Hands-on-practice: Seurat package for single cell RNA sequencing data analysis

單細胞RNA定序資料分析軟體簡介與實作：R package: Seurat

2023.7.27

國家衛生研究院 癌症研究所

研究助理　蔡芳榆

Install

- R
- Seurat

Import data

- 10x

QC

- Low-quality cells or empty droplets
- Cell doublets or multiplets

Clustering

- Normalizing the data
- Identification of highly variable features/genes
- Principal components analysis (PCA)
- Cluster the cells
- UMAP/tSNE

Cell Typing

- Finding differentially expressed features (cluster biomarkers)

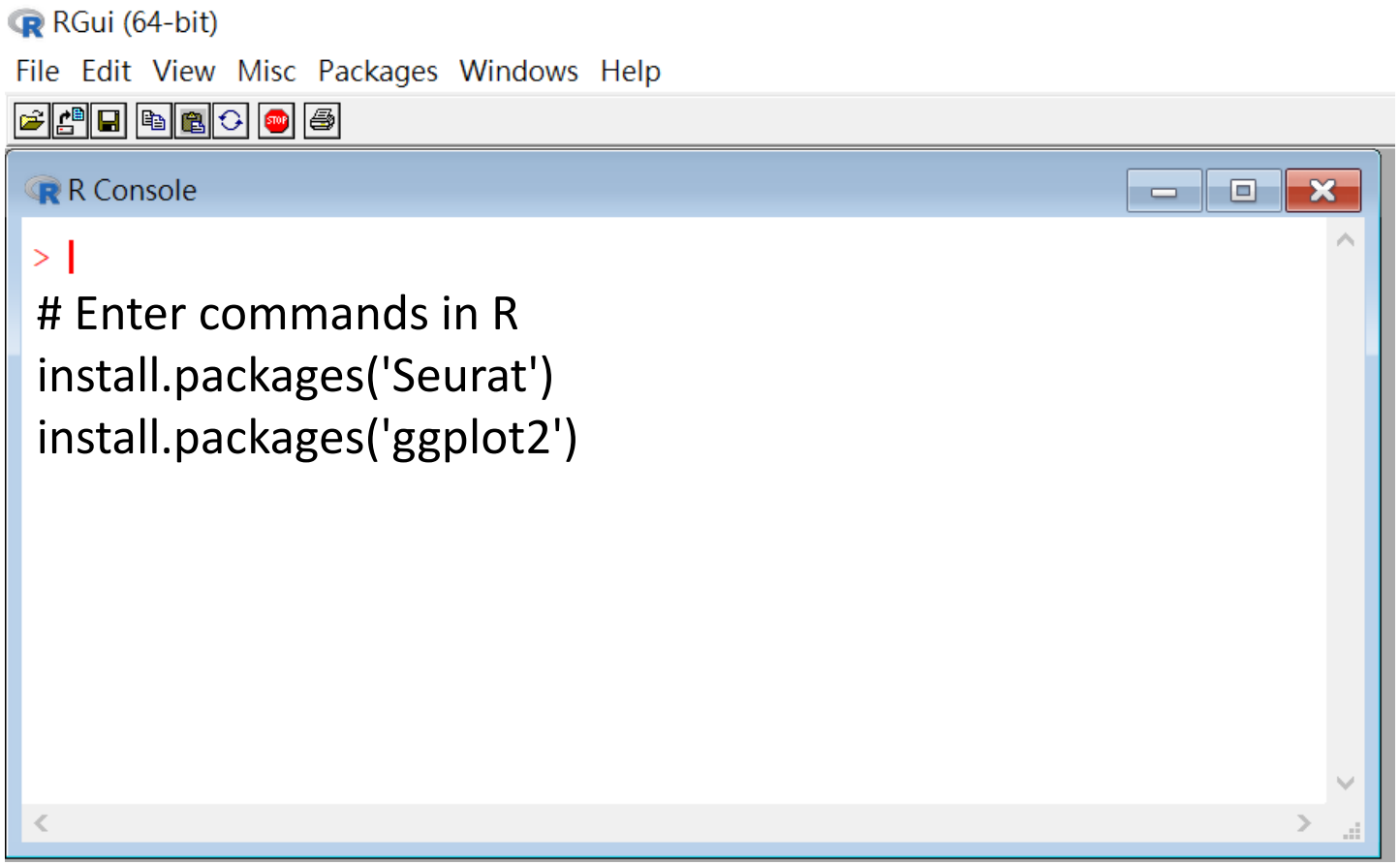
DEG

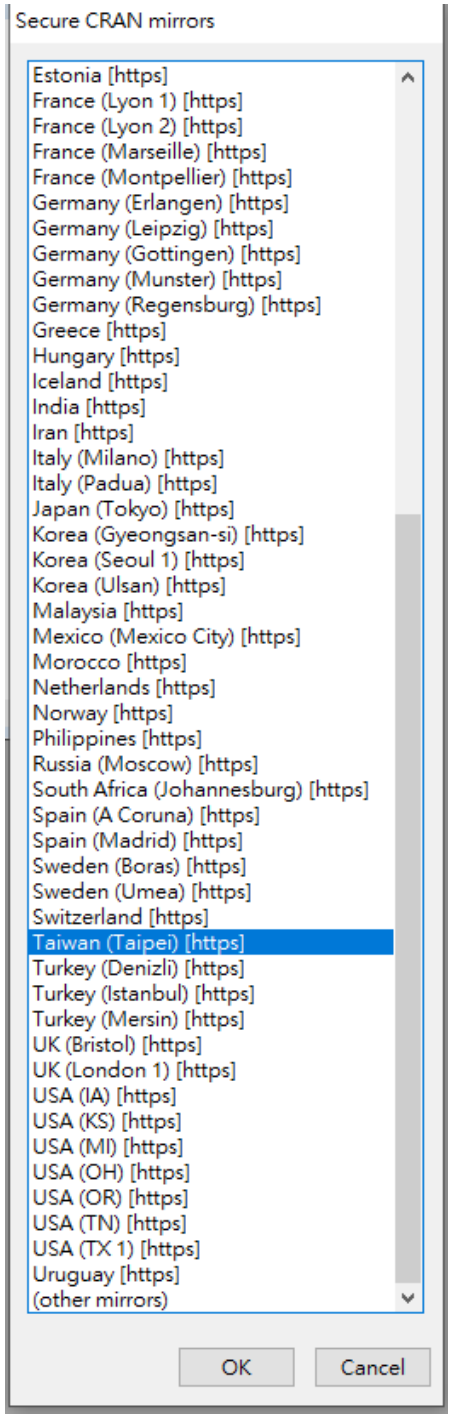
- Finding differentially expressed features (between conditions)

