

单细胞基础培训

生信帮

云南农业大学

云南省生物大数据重点实验室

2023年8月12日

目录



雲南農業大學
Yunnan Agricultural University

云南省生物大数据重点实验室



生信帮

第一章

单细胞简介

第二章

单细胞数据基础分析

第三章

单细胞数据高级分析



雲南農業大學
Yunnan Agricultural University

云南省生物大数据重点实验室



生信帮

第一章 单细胞简介

单细胞技术的发展史



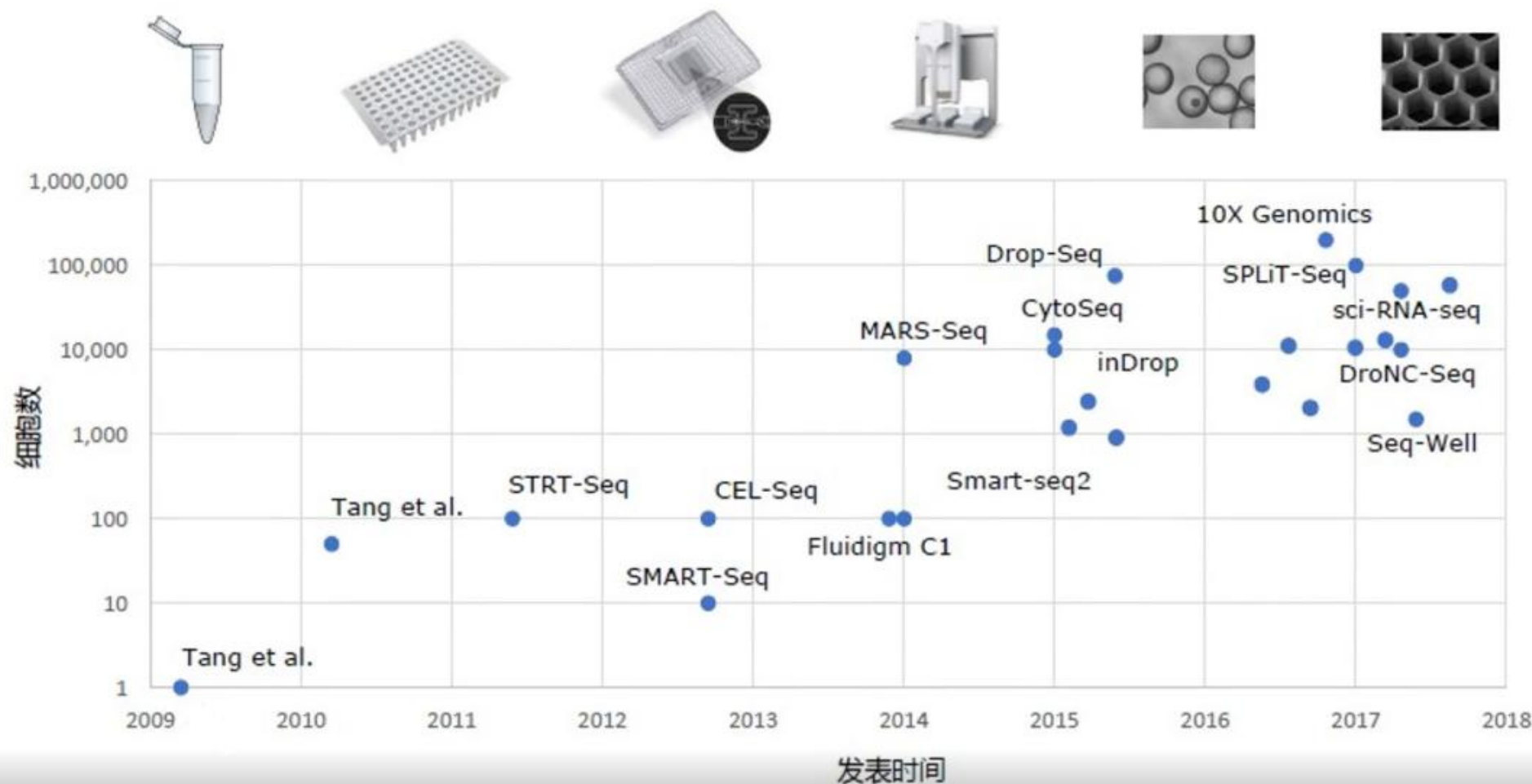
雲南農業大學
Yunnan Agricultural University

云南省生物大数据重点实验室



生信帮

2009汤富酬老师发表第一篇单细胞文章；2013 Nature Methods年度技术；2018 Science年度突破技术；2019 Nature Methods年度技术；2020 空间转录组Nature Methods年度技术



