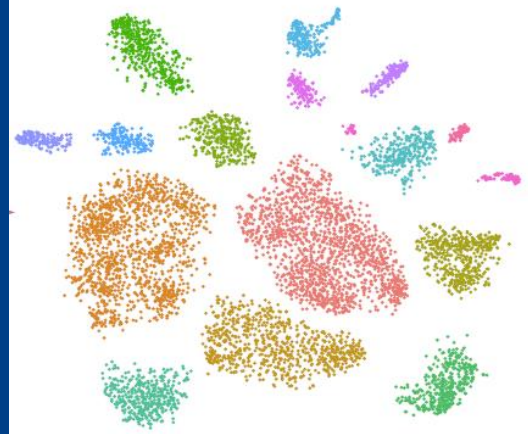
浙江大学
ZHEJIANG UNIVERSITY

单细胞转录组数据分析

——基于细胞类型的高级分析



褚琴洁　qinjiechu@zju.edu.cn

2023年10月30日

课件PPT

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

上节回顾（从FASTQ到基因-细胞表达矩阵）

方案一：UMI-tools + STAR + featureCounts

UMI-tools

STAR: ultrafast universal RNA-seq aligner

renkow[1], Chris Zaleski[1],
ingeras[1]
iences, Menlo Park, CA, USA

SUBREAD

Subread package: high-performance read alignment, quantification and mutation discovery

Step 1: get data

Step 2: Identify correct cell barcodes

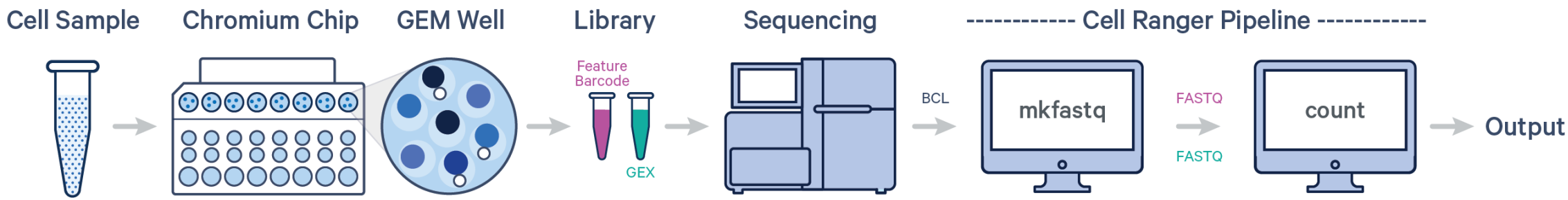
Step 3: Extract barcdoes and UMIs and add to read names

Step 4: Map reads

Step 5: Assign reads to genes

Step 6: Count UMIs per gene per cell

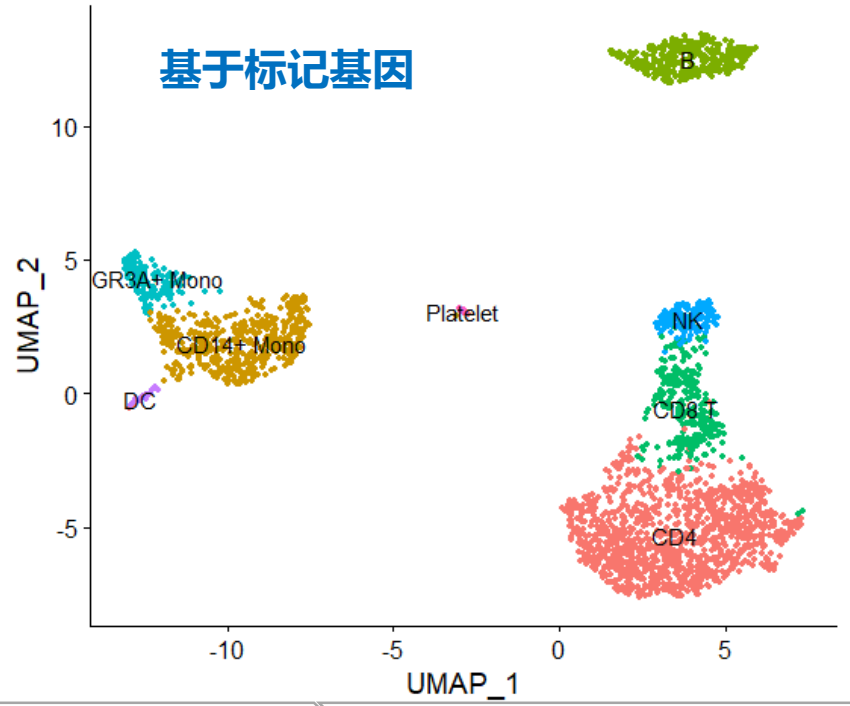
方案二：CellRanger



上节回顾（从表达矩阵到细胞类型注释）

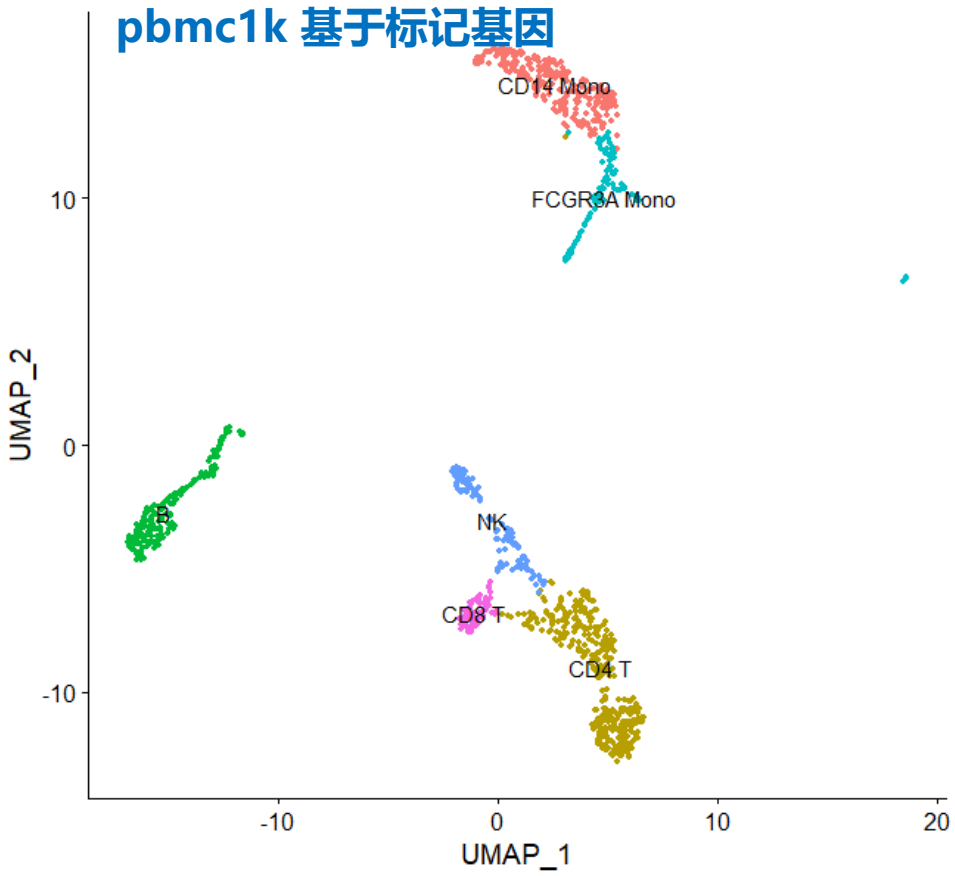
	Cell1	Cell2	…	CellN
Gene1	3	2	.	13
Gene2	2	3	.	1
Gene3	1	14	.	18
…	.	.	.	.
…	.	.	.	.
…	.	.	.	.
GeneM	25	0	.	0

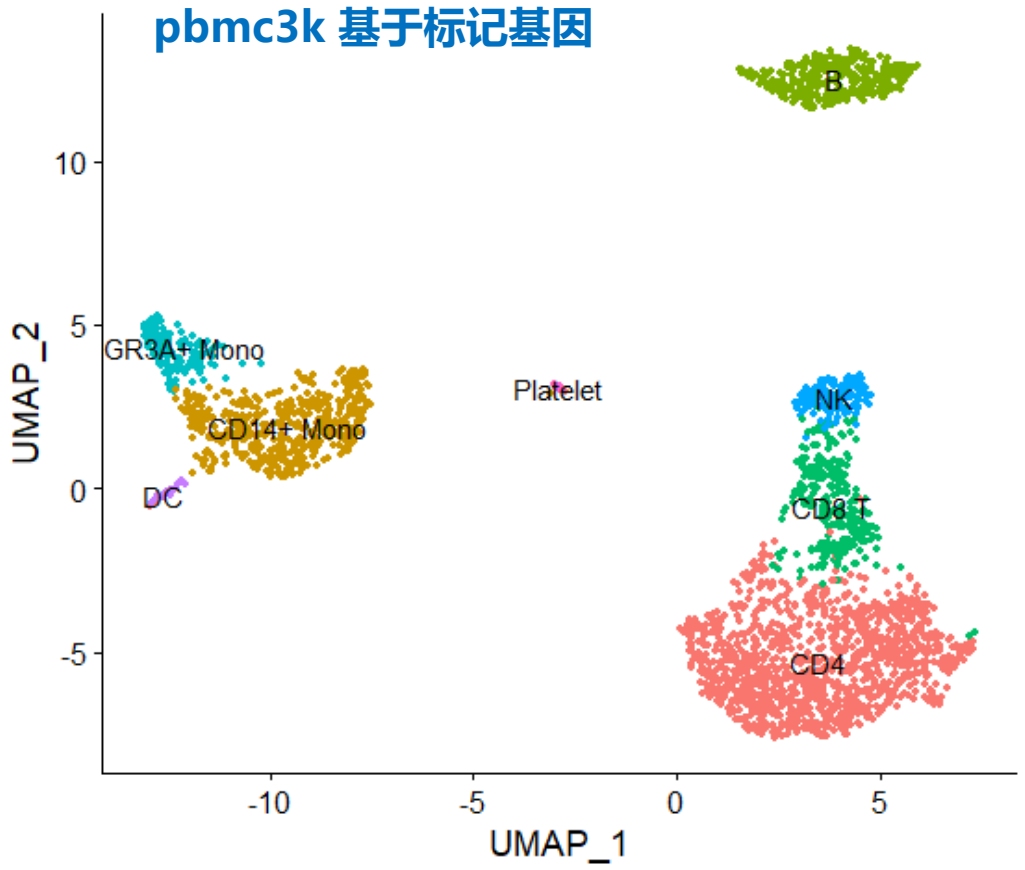
- 质控 Quality control
  - 基因数和UMI数、线粒体比例
  - 双细胞判断、去除空液滴、去除环境RNA、细胞周期判断（optional）
- 标准化 Normalization
- 特征基因选择 Feature selection
- 中心化 Scaling
- 降维 Dimensionality reduction
- 聚类 Cluster analysis
- 细胞类型注释 Cell type annotation



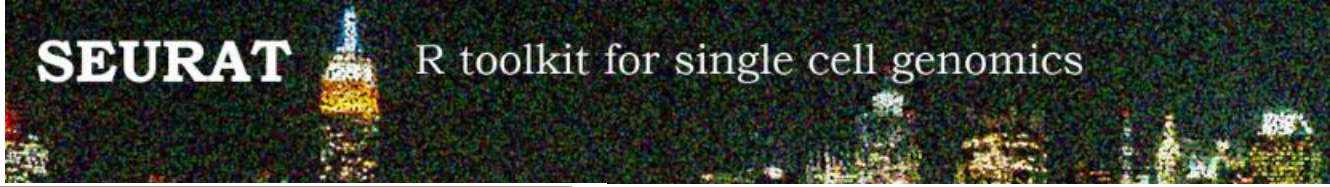
上节回顾（PBMC数据）

PBMC (peripheral blood mononuclear cell)，其主要细胞类型为血液里边具有单个核的细胞，主要包括淋巴细胞(T细胞、B细胞和NK细胞)，单核细胞，吞噬细胞，树突状细胞和其他少量细胞类型。



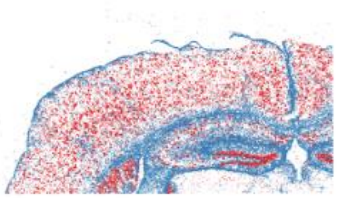


上节回顾（Seurat包介绍）



SATIJA LAB

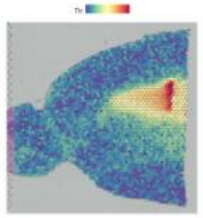
Analysis of spatial datasets (Imaging-based)



Learn to explore spatially-resolved data from multiplexed imaging technologies, including MERSCOPE, Xenium, CosMx SMI, and CODEX.

GO

Analysis of spatial datasets (Sequencing-based)



Learn to explore spatially-resolved transcriptomic data with examples from 10x Visium and Slide-seq v2.

GO

scRNA-seq Integration

Seurat v5

harmony

scvi-tools

Integrate scRNA-seq datasets using a variety of computational methods.

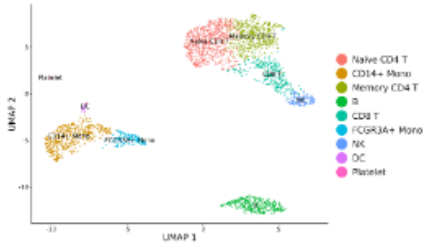
GO

Cross-modality Bridge Integration

Map scATAC-seq onto an scRNA-seq reference using a multi-omic bridge dataset.

GO

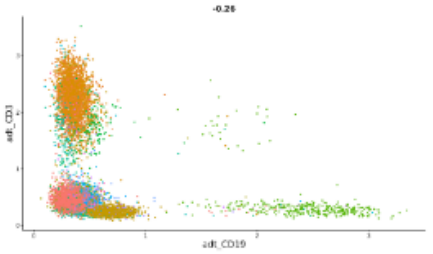
Guided tutorial — 2,700 PBMCs



A basic overview of Seurat that includes an introduction to common analytical workflows.

GO

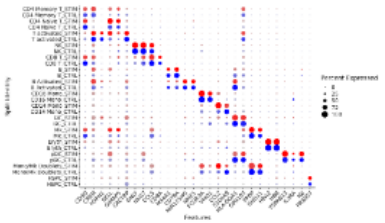
Multimodal analysis



An introduction to working with multi-modal datasets in Seurat.

GO

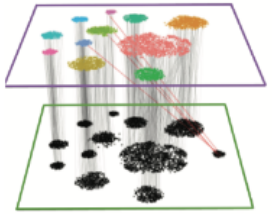
Introduction to scRNA-seq integration



An introduction to integrating scRNA-seq datasets in order to identify and compare shared cell types across experiments.

GO

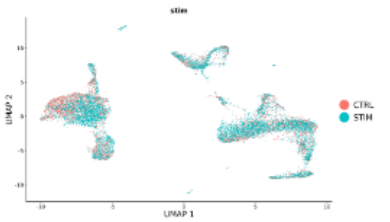
Mapping and annotating query datasets



Learn how to map a query scRNA-seq dataset onto a reference in order to automate the annotation and visualization of query cells.

GO

Fast integration using reciprocal PCA (RPCA)



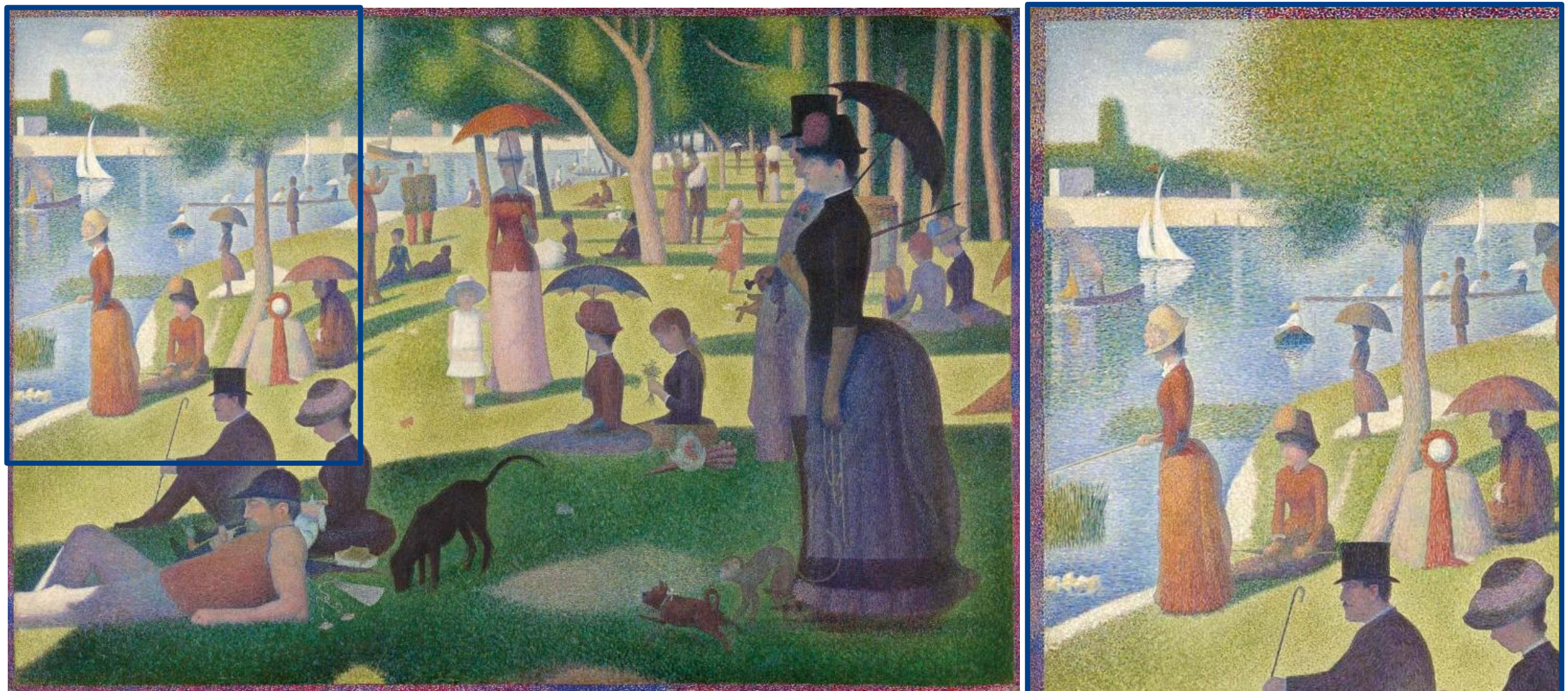
Identify anchors using the reciprocal PCA (rPCA) workflow, which performs a faster and more conservative integration.

GO

Georges Seurat

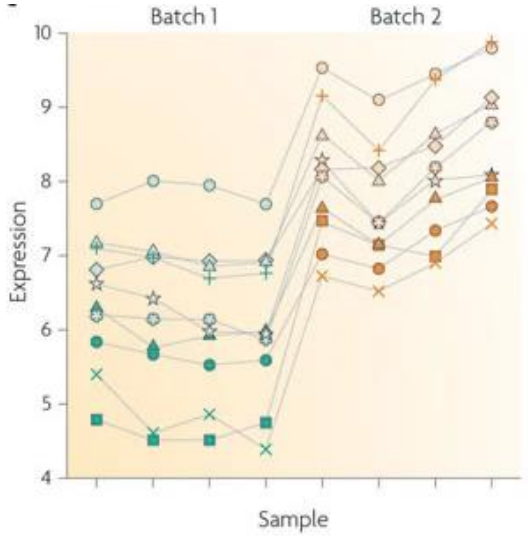
Painter, born December 2, 1859, Paris, France—died March 29, 1891, Paris

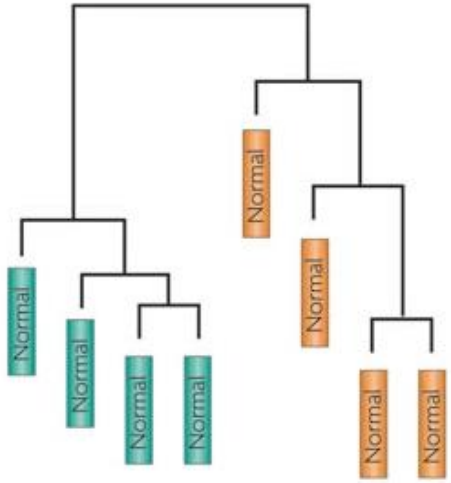
受过完整的美术学院教育，曾师从安格尔的学生亨利·莱曼（Henri Lehmann）学习古典主义绘画，后来又研究过卢浮宫中的大师作品，对光学和色彩理论特别关注并为之做了大量的实验。他的画作风格相当与众不同，Seurat的画**充满了细腻缤纷的小点**，当你靠近看，每一个点都充满著理性的笔触.



数据整合与批次效应

为什么要数据整合？什么是批次效应？批次效应会产生什么影响？是否要去除批次效应？





批次效应：生物学处理效应之外的其他因素导致的样本结果的波动

- 不同建库策略
- 不同实验操作者
- 不同测序平台
- 不同测序批次
- 不同公司或版本的试剂

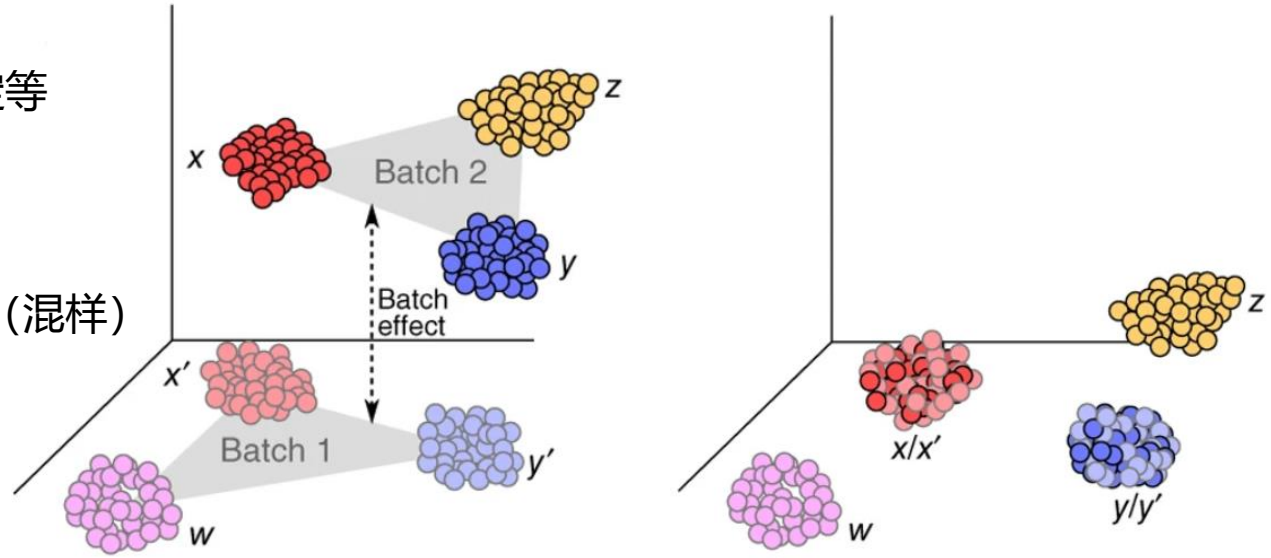
Leek et al., 2010

批次效应的影响

可能会引起分析结果的假阴性和假阳性、影响细胞亚群的鉴定等

如何解决?

- 尽量让一个项目的不同样本间没有或尽可能少的批次效应（混样）
- 通过生物信息学的方法，矫正批次效应的影响



假阴性

假设批次效应和处理效应不完全重叠

- 相当于扩大了组内差异
- 导致组间差异/组内差异的比值减少
- 降低了处理效应的显著性
- 即组间差异显著的基因减少

假阳性

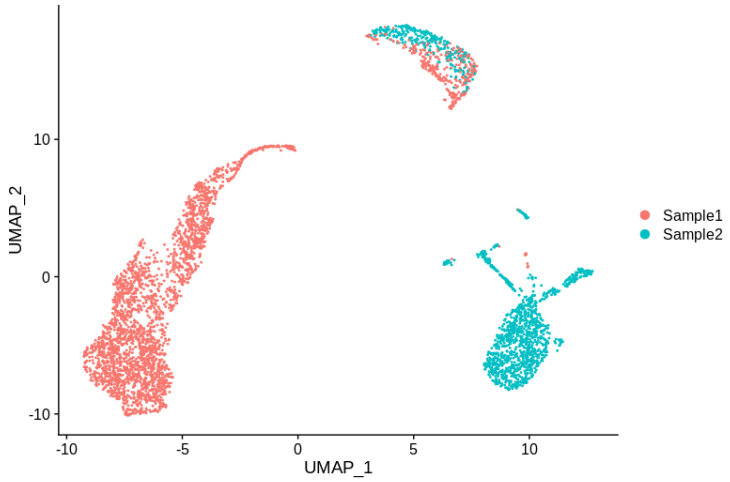
假设批次效应和处理效应完全重叠或近似重叠

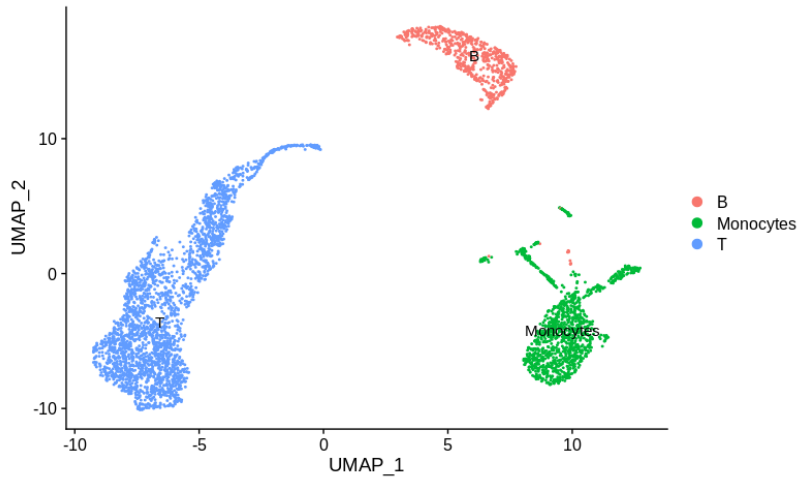
- 一般会整体加大组间的差异
- 难以区分差异是由于实验处理导致还是批次效应导致

Haghverdi et al., 2018

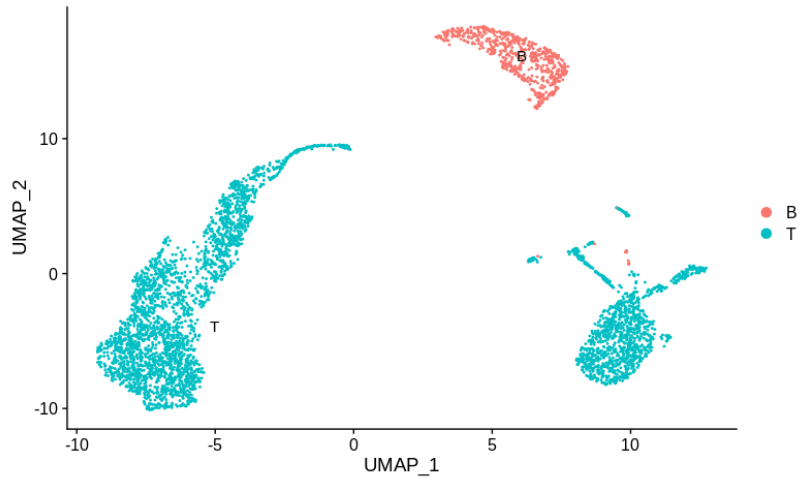
是否要去除批次效应?

- 发现批次效应和矫正批次效应从根本上说是生物学的，而不是技术的





- 无需进行批次效应的去除
- 研究为什么样本 1 中没有单核细胞，而样本 2 中没有T细胞



- 矫正之前，仍然需要更多信息
- 明确两个T细胞簇之间的差异
- 确定存在批次效应，上图中的差异超出了我们预期的T细胞组成或生物学差异；或者只对批次之间的相似性感兴趣。

