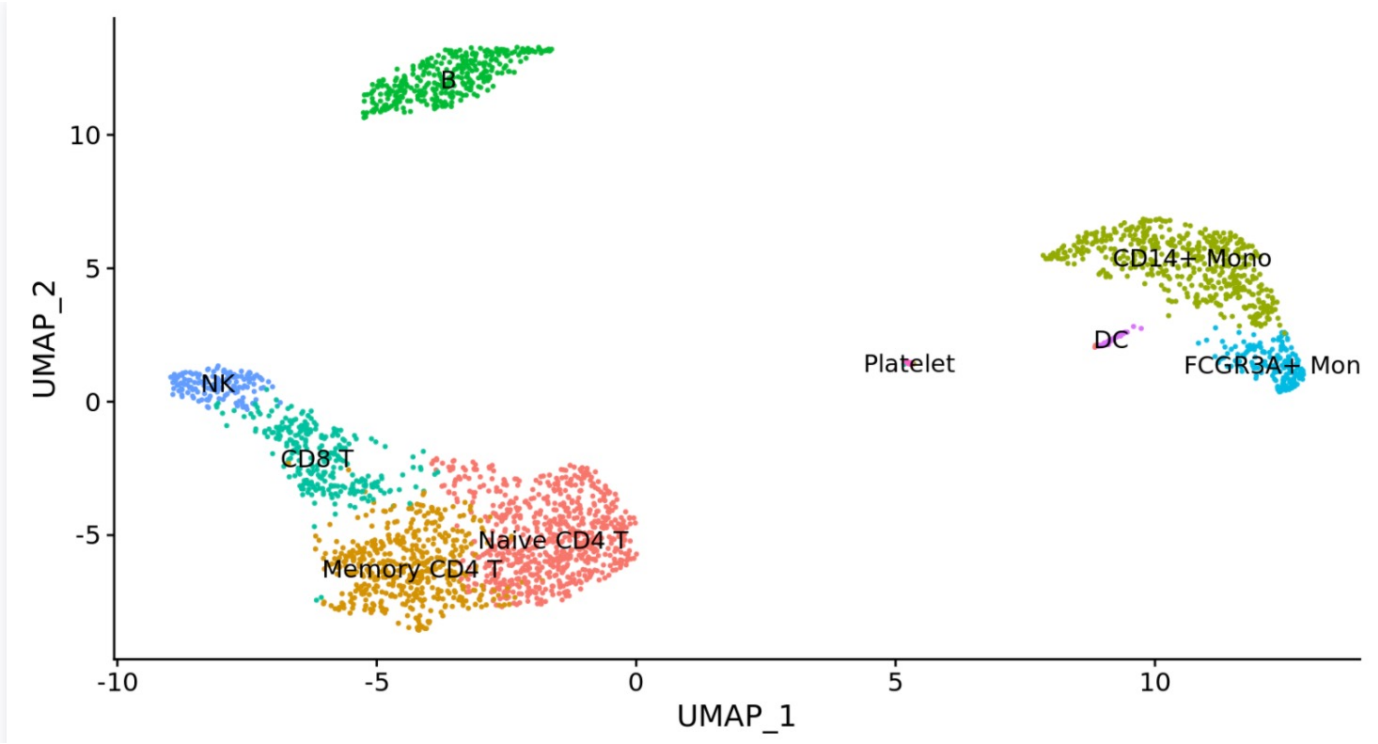


R：基本流程

8.识别细胞类型

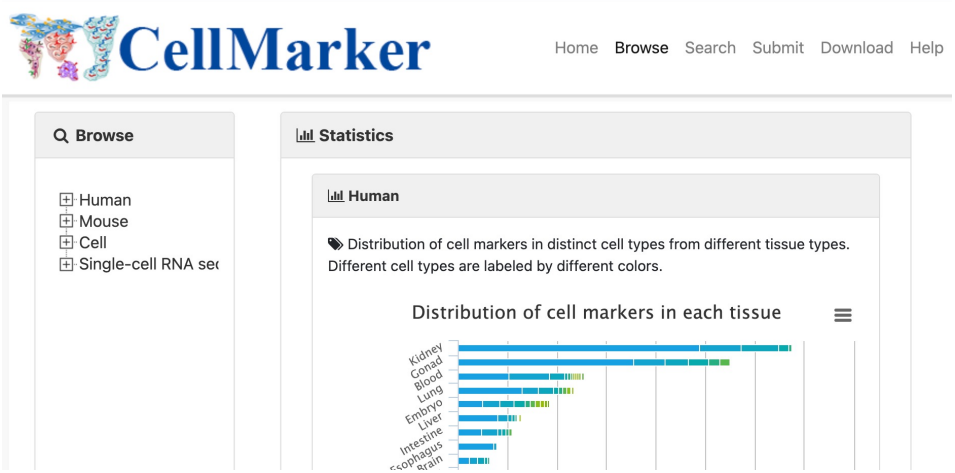
```
new.cluster.ids <- c("Naive CD4 T", "Memory CD4 T", "CD14+ Mono", "B", "CD8 T", "FCGR3A+ Mono",
    "NK", "DC", "Platelet")
names(new.cluster.ids) <- levels(pbmc)
pbmc <- RenameIdents(pbmc, new.cluster.ids)
DimPlot(pbmc, reduction = "umap", label = TRUE, pt.size = 0.5) + NoLegend()
```

Cluster ID	Markers	Cell Type
0	IL7R, CCR7	Naive CD4+ T
1	IL7R, S100A4	Memory CD4+
2	CD14, LYZ	CD14+ Mono
3	MS4A1	B
4	CD8A	CD8+ T
5	FCGR3A, MS4A7	FCGR3A+ Mono
6	GNLY, NKG7	NK
7	FCER1A, CST3	DC
8	PPBP	Platelet