13种单细胞技术对比

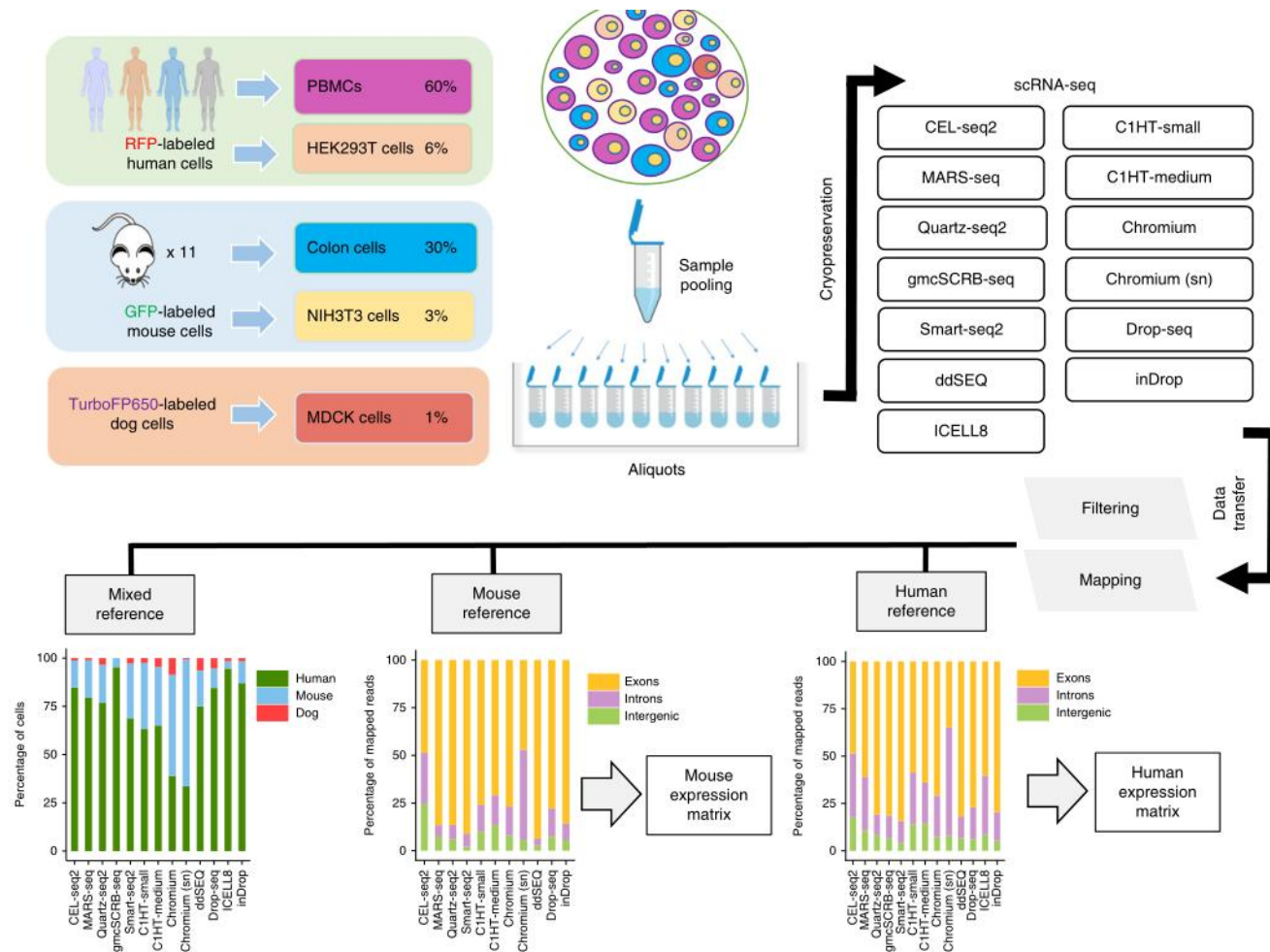


雲南農業大學
Yunnan Agricultural University



生信帮

云南省生物大数据重点实验室



13种单细胞RNA测序方法分别为：CEL-Seq2、MARS-Seq、Quartz-Seq2、gmcSCRB-Seq、Smart-Seq2、ddSEQ、ICELL8、C1HT-Small、C1HT-Medium、Chromium、Chromium(sn)、Drop-seq以及inDrop。