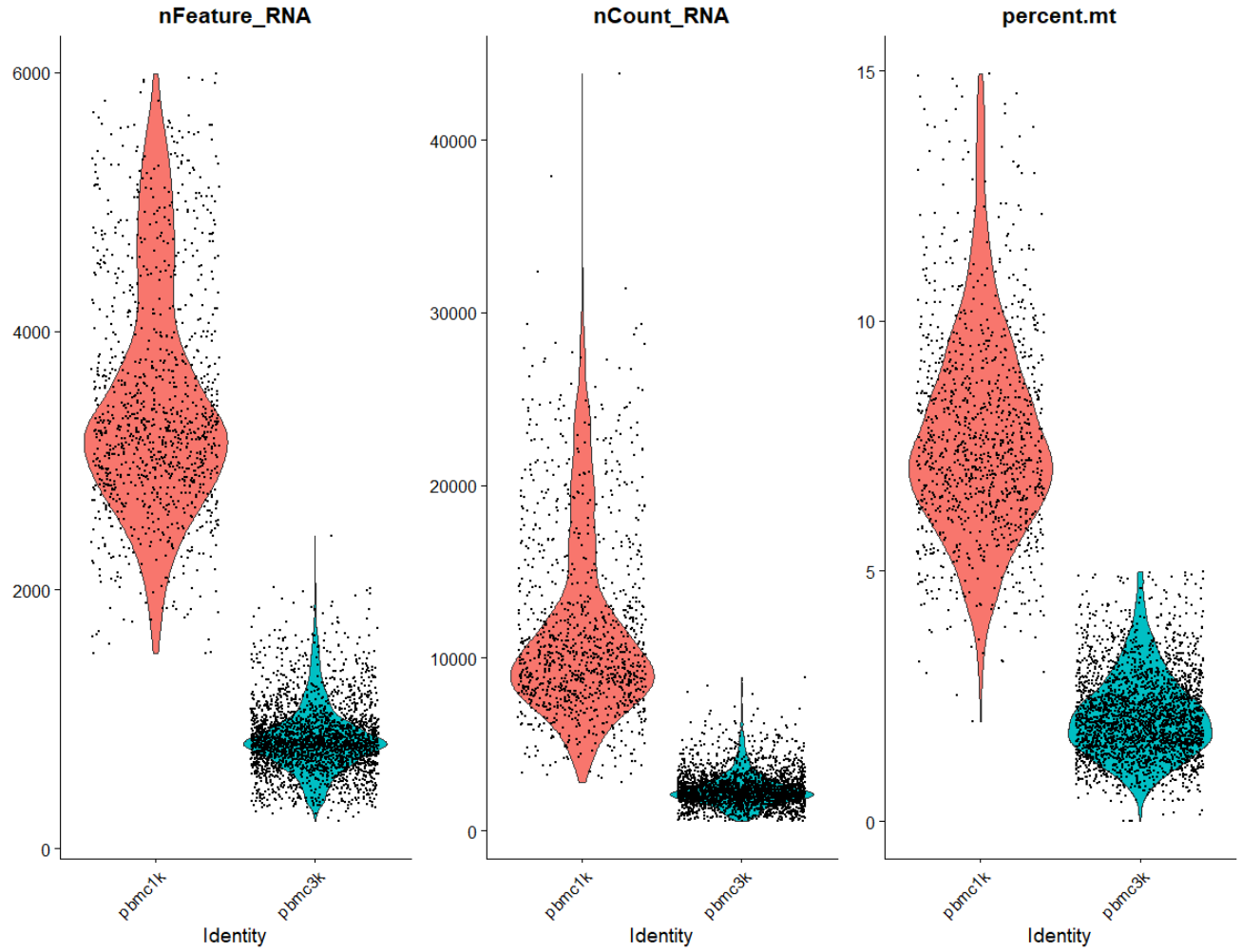
- 批次矫正方法不能明确到底怎么处理了数据，处理到什么程度
- 并不是所有的批次效应都可以或者应该被矫正

https://constantamateur.github.io/

整合 pbmc1k 和 pbmc3k 数据

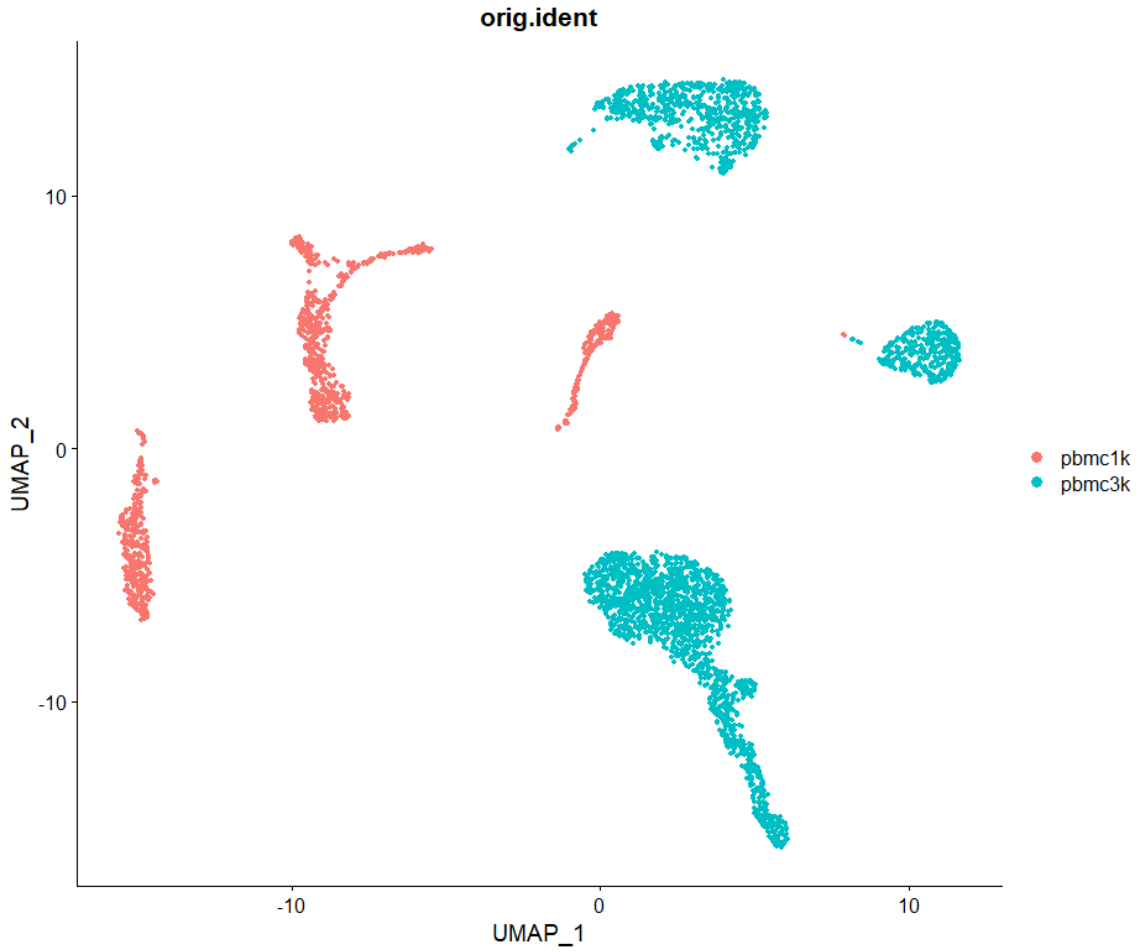
数据基本情况：

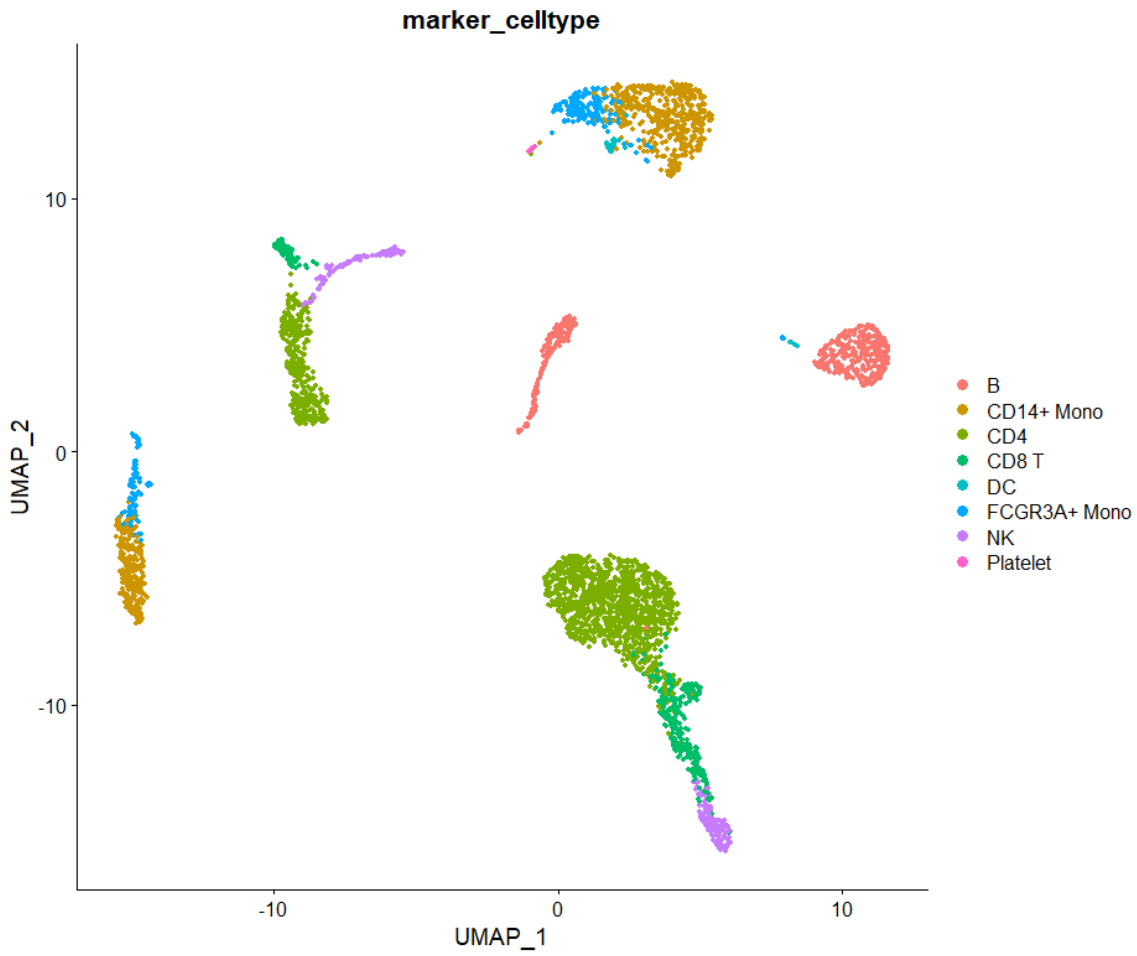
```
> pbmc1k
An object of class Seurat
23415 features across 1099 samples within 4 assays
Active assay: RNA (23148 features, 2000 variable features)
 3 other assays present: prediction.score.celltype.l1,
prediction.score.celltype.l2, predicted_ADT
 5 dimensional reductions calculated: pca, umap, tsne, ref.spca,
ref.umap

> pbmc3k
An object of class Seurat
13981 features across 2638 samples within 4 assays
Active assay: RNA (13714 features, 2000 variable features)
 3 other assays present: prediction.score.celltype.l1,
prediction.score.celltype.l2, predicted_ADT
 5 dimensional reductions calculated: pca, umap, tsne, ref.spca,
ref.umap

> pbmc_obj
An object of class Seurat
25823 features across 3737 samples within 4 assays
Active assay: RNA (25556 features, 0 variable features)
 3 other assays present: prediction.score.celltype.l1,
prediction.score.celltype.l2, predicted_ADT
```



整合 pbmc1k 和 pbmc3k 数据

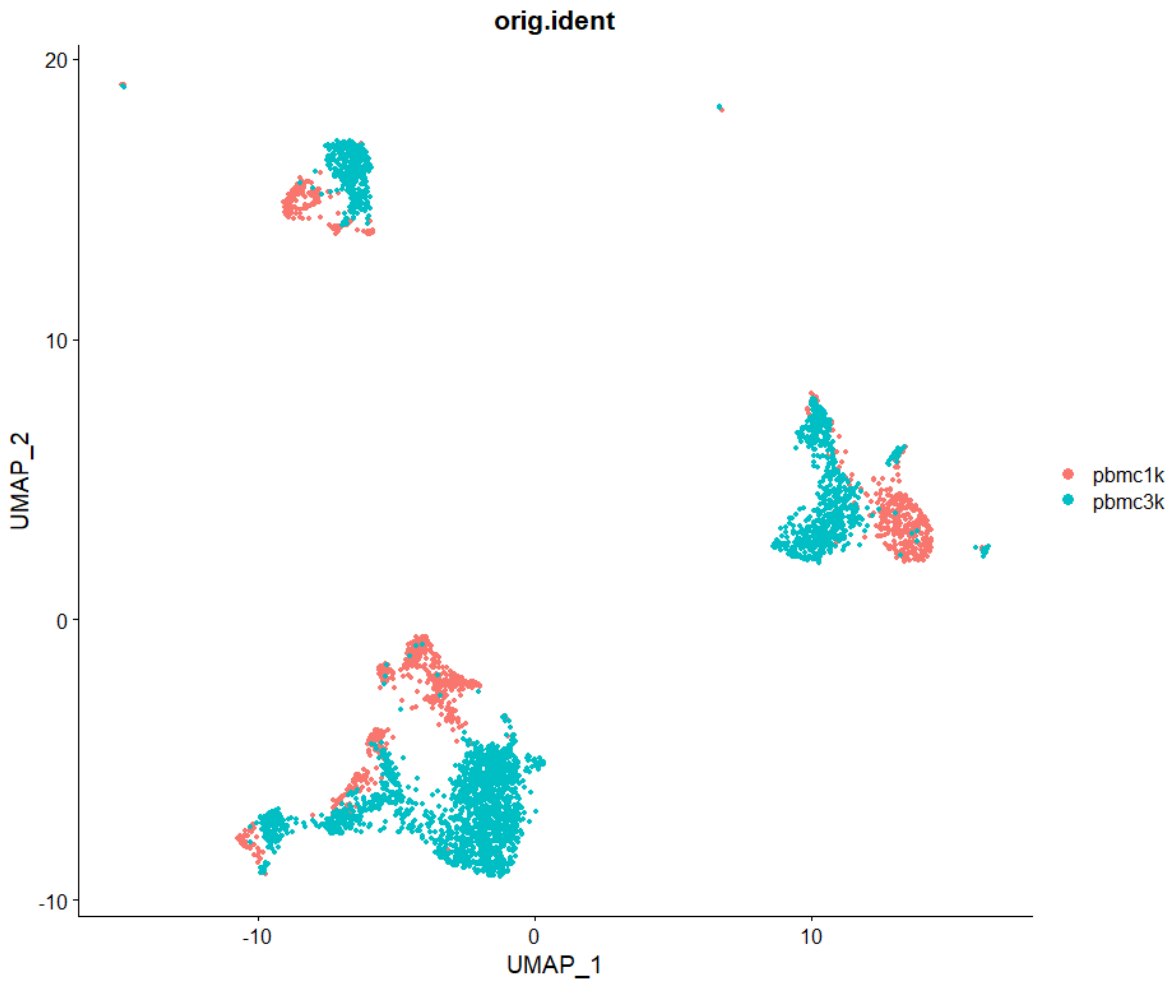
不考虑任何批次效应直接整合：存在明显的批次效应

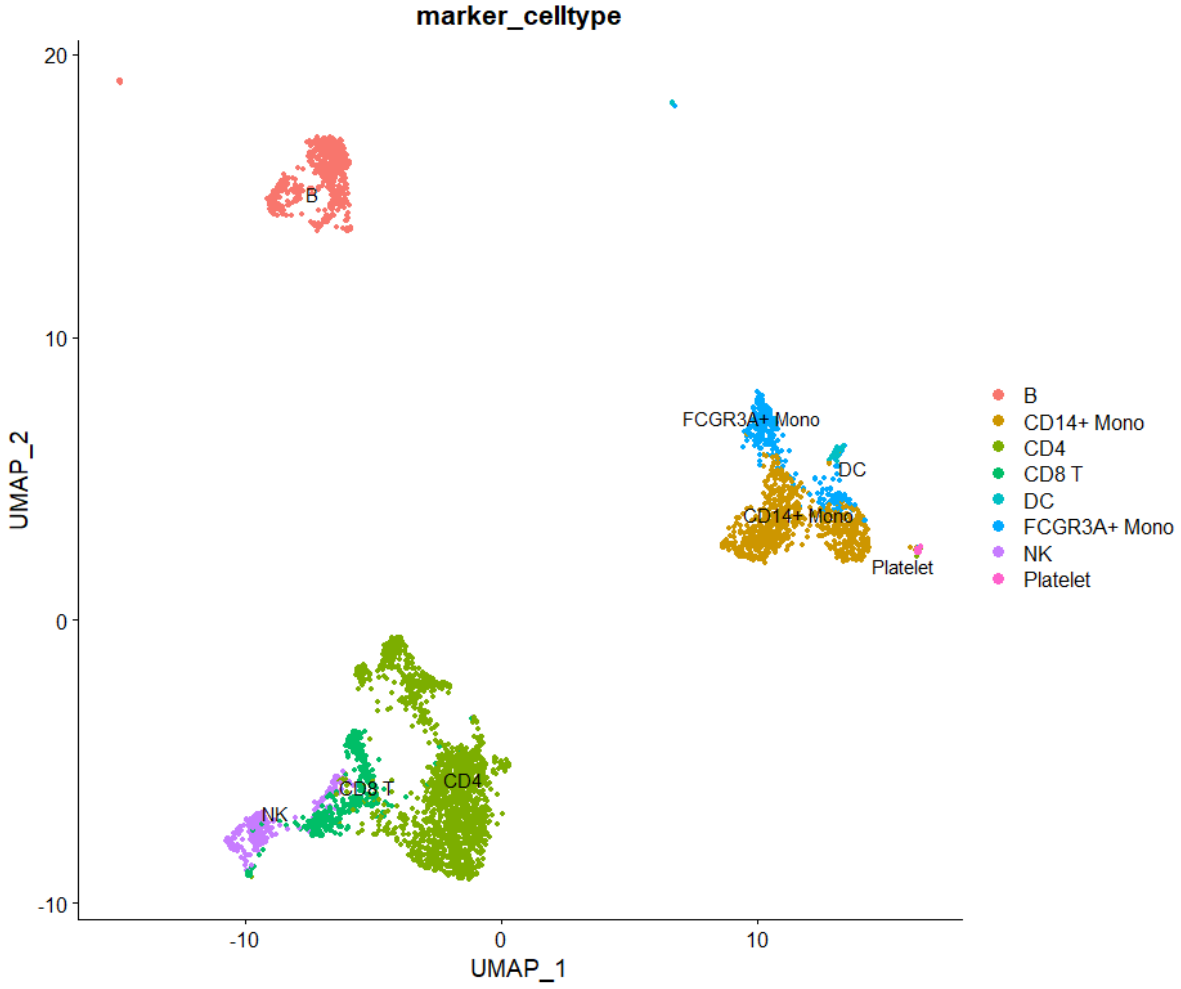




利用 Seurat 的RPCA整合数据

pbmc1k和pbmc3k数据示例（k.anchor = 5）





RPCA整合与简单合并在分析上的差异

简单合并

```
pbmc_obj <- NormalizeData(pbmc_obj, normalization.method = "LogNormalize", scale.factor = 10000)
pbmc_obj <- FindVariableFeatures(pbmc_obj, selection.method = "vst", nfeatures = 2000)
all.genes <- rownames(pbmc_obj)
pbmc_obj <- ScaleData(pbmc_obj, features = all.genes)
pbmc_obj <- RunPCA(pbmc_obj, features = VariableFeatures(object = pbmc_obj))
pbmc_obj <- FindNeighbors(pbmc_obj, dims = 1:10)
pbmc_obj <- FindClusters(pbmc_obj, resolution = 0.5)
pbmc_obj <- RunUMAP(pbmc_obj, dims = 1:10)
pbmc_obj <- RunTSNE(pbmc_obj, dims = 1:10)
```

RPCA整合

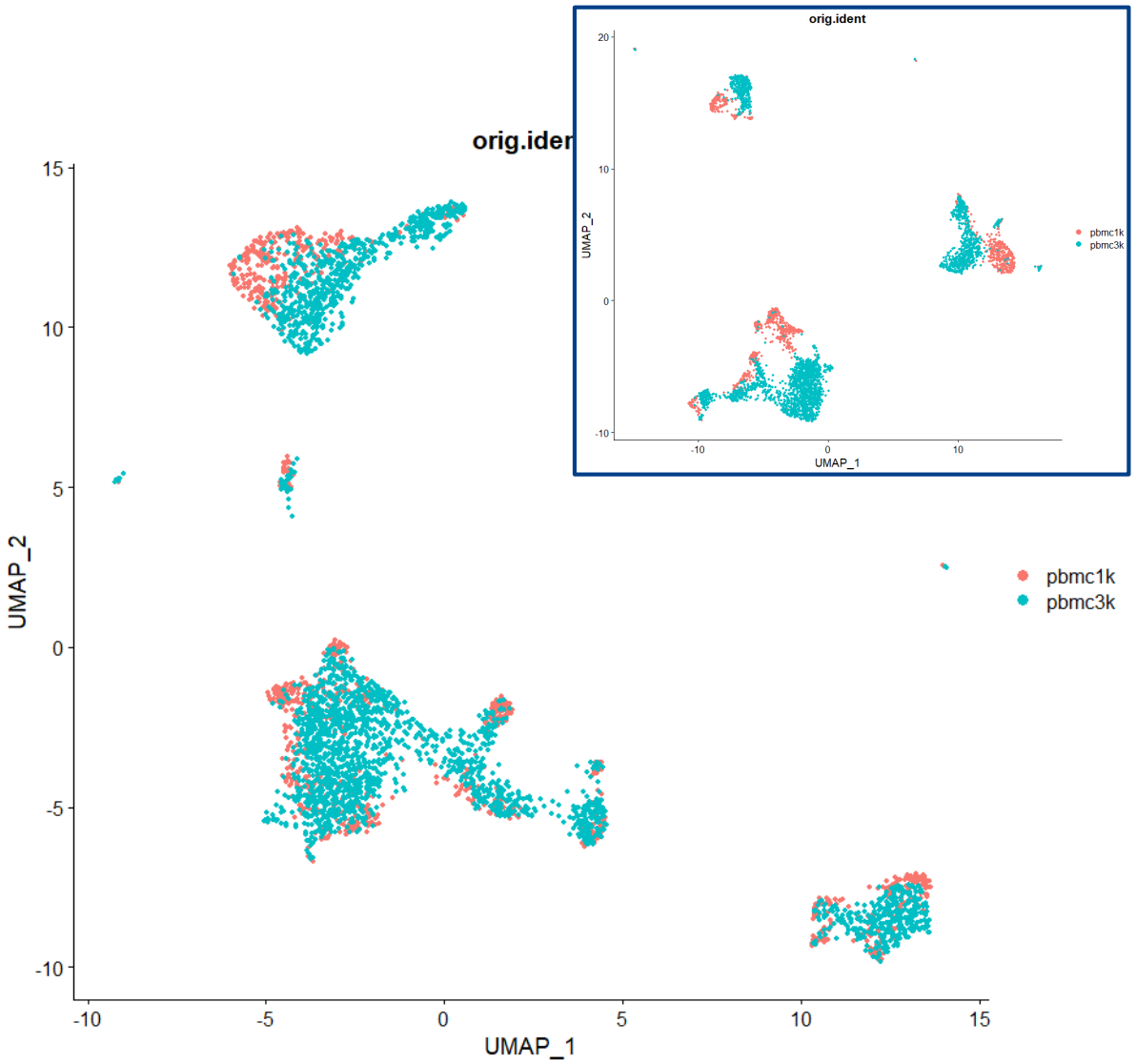
```
features <- SelectIntegrationFeatures(object.list = pbmc)
pbmc <- lapply(X = pbmc, FUN = function(x) {
  x <- ScaleData(x, features = features, verbose = FALSE)
  x <- RunPCA(x, features = features, verbose = FALSE)
})
```

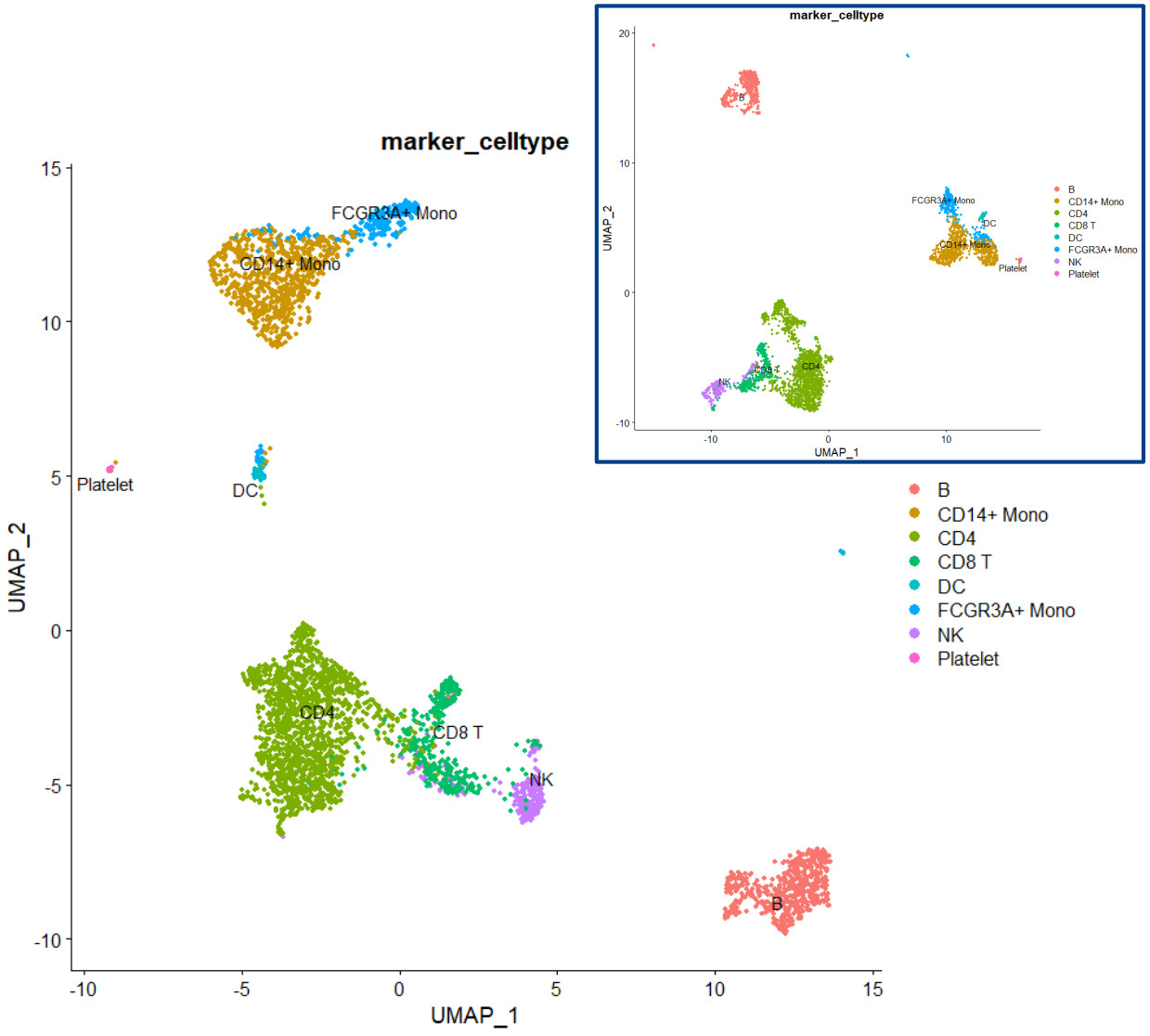
select features that are repeatedly variable across datasets for integration run PCA on each dataset using these features

```
anchors <- FindIntegrationAnchors(object.list = pbmc, anchor.features = features, reduction = "rpca")
combined <- IntegrateData(anchorset = anchors)

DefaultAssay(combined) <- "integrated"
```

利用 Seurat 的RPCA整合数据

pbmc1k和pbmc3k数据示例（FindIntegrationAnchors函数中k.anchor = 20来增加整合力度）



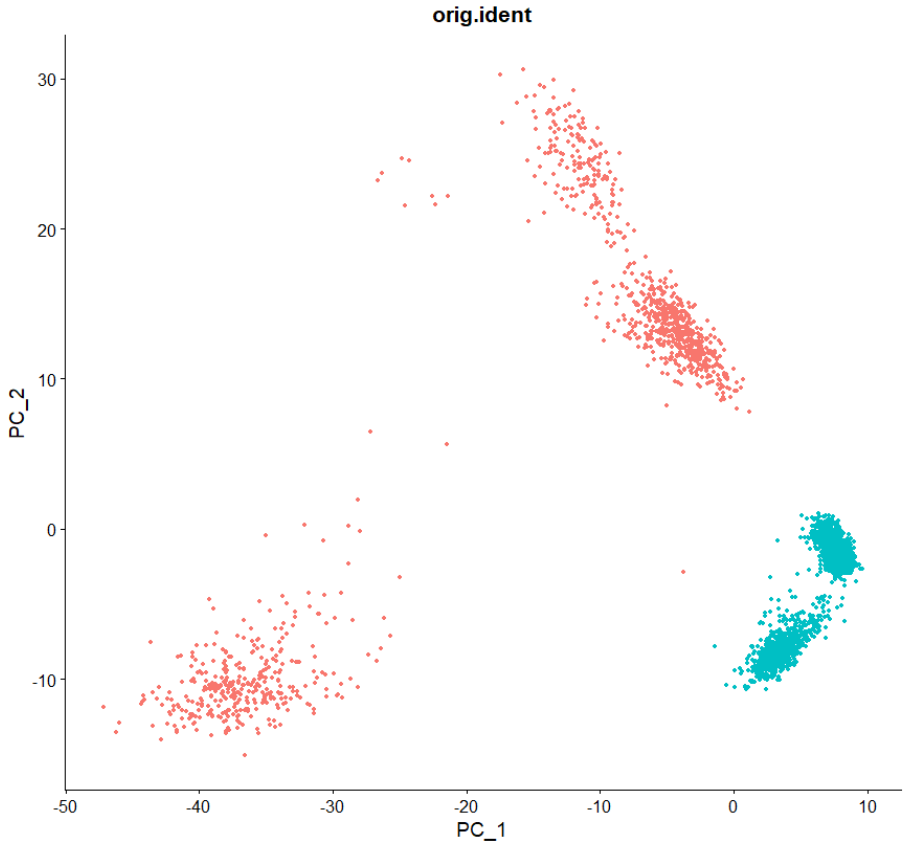


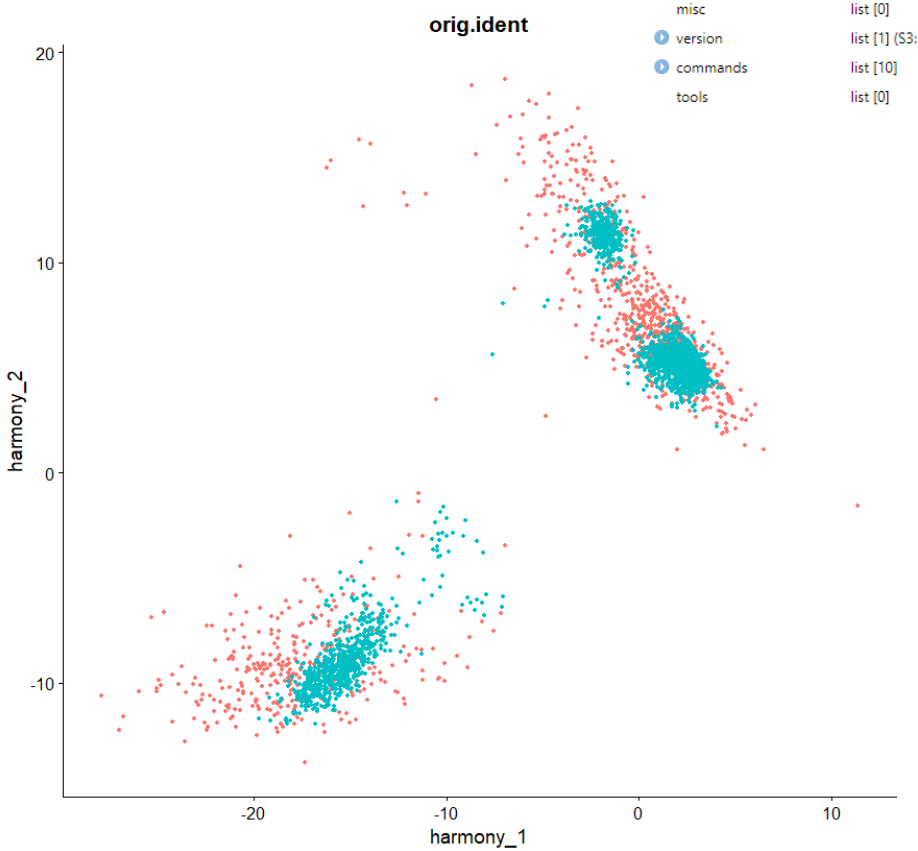
利用 Harmony 来整合数据

pbmc1k和pbmc3k数据示例

- 相比于RPCA，使用方法简单（一行代码），并且占用的内存少、运行速度快
- 在PCA之后运行，可以看做是另一种形式的降维

pbmc_obj	S4 [25556 x 3737] (SeuratObject:	S4 object of class Seurat
assays	list [4]	List of length 4
meta.data	list [3737 x 12] (S3: data.frame)	A data.frame with 3737 rows and 12 columns
active.assay	character [1]	'RNA'
active.ident	factor	Factor with 14 levels: "0", "1", "2", "3", "4", "5", ...
graphs	list [2]	List of length 2
neighbors	list [0]	List of length 0
reductions	list [4]	List of length 4
pca	S4 [3737 x 50] (SeuratObject::Din	S4 object of class DimReduc
harmony	S4 [3737 x 50] (SeuratObject::Din	S4 object of class DimReduc
umap	S4 [3737 x 2] (SeuratObject::Diml	S4 object of class DimReduc
tsne	S4 [3737 x 2] (SeuratObject::Diml	S4 object of class DimReduc
images	list [0]	List of length 0
project.name	character [1]	'pbmc_merge'
misc	list [0]	List of length 0
version	list [1] (S3: package_version, num	List of length 1
commands	list [10]	List of length 10
tools	list [0]	List of length 0