Install

R version 4.0 or greater is required (R-4.1.0)

Seurat 4.0.6





Import data

https://satijalab.org/seurat/articles/pbmc3k_tutorial.html

Example Data:
Peripheral Blood Mononuclear Cells (PBMC) freely available from 10X Genomics.
There are 2,700 single cells that were sequenced on the Illumina NextSeq 500.

- **10X**

```
rm(list=ls(all=TRUE))
library(Seurat)
setwd("C:/SC_Example")
test.data <- Read10X(data.dir = "filtered_feature_bc_matrix")
pbmc <- CreateSeuratObject(counts = test.data, project = "pbmc", min.cells = 0, min.features = 0)
pbmc
```

請依實際狀況做更改

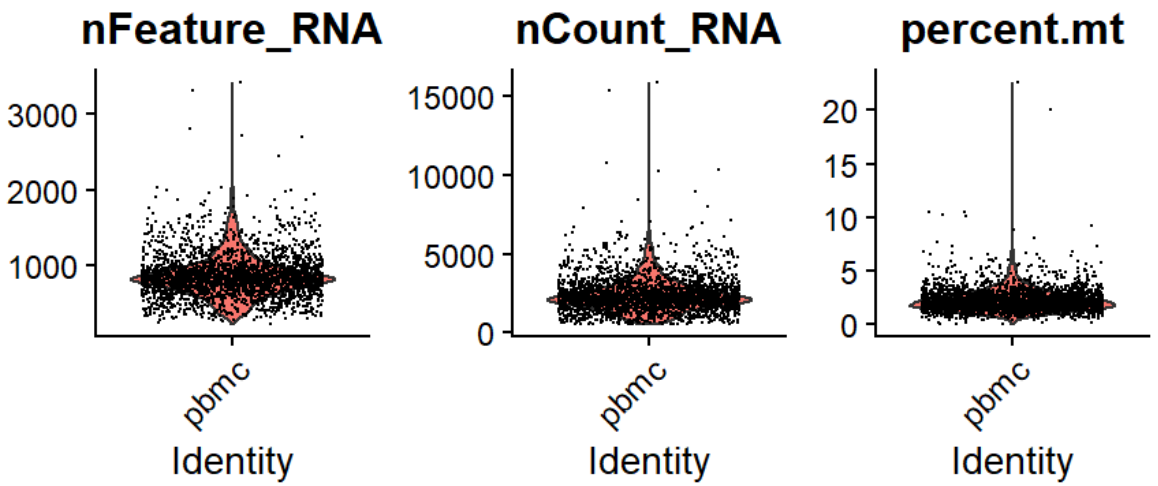
■ cellranger

保留gene
至少有幾個cell
表現

保留cell
至少有幾個gene
表現

◆**1.1**

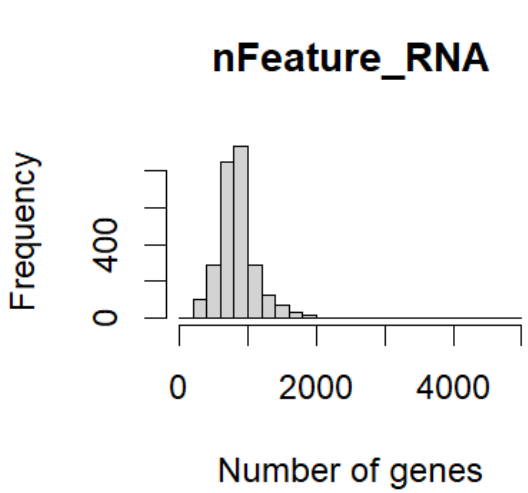
```
#The percentage of reads that map to the mitochondrial genome
pbmc[["percent.mt"]] <- PercentageFeatureSet(pbmc, pattern = "^MT-")
zz<- grep("^MT-", rownames(pbmc))
MTgene<- rownames(pbmc)[zz]
MTgene


# Visualize QC metrics as a violin plot
VlnPlot(pbmc, features = c("nFeature_RNA", "nCount_RNA", "percent.mt"), ncol = 3)
```