单细胞技术流程



雲南農業大學
Yunnan Agricultural University

云南省生物大数据重点实验室



生信帮

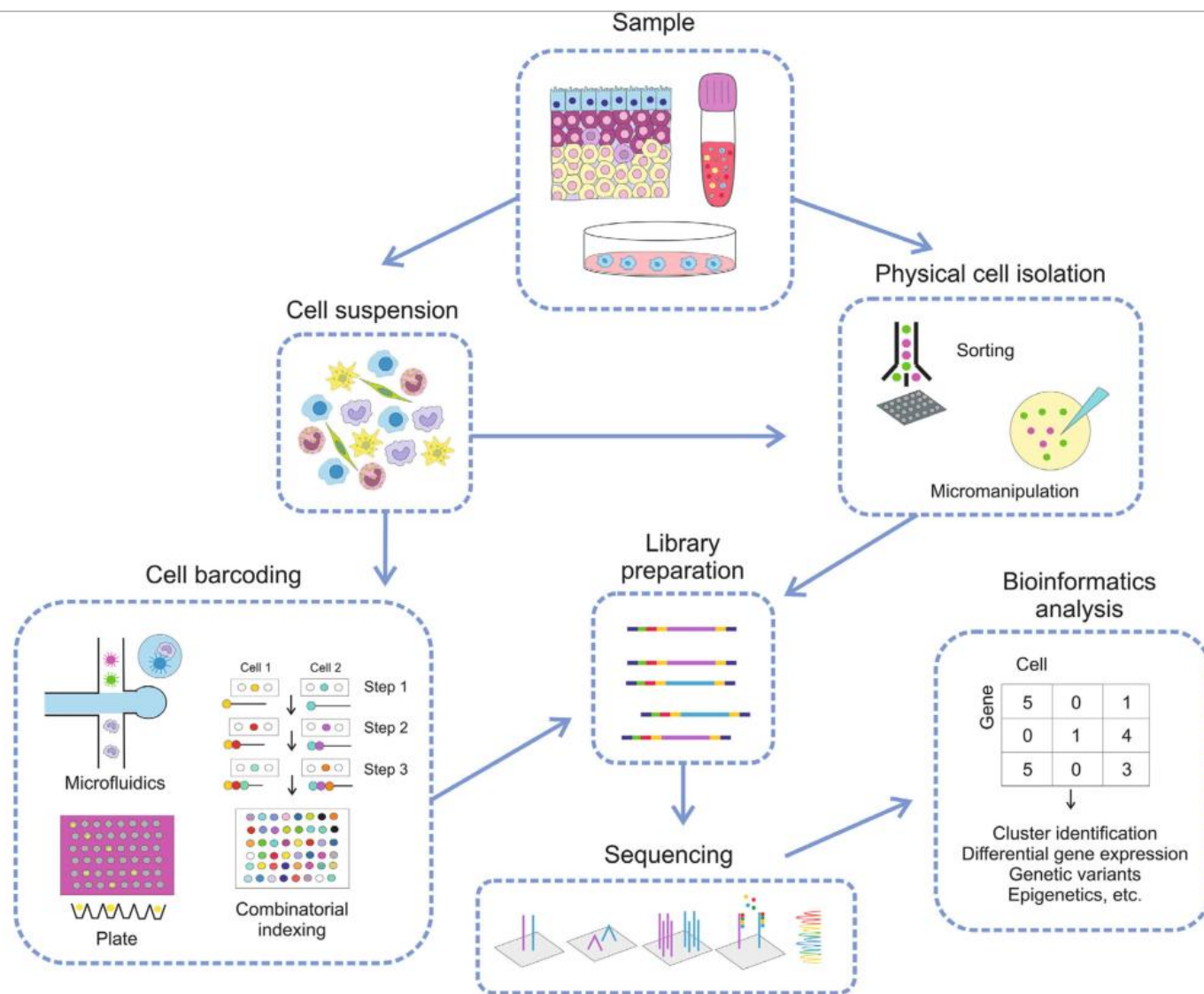


Fig. 1. Typical workflow for scRNA-seq.

主要过程包括单细胞分离、单细胞裂解、核酸扩增、高通量测序、数据处理和数据分析。

- 分离单细胞样品的方法包括有限稀释、显微操作、激光捕获显微切割（Laser capture microdissection, LCM）、荧光激活细胞分选（fluorescence activated cell sorters, FACS）和微流控技术
- 常见的细胞裂解方法，如物理、化学和酶法
- 测序方法：
 - 只允许对转录组的 5' 和 3' 末端进行测序
 - Smart-seq2 支持 cDNA 测序的全读长覆盖
 - 基于微孔板的 scRNA-seq 技术
 - 基于微流控的 scRNA-seq 方法