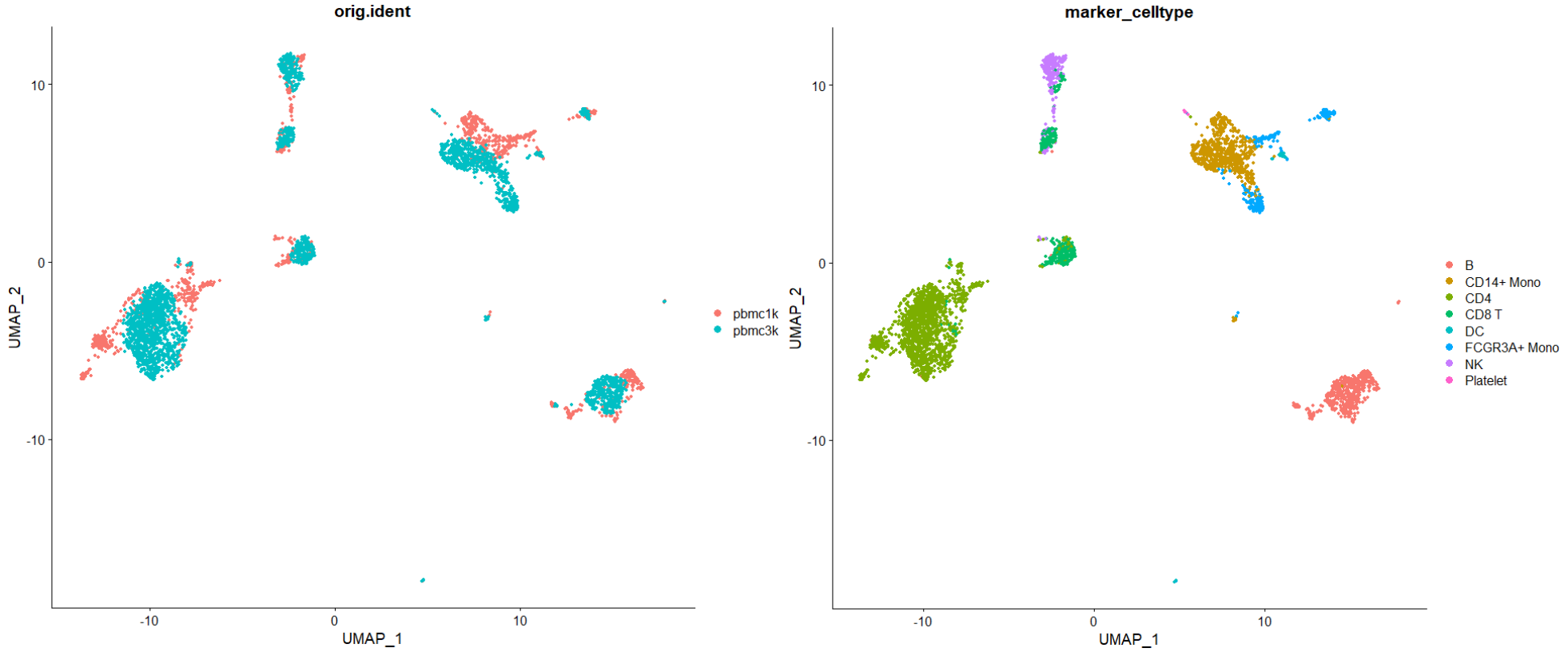




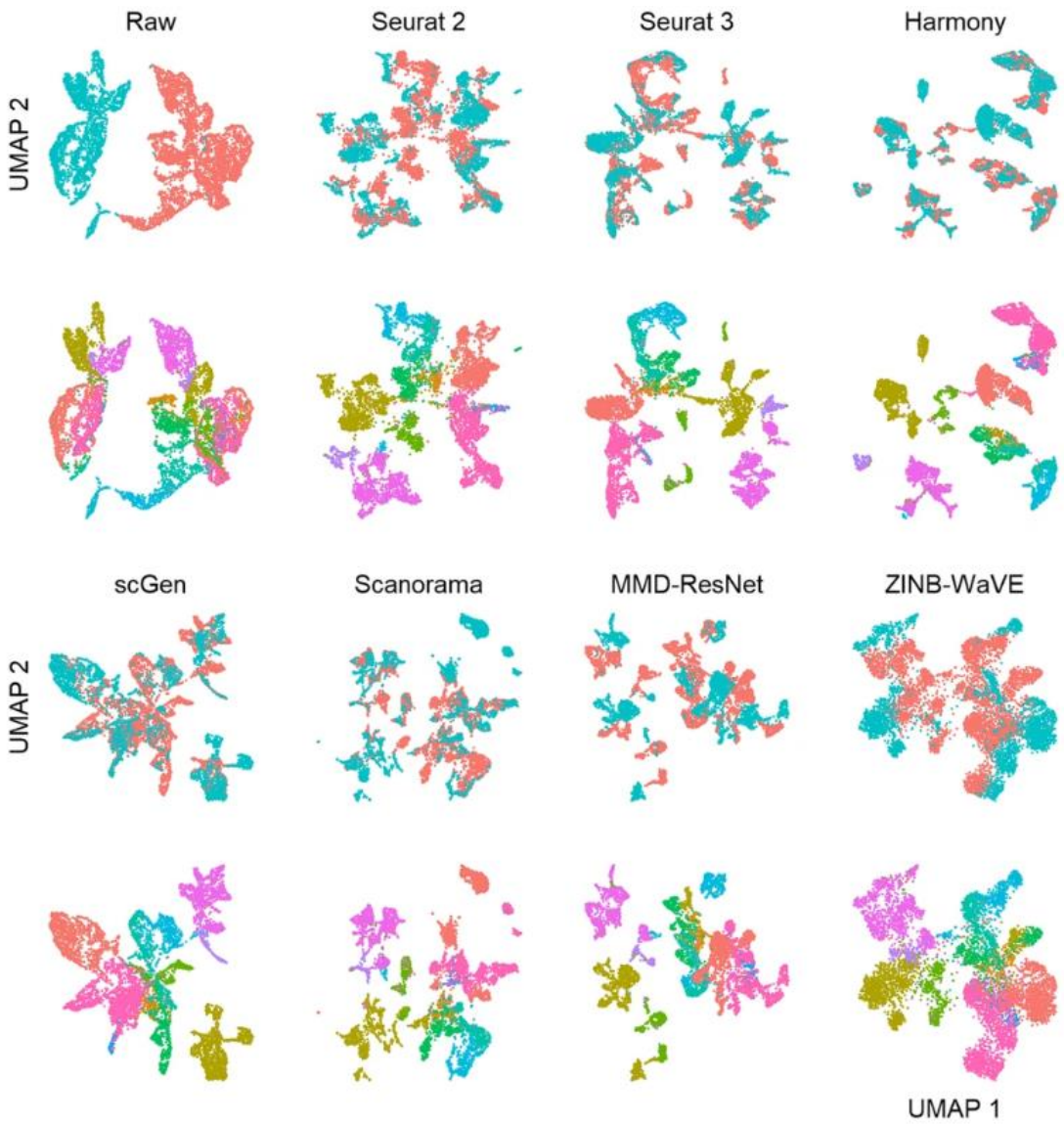
利用 Harmony 来整合数据

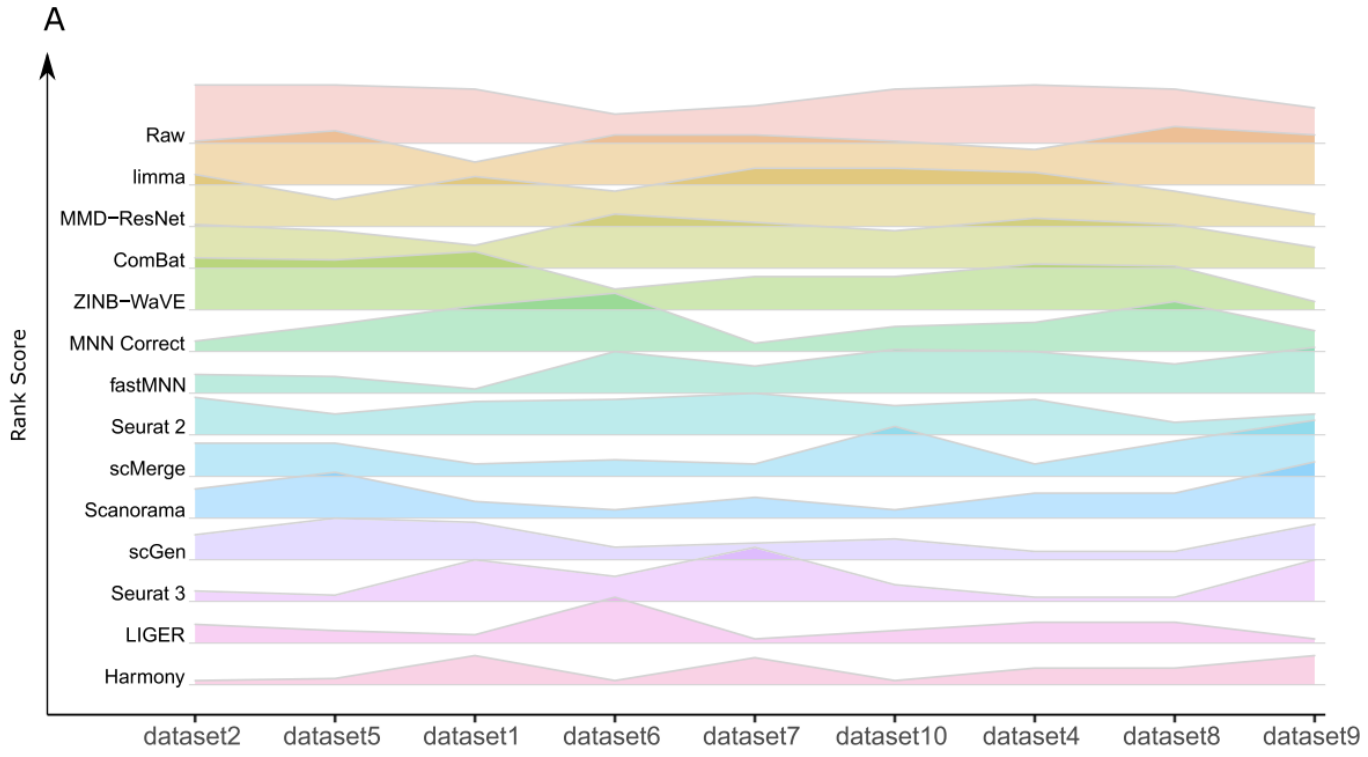
pbmc1k和pbmc3k数据示例（重要可调参数）

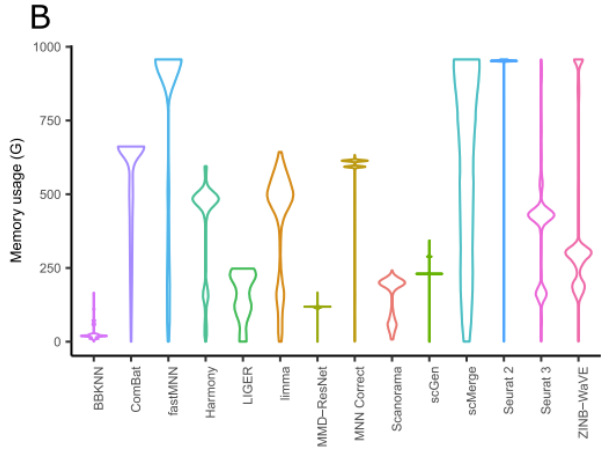
- group.by.vars：设置按哪个分组来整合
- max.iter.harmony：设置迭代次数，默认是10。运行RunHarmony结果会提示在迭代多少次后完成了收敛
- lambda：默认值是1，决定了Harmony整合的力度。lambda值调小，整合力度变大，反之。
- theta：Larger values of theta result in more diverse clusters
- dims.use参数：Which PCA dimensions to use for Harmony. By default, use all.

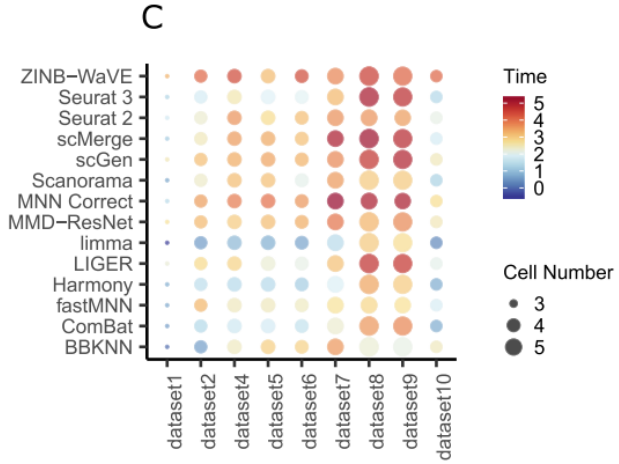


数据整合/批次矫正方法之间的差异





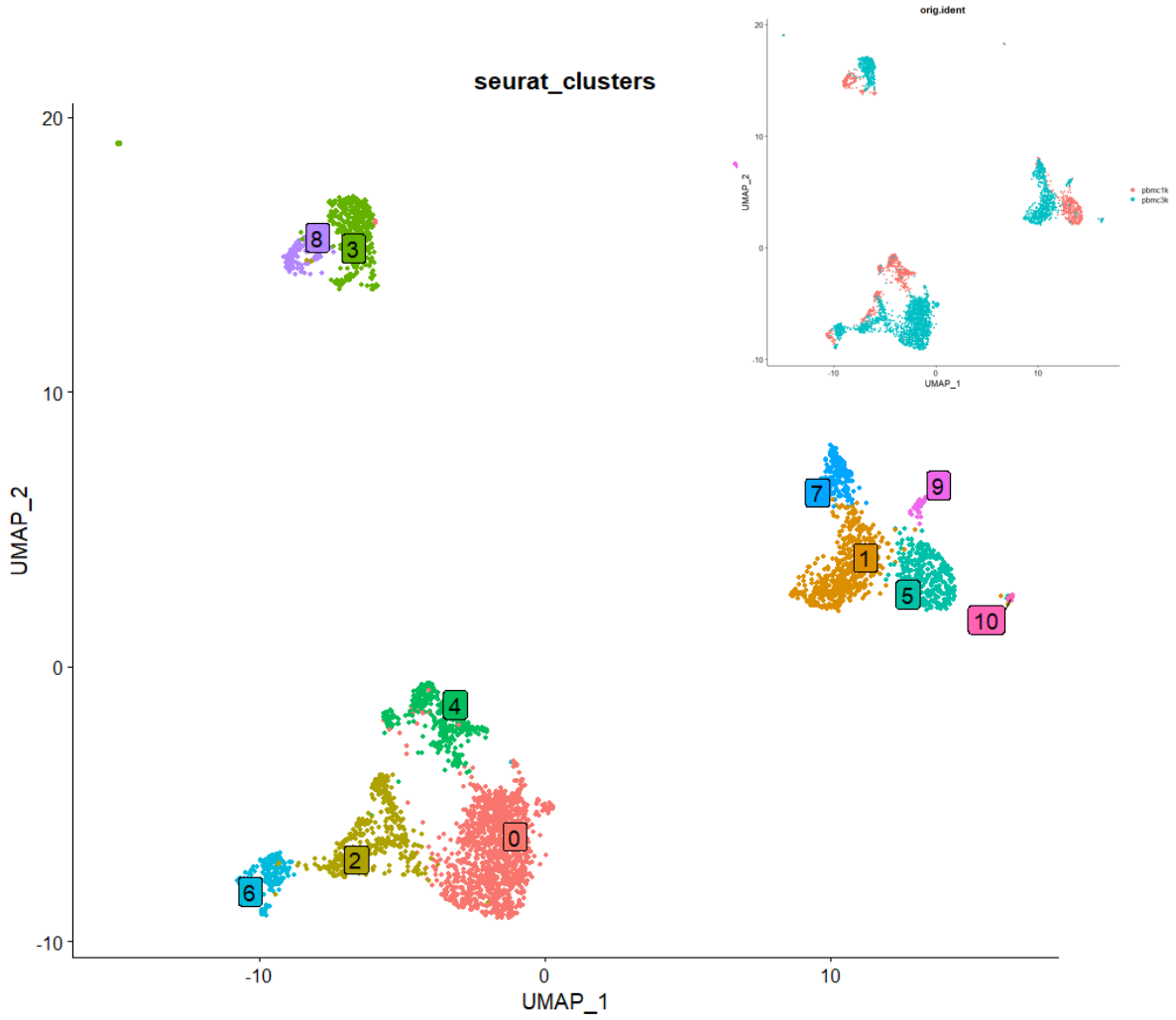


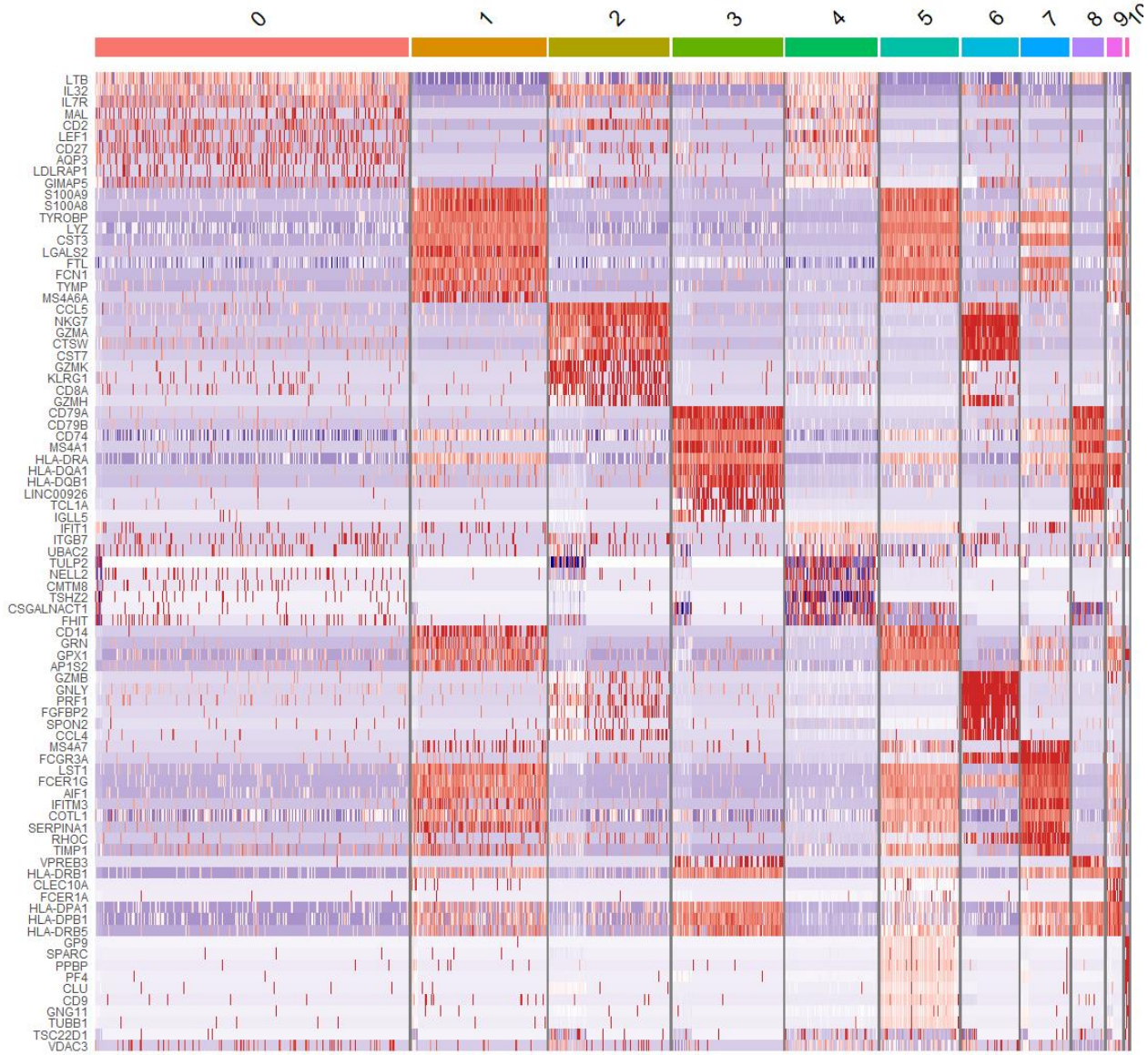


Tran et al., Genome Biology 2020

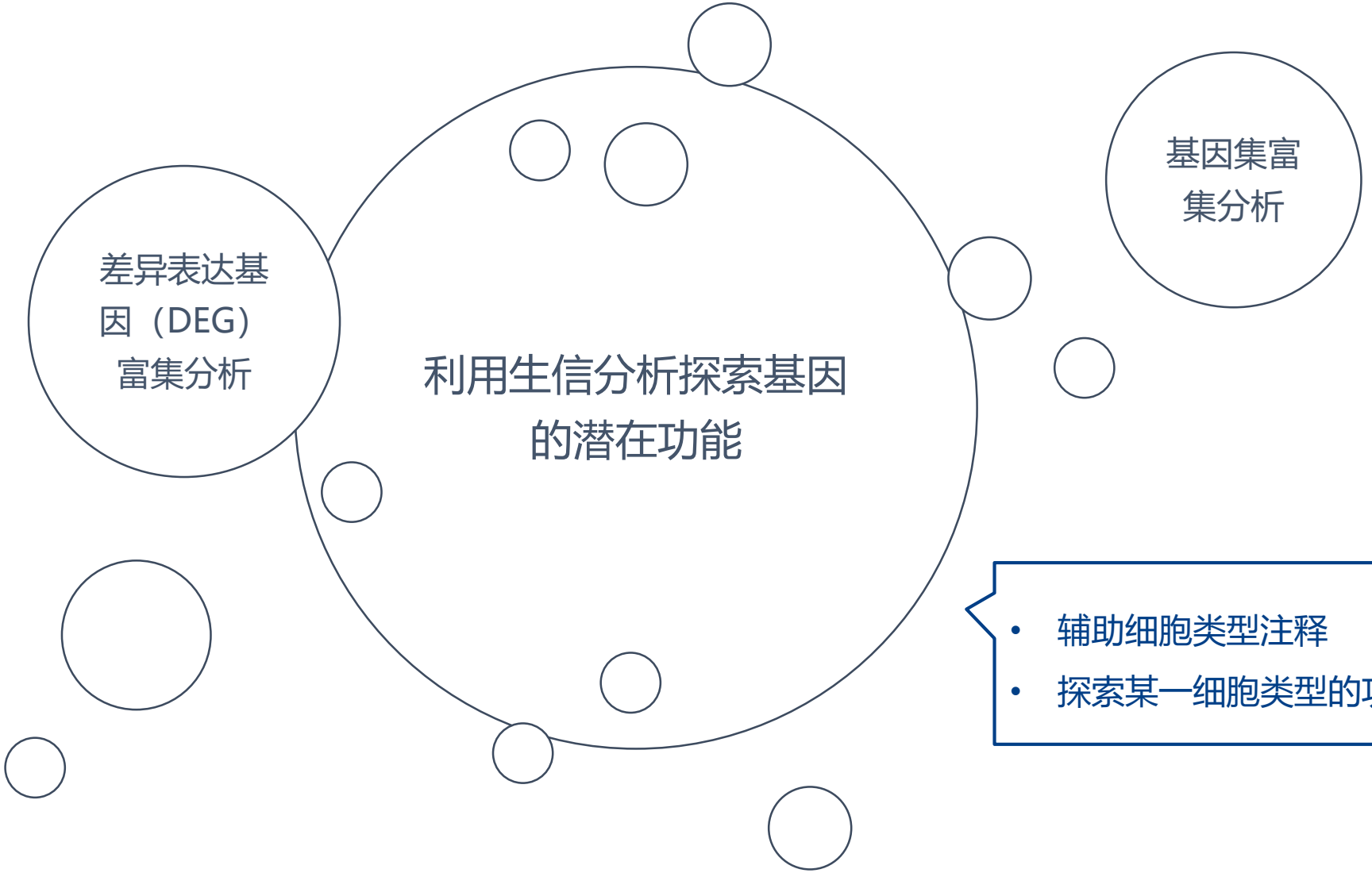
基因功能富集方法

pbmc1k和pbmc3k整合数据（使用RPCA默认参数的结果）





基因功能富集方法



- 辅助细胞类型注释
- 探索某一细胞类型的功能通路

基因注释数据库——GO

http://geneontology.org/

Current release 2023-10-09:

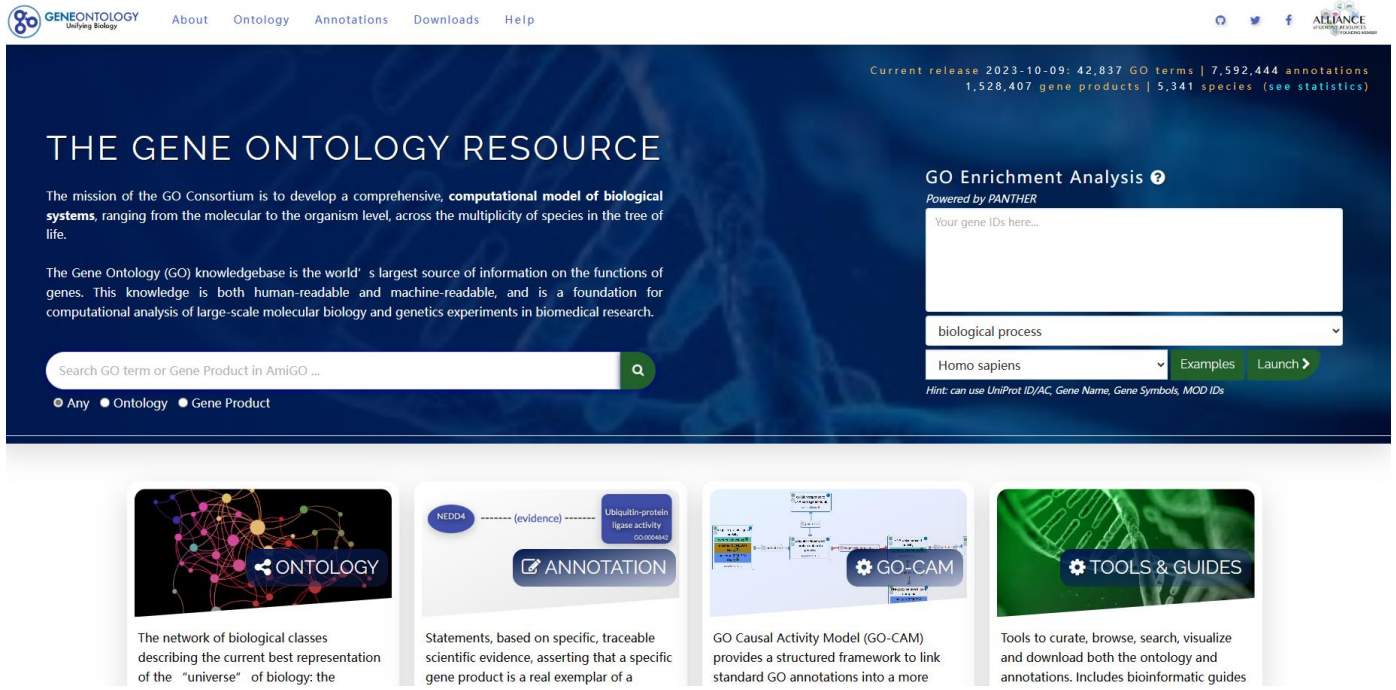
42,837 GO terms

7,592,444 annotations

1,528,407 gene products

5,341 species

GO（Gene Ontology）数据库由基因本体论联合会建立，该数据库将全世界所有与基因相关的研究结果进行分类汇总。对不同数据库中关于基因和基因产物的生物学术语进行标准化，对基因和蛋白功能进行统一的限定和描述。



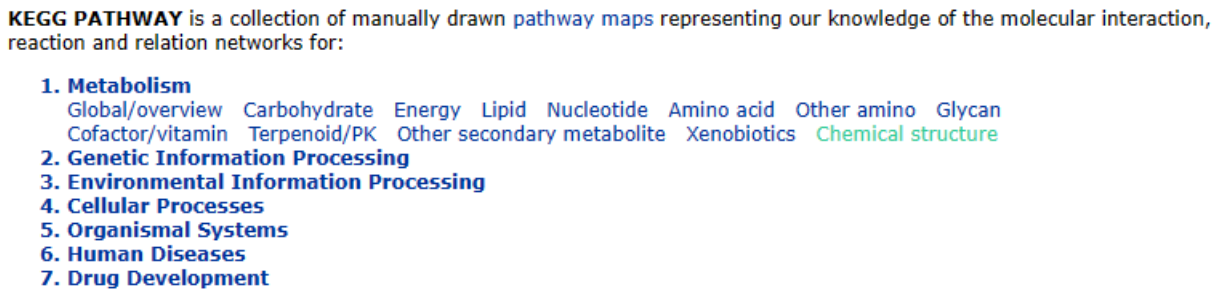
- 基因本体论定义了用来描述基因功能的概念/类，以及这些概念之间的关系。
- GO术语组织在一个有向无环图中，其中术语之间的边表示父子关系。
- 将功能分为三个方面：
  - BP（Biological Process，生物学过程）
  - CC（Cellular Component，细胞元件）
  - MF（Molecular Function，分子功能）

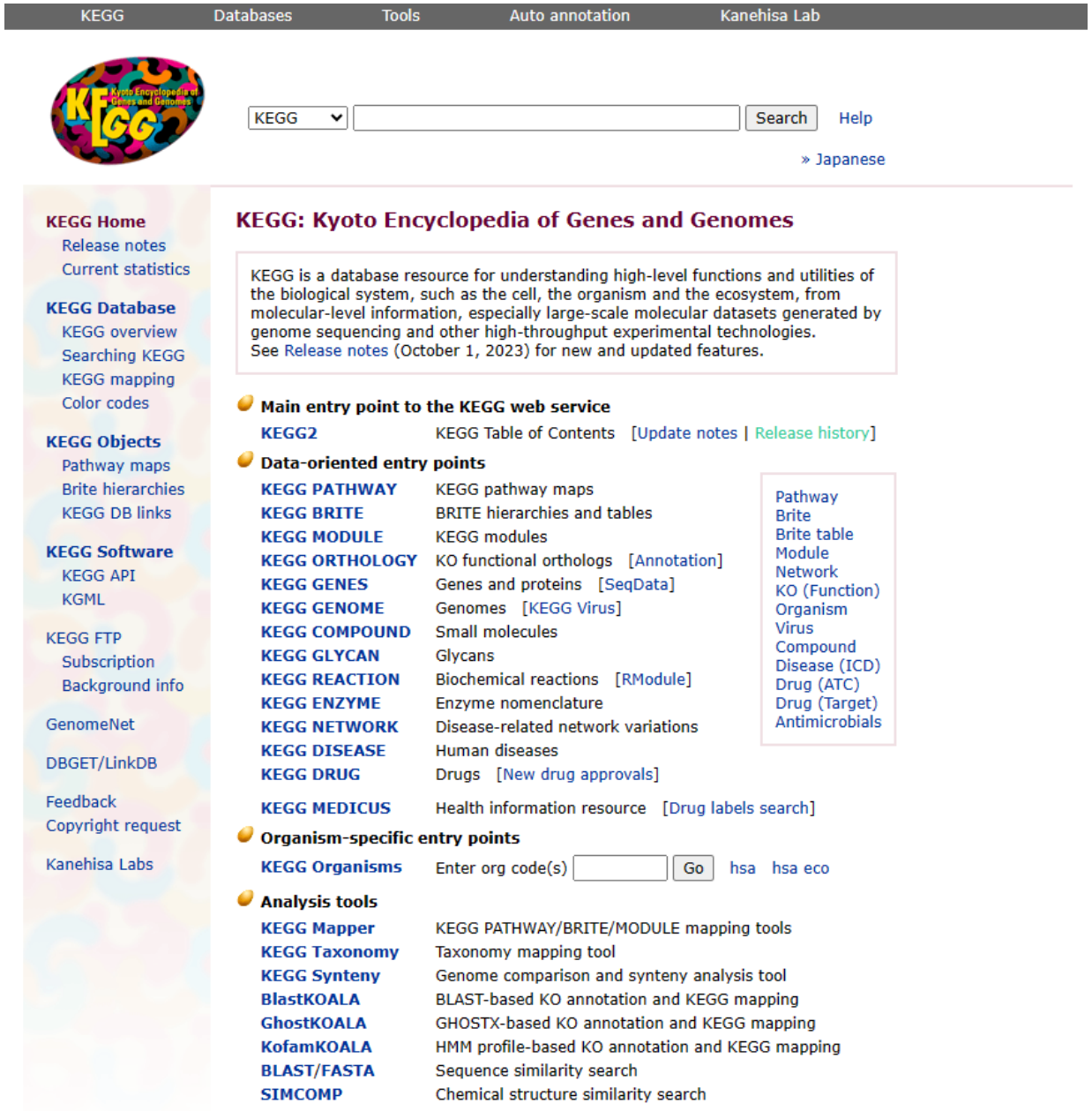
基因注释数据库——KEGG

http://www.genome.jp/kegg

- KEGG（Kyoto Encyclopedia of Genes and Genomes）是由日本京都大学和东京大学联合开发的数据库，是基因组测序和其他高通量实验技术生成的大规模分子数据集的整合和解读的参考知识库，可以用来查询代谢途径、酶（或编码酶的基因）、产物等，也可以通过BLAST比对查询未知序列的代谢途径信息。

- KEGG是一组人工绘制的代表分子相互作用和反应网络的通路图。
- 途径涵盖了广泛的生化过程，可分为7大类：新陈代谢、遗传和环境信息处理、细胞过程、机体系统、人类疾病和药物开发。