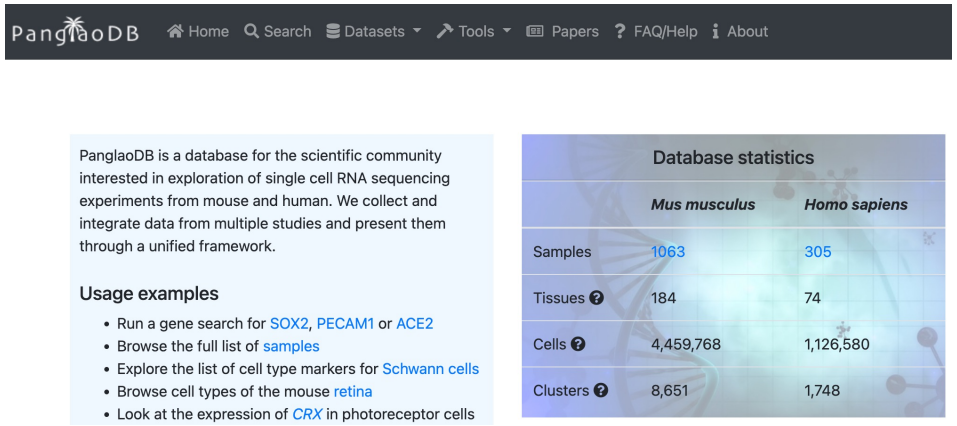


细胞类型鉴定

- 三种方法:
  - *利用marker基因查找网站进行注释*



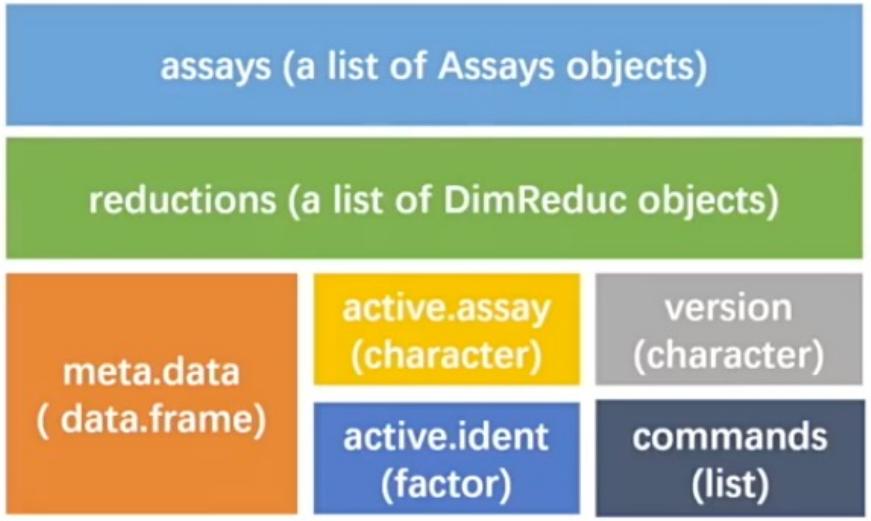
    - CellMarker：http://bio-bigdata.hrbmu.edu.cn/CellMarker/browse.jsp
    - PanglaoDB：A Single Cell Sequencing Resource For Gene Expression Data



  - *使用singler进行注释*
  - *根据已有的生物学知识或者文献，按照dotplot来注释。*

理解seurat对象

Seurat Object



The Seurat object is a class allowing for the storage and manipulation of multimodal single-cell data.

Assay Object

The Assay objects are designed to hold expression data of a single type, such as RNA-seq gene expression, CITE-seq ADTs, or cell hashtags.

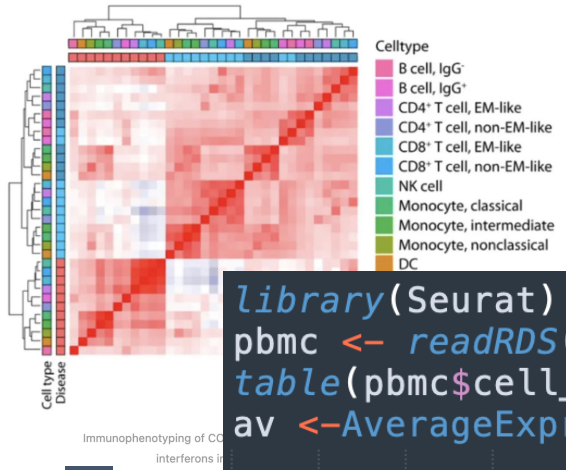
DimReduc Object

DimReduc objects represent transformations of the data contained within the Assay object(s) via various dimensional reduction techniques such as PCA.