snRNA-seq Vs scRNA-seq



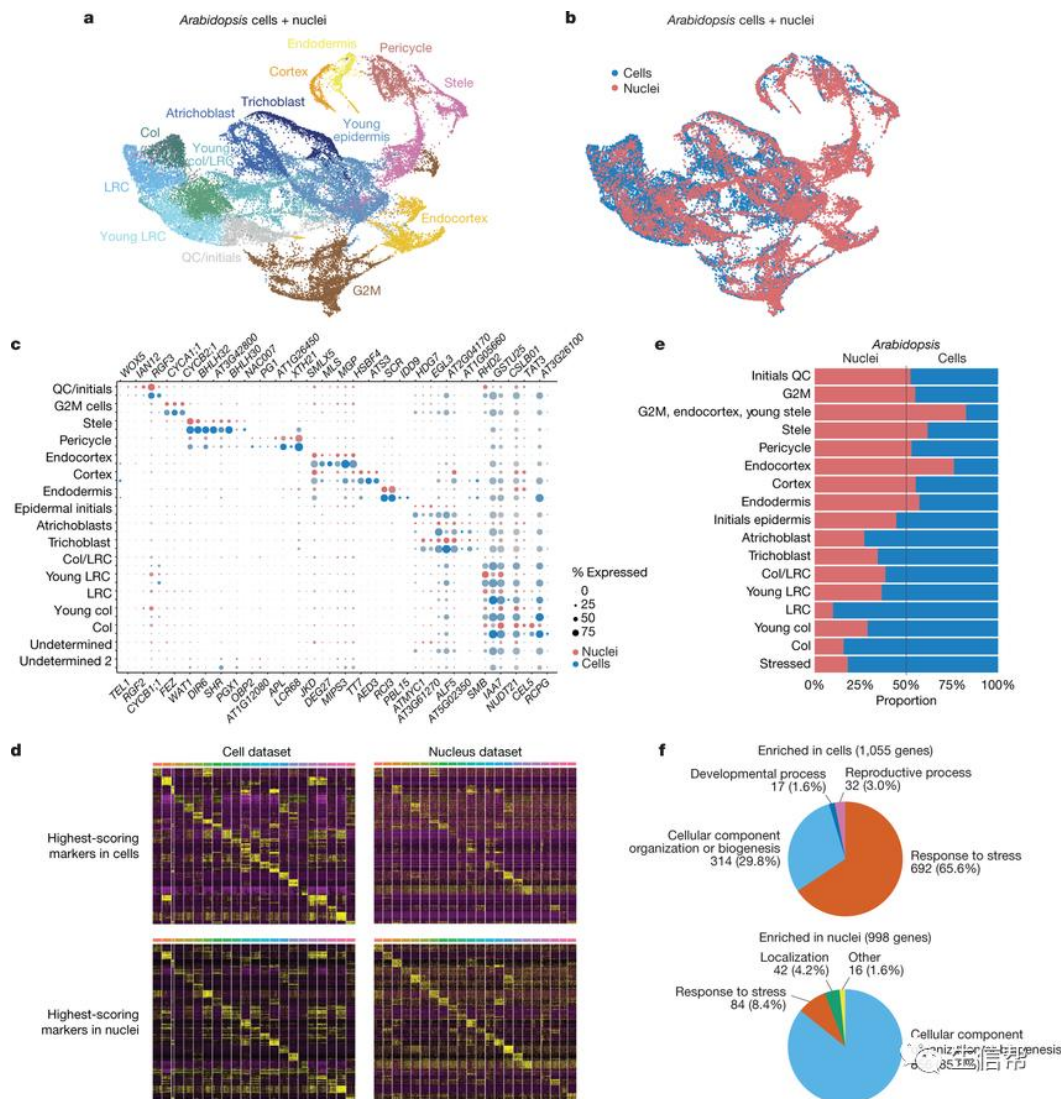
雲南農業大學
Yunnan Agricultural University



生信帮

云南省生物大数据重点实验室

- 检测到的基因平均数量在细胞分析中比核分析高2.7倍（拟南芥）和1.4倍（玉米），细胞核文库要求更多细胞数量，才能得到更精细的细胞类型。
- 核分析中的mRNA含量较低，但核分析仍检测到了细胞分析中的89%（拟南芥）和88%（玉米）的总基因数
- 细胞和核分析均能生成与主要细胞类型相对应的UMAP聚类
- 在拟南芥中，我们发现总基因数的14%（3218个基因）在细胞文库和核文库之间的聚类分析中呈差异表达。
- 细胞文库中显示了较高比例的与应激相关的基因。
- 细胞文库和核文库之间的大部分差异似乎与RNA在细胞器中的稳定性有关。例如，在核文库中富集的mRNA与已知在细胞质文库中具有较高降解速率的转录本显著重叠



10X单细胞简介

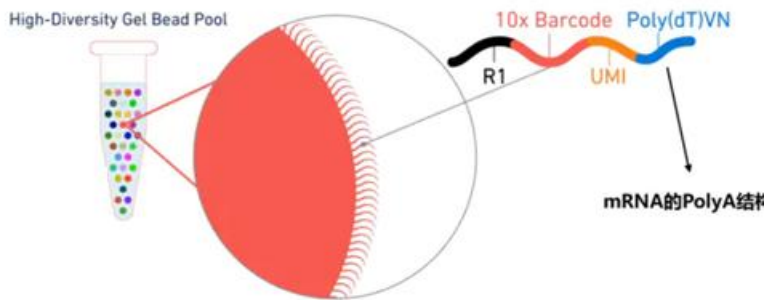
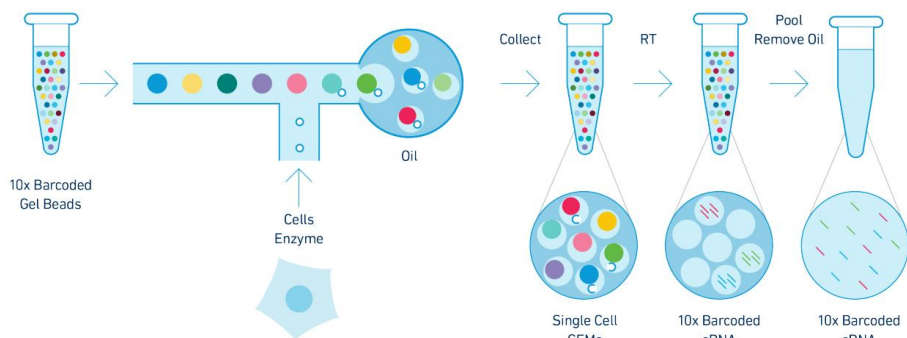


雲南農業大學
Yunnan Agricultural University

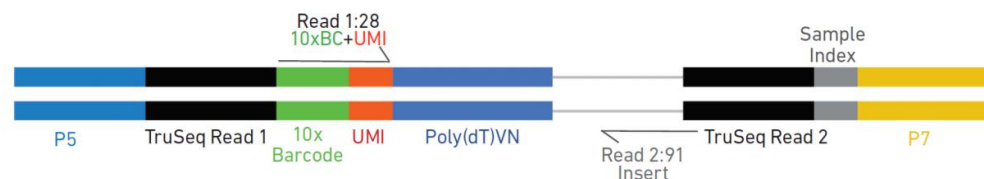


云南省生物大数据重点实验室

生信帮



Chromium Single Cell 3' Gene Expression Library



1. 导入凝胶微珠。
2. 第一纵向道输入细胞。
3. 第二纵向道输入油（理想情况就是一个油滴包裹一个凝胶微珠，一个凝胶微珠吸附一个细胞）。
4. 细胞裂解，利用3'端poly(A)碱基互补原理特定抓取mRNA构建转录文库。
5. 库检合格后，使用Illumina平台进行测序100G。
6. 数据产出：

R1_001.fastq.gz #barcod-umi信息

R2_001.fastq.gz #转录本序列

V2版本：26bp+98bp（16bp 10XBC+10bp UMI+98 bp read2）

V3版本：28bp+91bp（16bp 10XBC+12bp UMI+91 bp read2）

注：

一个油滴 (GEM)=一个单细胞+一个凝胶微珠=一个scRNA-Seq。)