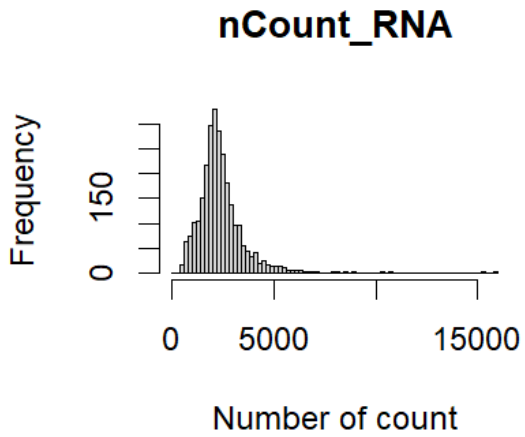


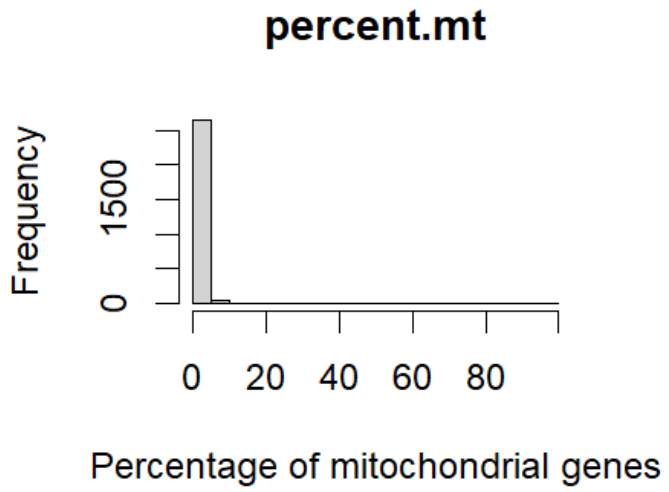
◆**1.1**

breaks: 0-5000間, 200為一條; xlab=X軸標籤; main= 圖的標圖

```
hist( unlist(pbmc@meta.data$nFeature_RNA), breaks=seq(0,5000,200), xlab="Number of genes", main="nFeature_RNA")
hist( unlist(pbmc@meta.data$nCount_RNA), breaks=seq(0,16000,200), xlab="Number of count", main="nCount_RNA")
hist( unlist(pbmc@meta.data$percent.mt), breaks=seq(0,100,5), xlab="Percentage of mitochondrial genes", main="percent.mt")
```








```
#Filter cells that have unique feature counts less than 200 and >5% mitochondrial counts
pbmc.1 <- subset(pbmc, subset = nFeature_RNA > 200  & nFeature_RNA < 2500  &  percent.mt < 5)
pbmc.1
#save data
saveRDS(pbmc.1, file = "pbmc step1.rds")
```

◆2.1 normalizing ,HVG, PCA

```
rm(list=ls(all=TRUE))
library(Seurat)
library(ggplot2)

setwd("C:/SC_Example")
pbmc.1<- readRDS("pbmc step1.rds")

#Normalizing the data
pbmc.1 <- NormalizeData(pbmc.1, normalization.method = "LogNormalize", scale.factor = 10000)

#Identification of highly variable features (feature selection)
pbmc.1 <- FindVariableFeatures(pbmc.1, selection.method = "vst", nfeatures = 2000)

#Scaling the data
all.genes <- rownames(pbmc.1)
pbmc.1 <- ScaleData(pbmc.1, features = all.genes)

#Perform linear dimensional reduction
pbmc.1 <- RunPCA(pbmc.1, features = VariableFeatures(object = pbmc.1),verbose=FALSE)
```

以cell 為單位

$$\log_e\left(\frac{\text{count}}{\text{Total count}} * 10000\right)+1$$

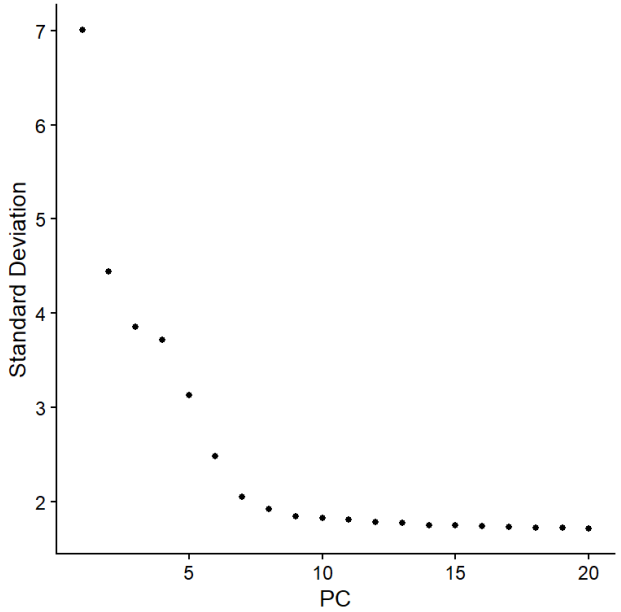
以gene 為單位 $\frac{\text{x-mean(x)}}{\text{sd(x)}}$

◆2.2 clustering, tSNE/UMAP

#Determine the 'dimensionality' of the dataset

ElbowPlot(pbmc.1, ndims = 20)

n.pcs<- 10



#Cluster the cells

pbmc.1 <- FindNeighbors(pbmc.1, dims =1:n.pcs)

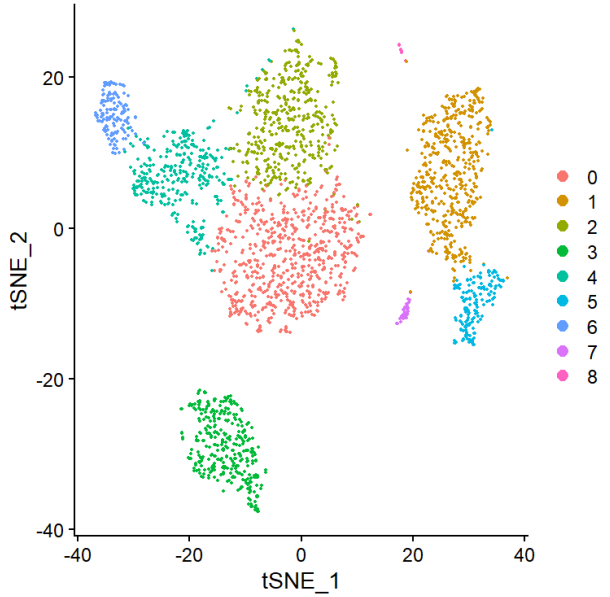
pbmc.1 <- FindClusters(pbmc.1, resolution = 0.5)

head(Idents(pbmc.1))