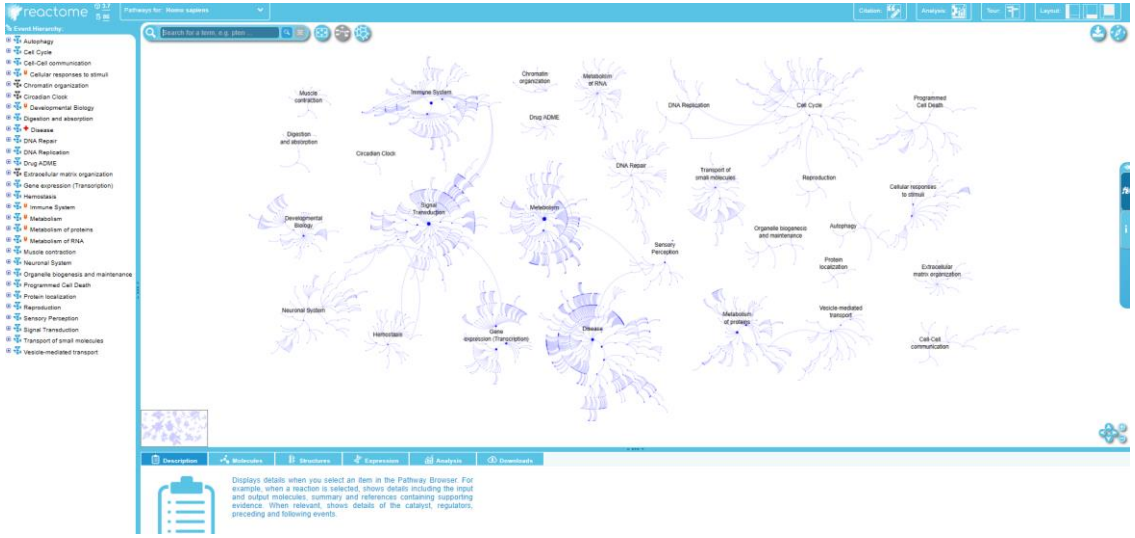


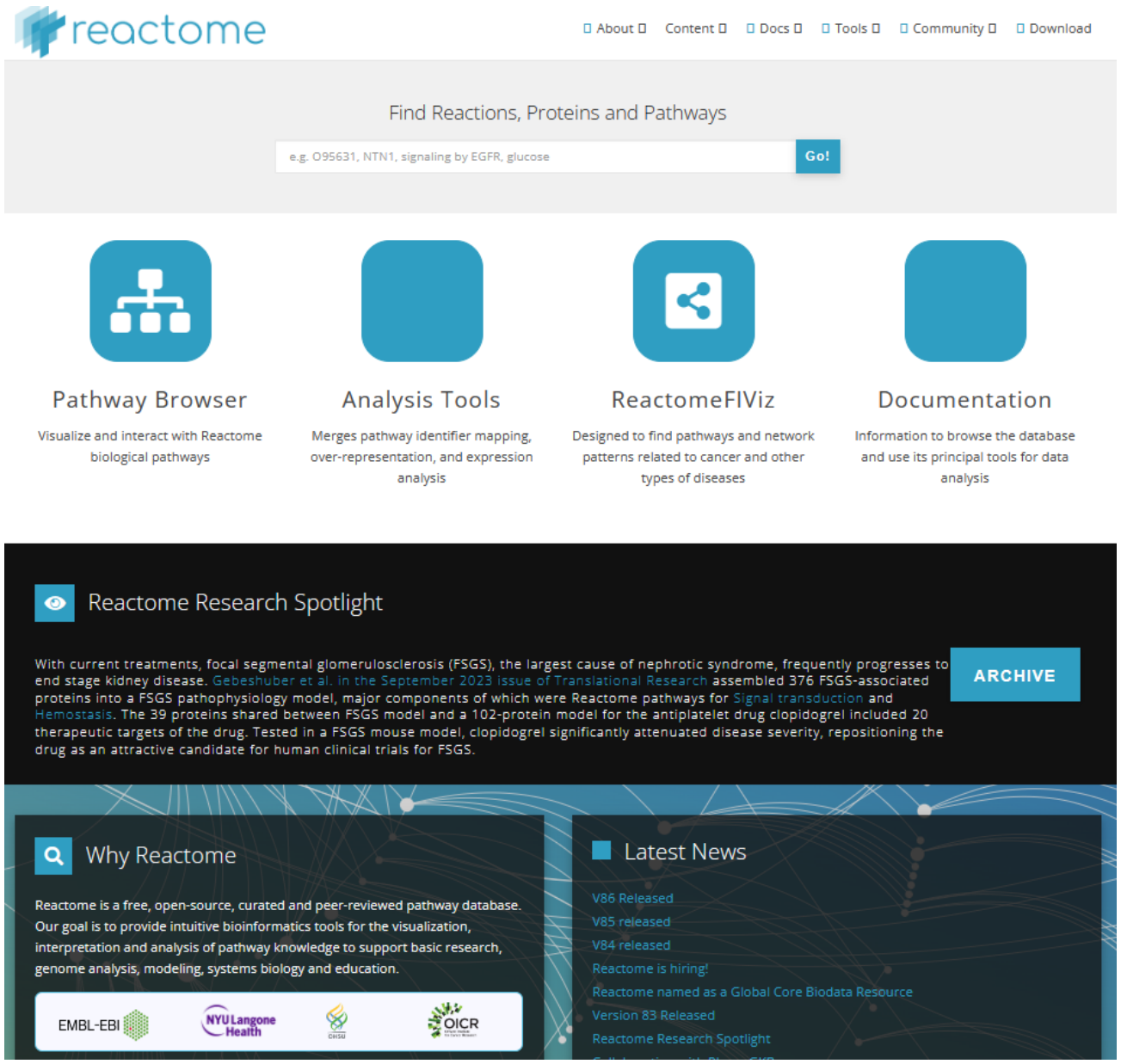


基因注释数据库——Reactome

https://reactome.org/

- Reactome数据库是一个免费开源的通路数据库，提供直观的生物信息学工具，用于可视化，解释和分析途径相关知识，以支持基础研究，基因组分析，建模，系统生物学研究等。
- 该数据库目前覆盖19个物种的通路研究，包括经典的代谢通路、信号转导、基因转录调控、细胞凋亡与疾病。数据库引用了100多个不同的生物信息学资源库，包括NCBI、Ensembl、UniProt、UCSC基因组浏览器、ChEBI小分子数据库和PubMed文献数据库等。





基因注释数据库——MsigDB

https://www.gsea-msigdb.org/gsea/msigdb/index.jsp

- MSigDB数据库是由Broad Institute研究所的科学家提出GSEA富集方法的同时提供的基因集数据库。
- 该数据库从位置、功能、代谢途径和靶标结合等多种角度出发，构建出许多的基因集合。目前包括H和C1-C8这九个系列的基因及，可供下载以及R软件包（msigdbr）载入，以用于富集分析。

- C7 (immunologic gene sets, 5219 gene sets)
  - IMMUNESIGDB (ImmuneSigDB gene sets, 4872 gene sets)
  - VAX (vaccine response gene sets, 347 gene sets)
- C8 (cell type signature gene sets, 830 gene sets)

Click on a gene set name to view its gene set page. Back to Top

GOLDRATH_EFF_VS_MEMORY_CD8_TCELL_DN	GSE21546_SAP1A_KO_VS_SAP1A_KO_AND_ELK1_KO_DP_THYMOCYTES_UP	GSE36476_YOUNG_VS_OLD_DONOR_MEMORY_CD4_TCELL_72H_TSST_ACT_UP
GOLDRATH_EFF_VS_MEMORY_CD8_TCELL_UP	GSE21546_UNSTIM_VS_ANTI_CD3_STIM_DP_THYMOCYTES_DN	GSE36476_YOUNG_VS_OLD_DONOR_MEMORY_CD4_TCELL_DN
GOLDRATH_NAIVE_VS_EFF_CD8_TCELL_DN	GSE21546_UNSTIM_VS_ANTI_CD3_STIM_DP_THYMOCYTES_UP	GSE36476_YOUNG_VS_OLD_DONOR_MEMORY_CD4_TCELL_UP
GOLDRATH_NAIVE_VS_EFF_CD8_TCELL_UP	GSE21546_UNSTIM_VS_ANTI_CD3_STIM_EL K1_KO_DP_THYMOCYTES_DN	GSE36527_CD62L_HIGH_CD69_NEG_VS_CD62L_LOW_CD69_POS_TREG_KLRG1_NEG_DN
GOLDRATH_NAIVE_VS_MEMORY_CD8_TCELL_DN	GSE21546_UNSTIM_VS_ANTI_CD3_STIM_EL K1_KO_DP_THYMOCYTES_UP	GSE36527_CD62L_HIGH_CD69_NEG_VS_CD62L_LOW_CD69_POS_TREG_KLRG1_NEG_UP
GOLDRATH_NAIVE_VS_MEMORY_CD8_TCELL_UP	GSE21546_UNSTIM_VS_ANTI_CD3_STIM_SA P1A_KO_AND_ELK1_KO_DP_THYMOCYTES_DN	GSE36527_CD62L_HIGH_VS_CD62L_LOW_TREG_CD69_NEG_KLRG1_NEG_DN
GSE10094_LCMV_VS_LISTERIA_IND_EFF_CD4_TCELL_DN	GSE21546_UNSTIM_VS_ANTI_CD3_STIM_SA P1A_KO_AND_ELK1_KO_DP_THYMOCYTES_UP	GSE36527_CD62L_HIGH_VS_CD62L_LOW_TREG_CD69_NEG_KLRG1_NEG_UP
GSE10094_LCMV_VS_LISTERIA_IND_EFF_CD4_TCELL_UP	GSE21546_UNSTIM_VS_ANTI_CD3_STIM_SA P1A_KO_DP_THYMOCYTES_DN	GSE36527_CD69_NEG_VS_POS_TREG_CD62L_LOS_KLRG1_NEG_DN
GSE10147_IL3_AND_HIVP17_VS_IL3_AND_CPG_STIM_PDC_DN	GSE21546_UNSTIM_VS_ANTI_CD3_STIM_SA P1A_KO_DP_THYMOCYTES_UP	GSE36527_CD69_NEG_VS_POS_TREG_CD62L_LOS_KLRG1_NEG_UP
GSE10147_IL3_AND_HIVP17_VS_IL3_AND_CPG_STIM_PDC_UP	GSE21546_WT_VS_ELK1_KO_ANTI_CD3_STIM_DP_THYMOCYTES_DN	GSE36826_NORMAL_VS_STAPH_AUREUS_INF_IL1R_KO_SKIN_DN
GSE10147_IL3_VS_IL3_AND_CPG_STIM_PDC_DN	GSE21546_WT_VS_ELK1_KO_ANTI_CD3_STIM_DP_THYMOCYTES_UP	GSE36826_NORMAL_VS_STAPH_AUREUS_INF_IL1R_KO_SKIN_UP
GSE10147_IL3_VS_IL3_AND_CPG_STIM_PDC_UP		GSE36826_NORMAL_VS_STAPH_AUREUS_INF_SKIN_DN
GSE10147_IL3_VS_IL3_AND_HIVP17_STIM_PDC_DN		
GSE10147_IL3_VS_IL3_AND_HIVP17_STIM_PDC_UP		

MSigDB Molecular Signatures Database

Human Collections

H **hallmark gene sets** are coherently expressed signatures derived by aggregating many MSigDB gene sets to represent well-defined biological states or processes.

C1 **positional gene sets** corresponding to human chromosome cytogenetic bands.

C2 **curated gene sets** from online pathway databases, publications in PubMed, and knowledge of domain experts.

C3 **regulatory target gene sets** based on gene target predictions for microRNA seed sequences and predicted transcription factor binding sites.

C4 **computational gene sets** defined by mining large collections of cancer-oriented microarray data.

C5 **ontology gene sets** consist of genes annotated by the same ontology term.

C6 **oncogenic signature gene sets** defined directly from microarray gene expression data from cancer gene perturbations.

C7 **immunologic signature gene sets** represent cell states and perturbations within the immune system.

C8 **cell type signature gene sets** curated from cluster markers identified in single-cell sequencing studies of human tissue.

Mouse Collections

MH **mouse-ortholog hallmark gene sets** are versions of gene sets in the MSigDB Hallmarks collection mapped to their mouse orthologs.

M1 **positional gene sets** corresponding to mouse chromosome cytogenetic bands.

M2 **curated gene sets** from online pathway databases, publications in PubMed, and knowledge of domain experts.

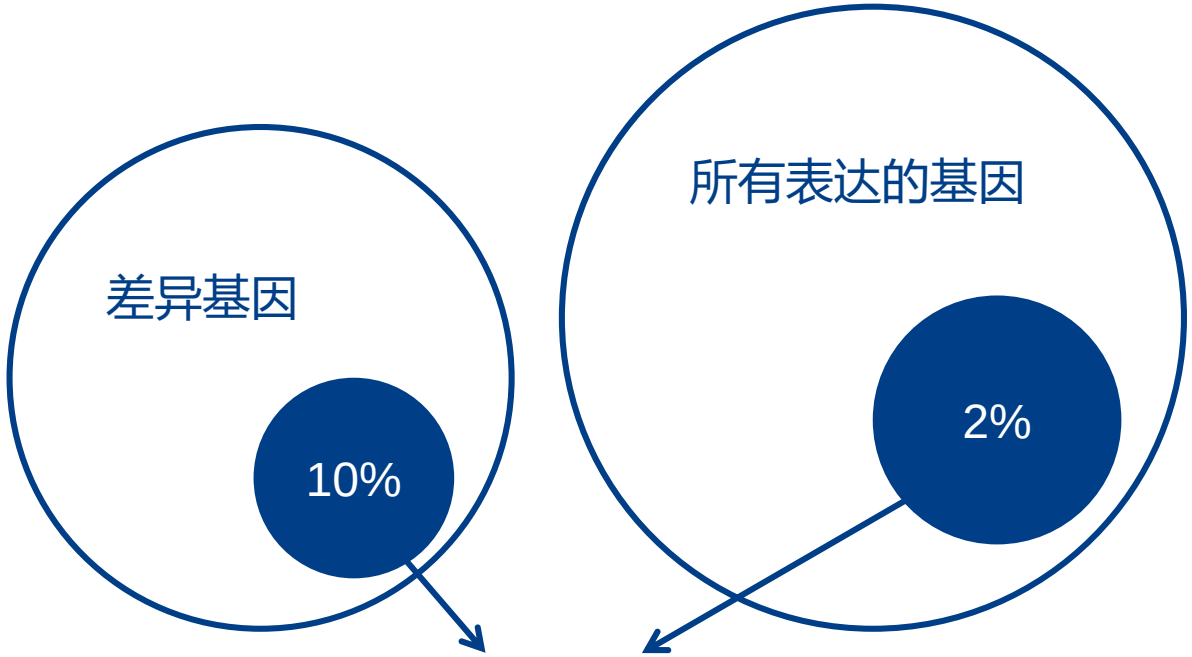
M3 **regulatory target gene sets** based on gene target predictions for microRNA seed sequences and predicted transcription factor binding sites.

M5 **ontology gene sets** consist of genes annotated by the same ontology term.

M8 **cell type signature gene sets** curated from cluster markers identified in single-cell sequencing studies of mouse tissue.

差异表达基因富集（GO分析）

GO富集分析的简单原理：前景基因和背景基因



$$P = 1 - \sum_{i=0}^{m-1} \frac{\binom{M}{i}\binom{N-M}{n-i}}{\binom{N}{n}}$$

显著性的计算：利用超几何检验

- N为所有Unigene中具有GO注释的基因数目；
- n为N中差异表达基因的数目；
- M为所有Unigene中注释为某特定GO term的基因数目；
- m为注释为某特定GO term的差异表达基因数目。

问题：计算10%与2%相比是否有显著差异?

差异表达基因富集（GO分析）

GO富集分析的实现：在线网站工具和本地R包（cluster3的差异基因，在线工具）

CC

MF

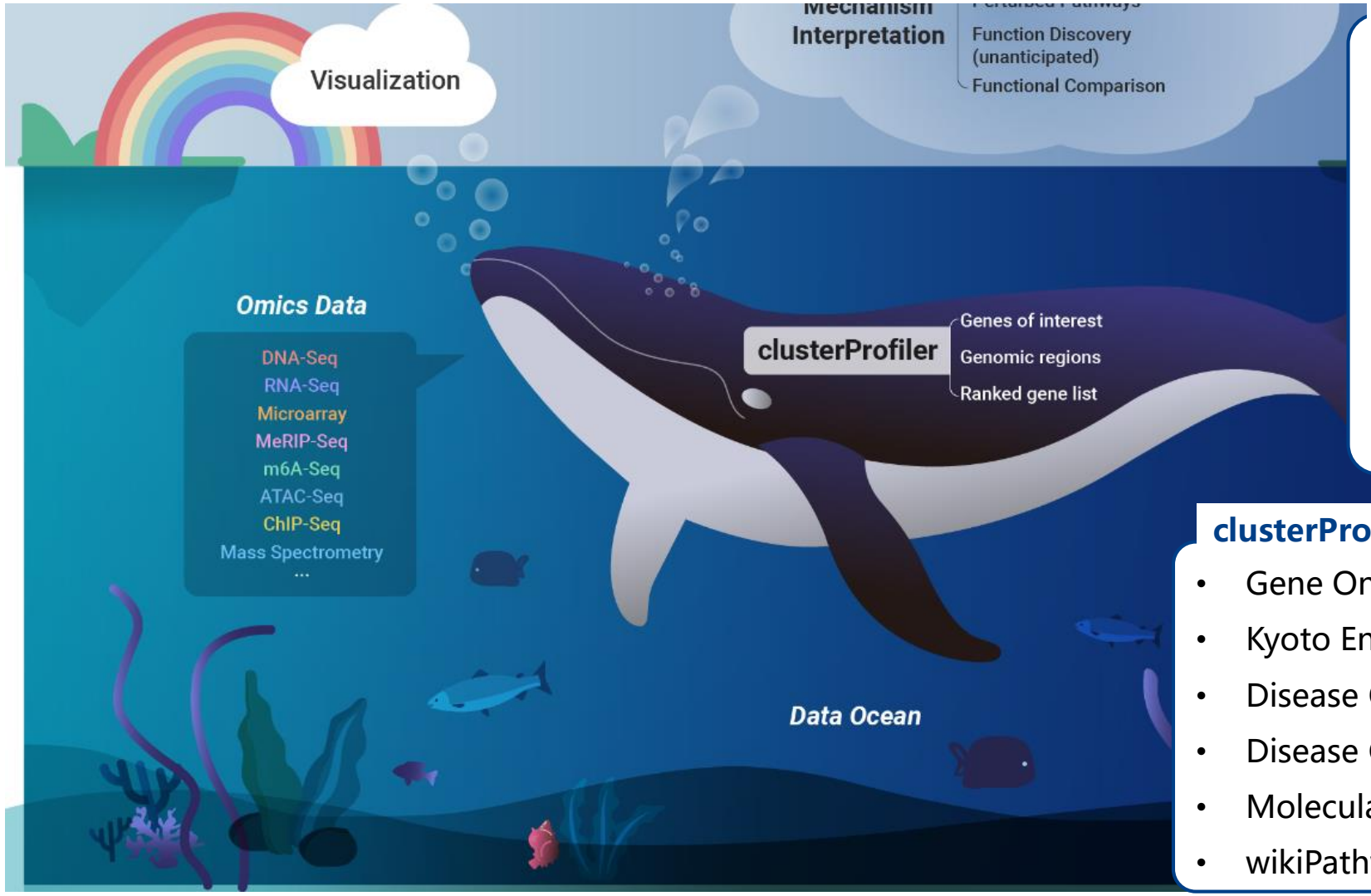
BP

GO biological process complete
- B cell receptor signaling pathway
  - antigen receptor-mediated signaling pathway
    - immune response-activating cell surface receptor signaling pathway
      - immune response-regulating cell surface receptor signaling pathway
        - immune response-regulating signaling pathway
          - response to stimulus
        - cell surface receptor signaling pathway
      - immune response-activating signaling pathway
        - activation of immune response

GO cellular component complete	Homo sapiens (REF) #	upload_1 (Hierarchy NEW!) #	expected	Fold Enrichment	+/-	raw P value	FDR
MHC class II protein complex	17	10	.02	> 100	+	2.03E-25	4.03E-22
↳MHC protein complex	25	10	.02	> 100	+	4.41E-24	4.37E-21
↳plasma membrane protein complex	729	13	.67	19.33	+	3.42E-15	5.21E-13
↳membrane protein complex	1351	13	1.25	10.43	+	8.24E-12	6.80E-10
↳protein-containing complex	6363	14	5.87	2.38	+	1.56E-04	5.94E-03
↳plasma membrane	5866	16	5.41	2.96	+	7.26E-07	3.42E-05
↳cell periphery	6350	16	5.86	2.73	+	2.35E-06	1.08E-04
B cell receptor complex	6	2	.01	> 100	+	2.25E-05	9.08E-04
lumenal side of endoplasmic reticulum membrane	29	9	.03	> 100	+	8.05E-21	5.32E-18
↳lumenal side of membrane	37	9	.03	> 100	+	5.42E-20	2.69E-17
↳organelle membrane	3720	10	3.43	2.91	+	6.95E-04	2.60E-02
↳side of membrane	612	13	.56	23.02	+	3.72E-16	7.37E-14
↳endoplasmic reticulum membrane	1182	9	1.09	8.25	+	3.80E-07	1.98E-05
↳endoplasmic reticulum subcompartment	1188	9	1.10	8.21	+	3.97E-07	2.01E-05
↳organelle subcompartment	1510	9	1.39	6.46	+	2.94E-06	1.26E-04
↳endoplasmic reticulum	2070	11	1.91	5.76	+	3.79E-07	2.03E-05
↳endomembrane system	4778	13	4.41	2.95	+	3.64E-05	1.41E-03
↳nuclear outer membrane-endoplasmic reticulum membrane network	1205	9	1.11	8.09	+	4.47E-07	2.21E-05
ER to Golgi transport vesicle membrane	63	9	.06	> 100	+	4.14E-18	1.64E-15
↳transport vesicle membrane	236	9	.22	41.33	+	3.41E-13	3.75E-11
↳transport vesicle	433	9	.40	22.53	+	6.76E-11	4.96E-09
↳cytoplasmic vesicle	2513	11	2.32	4.74	+	2.67E-06	1.20E-04
↳intracellular vesicle	2517	11	2.32	4.74	+	2.72E-06	1.20E-04
↳vesicle	3995	14	3.69	3.80	+	4.68E-07	2.26E-05
↳bounding membrane of organelle	2159	10	1.99	5.02	+	6.15E-06	2.54E-04
↳cytoplasmic vesicle membrane	1220	10	1.13	8.88	+	3.11E-08	1.81E-06
↳vesicle membrane	1238	10	1.14	8.75	+	3.57E-08	2.02E-06
↳coated vesicle membrane	202	9	.19	48.29	+	8.80E-14	1.09E-11
↳coated vesicle	315	9	.29	30.97	+	4.22E-12	3.64E-10
↳COPII-coated ER to Golgi transport vesicle	93	9	.09	> 100	+	1.10E-16	2.72E-14
clathrin-coated endocytic vesicle membrane	73	9	.07	> 100	+	1.42E-17	4.69E-15
↳clathrin-coated vesicle membrane	136	9	.13	71.72	+	2.86E-15	4.72E-13
↳clathrin-coated vesicle	218	9	.20	44.74	+	1.71E-13	1.99E-11
↳clathrin-coated endocytic vesicle	92	9	.08	> 100	+	1.00E-16	2.84E-14
↳endocytic vesicle	348	9	.32	28.03	+	1.01E-11	7.98E-10
↳endocytic vesicle membrane	197	9	.18	49.51	+	7.08E-14	9.34E-12

差异表达基因富集（GO分析）

GO富集分析的实现：R包clusterProfiler（cluster3的差异基因）



GO分析R语言代码

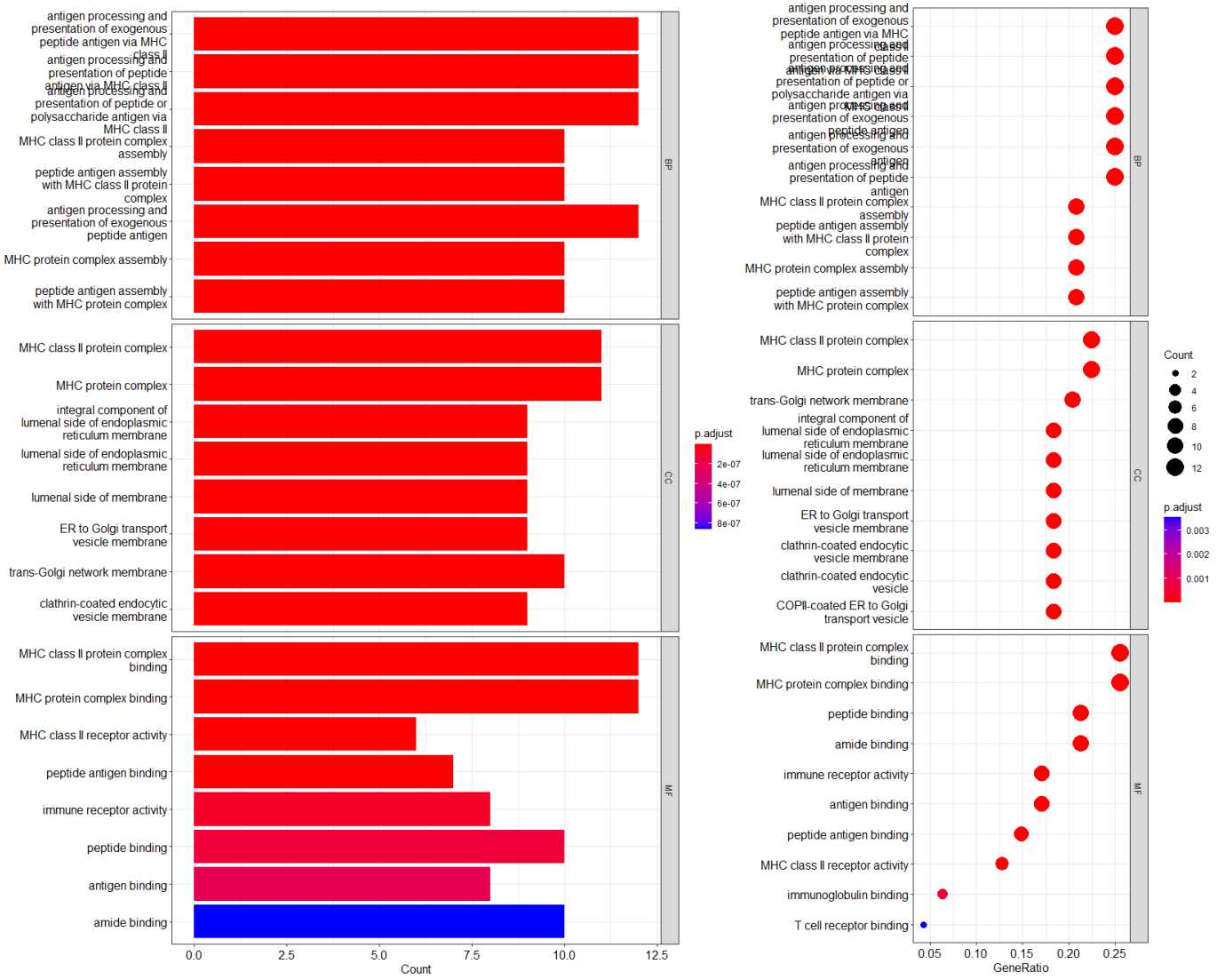
```
library(clusterProfiler)
GO <- enrichGO(gene$ENTREZID,
               OrgDb = GO_database,
               keyType = "ENTREZID
               ont = "ALL",
               pvalueCutoff = 0.05,
               qvalueCutoff = 0.05,
               readable = T)
```

clusterProfiler支持的基因集或基因通路数据库

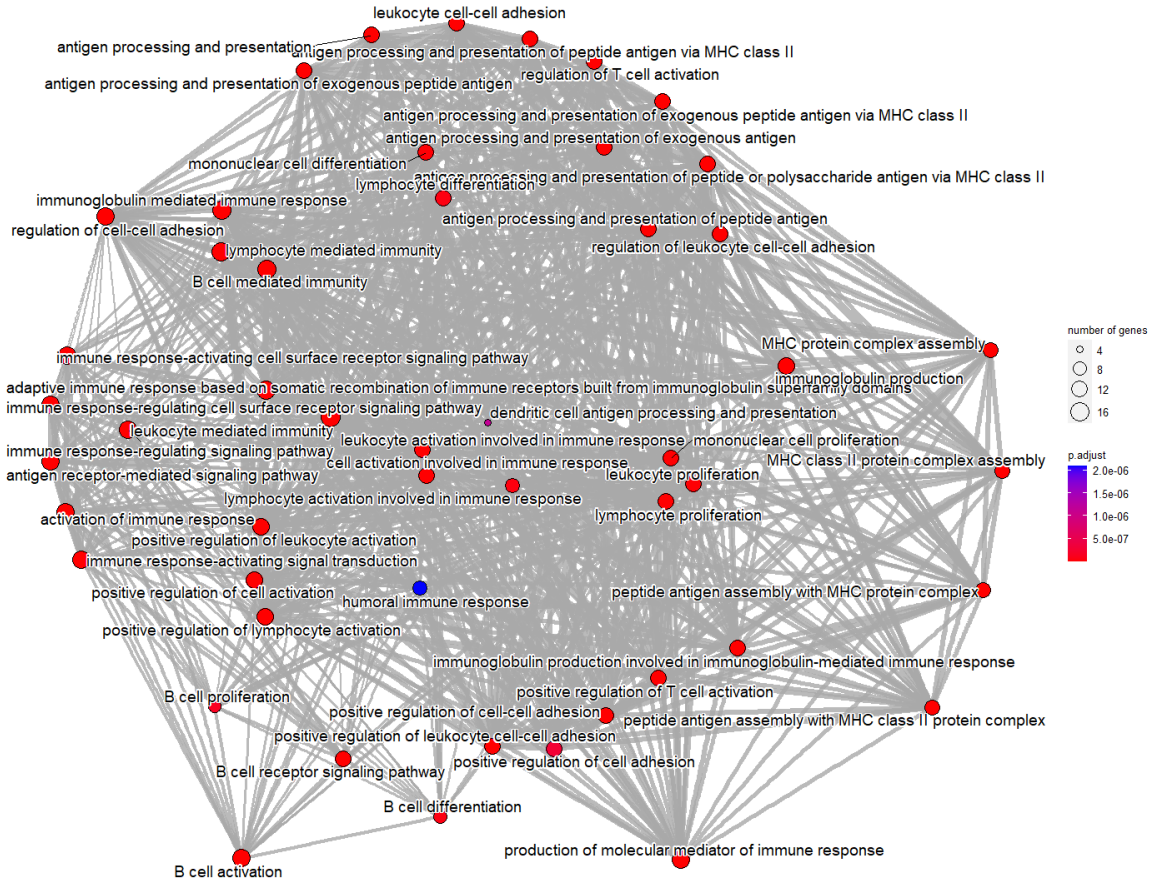
- Gene Ontology(GO)
- Kyoto Encyclopedia of Genes and Genomes(KEGG)
- Disease Ontology(DO)
- Disease Gene Network (DisGeNET)
- Molecular Signatures Database (MSigDb)
- wikiPathways

差异表达基因富集（GO分析）

GO富集分析的结果解释（cluster3的差异基因的GO富集分析结果）



通路间关联网络图

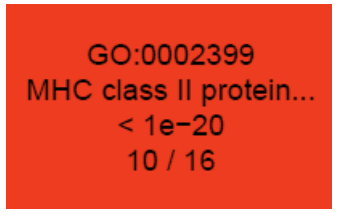


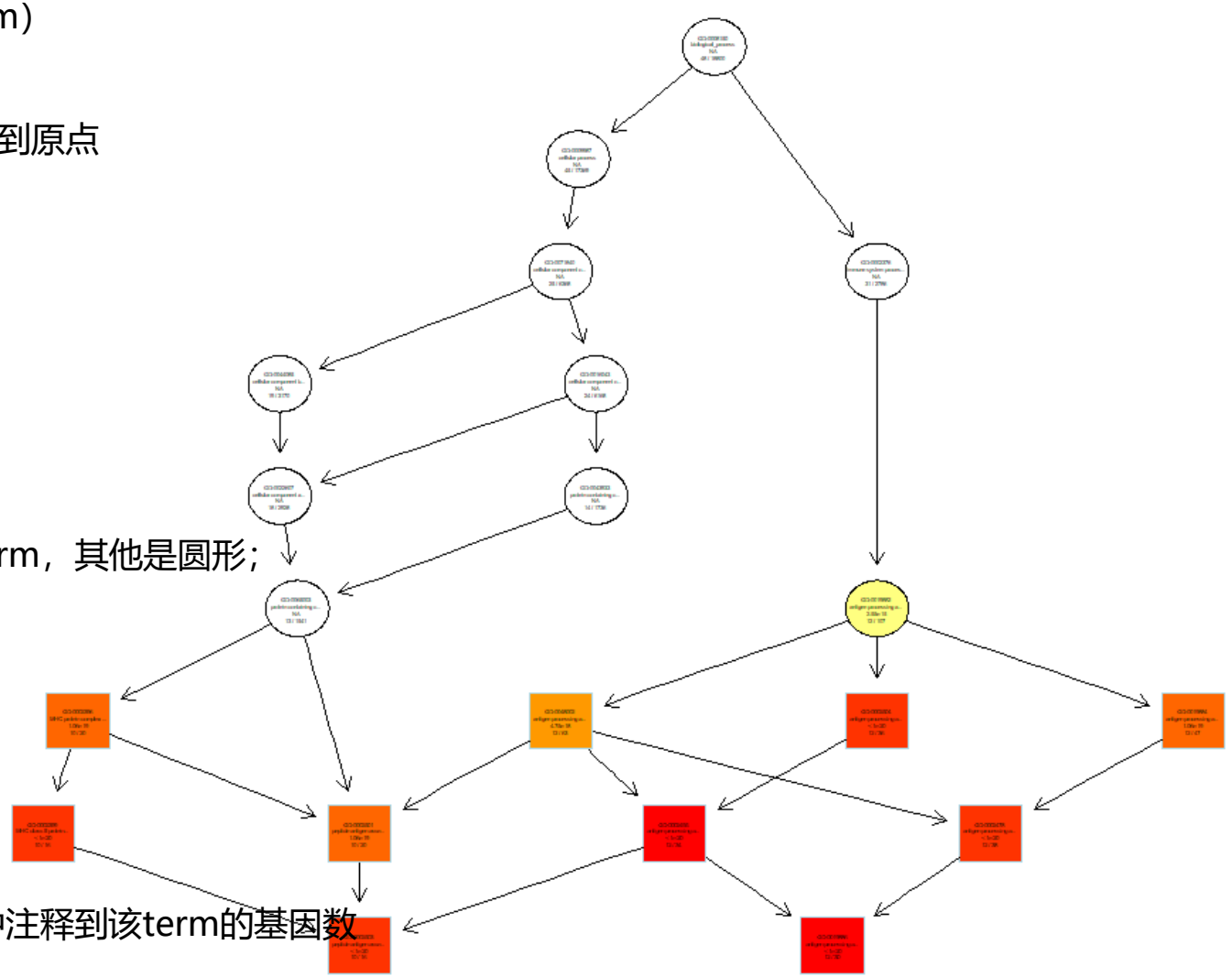
差异表达基因富集（GO分析）

GO富集分析的结果解释（cluster3的差异基因的GO富集分析结果）：有向无环图（Directed Acyclic Graphs，DAG）

注释系统中每一个节点就代表了一个基本描述单元（term）

- 有向指的是term之间的单向指向性关系
- 无环指的是从任何一点开始沿着规定的指向都不能回到原点



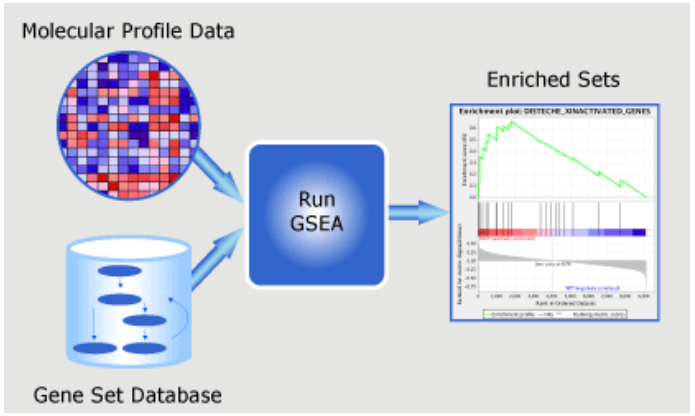


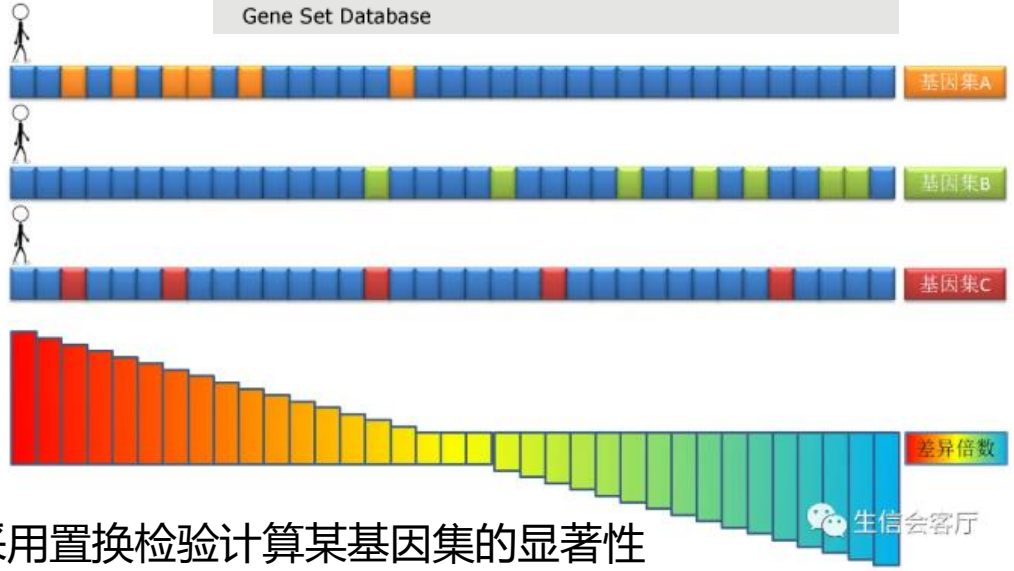
- 形状：方形是默认输出的显著性最高的前10个GO term，其他是圆形；
- 颜色：颜色越深，代表p值越小，富集越显著；
- 文字：图形里面的文字从上而下分别代表
  - GO term编号
  - GO term的文字描述
  - P值
  - 前景基因中注释到该term的基因数/背景基因中注释到该term的基因数

基因集富集（GSEA）

GSEA的原理（Gene Set Enrichment Analysis）：

- 先将基因在两种样品中的差异表达程度（如logFC）或者表型相关度进行排序（并不是计算差异基因）
- 然后判断来自功能注释等预定义的基因集或自定义的基因集是否倾向于落在有序列表的顶部或底部





采用置换检验计算某基因集的显著性

Enrichment Score (ES) Calculation

Start with ranked list (L) of genes that are in (Hit) or not in (Miss) a gene set (S), using fold change (FC) as example metric

Ranked List (L)	FC		Contribution to running sum for ES	Hits +|FC| / Σ	Misses -1/(N-N_H)	Running sum for ES
	15	Hit	+0.15	+0.15		0.15
	12	Hit	+0.12	+0.12		0.27
	10	Miss	-0.001		-0.001	0.269
	9	Hit	+0.09	+0.09		0.359
	8	Hit	+0.08	+0.08		0.439
	6	Miss	-0.001		-0.001	0.438
...	...	...	...	...	...	...