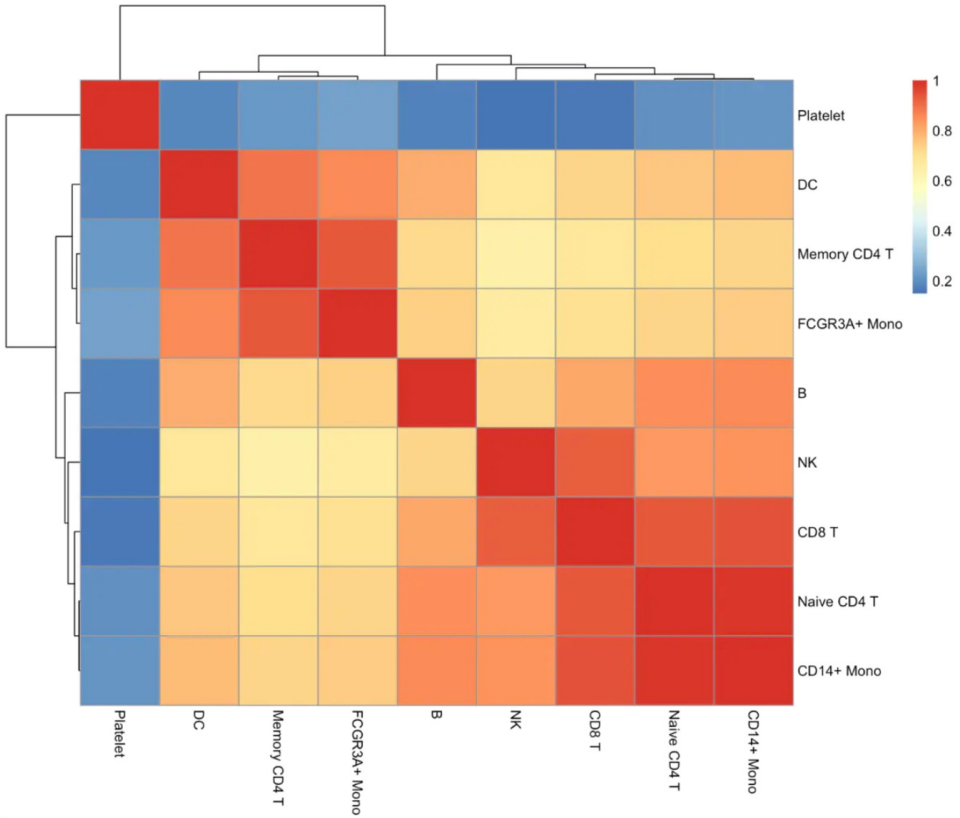
R：理解对象
常用函数

9.不同单细胞群之间的相关性分析

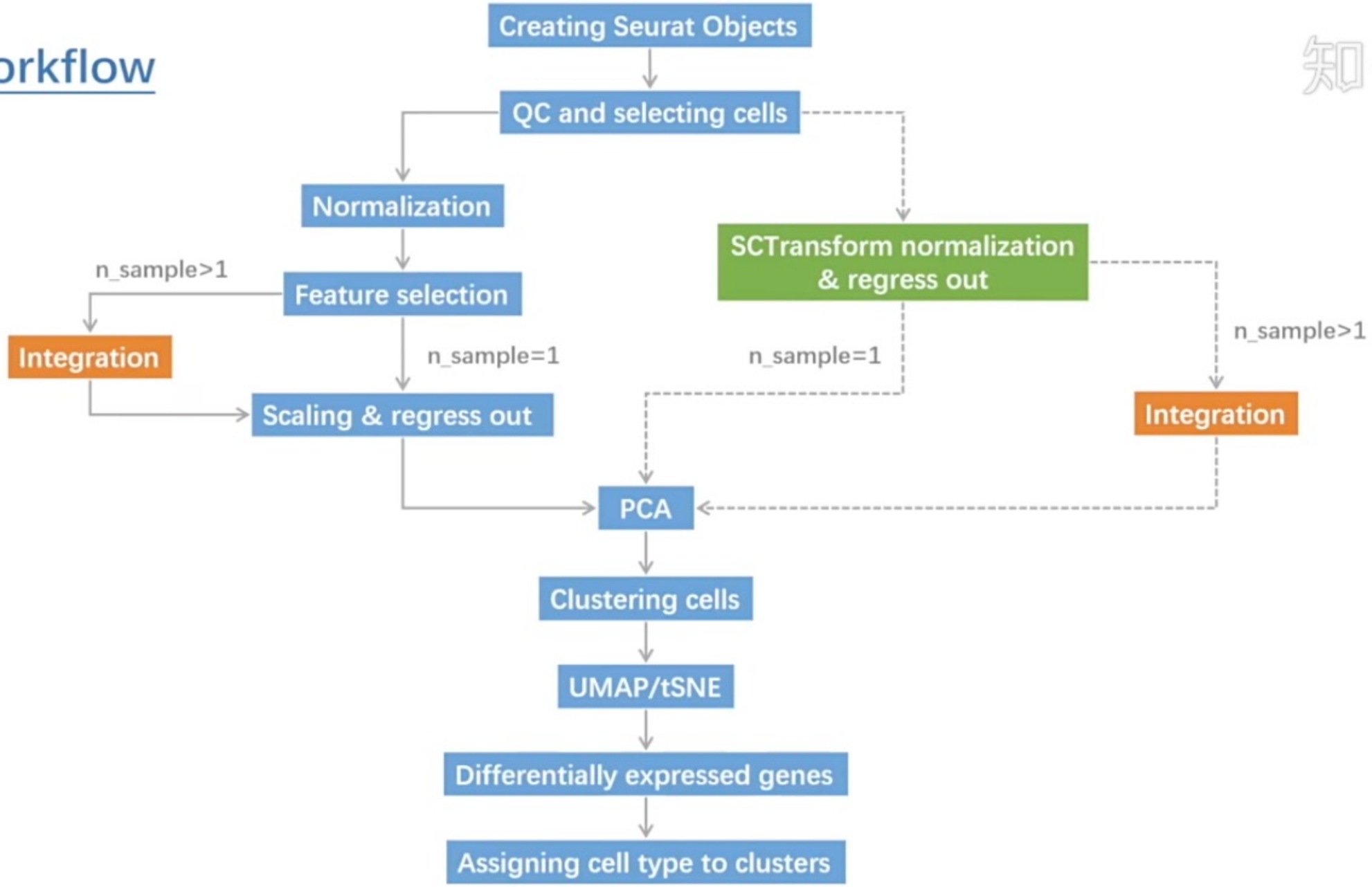
- 相关性分析是指对两个或多个具备相关性的变量元素进行分析，从而衡量两个变量因素的相关密切程度。（**Pearson，Spearman等**）