350万种Barcoded beads, 100万种UMI序列。测序后，Barcode标记细胞，UMI标记转录本并用于后续定量。

10X单细胞简介



雲南農業大學
Yunnan Agricultural University



云南省生物大数据重点实验室

生信帮

其技术核心是油滴包裹的凝胶珠（GEM），该系统有GEM形成后，细胞裂解，凝胶珠自动溶解释放大量barcode序列，随后 mRNA逆转录产生带Barcode和UMI信息的cDNA。油滴破碎，cDNA为模板进行PCR扩增，然后进行cDNA打断、加测序接头P5及测序引物R1等传统二代测序的建库过程。

➤ 优势

- 平台成本低，通量高；

➤ 缺点

- 对细胞总量要求较高；
- 为3'测序，基因检测率不及SMART-seq等全长测序；
- 操作步骤较多。

➤ 适用范围

- 直径 $<20\mu\text{m}$ 的细胞（受磁珠结合效率限制）。
- 大规模细胞样本，适用于分群、谱系等研究。

单细胞生物信息学分析内容



雲南農業大學
Yunnan Agricultural University

云南省生物大数据重点实验室



生信帮



Fig. 2. Approaches in bioinformatics analysis of scRNA-seq data.

植物单细胞研究

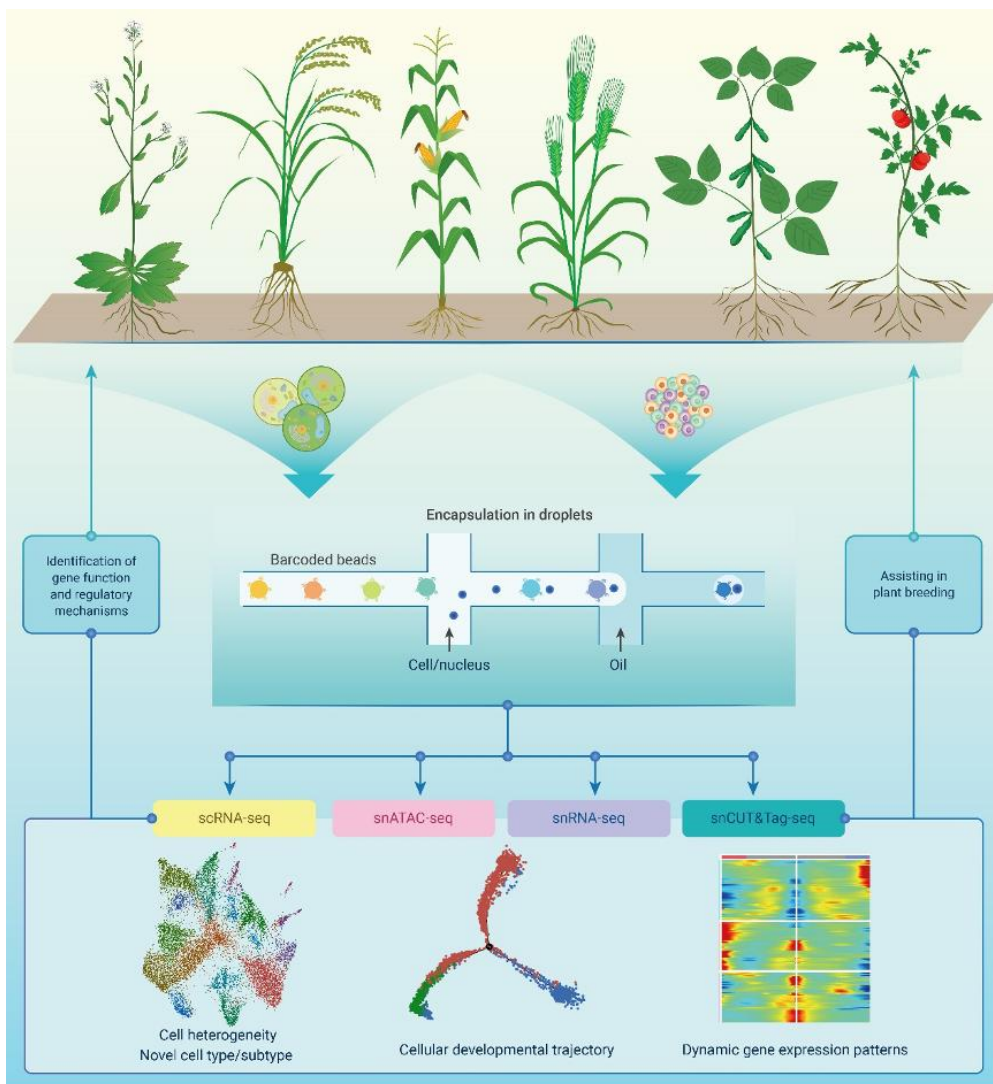


云南农业大学
Yunnan Agricultural University



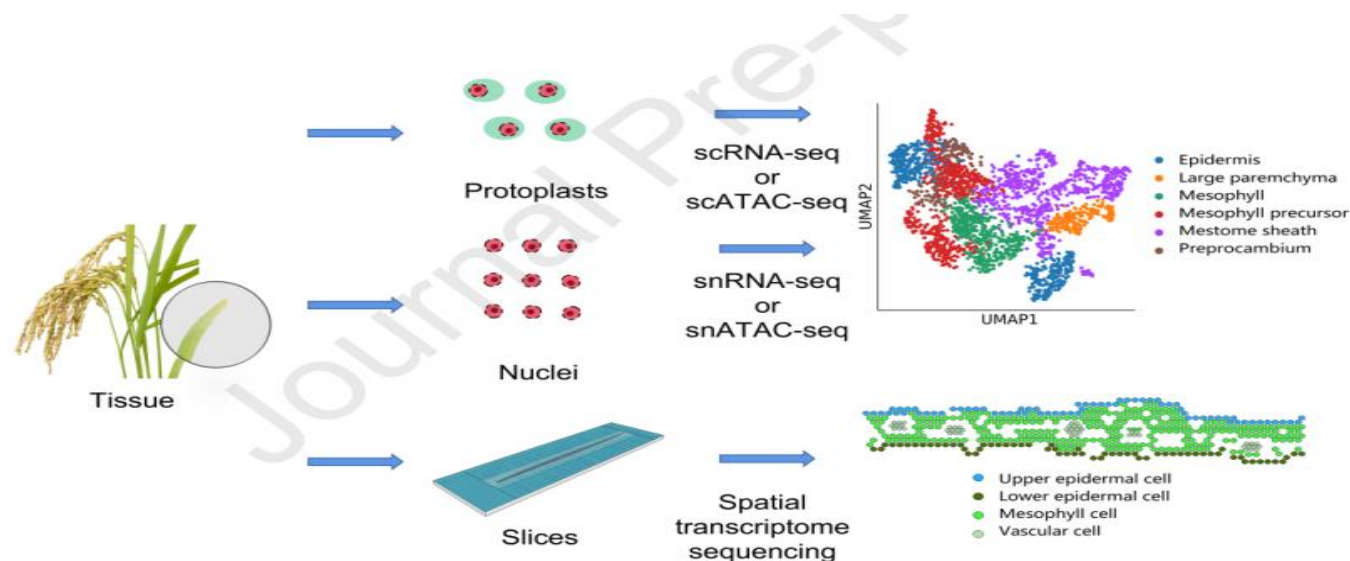
云南省生物大数据重点实验室

生信帮



植物单细胞研究起步相对较晚，发表的文献也比较少，植物单细胞研究主要集中在揭示细胞类型、基因表达谱、细胞分化和发育、细胞相互作用等方面。植物单细胞研究能帮助了解植物多样性和复杂性，推动植物科学和农业领域的研究进展。