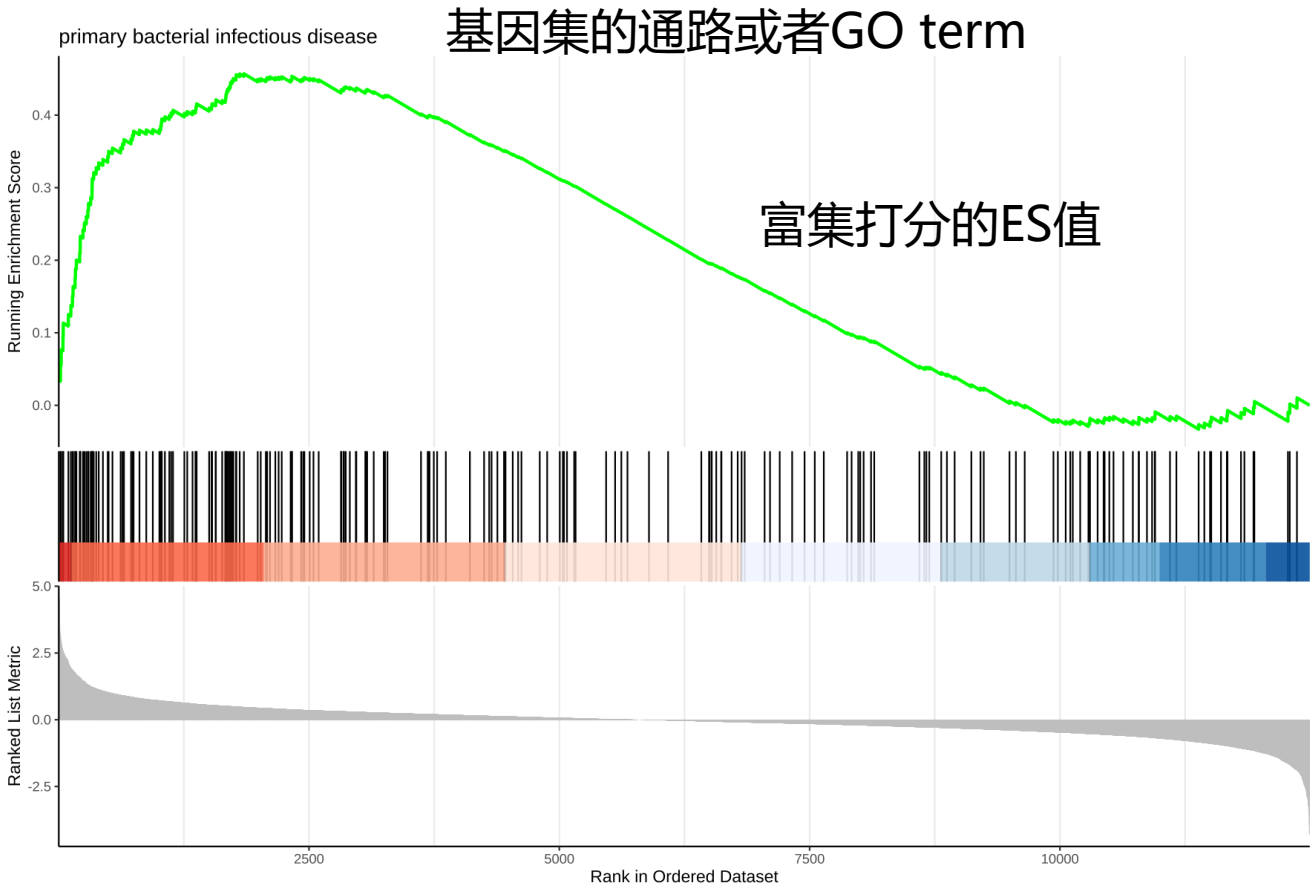
Hits: Genes ∈ S +|FC| / Σ
Misses: Genes ∉ S -1/(N-N_H)

Σ = sum of fold changes for genes in gene set (S) (e.g., 100)
N = no. of genes in the array (e.g., 1020)
N_H = no. of genes in the gene set (S) (e.g., 20)

ES(S) = value of maximum deviation from 0 of the running sum

基因集富集（GSEA）

GSEA的分析结果解释：



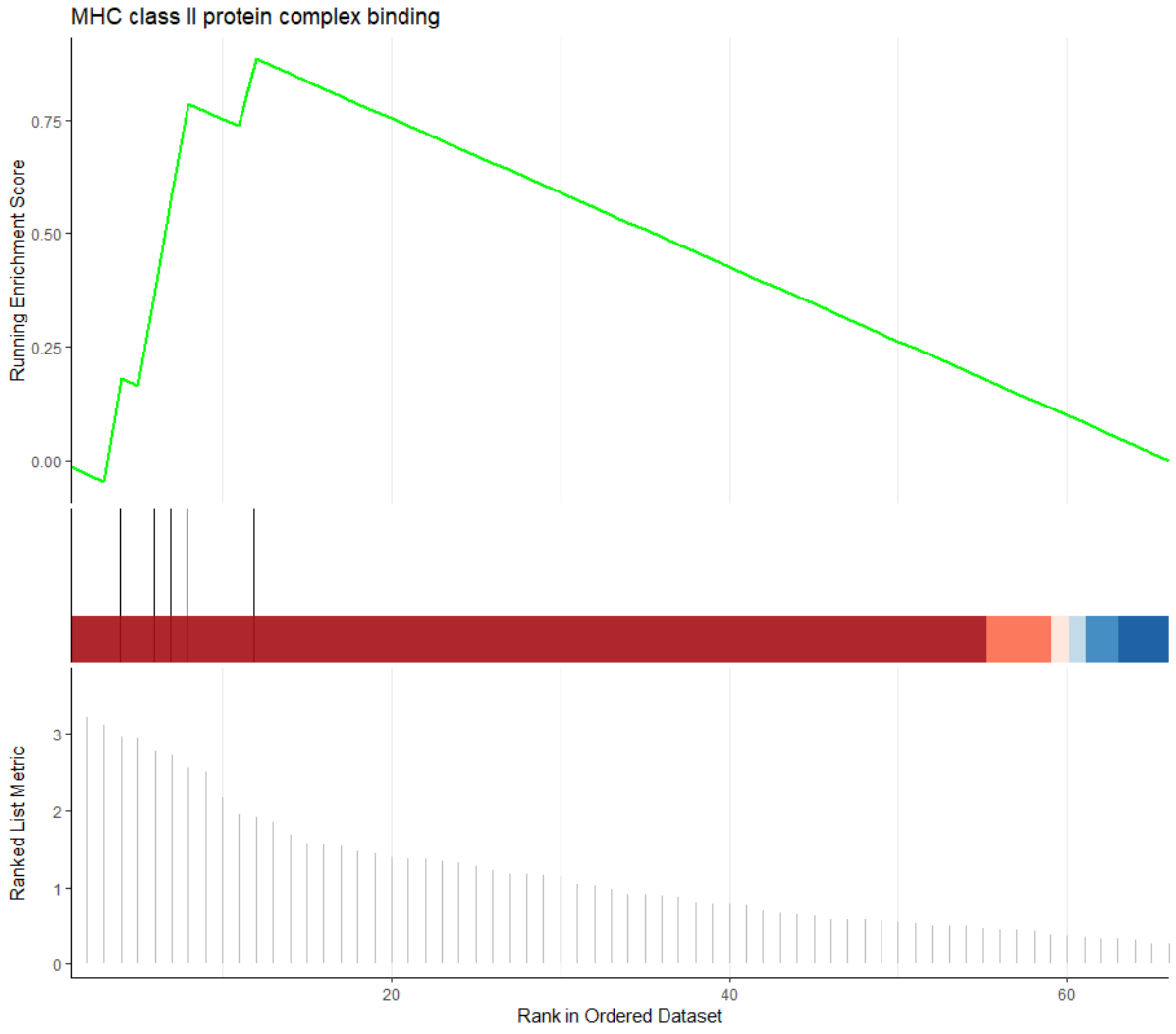
感兴趣的基因在基因集中所处的位置

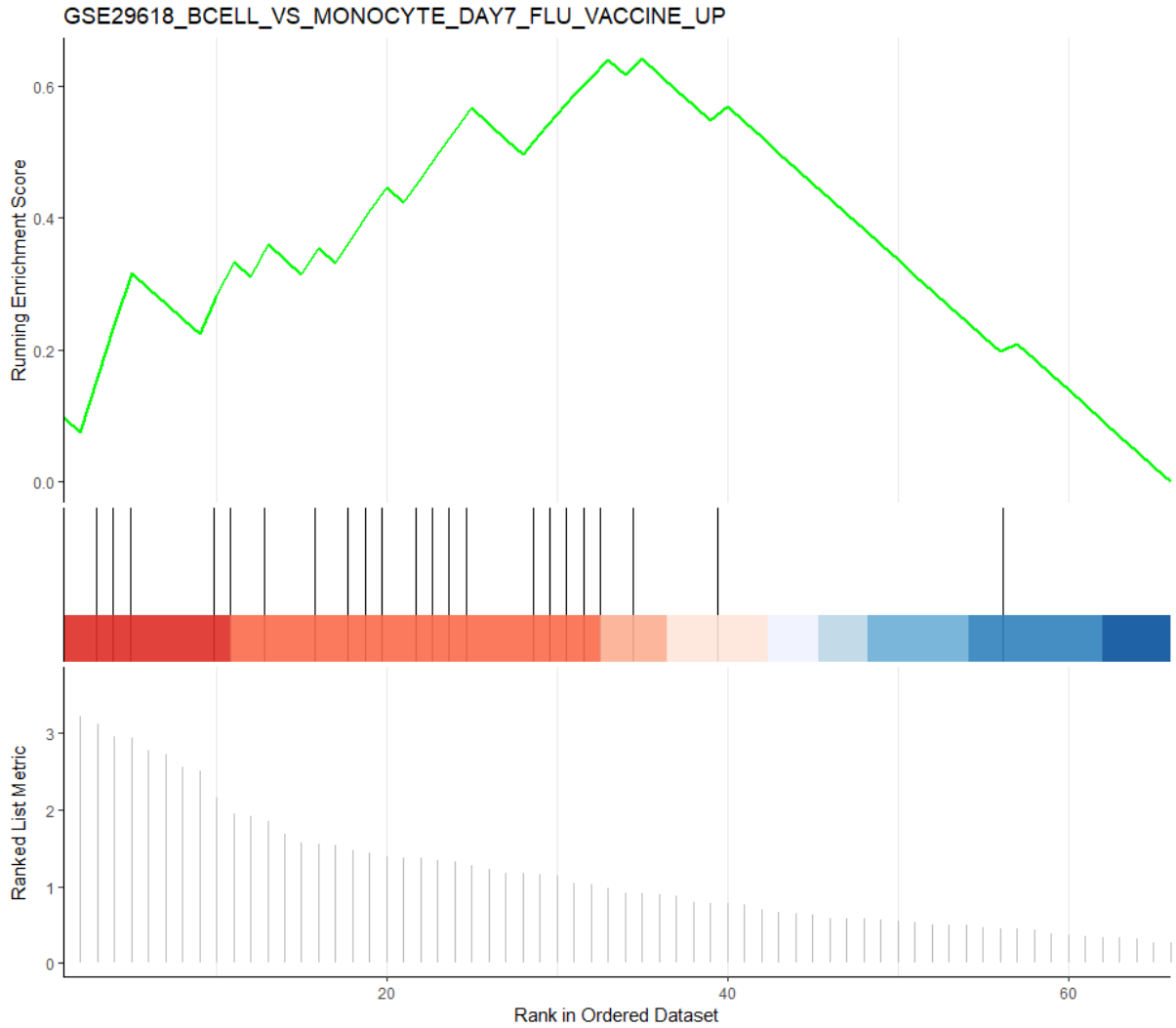
排序好的所有基因

基因表达与表型的关联，绝对值越大代表关联越强
数值大于0代表正相关，小于0代表负相关

基因集富集（GSEA）

GSEA富集分析的结果解释（cluster3的差异基因的GSEA富集分析结果）





什么是拟时序分析

拟时间序列（Pseudotime）分析，又称细胞轨迹（cell trajectory）分析，是通过构建细胞间的变化轨迹来重塑细胞随着时间的变化过程。从具体的分类分析和复杂程度来说，可以分为细胞轨迹分析和细胞谱系分析。

细胞轨迹分析

指的是简单模型的细胞变化轨迹分析，通常指的是细胞沿着某个过程有特定化的变化终点，轨迹具有简单树状结构，一端是“根”，另一端是“叶”

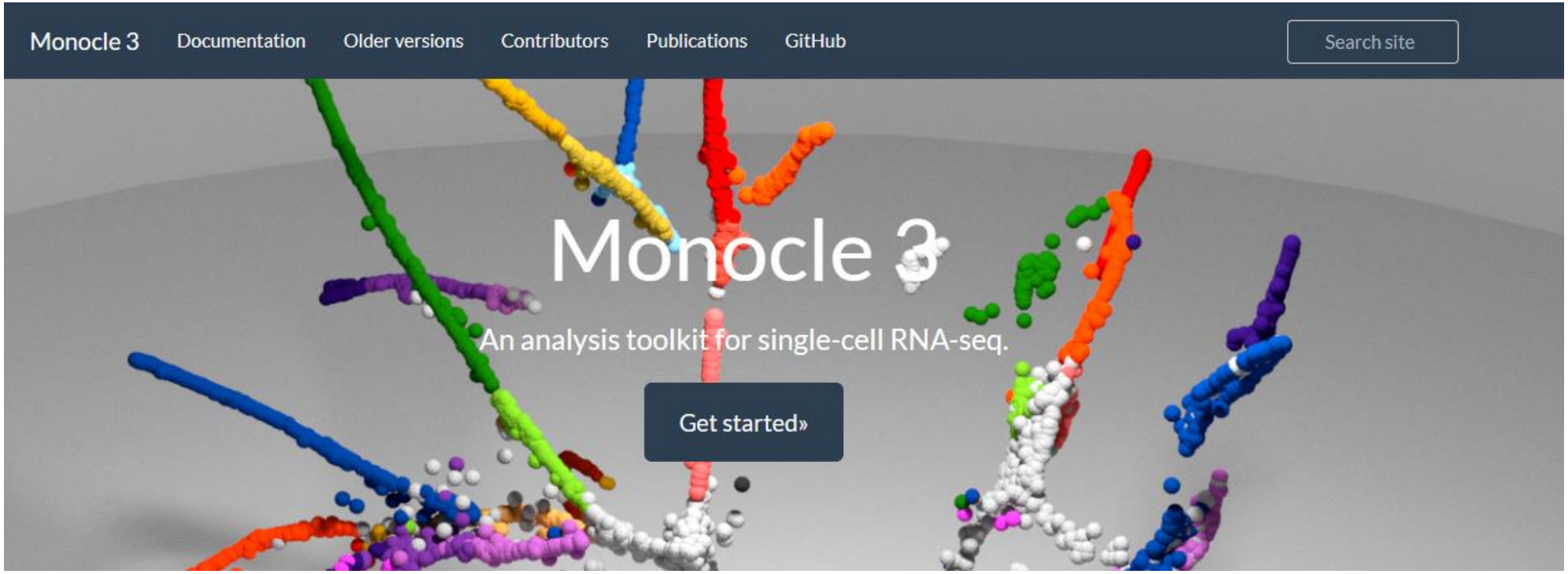
细胞谱系分析

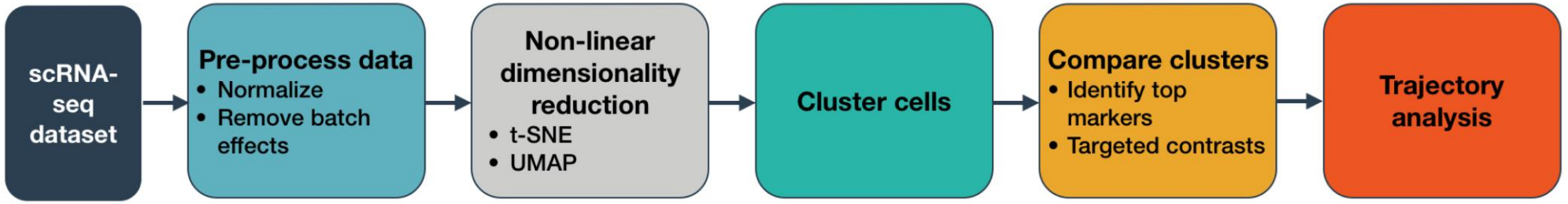
通常指的是某类祖源细胞，在特定条件下，有多个发育轨迹和命运，变化过程类似复杂树状分支变化过程

伪时间是一个抽象的分化单位：它只是一个cell到轨迹起点的距离，沿着最短路径测量。轨迹的总长度是由细胞从起始状态移动到结束状态所经历的总转录变化量来定义的。

什么是拟时序分析

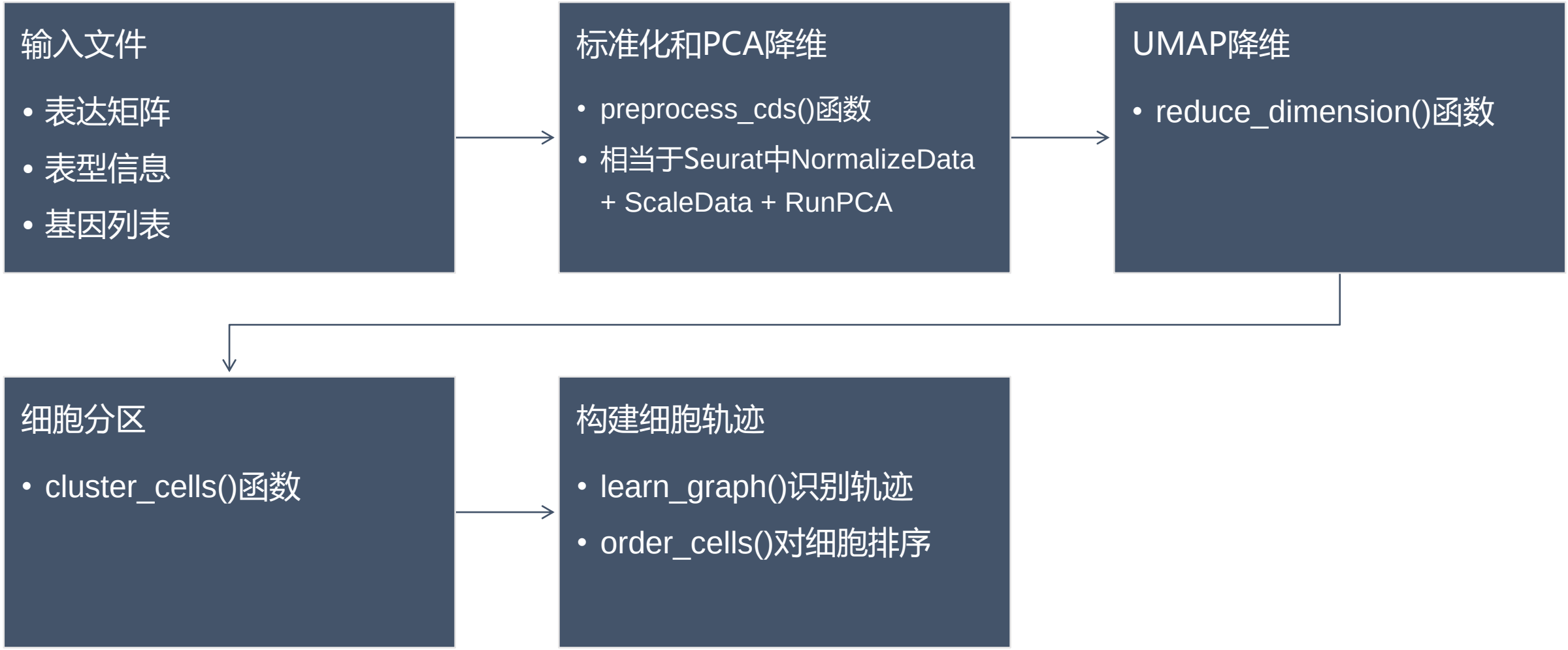
Monocle 软件来进行拟时序分析， Monocle 可以进行细胞的聚类分群和计数、重建单细胞轨迹、差异表达分析。





Monocle3使用

分析流程示意图



Monocle3使用

（1）输入文件：表达矩阵信息、表型信息和基因信息

数据集使用的是pbmc3k的数据集，由于pbmc都是分化成熟的免疫细胞，理论上并不存在直接的分化关系，因此不适合用来做拟时轨迹分析。这里仅作为学习演示。

```
> data
13714 x 2638 sparse Matrix of class "dgCMatrix"
  [[ suppressing 51 column names 'AAACATACAACCAC-1', 'AAACATTGAGCTAC-1', 'AAACATTGATCAGC-1' ... ]]
  [[ suppressing 51 column names 'AAACATACAACCAC-1', 'AAACATTGAGCTAC-1', 'AAACATTGATCAGC-1' ... ]]

AL627309.1    . . . . . .
AP006222.2    . . . . . .
RP11-206L10.2 . . . . . .
RP11-206L10.9 . . . . . .
LINC00115     . . . . . .
NOC2L         . . . . . .
KLHL17        . . . . . .
PLEKHN1       . . . . . .
RP11-54O7.17  . . . . . .
HES4          . . . . . .
```

13714 genes，2638 cells

```
> head(cell_metadata)
               orig.ident nCount_RNA nFeature_RNA percent.mt RNA_snn_res.0.5 seurat_clusters  singleR predicted.celltype.l1.score
AAACATACAACCAC-1     pbmc3k       2419          779  3.0177759               2               2  T_cells                   0.8620213
AAACATTGAGCTAC-1     pbmc3k       4903         1352  3.7935958               3               3   B_cell                   1.0000000
AAACATTGATCAGC-1     pbmc3k       3147         1129  0.8897363               2               2  T_cells                   0.9692498
AAACCGTGCTTCCG-1     pbmc3k       2639          960  1.7430845               1               1 Monocyte                   1.0000000
AAACCGTGTATGCG-1     pbmc3k        980          521  1.2244898               6               6  NK_cell                   1.0000000
AAACGCACTGGTAC-1     pbmc3k       2163          781  1.6643551               2               2  T_cells                   0.9202275
               predicted.celltype.l1 predicted.celltype.l2.score predicted.celltype.l2 marker_celltype
AAACATACAACCAC-1                CD8 T                   0.6878183              CD8 TCM              CD4
AAACATTGAGCTAC-1                    B                   0.5139331            B memory                B
AAACATTGATCAGC-1                CD4 T                   0.8255622              CD4 TCM              CD4
AAACCGTGCTTCCG-1                 Mono                   0.7224629            CD14 Mono       CD14+ Mono
AAACCGTGTATGCG-1                   NK                   1.0000000                   NK               NK
AAACGCACTGGTAC-1                CD4 T                   0.5237426              CD4 TCM              CD4
```

2638 cells

```
> head(gene_annotation)
              gene_short_name
AL627309.1          AL627309.1
AP006222.2          AP006222.2
RP11-206L10.2    RP11-206L10.2
RP11-206L10.9    RP11-206L10.9
LINC00115           LINC00115
NOC2L                   NOC2L
```

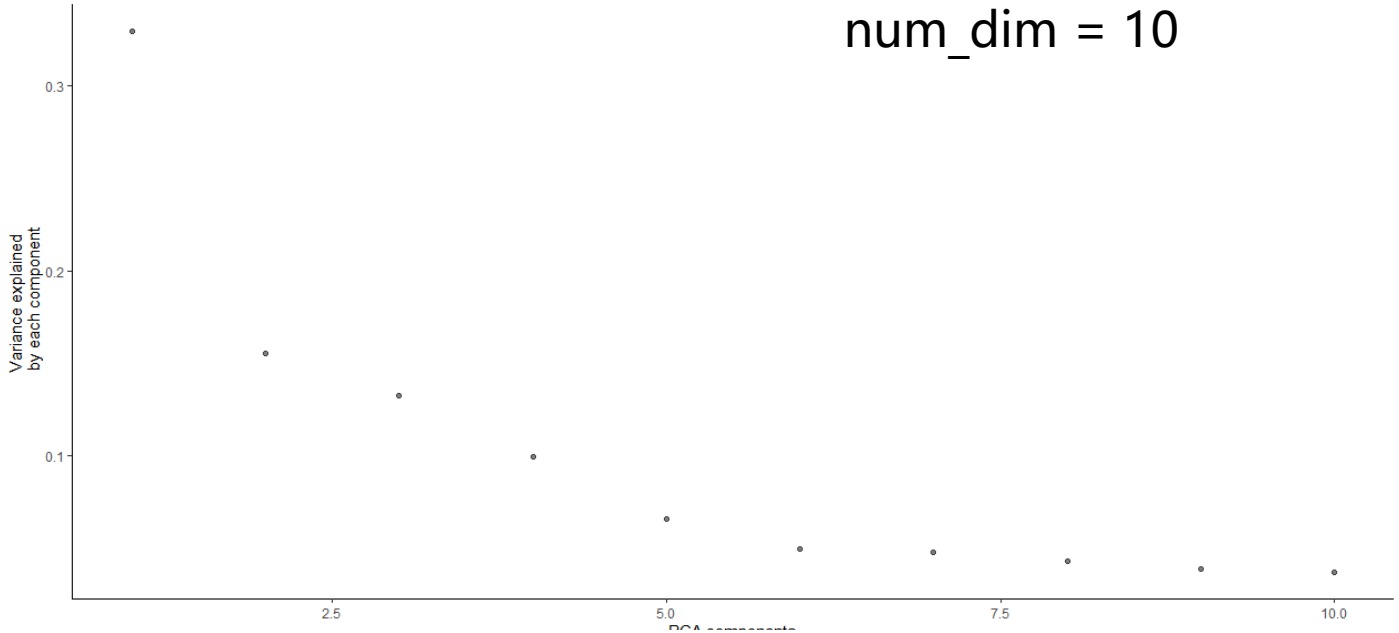
13714 genes

Monocle3使用

（2）标准化和PCA降维

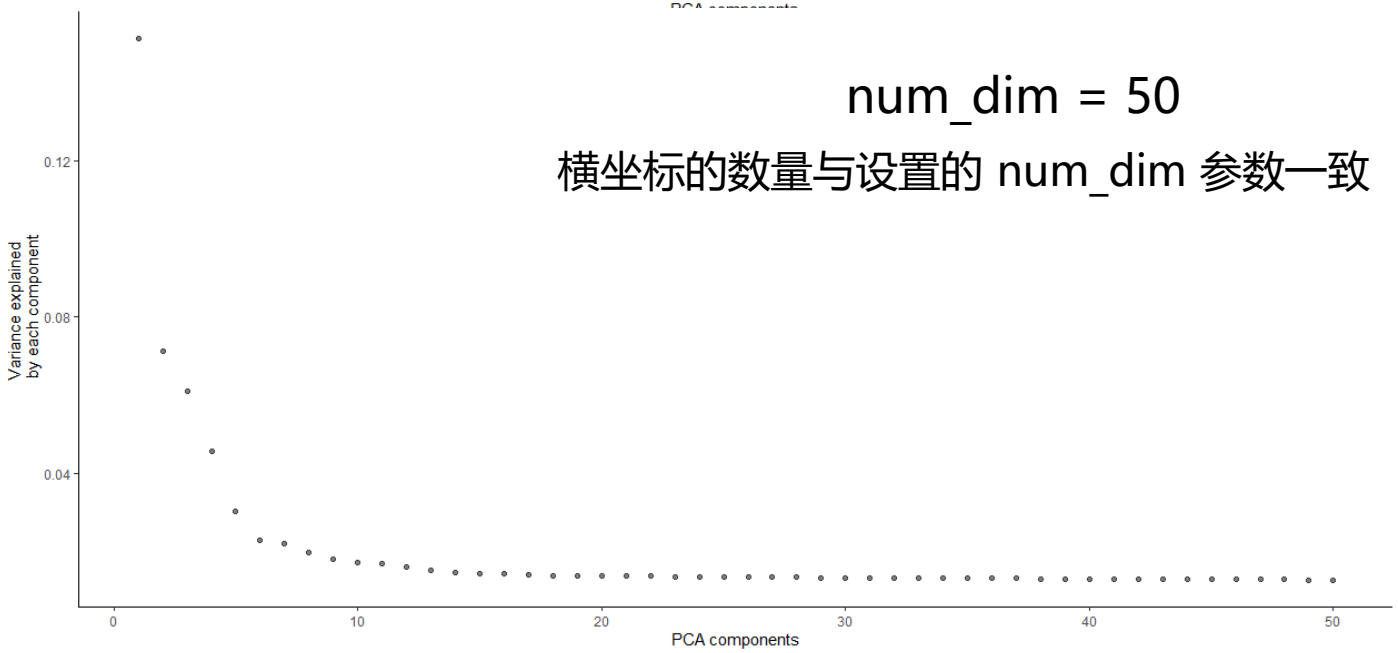
```
preprocess_cds(
  cds,
  method = c("PCA", "LSI"),
  num_dim = 50,
  norm_method = c("log", "size_only", "none"),
  use_genes = NULL,
  pseudo_count = NULL,
  scaling = TRUE,
  verbose = FALSE,
  build_nn_index = FALSE,
  nn_control = list()
)
```