#Run non-linear dimensional reduction (UMAP/tSN

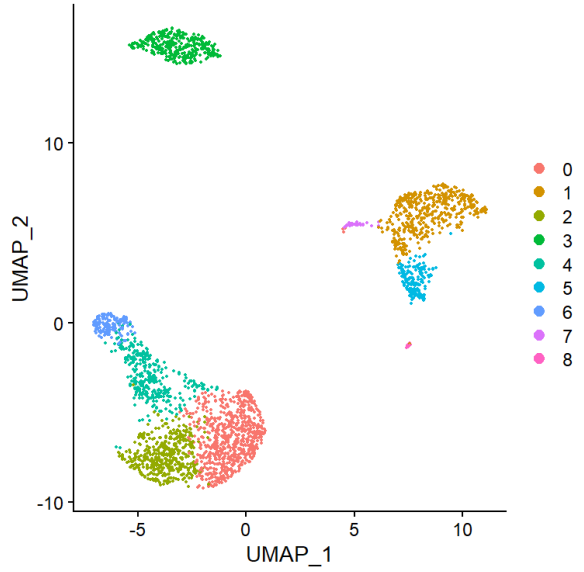
pbmc.1 <- RunTSNE(pbmc.1, dims = 1:n.pcs)

DimPlot(pbmc.1, reduction = "tsne")



pbmc.1 <- RunUMAP(pbmc.1, dims = 1:n.pcs)

DimPlot(pbmc.1, reduction = "umap")



◆3.1 DEG (cluster marker)

```
## Find markers for each cluster
pbmc.markers <- FindAllMarkers(object =pbmc.1, test.use = "wilcox", only.pos = TRUE, min.pct =0.1, logfc.threshold = 0.25)
pbmc.markers[1,]
write.table(pbmc.markers,"DEG_cluster.txt",sep="\t", row.names=F,col.names=T)
```

```
> pbmc.markers[1:10,]
              p_val avg_log2FC pct.1 pct.2     p_val_adj cluster  gene
RPS12 1.647066e-146  0.7347797 1.000 0.991 5.392164e-142       0 RPS12
RPS6  2.677487e-145  0.6854074 1.000 0.995 8.765556e-141       0  RPS6
RPL32 8.402925e-143  0.6318857 0.999 0.995 2.750950e-138       0 RPL32
RPS14 1.186082e-136  0.6377198 1.000 0.994 3.882994e-132       0 RPS14
RPS27 3.122041e-136  0.7163892 0.999 0.992 1.022094e-131       0 RPS27
RPS25 1.493222e-125  0.7562751 0.997 0.975 4.888509e-121       0 RPS25
RPL31 9.668542e-124  0.7651704 0.999 0.963 3.165287e-119       0 RPL31
RPL9  1.870069e-121  0.7575762 0.999 0.970 6.122232e-117       0  RPL9
RPS3  1.059537e-116  0.5956159 1.000 0.994 3.468713e-112       0  RPS3
RPS3A 1.228037e-116  0.8094488 0.999 0.974 4.020347e-112       0 RPS3A
```

pct.1: The percentage of cells where the gene is detected in the first group
cluster 那一群有表現的比例

pct.2: The percentage of cells where the gene is detected in the seconnd group
非cluster 那一群有表現的比例