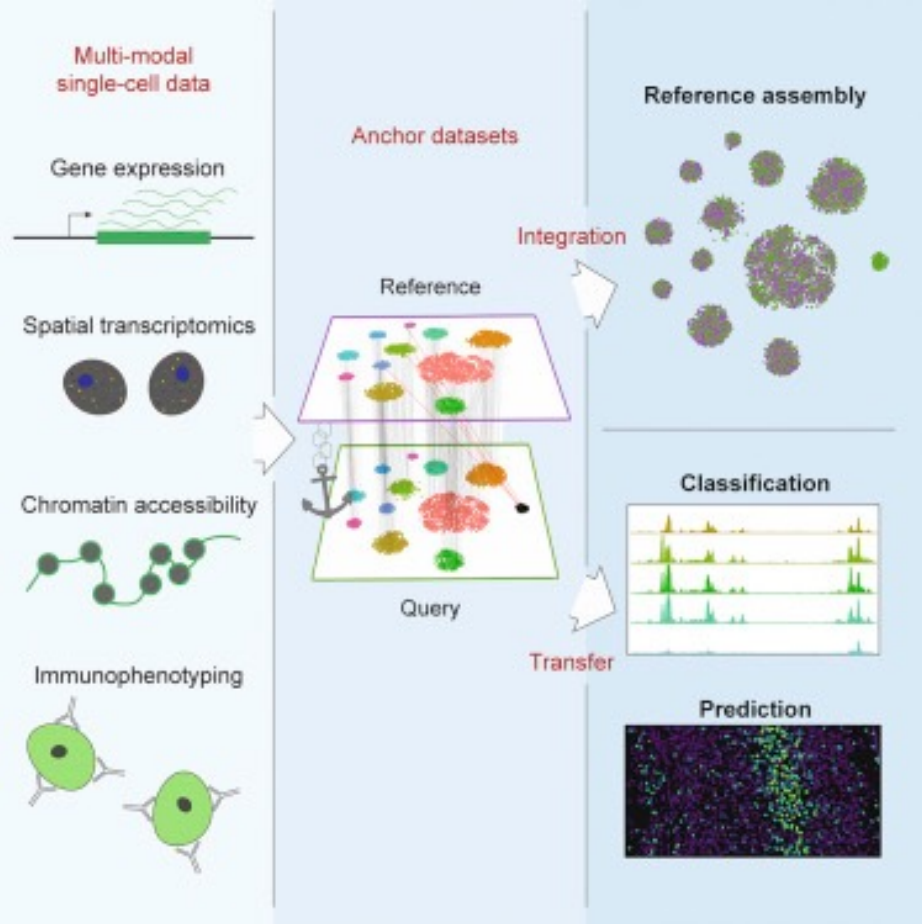

```r
library(Seurat)
pbmc <- readRDS("pbmc.rds")
table(pbmc$cell_type)
av <-AverageExpression(pbmc,
                       group.by = "cell_type",
                       assays = "RNA")
av=av[[1]]
head(av)

#选出标准差最大的1000个基因
cg=names(tail(sort(apply(av, 1, sd)),1000))
#查看这1000个基因在各细胞群中的表达矩阵
View(av[cg,])
#查看细胞群的相关性矩阵
View(cor(av[cg,],method = 'spearman'))
#pheatmap绘制热图
pheatmap::pheatmap(cor(av[cg,],method = 'spearman')) #默认是Pearson
```