num_dim = 10



num_dim = 50

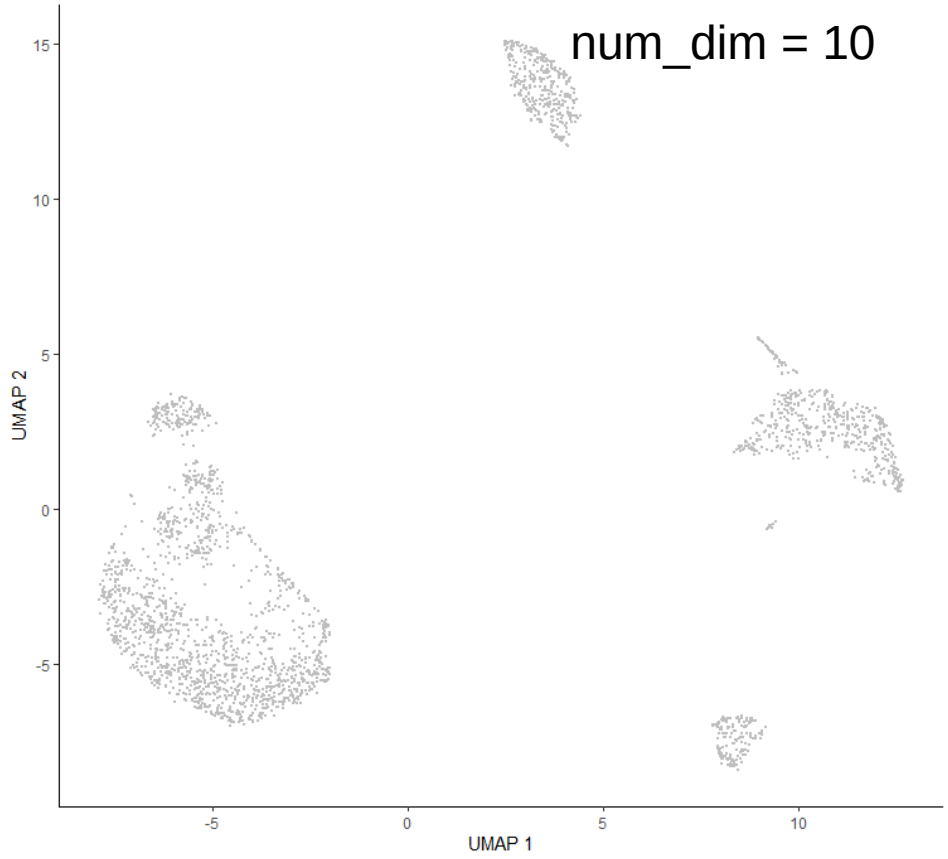
横坐标的数量与设置的 num_dim 参数一致

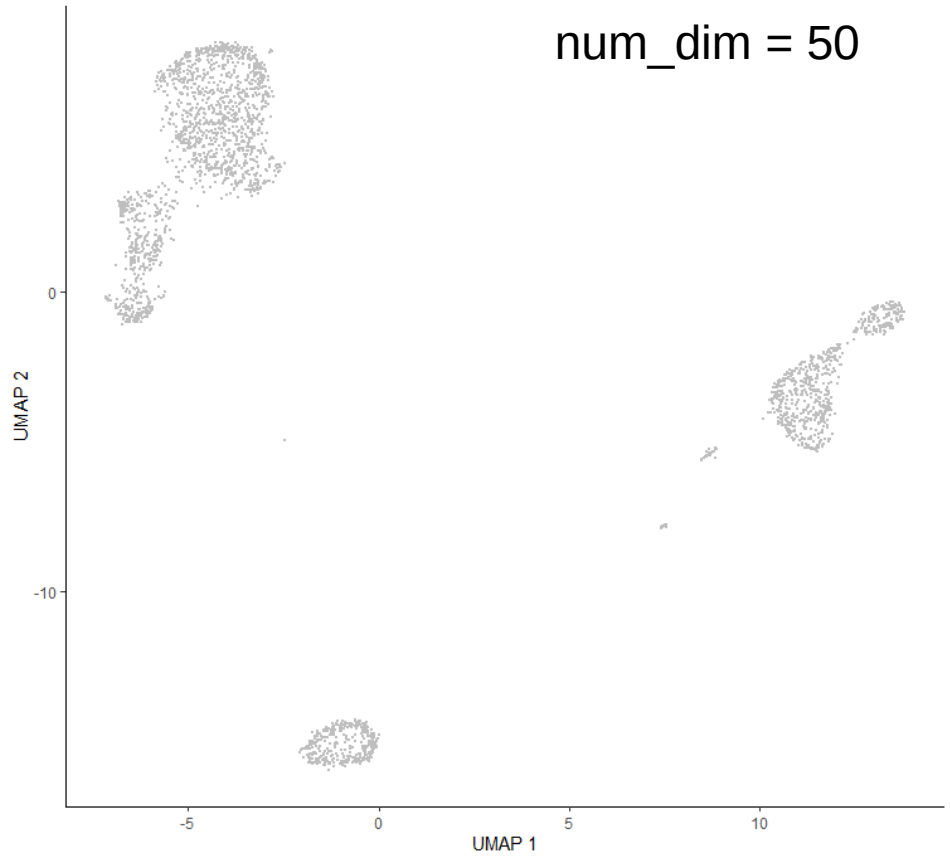


Monocle3使用

（3）UMAP降维可视化

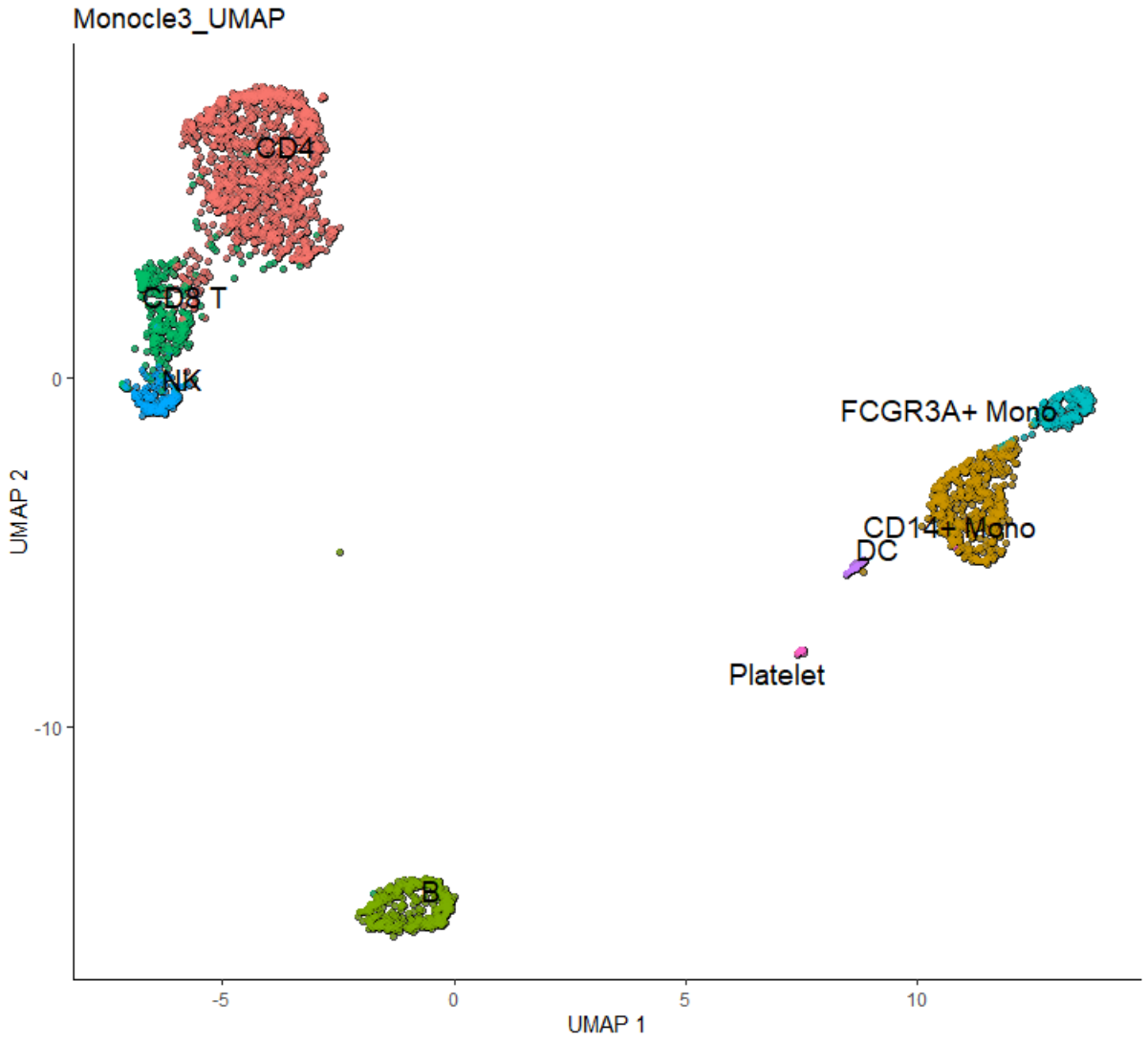
```
reduce_dimension(
  cds,
  max_components = 2,
  reduction_method = c("UMAP", "tSNE", "PCA", "LSI", "Aligned"),
  preprocess_method = NULL,
  umap.metric = "cosine",
  umap.min_dist = 0.1,
  umap.n_neighbors = 15L,
  umap.fast_sgd = FALSE,
  umap.nn_method = "annoy",
  verbose = FALSE,
  cores = 1,
  build_nn_index = FALSE,
  nn_control = list(),
  ...
)
```

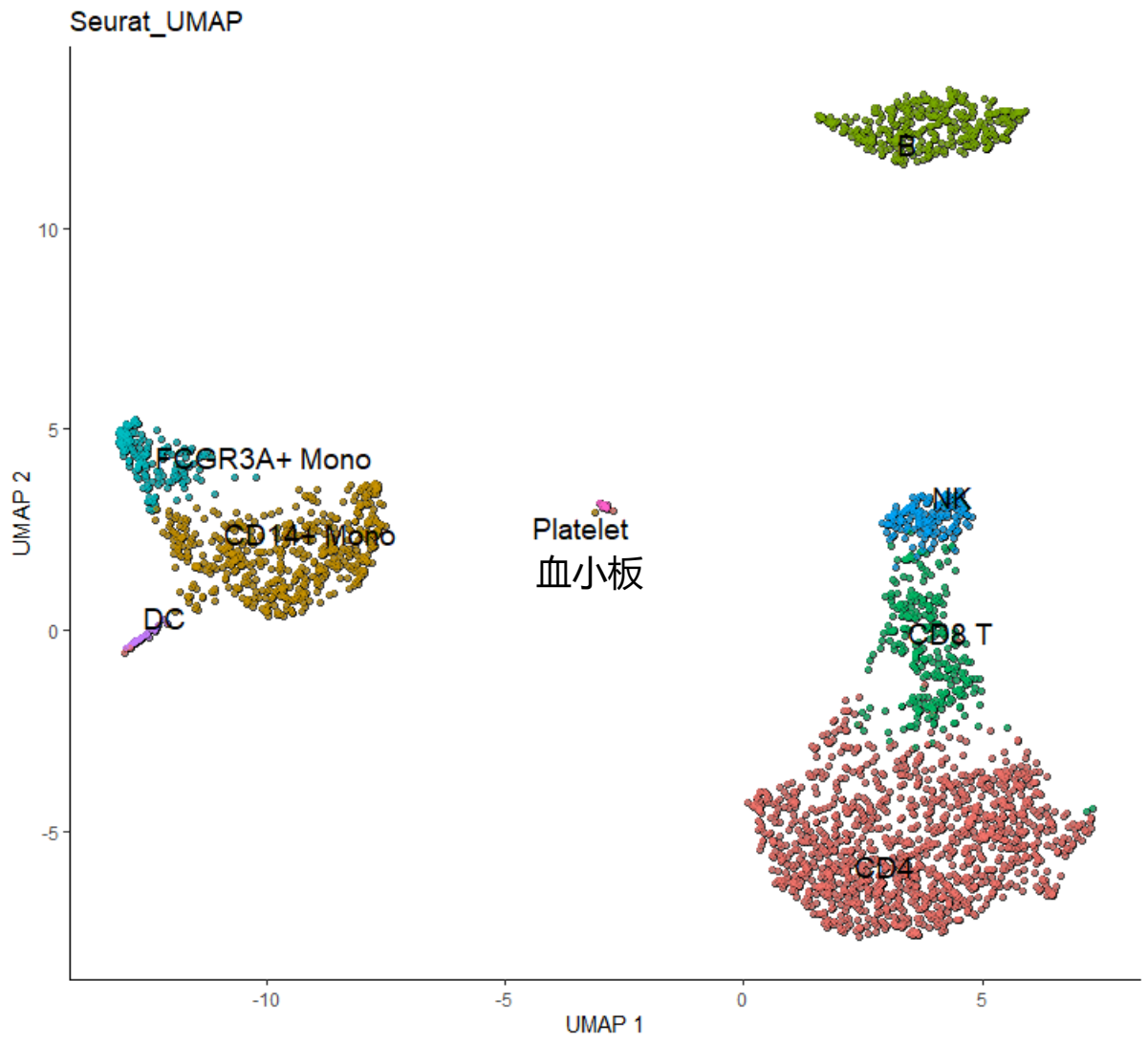




Monocle3使用

（3）UMAP降维可视化（num_dim = 50）： Monocle3与Seurat聚类的差异

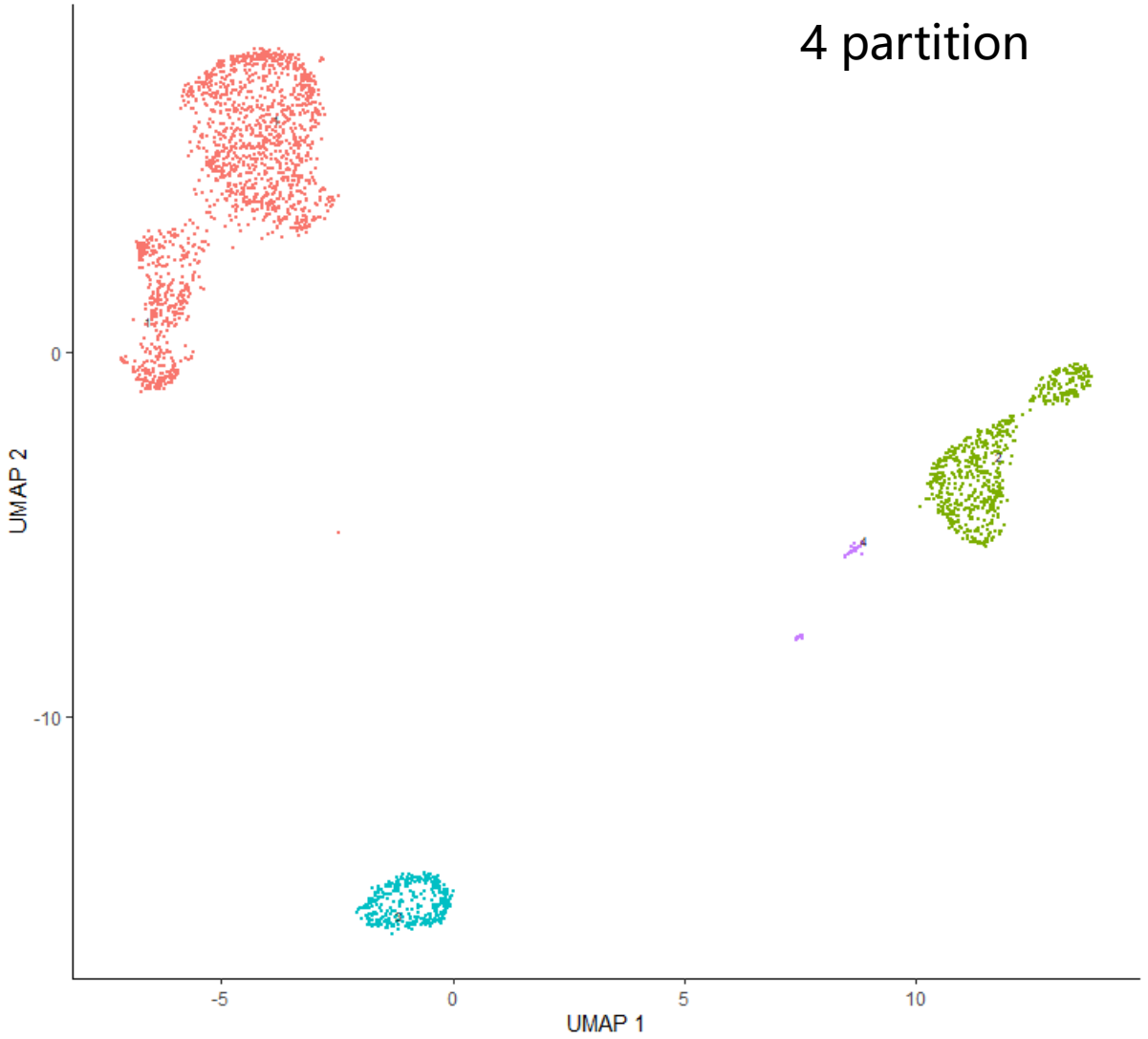




Monocle3使用

（4）细胞分区

```
cluster_cells(
  cds,
  reduction_method = c("UMAP", "tSNE", "PCA", "LSI", "Aligned"),
  k = 20,
  cluster_method = c("leiden", "louvain"),
  num_iter = 2,
  partition_qval = 0.05,
  weight = FALSE,
  resolution = NULL,
  random_seed = 42,
  verbose = FALSE,
  nn_control = list(),
  ...
)
```



Monocle能够通过其聚类程序了解细胞何时应该放置在相同的轨迹中，而不是放置在不同的轨迹中。

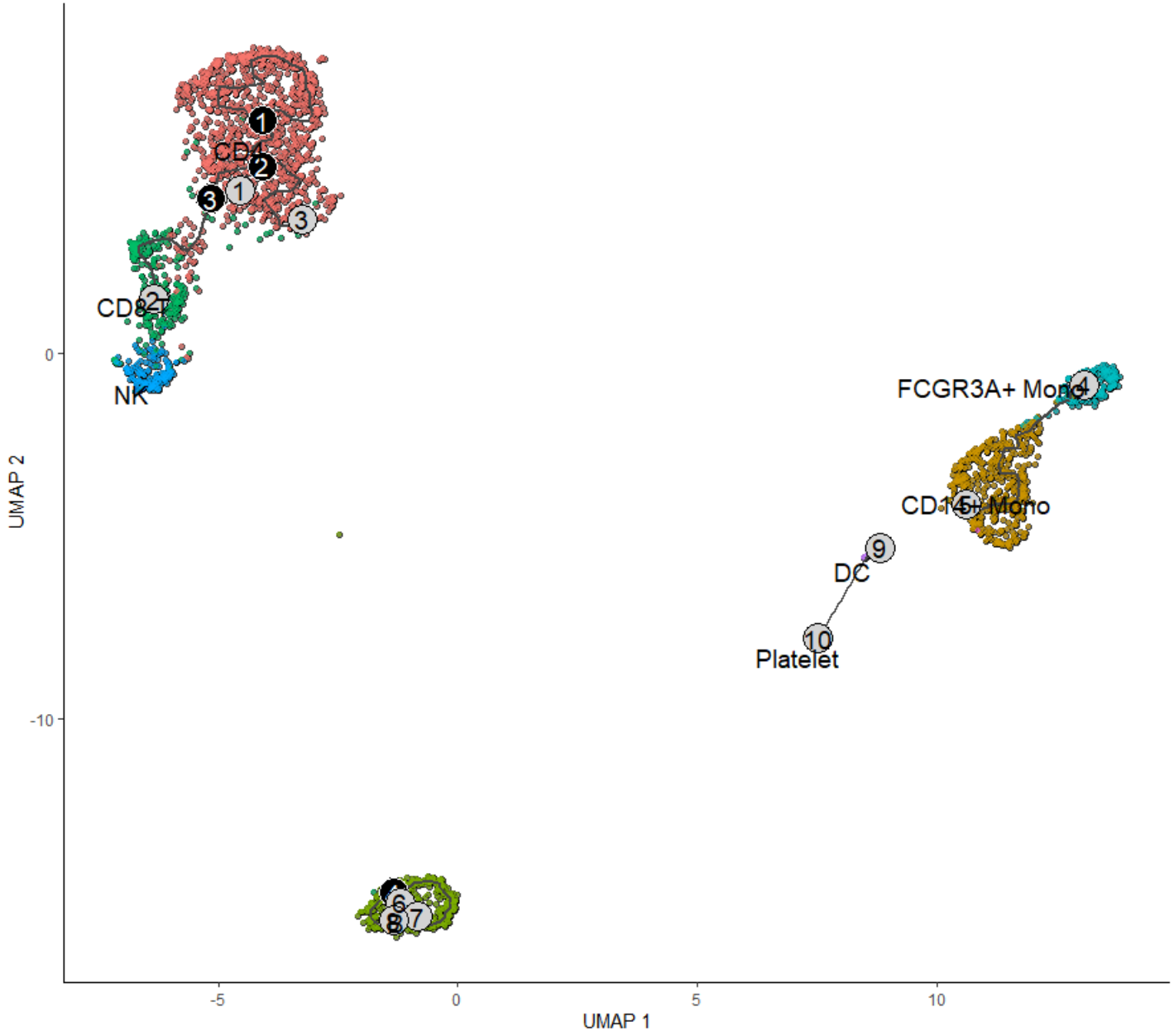
- 一个数据中可能存在多个不同的轨迹
- 不同分区的细胞会进行单独的轨迹分析

Monocle3使用

（5）构建细胞轨迹（learn_graph函数识别轨迹）

```
learn_graph(
  cds,
  use_partition = TRUE,
  close_loop = TRUE,
  learn_graph_control = NULL,
  verbose = FALSE
)
```

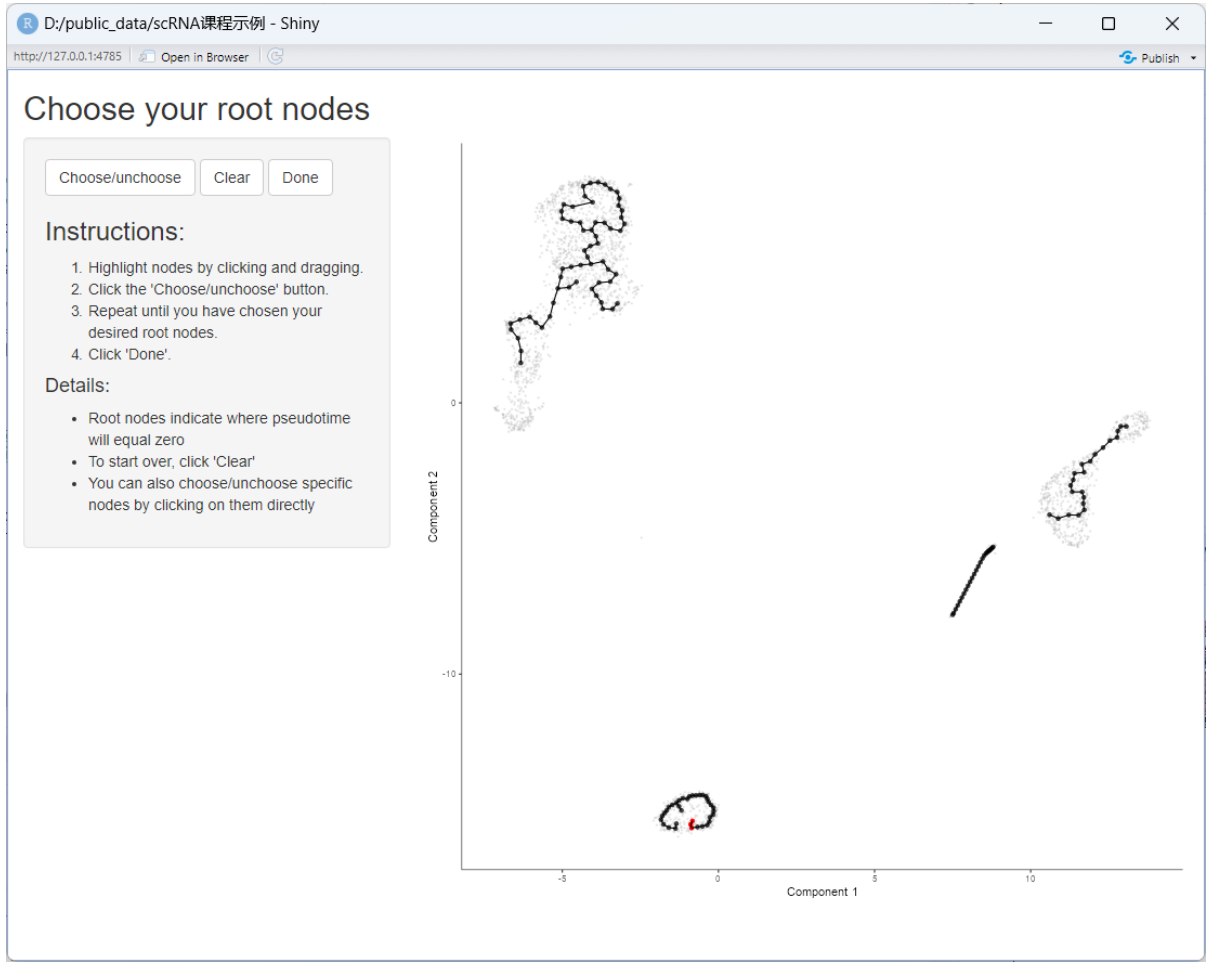
- 黑色的线显示的是graph的结构。
- 数字带白色圆圈表示不同的结局。
- 数字带黑色圆圈代表分叉点，从这个点开始，细胞可以有多个结局。

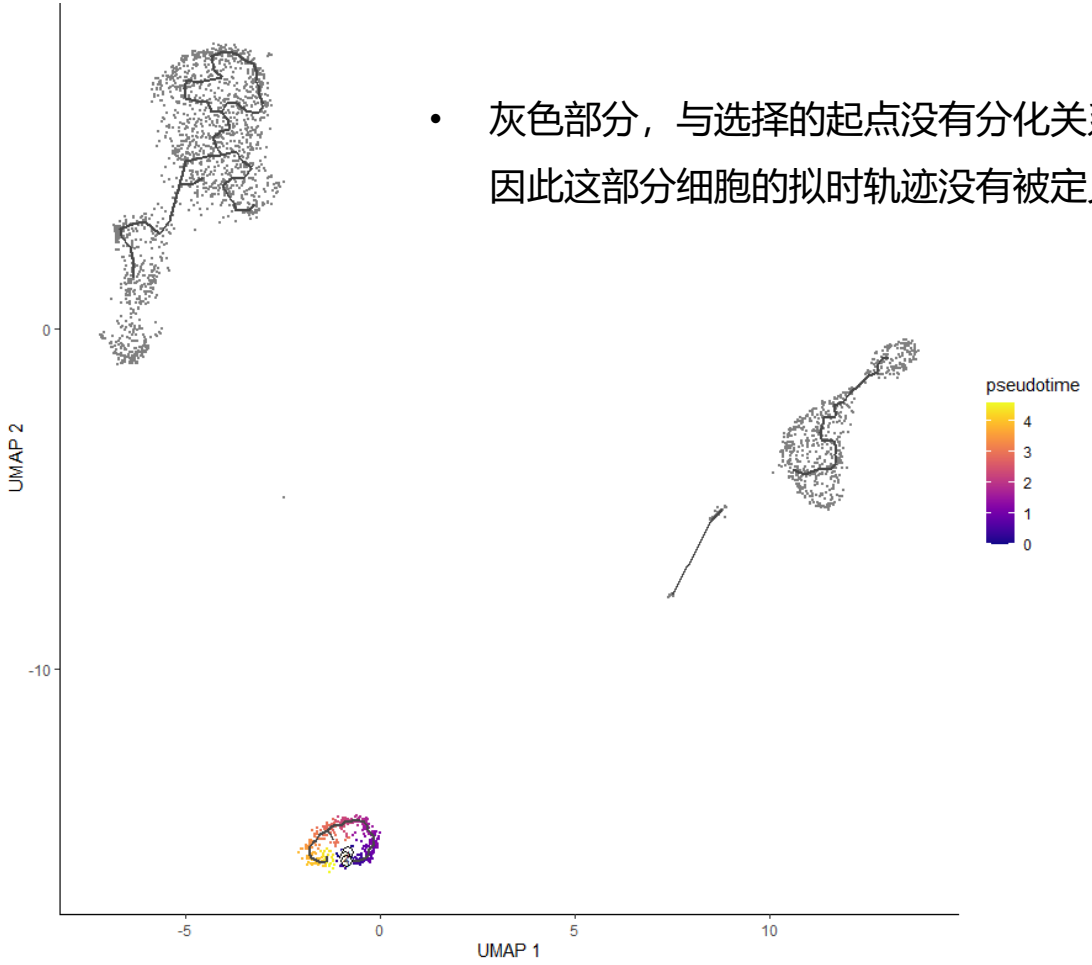


Monocle3使用

（5）构建细胞轨迹（order_cells函数对细胞排序）

- 为了对细胞进行排序，我们需要在Monocle中指定某条轨迹的起始点，也就是需要指定轨迹的roots
- 手动在图上选择一个位置，然后点击Done（比如图上的红点，可以选择多个位置）





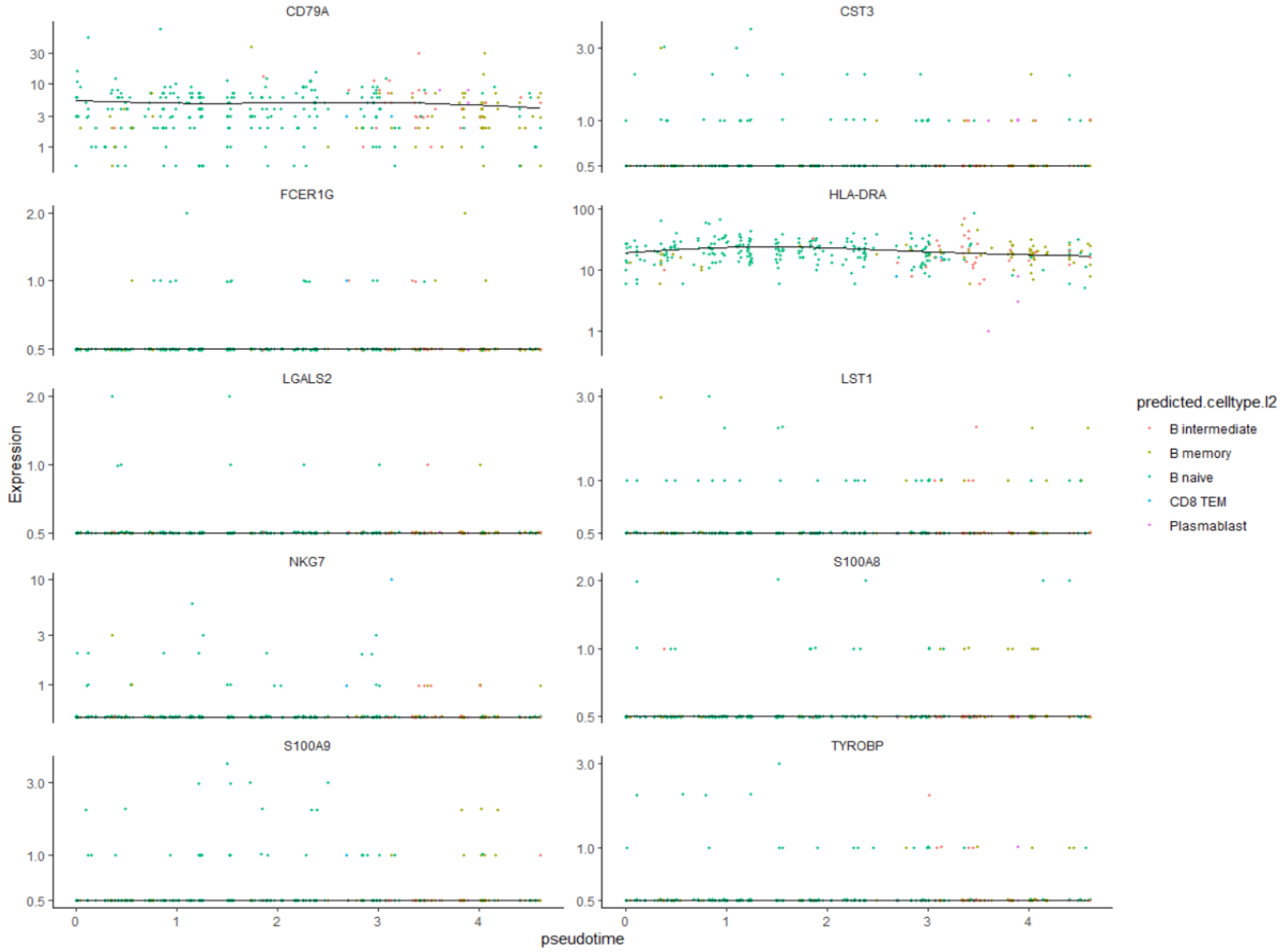
- 灰色部分，与选择的起点没有分化关系，因此这部分细胞的拟时轨迹没有被定义

Monocle3使用

拟时轨迹差异基因

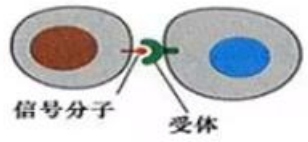
- graph_test()函数计算莫兰指数（morans_I）
- Moran的I统计量是一种多向多维空间自相关的度量。统计数据通过最近邻图对数据点之间的空间关系进行编码，使其特别适合于分析大型scRNA-seq数据集。
- Moran指数的范围为-1到1之间，0表示在空间中不相关，而1表示高度正相关，小于0的Moran指数一般都没有统计学意义。

	gene_short_name	p_value	morans_test_statistic	morans_I	status	q_value
TYROBP	TYROBP	0	145.12906	0.8468988	OK	0
S100A8	S100A8	0	142.23584	0.8297113	OK	0
S100A9	S100A9	0	141.09747	0.8231831	OK	0
CST3	CST3	0	139.01611	0.8111844	OK	0
CD79A	CD79A	0	136.64296	0.7966874	OK	0
LGALS2	LGALS2	0	135.18933	0.7885162	OK	0
FCER1G	FCER1G	0	133.23515	0.7774292	OK	0
NKG7	NKG7	0	132.33076	0.7719190	OK	0
LST1	LST1	0	129.34612	0.7546936	OK	0
HLA-DRA	HLA-DRA	0	129.19145	0.7538779	OK	0

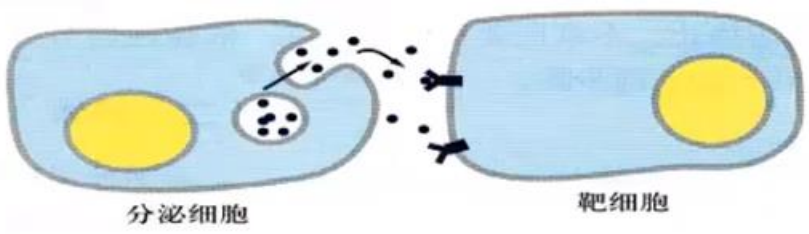


什么是细胞通讯?