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

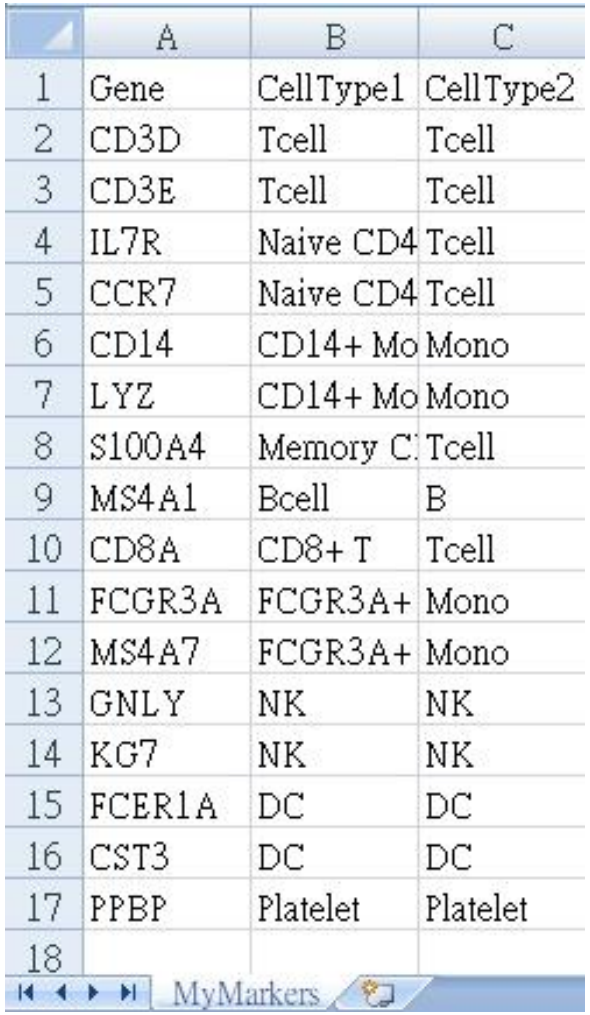
◆3.1 DEG (cluster marker)

```
> pbmc.markers3
               p_val avg_log2FC pct.1 pct.2     p_val_adj cluster   gene                CellType CellType.2
CCR7    5.253669e-83  1.3237568 0.440 0.111  1.719946e-78       0   CCR7             Naive CD4+ T      Tcell
CD3D    1.136915e-78  0.9315162 0.853 0.403  3.722033e-74       0   CD3D                   Tcell      Tcell
CD3E    2.246031e-53  0.8427934 0.732 0.399  7.353057e-49       0   CD3E                   Tcell      Tcell
IL7R    1.670053e-37  0.7391429 0.608 0.330  5.467421e-33       0   IL7R Naive CD4+ T;Memory CD4+      Tcell
CD14   7.715470e-291  2.8084634 0.664 0.029 2.525891e-286       1   CD14               CD14+ Mono      Mono
CST3   2.819410e-267  2.9914765 0.992 0.266 9.230184e-263       1   CST3                       DC        DC
LYZ    6.170586e-266  4.5059040 1.000 0.517 2.020127e-261       1    LYZ               CD14+ Mono      Mono
S100A4 2.785551e-178  1.6542349 1.000 0.771 9.119336e-174       1 S100A4             Memory CD4+      Tcell
MS4A7   1.922031e-24  0.5028249 0.259 0.087  6.292344e-20       1  MS4A7             FCGR3A+ Mono      Mono
IL7R1   3.974449e-65  1.1875370 0.752 0.330  1.301155e-60       2   IL7R Naive CD4+ T;Memory CD4+      Tcell
CD3D1   1.061409e-64  0.9164514 0.914 0.438  3.474842e-60       2   CD3D                   Tcell      Tcell
CD3E1   9.609833e-55  0.8865347 0.840 0.412  3.146067e-50       2   CD3E                   Tcell      Tcell
MS4A1   0.000000e+00  3.3773568 0.855 0.053  0.000000e+00       3  MS4A1                  Bcell          B
CD8A   1.684884e-102  1.8850322 0.468 0.069  5.515974e-98       4   CD8A                 CD8+ T      Tcell
CD3D2   5.563834e-50  1.1056842 0.849 0.474  1.821488e-45       4   CD3D                   Tcell      Tcell
CD3E2   1.250362e-21  0.7290685 0.711 0.454  4.093437e-17       4   CD3E                   Tcell      Tcell
GNLY    4.276716e-16  1.1755342 0.329 0.159  1.400111e-11       4   GNLY                      NK         NK
IL7R2   5.562830e-05  0.4698936 0.465 0.394  1.000000e+00       4   IL7R Naive CD4+ T;Memory CD4+      Tcell
MS4A71 3.077954e-186  2.7172865 0.812 0.073 1.007661e-181       5  MS4A7             FCGR3A+ Mono      Mono
FCGR3A 2.201148e-184  3.3134927 0.981 0.135 7.206119e-180       5 FCGR3A             FCGR3A+ Mono      Mono
CST31   9.350122e-74  1.8845448 1.000 0.359  3.061043e-69       5   CST3                       DC        DC
S100A41 2.132313e-60  1.4841389 1.000 0.801  6.980766e-56       5 S100A4             Memory CD4+      Tcell
GNLY1  2.195712e-172  4.8802764 0.964 0.136 7.188323e-168       6   GNLY                      NK         NK
FCGR3A1 2.463728e-115 2.3504823 0.886 0.147 8.065752e-111       6 FCGR3A             FCGR3A+ Mono      Mono
FCER1A  3.748803e-251  3.8679707 0.829 0.010 1.227283e-246       7 FCER1A                   DC        DC
CST32   1.696651e-23  2.5123174 1.000 0.390  5.554496e-19       7   CST3                   DC        DC
LYZ1    4.326096e-13  1.7601162 0.971 0.600  1.416277e-08       7    LYZ               CD14+ Mono      Mono
PPBP   3.684548e-110  8.5699598 1.000 0.024 1.206247e-105       8   PPBP                 Platelet  Platelet
```

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

◆3.1 DEG (cluster marker)

```
markers<- read.table("MyMarkers.csv",sep=",", header=T,fill=T, quote = "", check.names=F )
head(markers)



pbmc.markers2<- pbmc.markers[,c(1:7,7,7)]
colnames(pbmc.markers2)[8]<- "CellType"
for(i in 1:dim(markers)[1]){
xx<- pbmc.markers2[,7] %in% markers[i,1]
pbmc.markers2[xx,8]<- markers[i,2]
pbmc.markers2[xx,9]<- markers[i,3]
}

xx<- pbmc.markers2[,7] %in% markers[,1]
pbmc.markers3<- pbmc.markers2[xx,]
pbmc.markers3

write.table(pbmc.markers3,"DEG_cluster_marker.txt",sep="\t", row.names=F,col.names=T)
```



	A	B	C
1	Gene	CellType1	CellType2
2	CD3D	Tcell	Tcell
3	CD3E	Tcell	Tcell
4	IL7R	Naive CD4	Tcell
5	CCR7	Naive CD4	Tcell
6	CD14	CD14+ Mo	Mono
7	LYZ	CD14+ Mo	Mono
8	S100A4	Memory C	Tcell
9	MS4A1	Bcell	B
10	CD8A	CD8+ T	Tcell
11	FCGR3A	FCGR3A+	Mono
12	MS4A7	FCGR3A+	Mono
13	GNLY	NK	NK
14	KG7	NK	NK
15	FCER1A	DC	DC
16	CST3	DC	DC
17	PPBP	Platelet	Platelet
18			