Workflow


Creating Seurat Objects
QC and selecting cells
Normalization
SCTransform normalization & regress out
n_sample>1
Feature selection
Integration
n_sample=1
n_sample=1
n_sample>1
Scaling & regress out
Integration
PCA
Clustering cells
UMAP/tSNE
Differentially expressed genes
Assigning cell type to clusters

三、利用Seurat数据的整合功能分析多样本数据



Stuart T, et. al.,Cell. 2019

R：整合

Highlights:

- Seurat v3 identifies correspondences between cells in different experiments
- These “anchors” can be used to harmonize datasets into a single reference
- Reference labels and data can be projected onto query datasets
- Extends beyond RNA-seq to single-cell protein, chromatin, and spatial data

Four broad steps:

- data preprocessing and feature selection,
- dimension reduction and identification of “anchor” correspondences between datasets,
- filtering, scoring, and weighting of anchor correspondences,
- data matrix correction, or data transfer across experiments.

```r
library(Seurat)
library(SeuratData)
data("panc8")
pancreas.list <- SplitObject(panc8, split.by = "tech")
pancreas.list <- pancreas.list[c("celseq", "celseq2", "fluidigmc1", "smartseq2")]
#来看下数据结构，pancrease.list包含了4个seurat对象
#流程
#在寻找anchor之前先进行标准化，找出各个数据集自身的差异表达基因
for (i in 1:length(pancreas.list)) {
  pancreas.list[[i]] <- NormalizeData(pancreas.list[[i]], verbose = FALSE)
  pancreas.list[[i]] <- FindVariableFeatures(pancreas.list[[i]], selection.method = "vst",
                                             nfeatures = 2000, verbose = FALSE)
}
#寻找anchor，这里只用了其中的三个数据
reference.list <- pancreas.list[c("celseq", "celseq2", "smartseq2")]
pancreas.anchors <- FindIntegrationAnchors(object.list = reference.list, anchor.features = 2000, dims = 1:30)
#dims取了默认参数，也可以进行调整，官网推荐值在10-50之间
#利用这样的anchor去矫正批次和技术差异（相当于找到了共同的锚点，用这样的锚点去矫正表达水平）
pancreas.integrated <- IntegrateData(anchorset = pancreas.anchors, dims = 1:30)

#将seurat对象的默认assay设置成整合后的表达矩阵
DefaultAssay(pancreas.integrated) <- "integrated"
#归一化，注意，这里只有2000个基因
pancreas.integrated <- ScaleData(pancreas.integrated, verbose = FALSE)
#PCA
pancreas.integrated <- RunPCA(pancreas.integrated, npcs = 30, verbose = FALSE)
#使用umap根据pca降维的情况，取其前30个（总共也就30个）PC作为降维的高维向量
pancreas.integrated <- RunUMAP(pancreas.integrated, reduction = "pca", dims = 1:30)
#可视化，分别以技术和细胞类型作为标签
p1 <- DimPlot(pancreas.integrated, reduction = "umap", group.by = "tech")
p2 <- DimPlot(pancreas.integrated, reduction = "umap", group.by = "celltype", label = TRUE,
              repel = TRUE)   NoLegend()
plot_grid(p1, p2)
```