植物单细胞研究目前的难点在于专门针对植物数据分析工具较少，细胞类型marker不明确，相应的分析数据库比较少等。



植物单细胞文献发表数量

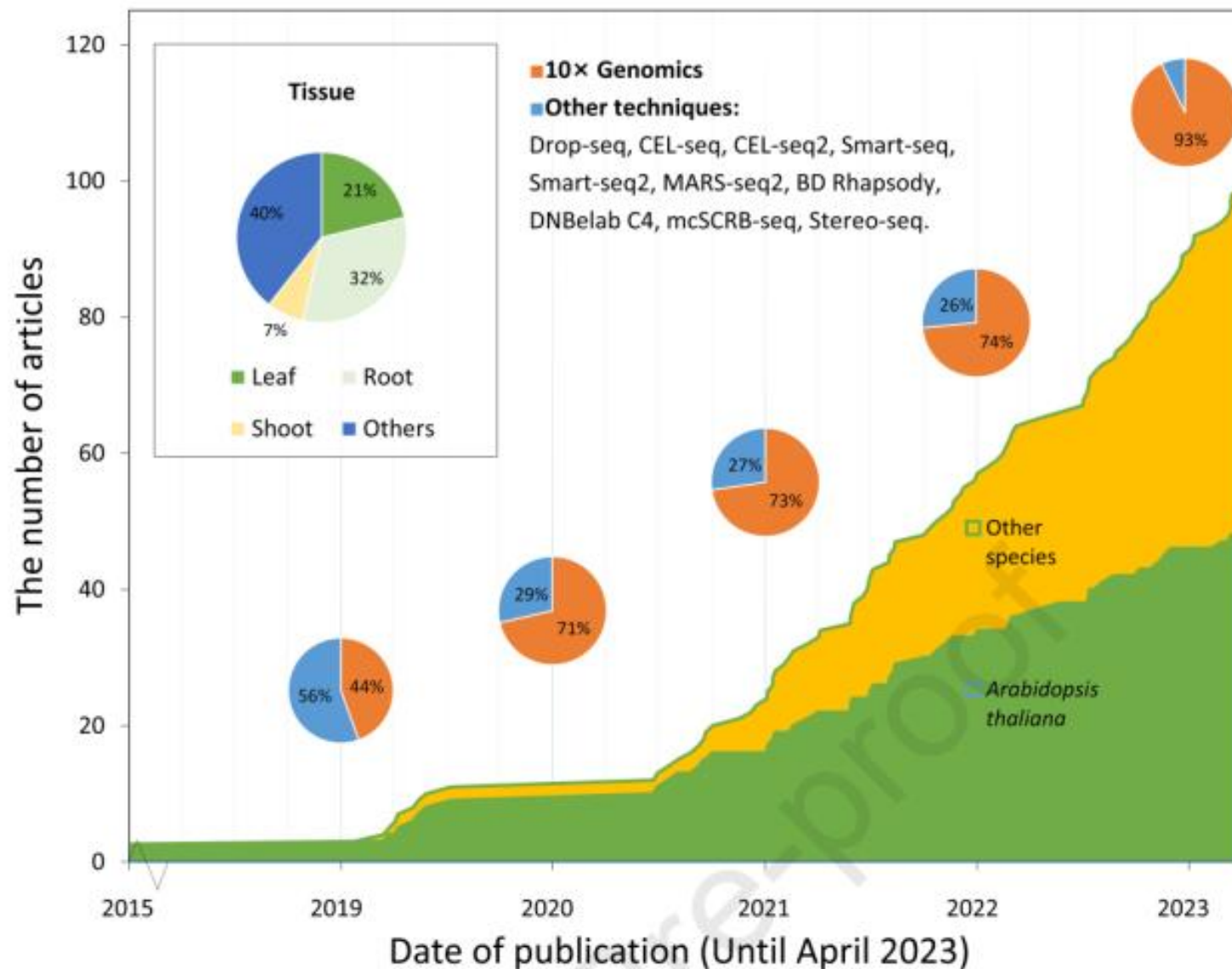


雲南農業大學
Yunnan Agricultural University

云南省生物大数据重点实验室



生信帮





雲南農業大學
Yunnan Agricultural University

云南省生物大数据重点实验室



生信帮

第二章 单细胞基础分析

单细胞分析流程图

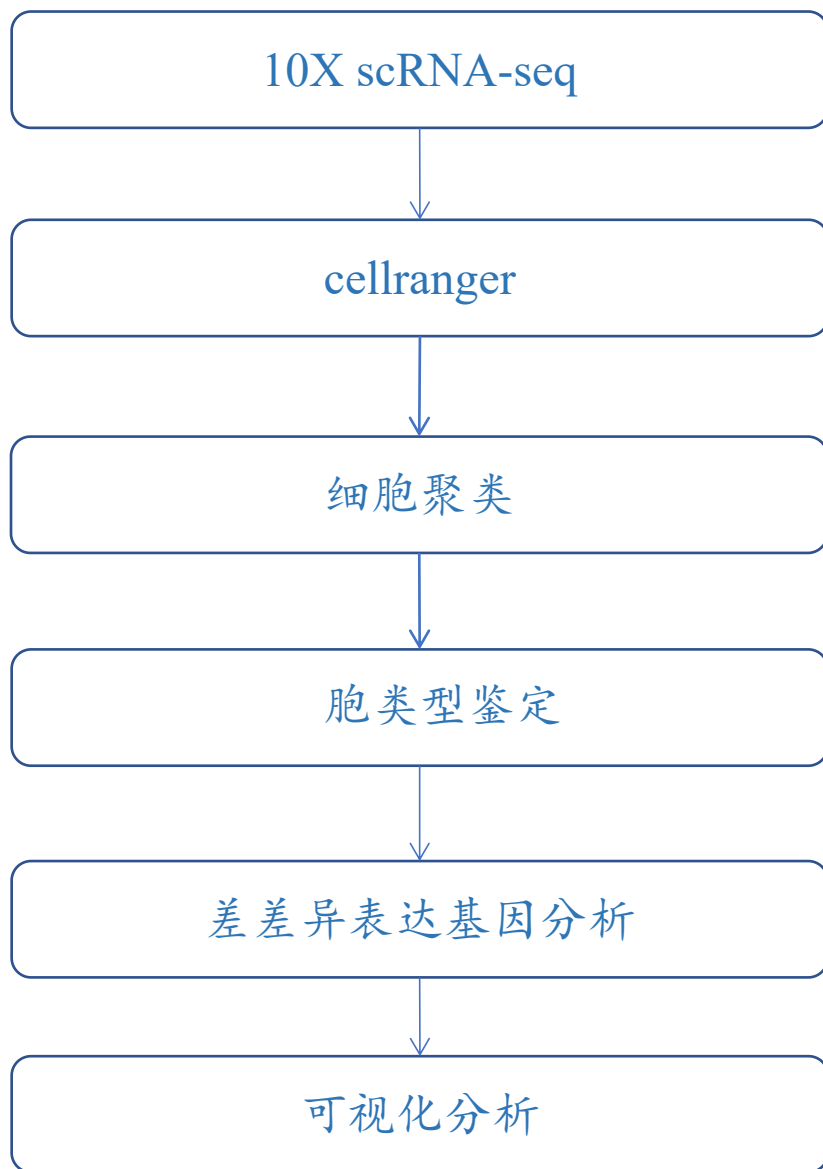


雲南農業大學
Yunnan Agricultural University



云南省生物大数据重点实验室

生信帮



1、数据预处理：

- 质量控制：对原始测序数据进行质量评估和过滤，去除低质量的细胞和读取。
- 读取对齐：将清洗后的测序读取与基因组参考序列进行比对，以确定每个细胞的基因表达水平。
- 特征选择：选择具有较高表达变异性的基因作为分析的特征基因。

2、细胞聚类：

- 降维：使用降维方法（如主成分分析（PCA）或流形学习）将高维度的数据降维到二维或三维空间。
- 聚类：应用聚类算法（如K均值聚类或层次聚类）将降维后的细胞划分为不同的细胞群集。

3、细胞类型鉴定：

- 基于参考数据集：将每个细胞的基因表达模式与已知的细胞类型的参考数据集进行比较，以确定细胞类型。
- 基于基因标记：利用已知的细胞类型特异性基因标记，判断细胞所属的类型。