MyMarkers

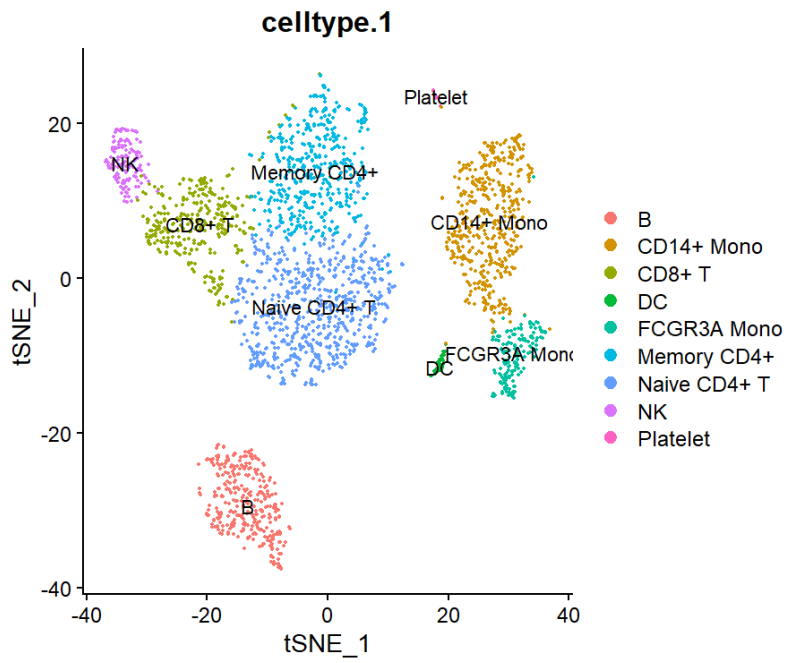
◆**3.2 Cell typing**

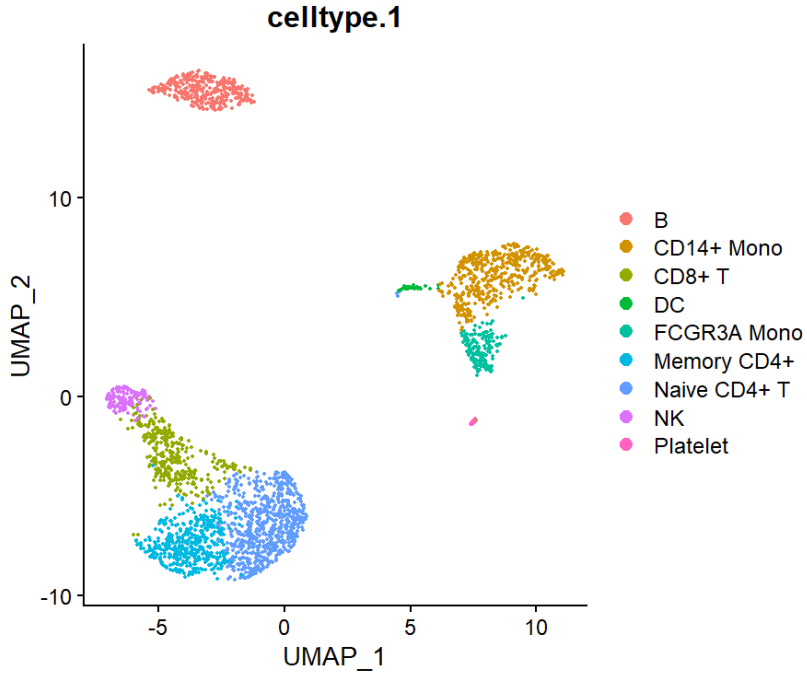
```
celltype<- rep("unknown", length(rownames(pbmc.1@meta.data)) )
celltype[Idents(pbmc.1) %in% c("0")]<- "Naive CD4+ T"
celltype[Idents(pbmc.1) %in% c("1")]<- "CD14+ Mono"
celltype[Idents(pbmc.1) %in% c("2")]<- "Memory CD4+"
celltype[Idents(pbmc.1) %in% c("3")]<- "B"
celltype[Idents(pbmc.1) %in% c("4")]<- "CD8+ T"
celltype[Idents(pbmc.1) %in% c("5")]<- "FCGR3A Mono"
celltype[Idents(pbmc.1) %in% c("6")]<- "NK"
celltype[Idents(pbmc.1) %in% c("7")]<- "DC"
celltype[Idents(pbmc.1) %in% c("8")]<- "Platelet"
pbmc.1@meta.data$celltype.1<- celltype
table(pbmc.1@meta.data$celltype.1)


saveRDS(pbmc.1, file = "pbmc step2.rds") #new
```

◆3.3 plots

DimPlot(pbmc.1, reduction = "tsne", label = TRUE, group.by="celltype.1")

DimPlot(pbmc.1, reduction = "umap", label = FALSE, group.by="celltype.1")