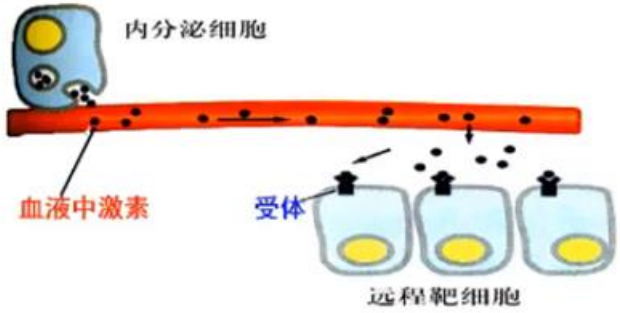
细胞通讯 (cell communication)：细胞识别与之相接触的细胞或者识别周围环境中存在的各种信号，并将其转化为细胞内信号进行传递，从而改变细胞内的代谢过程，影响细胞的生长发育，甚至诱导细胞的死亡。



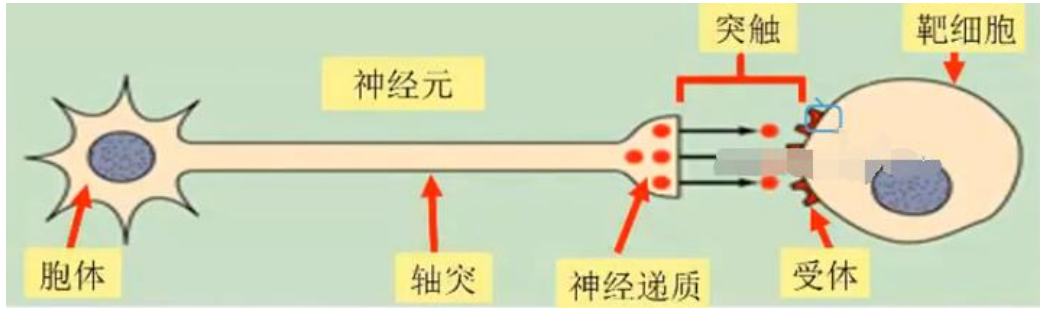
近分泌



旁分泌



内分泌



突触型

自分泌

什么是细胞通讯?

- 细胞通讯分析主要通过统计不同细胞类型中受体和配体的表达及配对情况，结合分子信息数据库推断不同细胞间的相互作用
- 利用scRNA-seq分析的细胞通讯，仅限于蛋白质配体-受体复合物介导的细胞间通讯
- 分析基础是基因表达数据和配体-受体数据库
- 揭示发育过程中各类细胞的相互作用、探索肿瘤免疫微环境、挖掘疾病潜在的治疗靶点。

名称	描述	配受体库	编程语言
CellphoneDB	是公开的人工矫正的，储存受体、配体以及两种相互作用的数据库。此外，还考虑了结构组成，能够描述异构复合物。	配体-受体+多聚体	Python
iTALK	通过平均表达量方式，筛选高表达的配体和受体，根据结果作圈图	配体-受体	R
CellChat	将细胞的表达数据作为输入，结合配受体及其辅助因子的相互作用来模拟细胞通讯。	配体-受体+多聚体+辅助因子	R
NicheNet	通过将相互作用的细胞表达数据与信号和基因调控网络的先验知识相结合来预测相互作用细胞之间的配体-靶标联系方式。	配体-受体+信号通路	R
CellCall	CellCall 是一个通过整合细胞内和细胞间信号来推断细胞间通讯网络和内部调节信号的工具包	配体-受体+TF因子	R

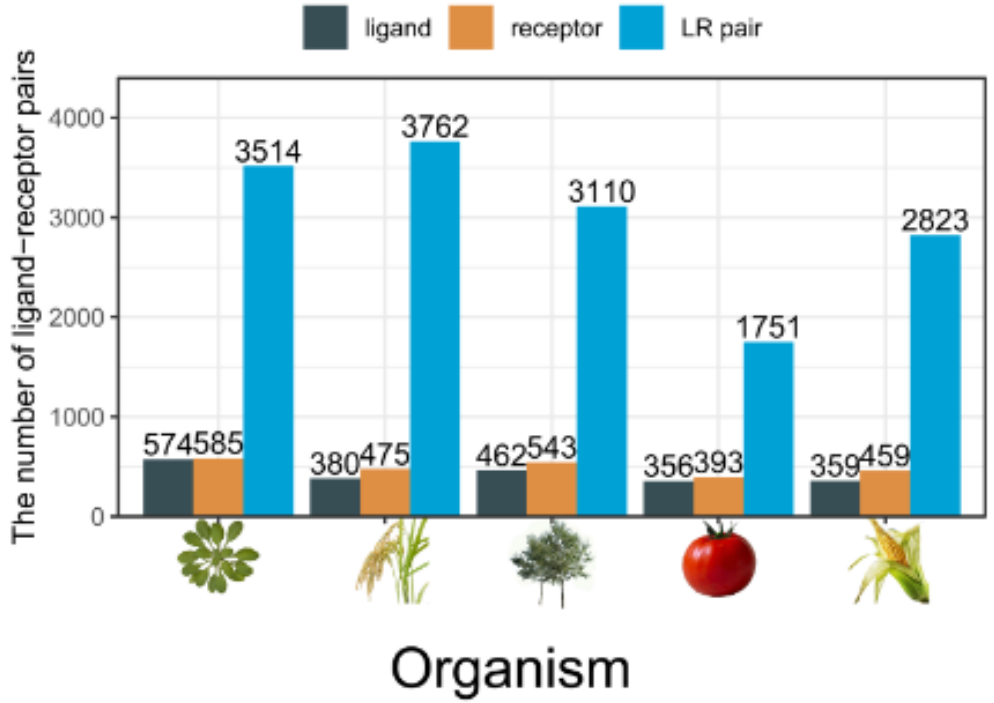
受体-配体数据库

- CellChatDB 和 PlantPhoneDB

CellChatDB

Ligand-Receptor Interaction Explorer

CellChatDB considers multimeric ligand-receptor complexes, and the effects of soluble and membrane-bound stimulatory and inhibitory cofactors.

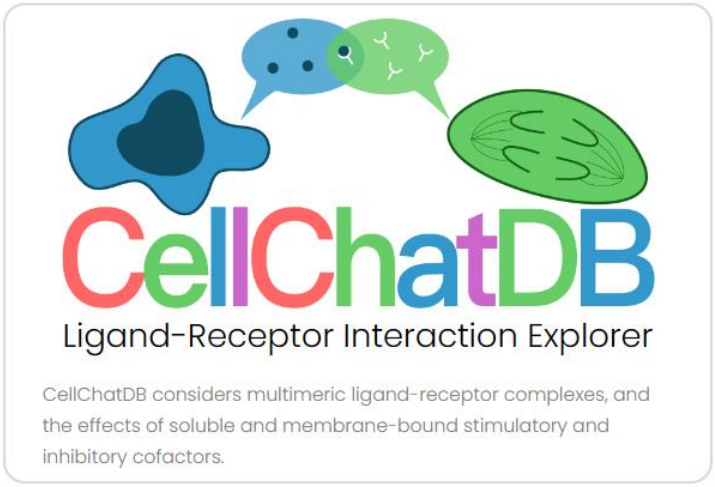


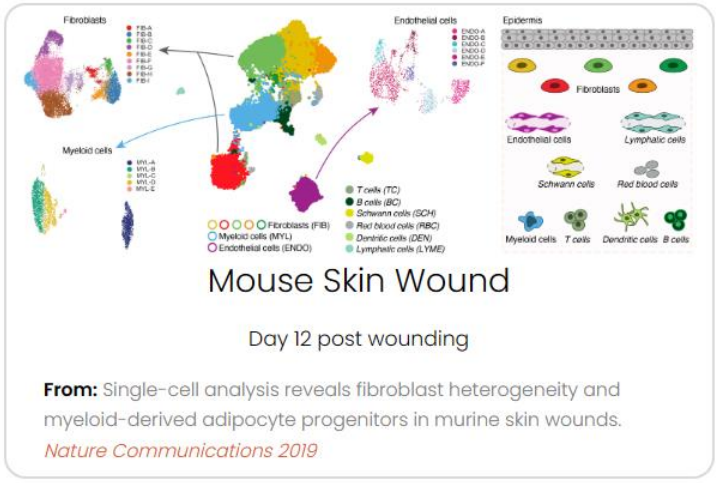
CellChatDB考虑了已知的配体受体复合物的组成，包括配体和受体的多聚复合物，以及几种辅助因子类型:可溶性激动剂、拮抗剂、共刺激和共抑制膜结合受体。CellChatDB包含**2021种**已验证的L-R 对，包括60%的分泌相互作用（secreting interactions）。此外，有48%的相互作用涉及异质分子复合物。

PlantPhoneDB收集了5个物种29个单细胞转录组数据集，总计约560,000细胞。对于拟南芥，利用关键词（secreted和cell membrane）搜索uniprot，从而鉴定可能的配体和受体信息。除此之外，还在TAIR、PlantSecKB、BioGRID、Interactome、IntAct、plant.MAP、STRING提取相关的PPI信息。对于非模式生物，则是利用工具进行配体和受体的预测。

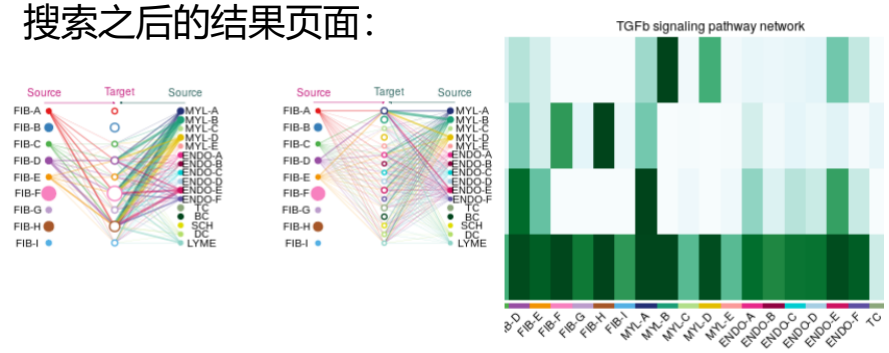
CellChat使用

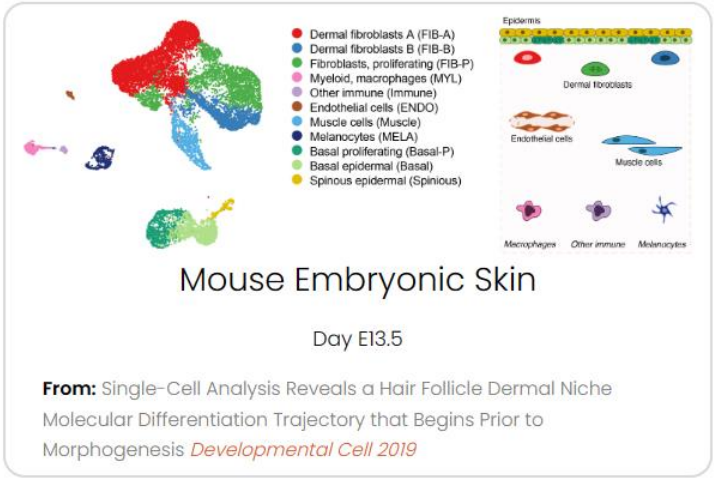
- 在线浏览网站：CellChat http://www.cellchat.org/

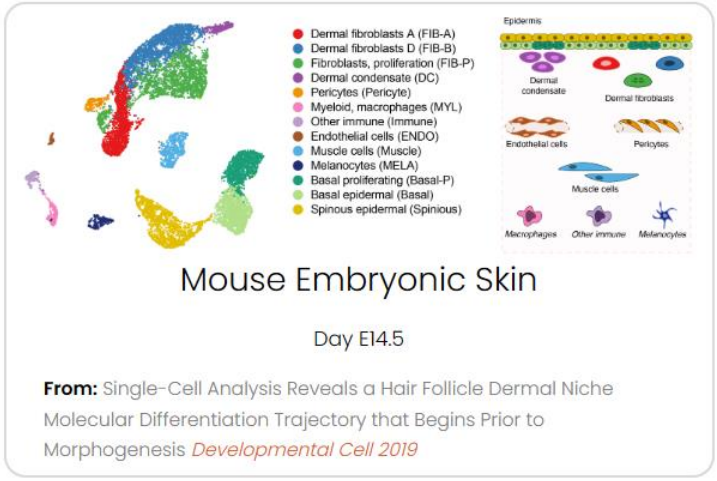


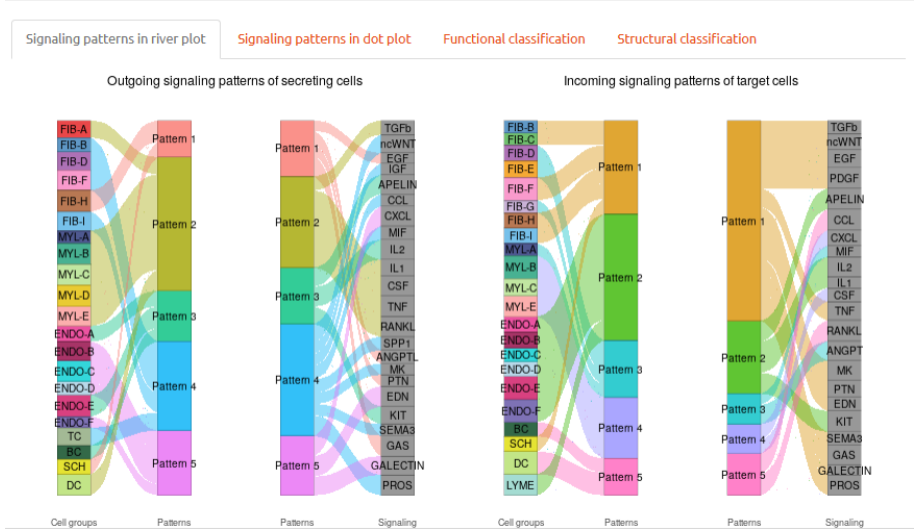


搜索之后的结果页面：





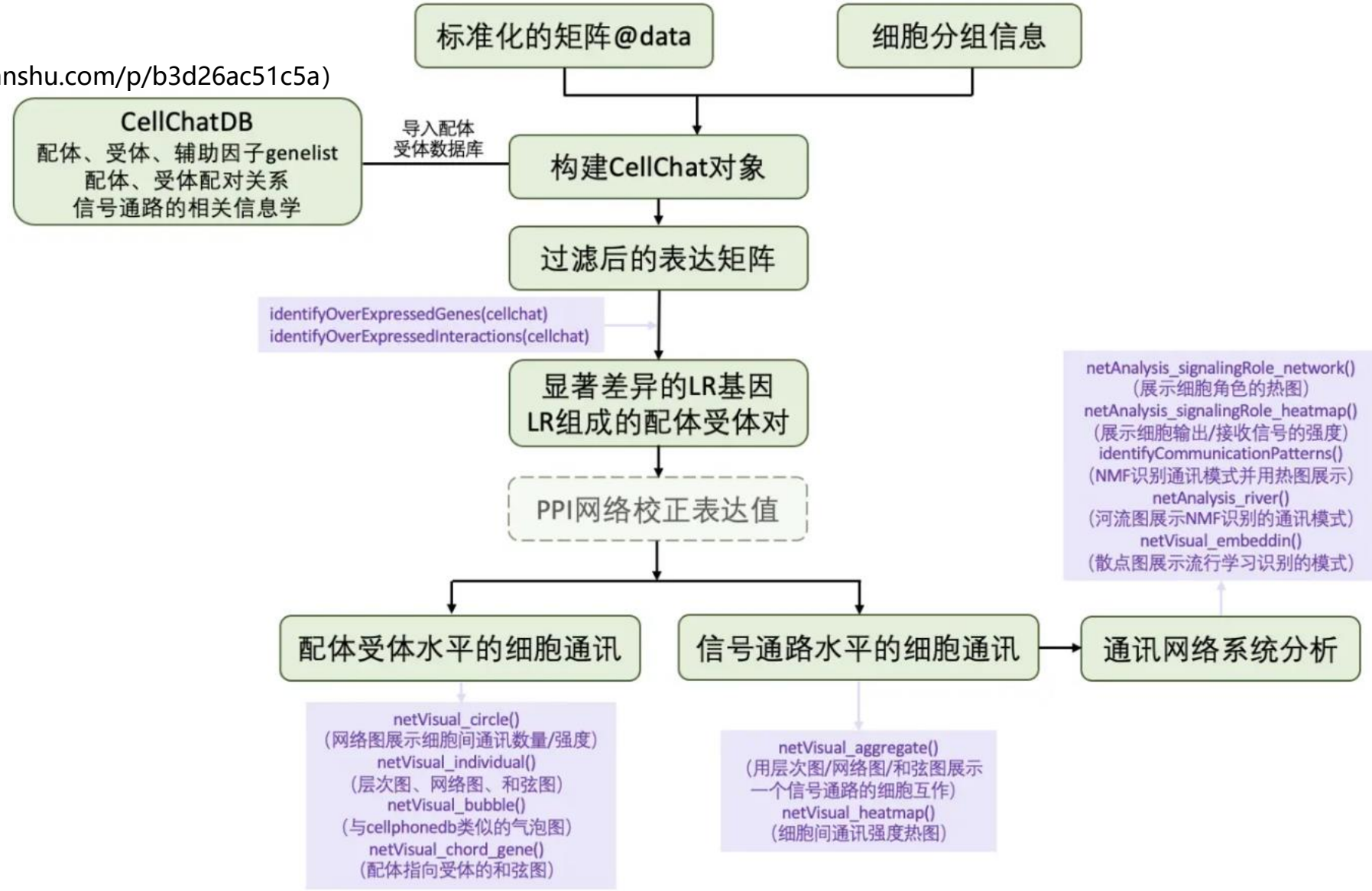




CellChat使用

- 整体分析流程

(来源于 https://www.jianshu.com/p/b3d26ac51c5a)



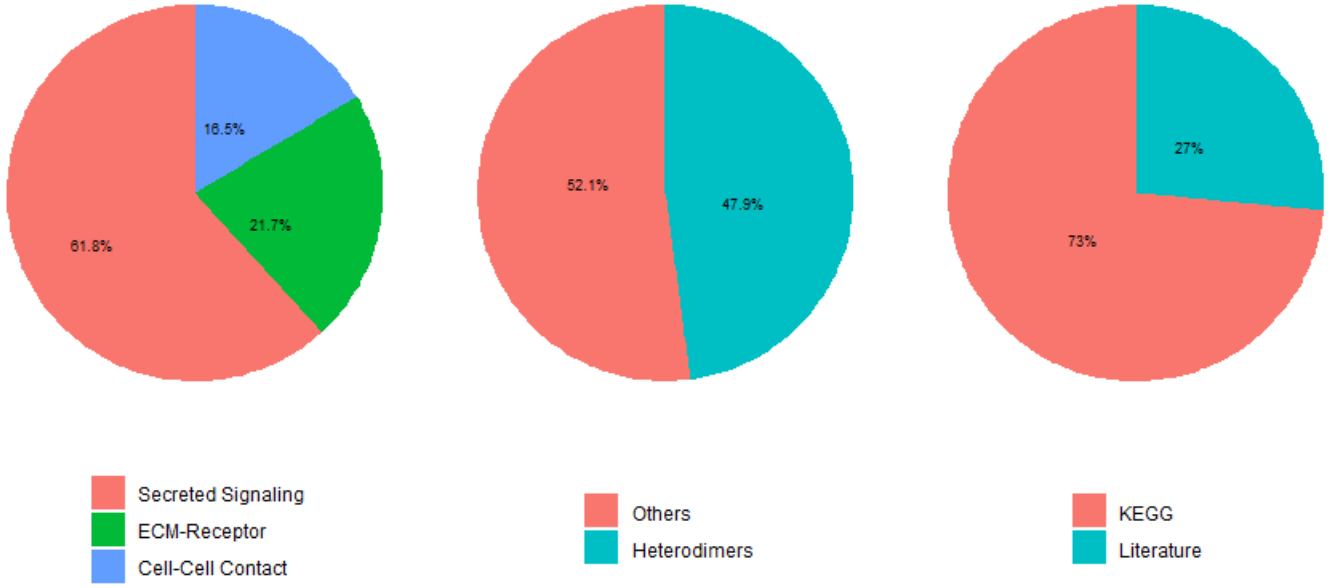
CellChat使用

（1）输入数据和受配体数据库准备

```
> pbmc3k
An object of class Seurat
13981 features across 2638 samples within 4 assays
Active assay: RNA (13714 features, 2000 variable features)
 3 other assays present: prediction.score.celltype.l1, prediction.score.celltype.l2, predicted_ADT
 5 dimensional reductions calculated: pca, umap, tsne, ref.spca, ref.umap
13714 x 2638 sparse Matrix of class "dgCMatrix"
  [[ suppressing 69 column names 'AAACATACAACCAC-1', 'AAACATTGAGCTAC-1', 'AAACATTGATCAGC-1' ... ]]
  [[ suppressing 69 column names 'AAACATACAACCAC-1', 'AAACATTGAGCTAC-1', 'AAACATTGATCAGC-1' ... ]]

AL627309.1    . . . . . .       . . . . . .       . . . . . .       . . . . . .       . . . . . .
AP006222.2    . . . . . .       . . . . . .       . . . . . .       . . . . . .       . . . . . .
RP11-206L10.2 . . . . . .       . . . . . .       . . . . . .       . . . . . .       . . . . . .
RP11-206L10.9 . . . . . .       . . . . . .       . . . . . .       . . . . . .       . . . . . .
LINC00115     . . . . . .       . . . . . .       . . . . . .       . . . . . .       . . . . . .
NOC2L         . . . . . 1.646272 . . . . . 1.398186 . . . . . 1.89939 . . . . . 1.36907 1.721224 . . . . . 1.568489
KLHL17        . . . . . .       . . . . . .       . . . . . .       . . . . . .       . . . . . .

AL627309.1    .       . .       . . .       . . . . . .....
AP006222.2    .       . .       . . .       . . . . . .....
RP11-206L10.2 .       . .       . . .       . . . . . .....
RP11-206L10.9 .       . .       . . .       . . . . . .....
LINC00115     .       . .       . . .       . . . . . .....
NOC2L         1.678814 . 1.253835 . . 3.791113 . . . . . .....
KLHL17        .       . .       . . .       . . . . . .....
```

- CellChat分析的输入是标准均一化的表达矩阵
- 需要有细胞的表型信息（metadata）

- 人类中的CellChatDB包含1,939个经过验证的分子相互作用，包括61.8%的旁分泌/自分泌信号相互作用，21.7%的细胞外基质（ECM） - 受体受体相互作用和16.5%的细胞细胞接触相互作用。
- CellChat中，可以先择特定的信息描述细胞间的相互作用，可以理解为从特定的侧面来刻画细胞间相互作用，比用一个大的配体库又精细了许多。



CellChat使用

（2）预处理

- 首先在一个细胞组中识别过表达的配体或受体，然后将基因表达数据投射到蛋白-蛋白相互作用(PPI)网络上

在矩阵的所有的基因中提取signaling gene（13714个基因，过滤后270个），结果保存在data.signaling

```
cellchat <- subsetData(cellchat)
```

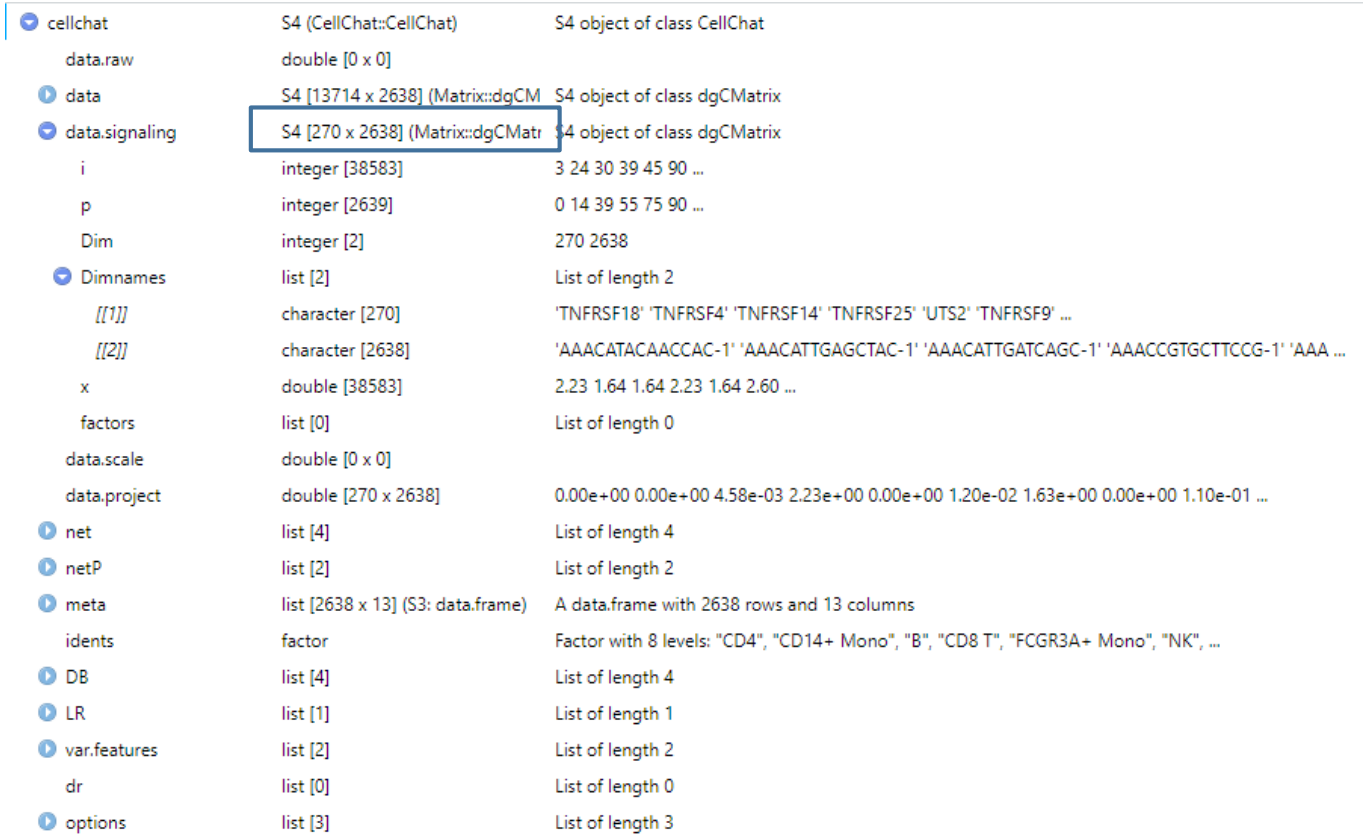
```
cellchat <- identifyOverExpressedGenes(cellchat)
```

相当于suerat中的FindAllMarkers，找每个细胞群中高表达的配体受体

```
cellchat <- identifyOverExpressedInteractions(cellchat)
```

```
cellchat <- projectData(cellchat, PPI.human)
```

找到配体受体关系后，projectData将配体受体对的表达值投射到PPI上，来对@data.signaling中的表达值进行校正。结果保存在@data.project



CellChat使用

（3）推断细胞通讯网络：通过为每个相互作用分配一个概率值并进行置换检验来推断生物意义上的细胞-细胞通信。

- 配体-受体水平：通过计算与每个信号通路相关的所有配体-受体相互作用的通信概率来推断信号通路水平上的通信概率
- 信号通路水平：通过计算链路的数量或汇总通信概率来计算细胞间的聚合通信网络

	source	target	ligand	receptor	prob	pval	interaction_name	interaction_name_2	pathway_name	annotation	evidence
1	CD14+ Mono	CD14+ Mono	TGFB1	TGFbR1_R2	4.949788e-06	0.00	TGFB1_TGFBR1_TGFBR2	TGFB1 - (TGFBR1+TGFBR2)	TGFb	Secreted Signaling	KEGG: hsa04350
2	CD8 T	CD14+ Mono	TGFB1	TGFbR1_R2	3.216494e-07	0.03	TGFB1_TGFBR1_TGFBR2	TGFB1 - (TGFBR1+TGFBR2)	TGFb	Secreted Signaling	KEGG: hsa04350
3	FCGR3A+ Mono	CD14+ Mono	TGFB1	TGFbR1_R2	1.871045e-06	0.00	TGFB1_TGFBR1_TGFBR2	TGFB1 - (TGFBR1+TGFBR2)	TGFb	Secreted Signaling	KEGG: hsa04350
4	NK	CD14+ Mono	TGFB1	TGFbR1_R2	1.793278e-06	0.00	TGFB1_TGFBR1_TGFBR2	TGFB1 - (TGFBR1+TGFBR2)	TGFb	Secreted Signaling	KEGG: hsa04350
5	Platelet	CD14+ Mono	TGFB1	TGFbR1_R2	6.517341e-07	0.00	TGFB1_TGFBR1_TGFBR2	TGFB1 - (TGFBR1+TGFBR2)	TGFb	Secreted Signaling	KEGG: hsa04350
6	CD14+ Mono	CD8 T	TGFB1	TGFbR1_R2	2.656984e-06	0.00	TGFB1_TGFBR1_TGFBR2	TGFB1 - (TGFBR1+TGFBR2)	TGFb	Secreted Signaling	KEGG: hsa04350
7	FCGR3A+ Mono	CD8 T	TGFB1	TGFbR1_R2	1.004355e-06	0.00	TGFB1_TGFBR1_TGFBR2	TGFB1 - (TGFBR1+TGFBR2)	TGFb	Secreted Signaling	KEGG: hsa04350
8	NK	CD8 T	TGFB1	TGFbR1_R2	9.626113e-07	0.00	TGFB1_TGFBR1_TGFBR2	TGFB1 - (TGFBR1+TGFBR2)	TGFb	Secreted Signaling	KEGG: hsa04350
9	Platelet	CD8 T	TGFB1	TGFbR1_R2	3.498567e-07	0.00	TGFB1_TGFBR1_TGFBR2	TGFB1 - (TGFBR1+TGFBR2)	TGFb	Secreted Signaling	KEGG: hsa04350
10	CD14+ Mono	FCGR3A+ Mono	TGFB1	TGFbR1_R2	4.209379e-06	0.00	TGFB1_TGFBR1_TGFBR2	TGFB1 - (TGFBR1+TGFBR2)	TGFb	Secreted Signaling	KEGG: hsa04350