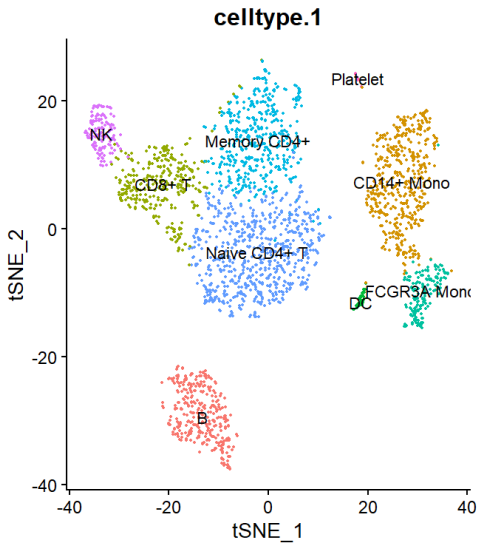


◆3.3 plots

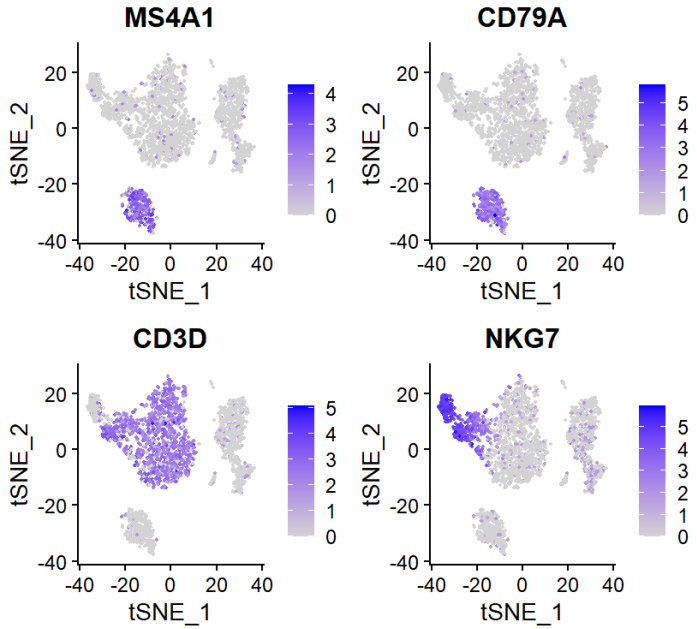
FeaturePlot(pbmc.1, features = c("MS4A1","CD79A", "CD3D", "NKG7"),reduction = "tsne")

VlnPlot(pbmc.1, features = c ("MS4A1","CD79A", "CD3D", "NKG7"),ncol=2, group.by="celltype.1")

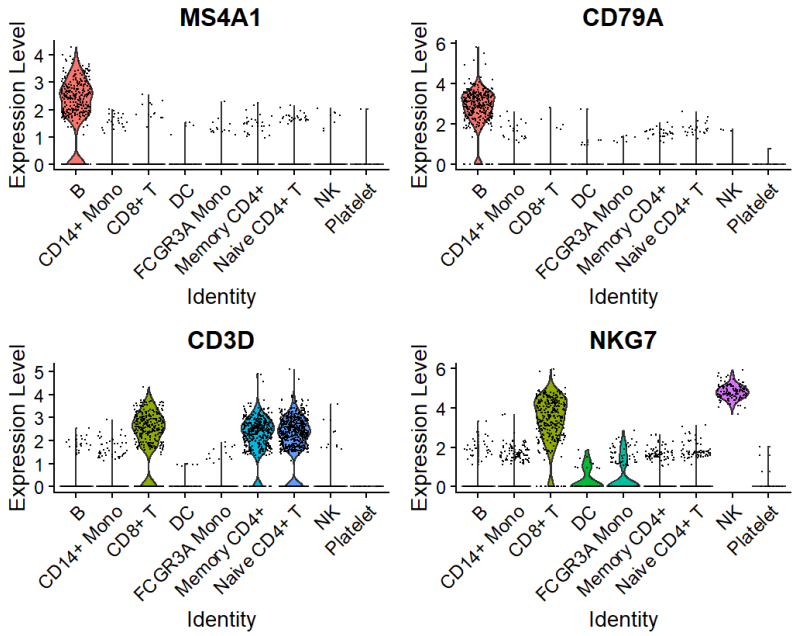
DotPlot(pbmc.1, features = c("MS4A1","CD79A", "CD3D", "NKG7")) +coord_flip()