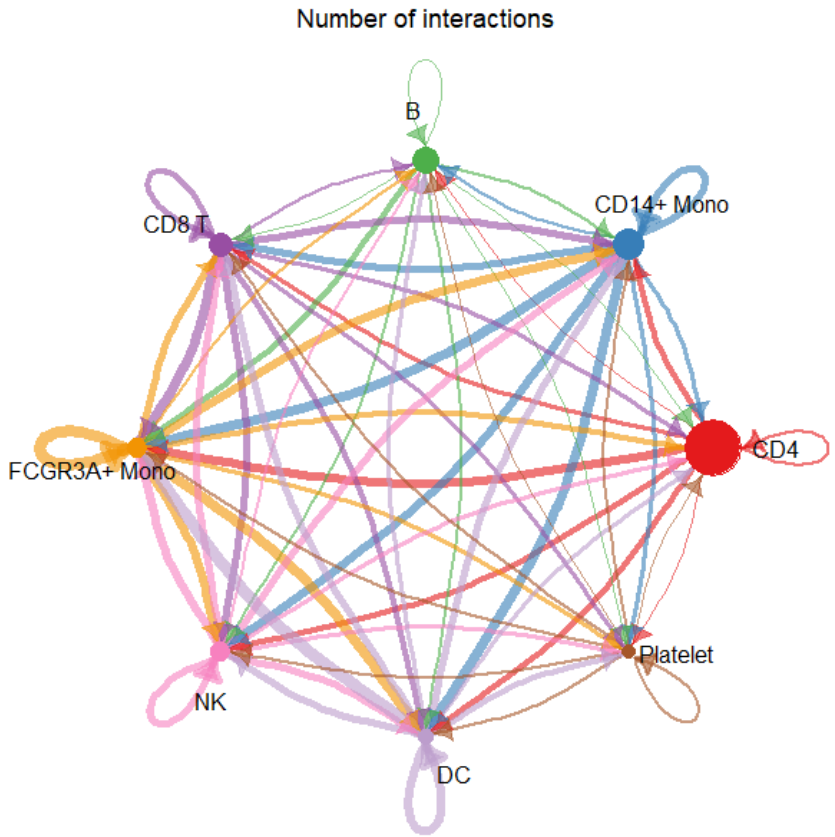
配体-受体水平细胞通讯网络

	source	target	pathway_name	prob	pval
1	B	B	LT	5.292116e-06	0.000000000
2	B	B	MIF	3.077386e-04	0.000000000
3	B	CD14+ Mono	ANNEXIN	2.326018e-07	0.000000000
4	B	CD14+ Mono	BAFF	6.727718e-08	0.003333333
5	B	CD14+ Mono	COMPLEMENT	4.432185e-10	0.030000000
6	B	CD14+ Mono	IFN-II	1.176846e-09	0.000000000
7	B	CD14+ Mono	IL1	4.389146e-10	0.000000000
8	B	CD14+ Mono	LT	7.547563e-05	0.000000000
9	B	CD14+ Mono	MIF	1.420356e-06	0.000000000
10	B	CD4	FLT3	2.048406e-08	0.000000000

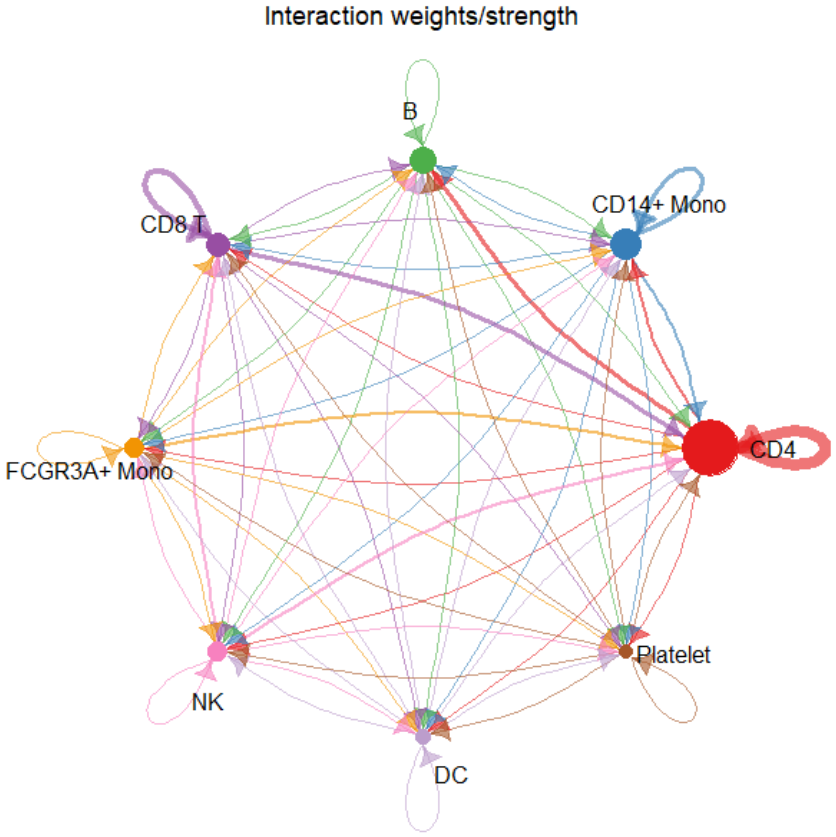
信号通路水平的细胞通讯网络

CellChat使用

（4）细胞相互关系可视化：细胞互作数量与强度统计分析



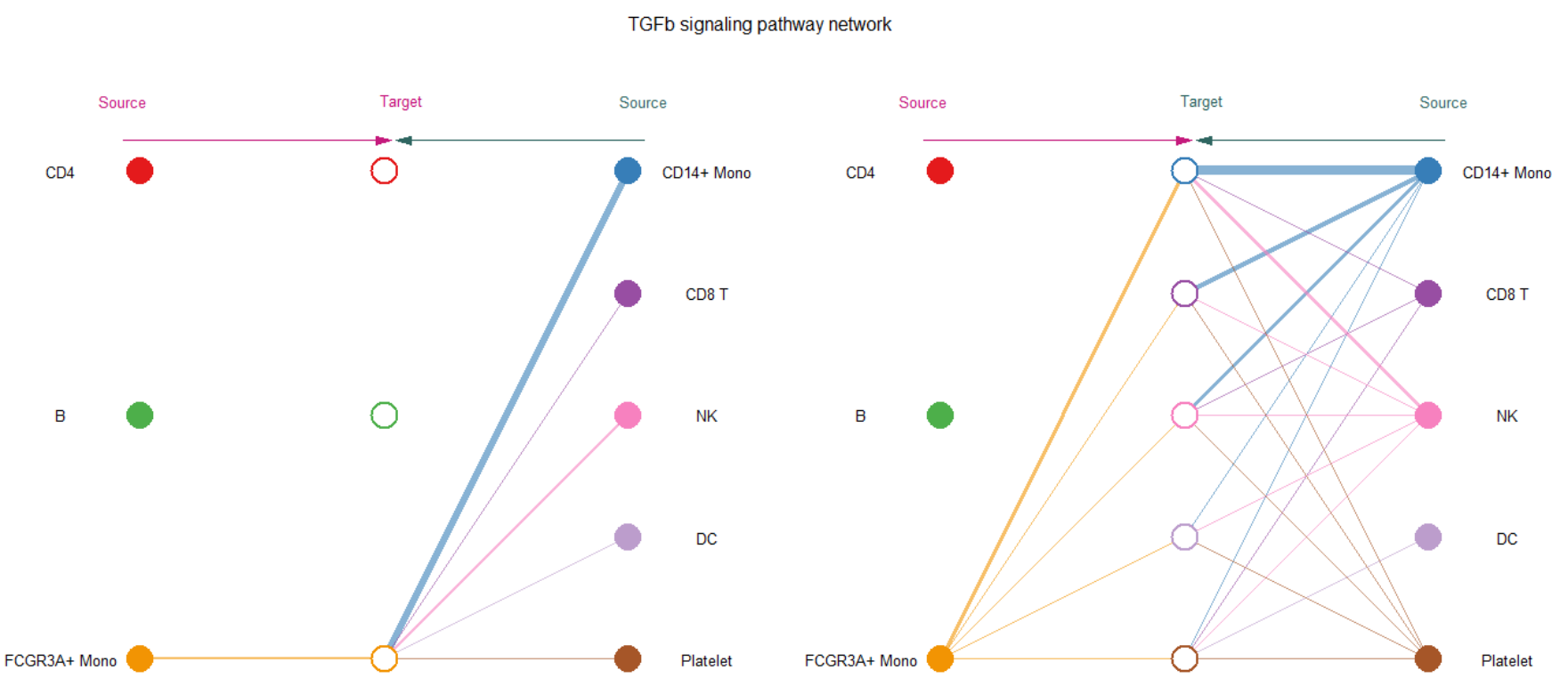
外周各种颜色圆圈的大小表示细胞的数量，圈越大，细胞数越多。发出箭头的细胞表达配体，箭头指向的细胞表达受体。配体-受体对越多，线越粗。

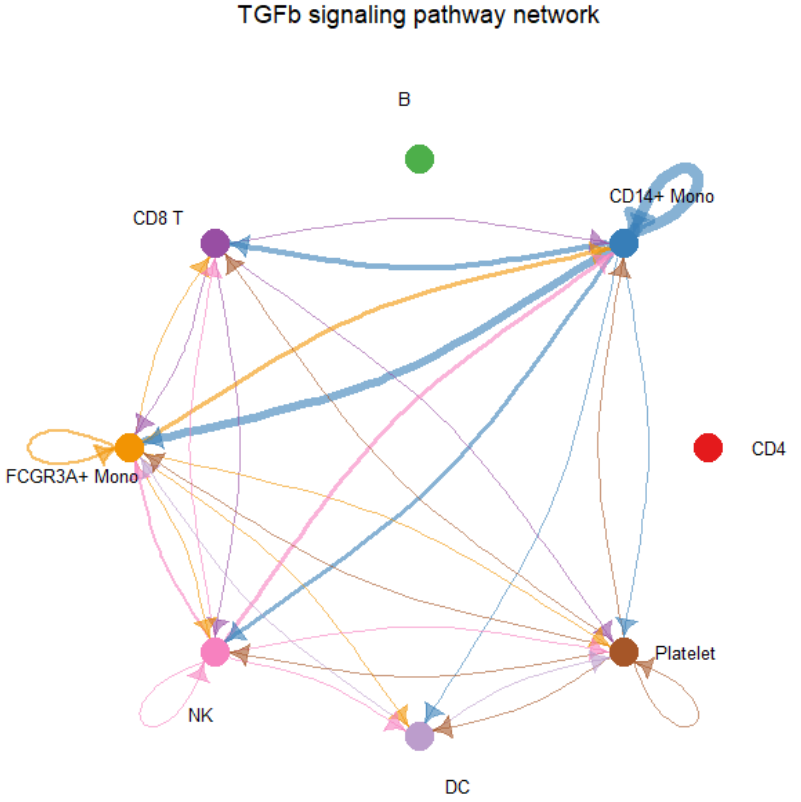


互作的概率/强度值（强度就是概率值相加）

CellChat使用

（4）细胞相互关系可视化：单个信号通路或配体-受体介导的细胞互作可视化（层次图、网络图）



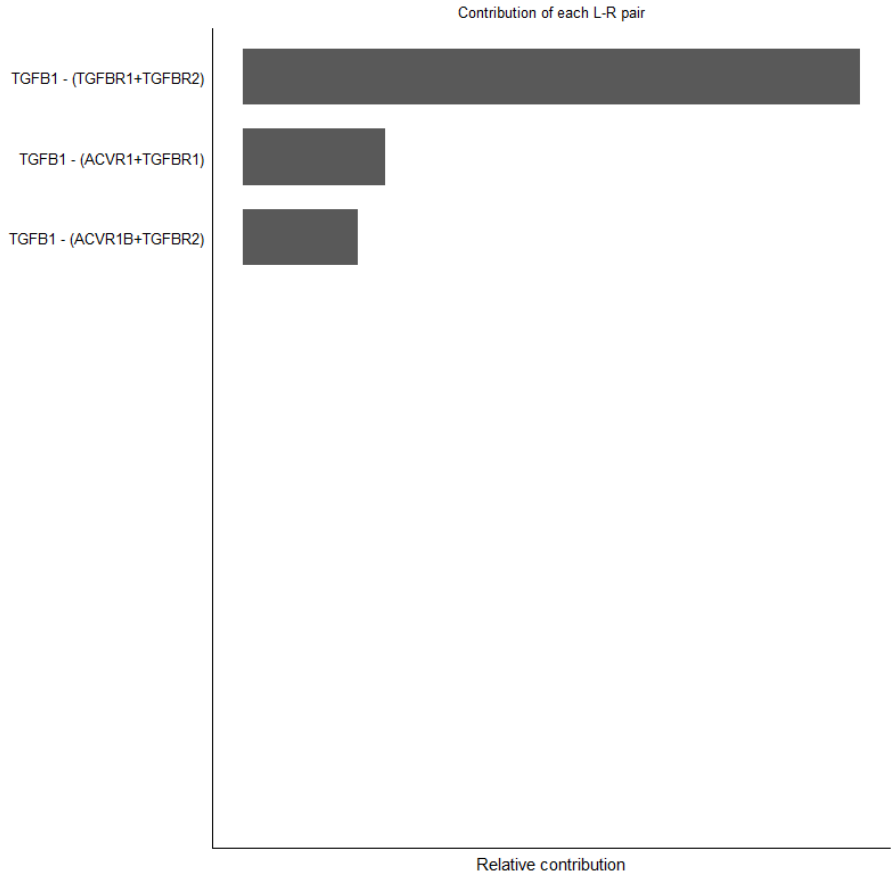


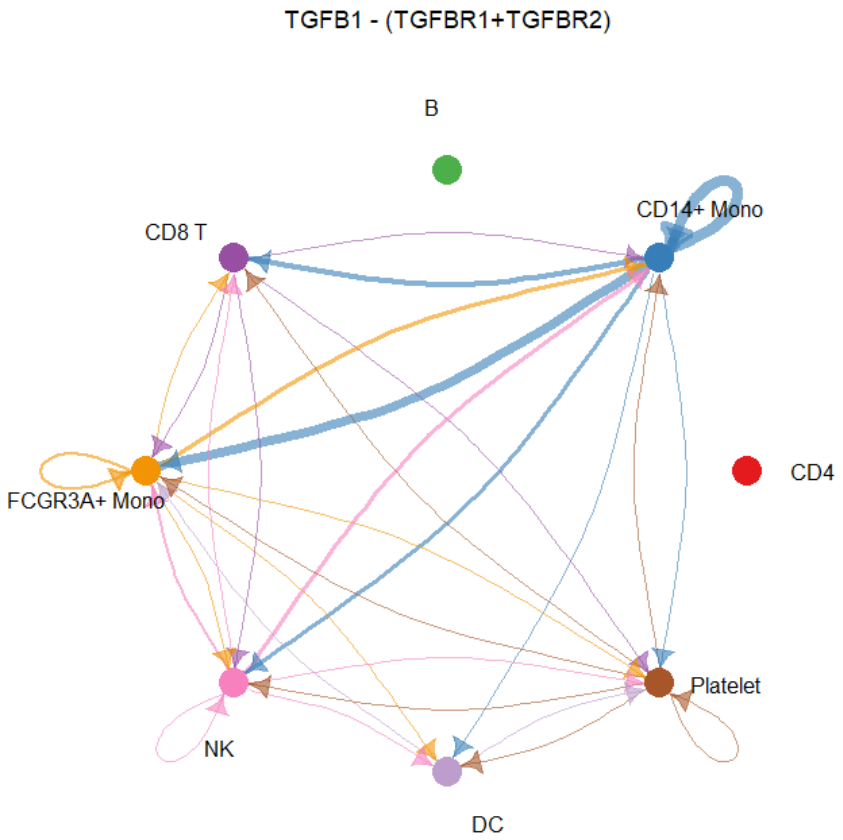
- 在层次图中，实体圆和空心圆分别表示源和目标
- 圆的大小与每个细胞组的细胞数成比例
- 线越粗，互作信号越强
- 左图中间的target是我们选定的靶细胞
- 右图是选中的靶细胞之外的另外一组放在中间看互作

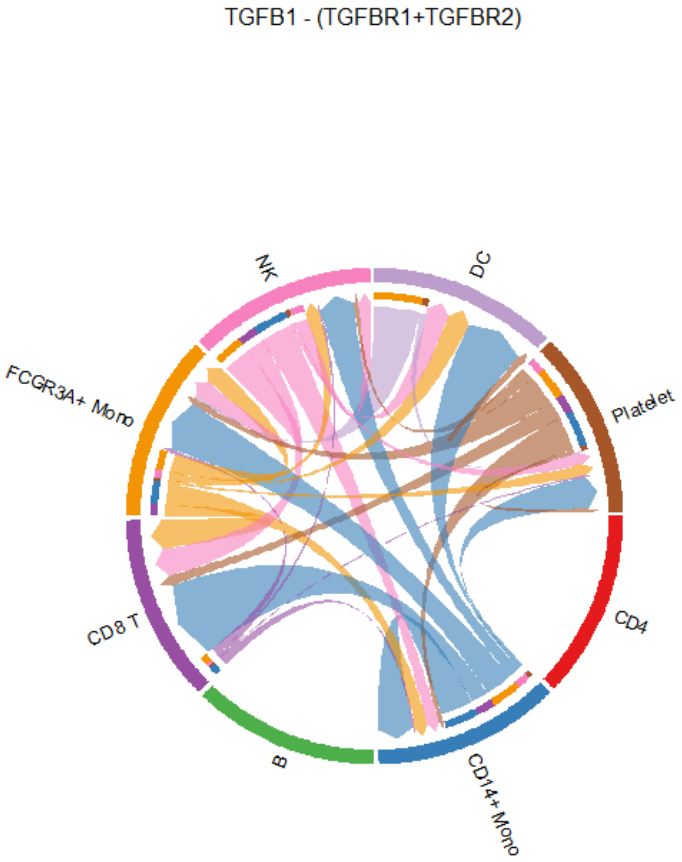
- 网络图

CellChat使用

（4）细胞相互关系可视化：配体-受体层级的可视化(计算各个配体-受体对对信号通路的贡献)



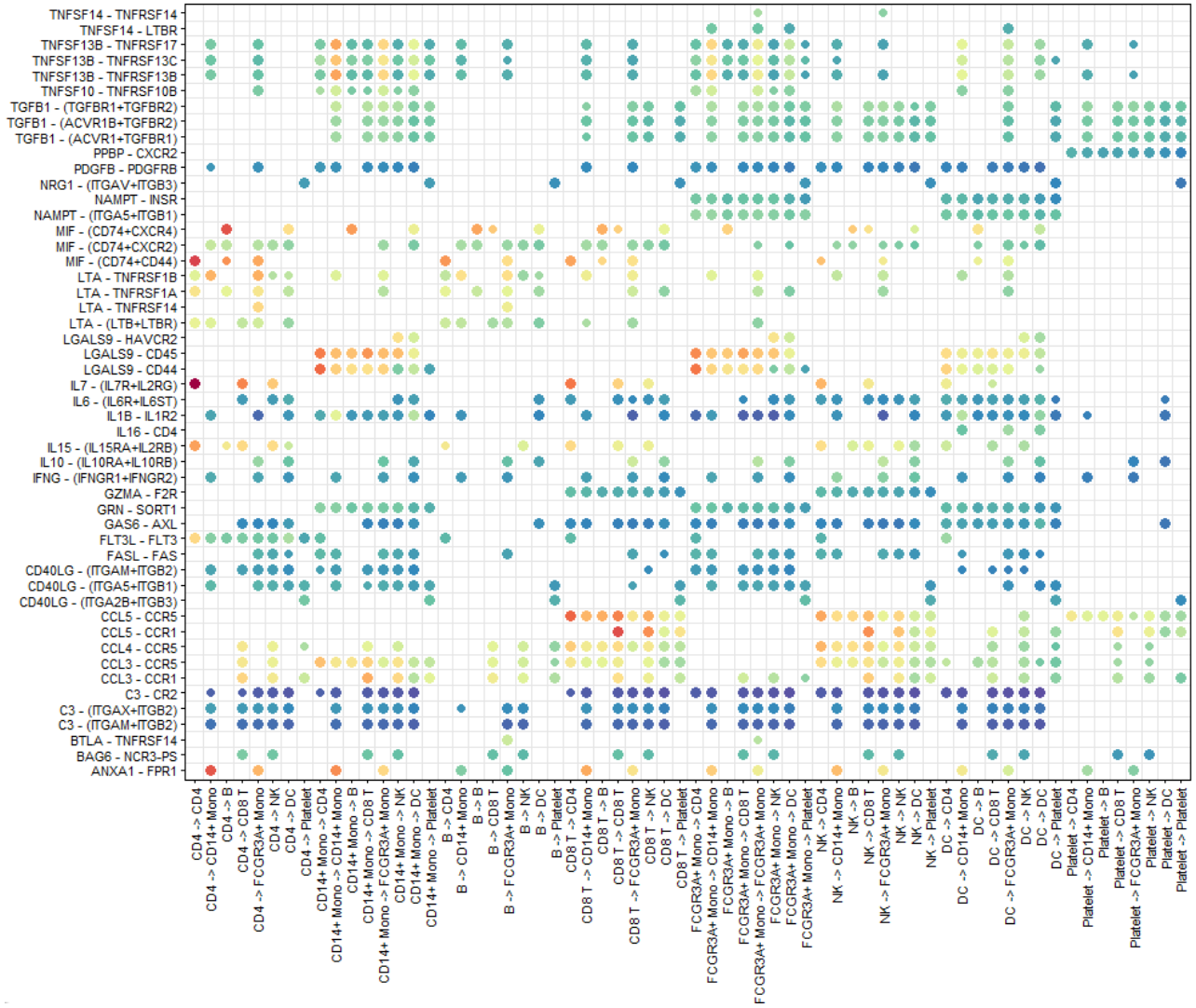


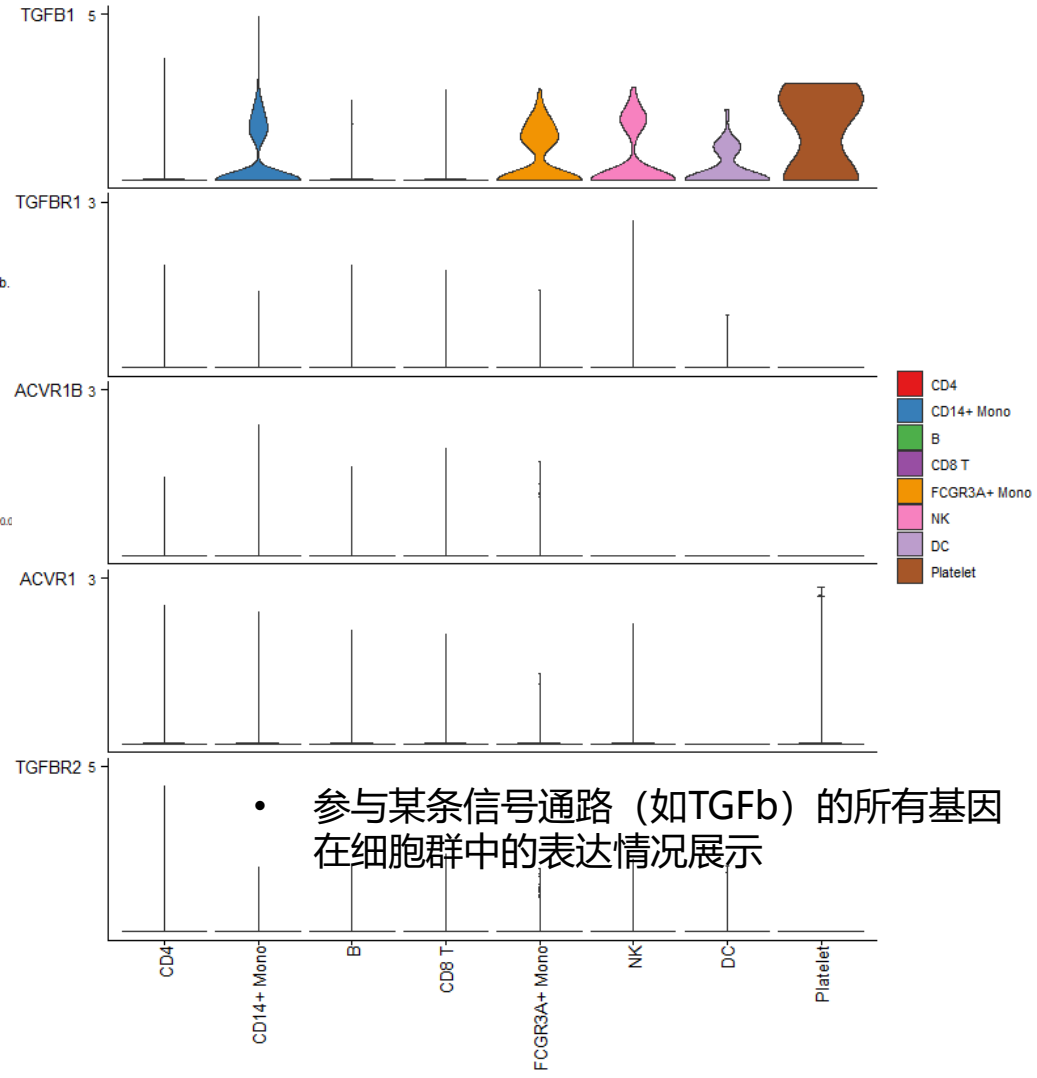


- 提取对TGFb通路有贡献的所有配体受体对
- 展示对TGFb通路贡献最大的配体受体对

CellChat使用

（4）细胞相互关系可视化：多个配体-受体介导的细胞互作关系可视化





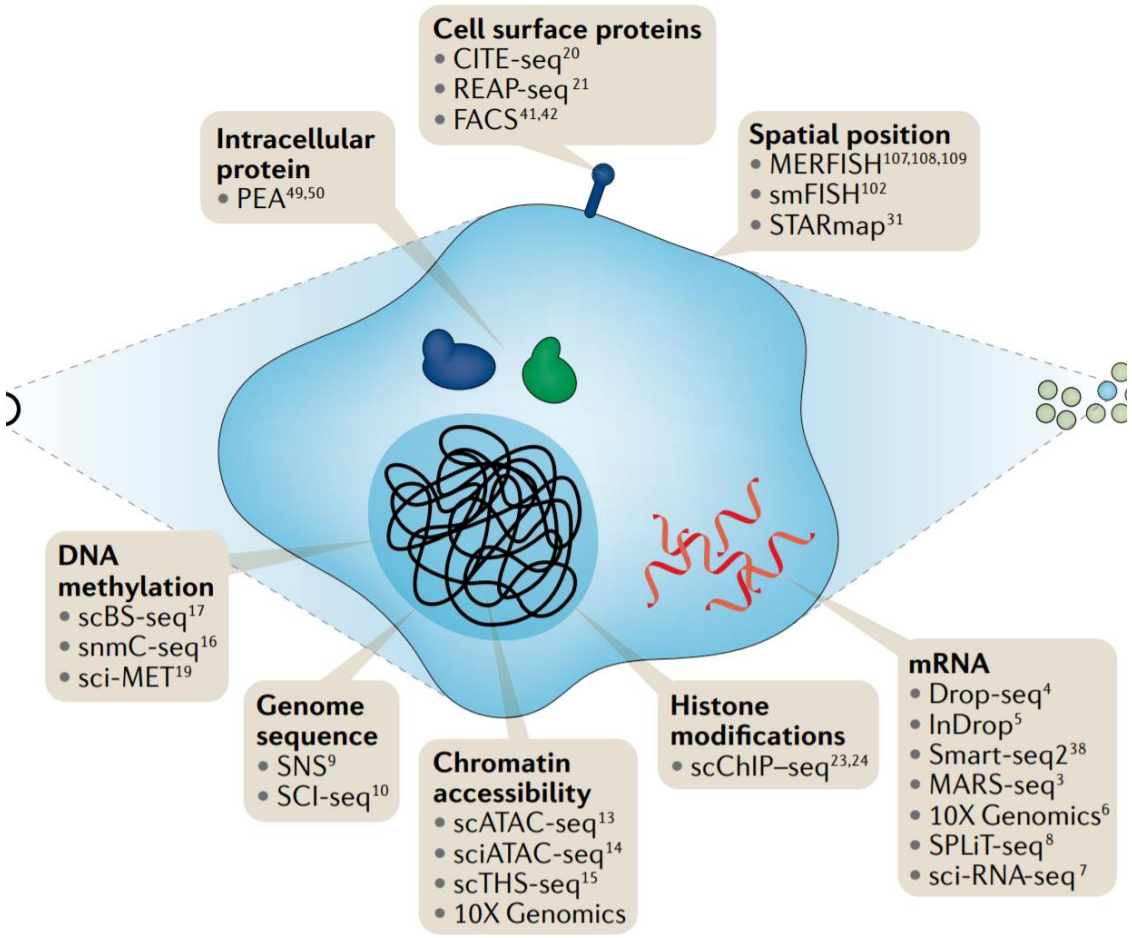
- 参与某条信号通路（如TGFb）的所有基因在细胞群中的表达情况展示

本节小结

- 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
- ……

拓展：单细胞多组学

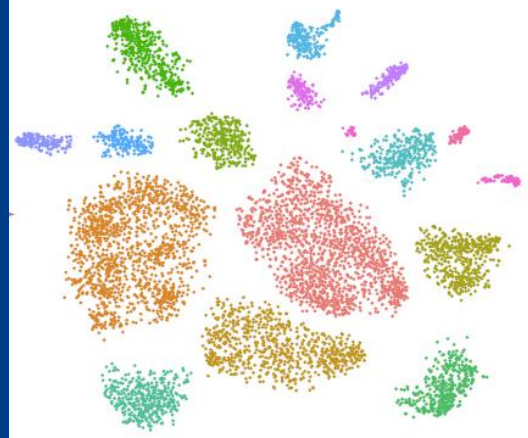
- 不同组学的单细胞测序技术



Data types	Method name	Feature throughput	Cell throughput
Unimodal			
mRNA	Drop-seq	Whole transcriptome	1,000–10,000
	InDrop	Whole transcriptome	1,000–10,000
	10X Genomics	Whole transcriptome	1,000–10,000
	Smart-seq2	Whole transcriptome	100–300
	MARS-seq	Whole transcriptome	100–300
	CEL-seq	Whole transcriptome	100–300
	SPLiT-seq	Whole transcriptome	≥50,000
	sci-RNA-seq	Whole transcriptome	≥50,000
Genome sequence	SNS	Whole genome	10–100
	SCI-seq	Whole genome	10,000–20,000
Chromatin accessibility	scATAC-seq	Whole genome	1,000–2,000
	sciATAC-seq	Whole genome	10,000–20,000
	scTHS-seq	Whole genome	10,000–20,000
DNA methylation	scBS-seq	Whole genome	5–20
	snmC-seq	Whole genome	1,000–5,000
	sci-MET	Whole genome	1,000–5,000
	scRRBS	Reduced representation genome	1–10
Histone modifications	scChIP–seq	Whole genome + single modification	1,000–10,000
Chromosome conformation	scHi-C-seq	Whole genome	1–10
Multimodal			
Histone modifications + spatial	NA	Single locus + single modification	10–100
mRNA + lineage	scGESTALT	Whole transcriptome	1,000–10,000
	ScarTrace	Whole transcriptome	1,000–10,000
	LINNAEUS	Whole transcriptome	1,000–10,000
Lineage + spatial	MEMOIR	NA	10–100
mRNA + spatial	osmFISH	10–50 RNAs	1,000–5,000
	STARmap	20–1,000 RNAs	100–30,000
	MERFISH	100–1,000 RNAs	100–40,000
	seqFish	125–250 RNAs	100–20,000
mRNA + cell surface protein	CITE-seq	Whole transcriptome + proteins	1,000–10,000
	REAP-seq	Whole transcriptome + proteins	1,000–10,000
mRNA + chromatin accessibility	sci-CAR	Whole transcriptome + whole genome	1,000–20,000
mRNA + DNA methylation	scM&T-seq	Whole genome	50–100
mRNA + genomic DNA	G&T-seq	Whole genome + whole transcriptome	50–200
mRNA + intracellular protein	NA	96 mRNAs + 38 proteins	50–100
		82 mRNAs + 75 proteins	50–200
DNA methylation + chromatin accessibility	scNOMe-seq	Whole genome	10–20

Tim Stuart and Rahul Satija, Nature Reviews, 2019

浙江大学
ZHEJIANG UNIVERSITY



单细胞转录组数据分析

——基于细胞类型的高级分析

褚琴洁　qinjiechu@zju.edu.cn

2023年10月30日