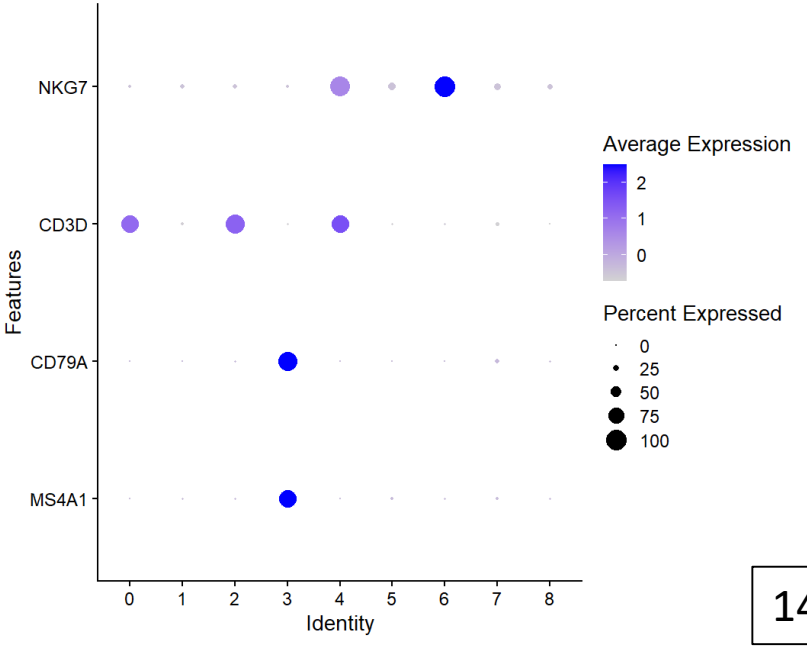
Feature Plot



Violin Plot



Dot Plot



◆4. DEG

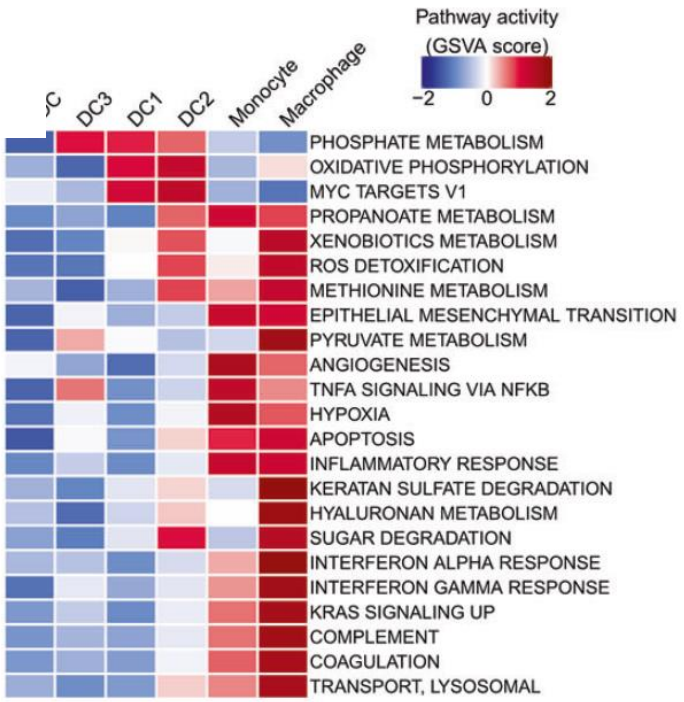
```
cellIDs<- rownames(pbmc.1@meta.data)


cell.use1<- cellIDs[pbmc.1@meta.data$celltype.1=="B"]
cell.use2<- cellIDs[pbmc.1@meta.data$celltype.1=="CD8+ T"]
pbmc.DEG <- FindMarkers(object =pbmc.1, ident.1=cell.use1, ident.2=cell.use2,
        test.use = "wilcox",only.pos = FALSE, min.pct =0.1,logfc.threshold = 0.25)
dim(pbmc.DEG)
head(pbmc.DEG)
write.table(pbmc.DEG,"DEG_condition.txt",sep="\t", row.names=F,col.names=T)
```

```
> head(pbmc.DEG)
                p_val avg_log2FC pct.1 pct.2     p_val_adj
HLA-DRA 5.447953e-114   4.791797 1.000 0.317 1.783551e-109
CD79A   4.550834e-111   4.390449 0.936 0.015 1.489852e-106
CCL5    3.204954e-110  -4.680050 0.142 0.960 1.049238e-105
CD74    2.722374e-109   3.466988 1.000 0.837 8.912508e-105
NKG7    4.969189e-109  -4.907363 0.087 0.945 1.626813e-104
IL32    4.221398e-103  -3.766465 0.099 0.926  1.382001e-98
```

- 後續可再進行…

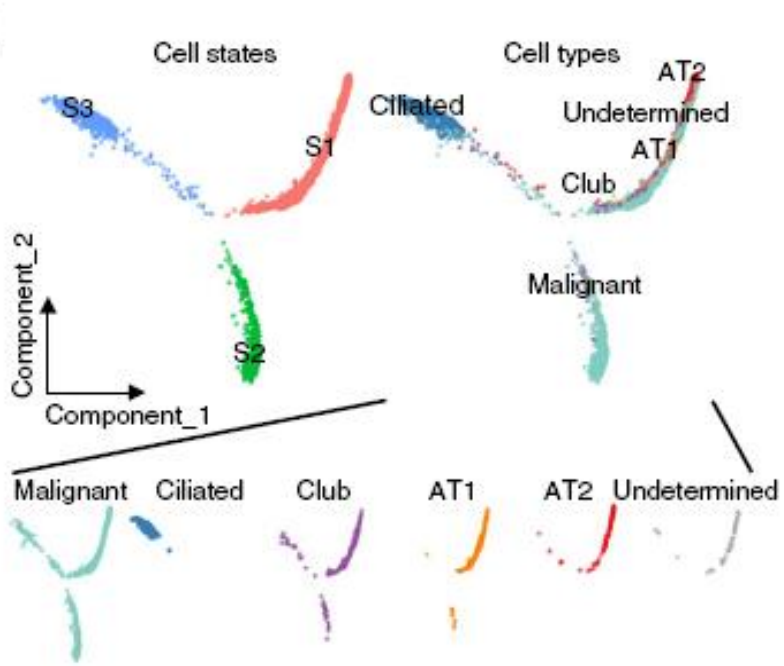
補充

➢ **gene set /pathway analysis**



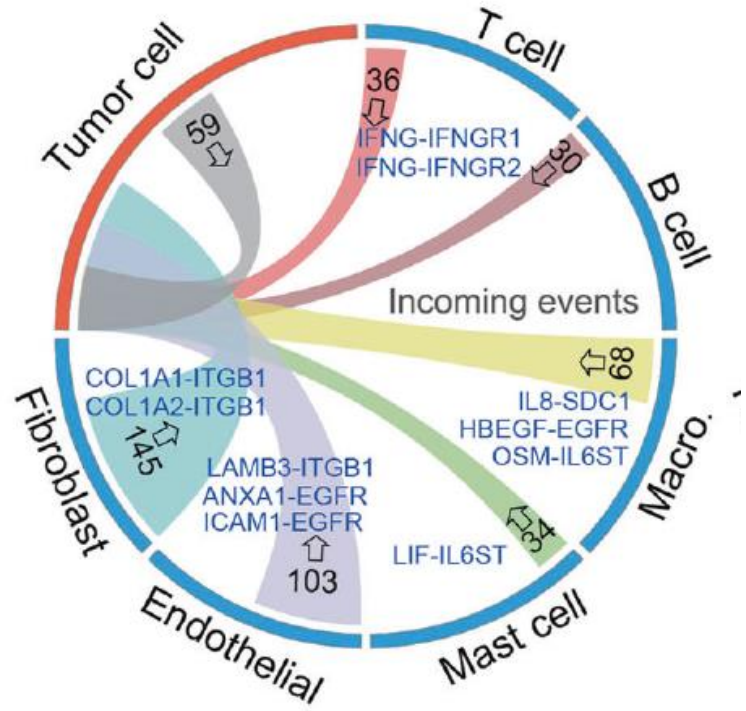
GSVA
(gene set variation analysis)

➢ **Trajectory**



monocle

➢ **Cell-cell interaction**



Cell Chat

Reference

- Tools for Single Cell Genomics

  https://satijalab.org/seurat/index.html

- Seurat – Guided Clustering Tutorial (2,700 PBMCs)

  https://satijalab.org/seurat/articles/pbmc3k_tutorial.html

END