4、差异表达基因分析：

- 基于群集：对每个细胞群集进行差异表达分析，找出在不同细胞群集中具有显著差异表达的基因。
- 单一细胞水平：对每个细胞进行差异表达分析，鉴定在不同条件（如不同时间点或处理）下发生显著变化的基因。

5、可视化和解释分析结果：

- 绘制降维图：利用可视化工具（如t-SNE、UMAP或PCA）将细胞群集在二维或三维空间中可视化，展示不同细胞类型和差异表达基因的分布。
- 功能富集分析：根据差异表达基因结果，进行功能富集分析，以确定在特定细胞类型或条件下与特定功能或通路相关的基因集合。

1. 单细胞数据定量之CellRanger



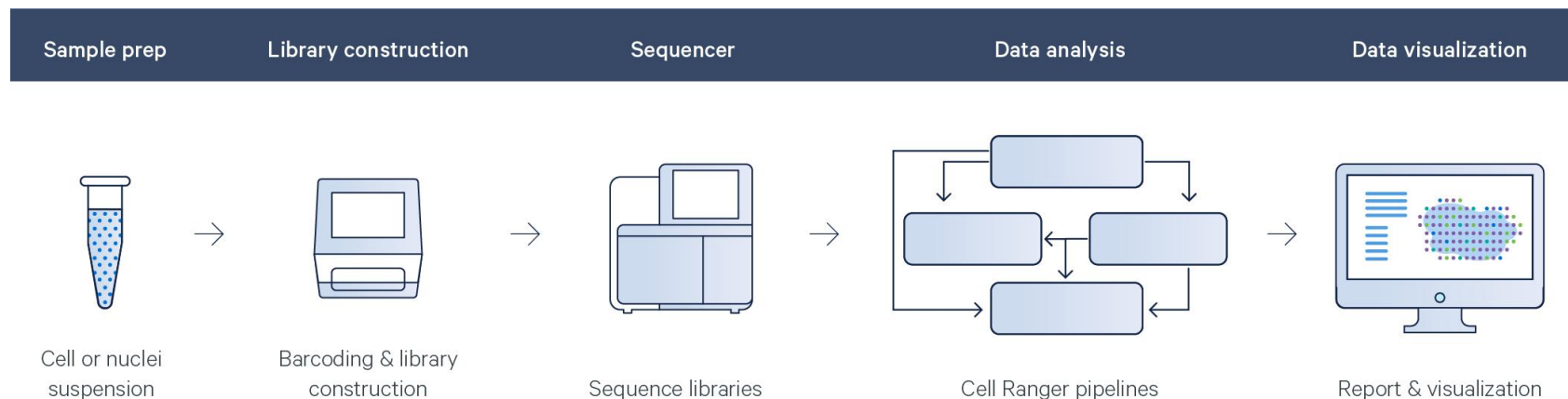
雲南農業大學
Yunnan Agricultural University



生信帮

云南省生物大数据重点实验室

完成测序后使用10X官方的软件CellRanger对单细胞数据进行初步的过滤和定量



1. 单细胞数据定量之CellRanger



雲南農業大學
Yunnan Agricultural University

云南省生物大数据重点实验室



生信帮

CellRanger的结果

Outputs:

- Run summary HTML:
- Run summary CSV:
- BAM:
- BAM index:
- Filtered gene-barcode matrices MEX:
- Filtered gene-barcode matrices HDF5:
- Unfiltered gene-barcode matrices MEX:
- Unfiltered gene-barcode matrices HDF5:
- Secondary analysis output CSV:
- Per-molecule read information:
- Loupe Cell Browser file:

1. 其中web_summary.html是分析结果的网页报告，包含质控结果和cellranger的聚类结果等，主要评估样本测序质量
2. Filtered_feature_bc_matrix 文件目录下是过滤后的表达量结果，有三个文件分别存储表达量，feature，barcode信息，后续分析使用
3. Filtered_feature_bc_matrix.h5 h5格式，内容和Filtered_feature_bc_matrix 内容相同

1. 单细胞数据定量之CellRanger



雲南農業大學
Yunnan Agricultural University



生信帮

云南省生物大数据重点实验室

Summary

Gene Expression

报告包含的两部分内容

11,178

Estimated Number of Cells

细胞数，
每个细胞的reads数

35,712

Mean Reads per Cell

3,954

Median Genes per Cell

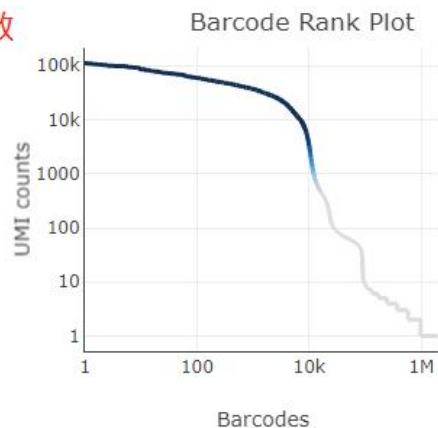
Sequencing

这个问号可以点击，产看说明

Number of Reads	399,192,322
Number of Short Reads Skipped	0
Valid Barcodes	98.3%
Valid UMIs	100.0%
Sequencing Saturation	34.2%
Q30 Bases in Barcode	96.6%
Q30 Bases in RNA Read	92.0%
Q30 Bases in UMI	95.5%

测序质控结果

Cells



Estimated Number of Cells	11,178
Fraction Reads in Cells	92.9%
Mean Reads per Cell	35,712
Median UMI Counts per Cell	13,112
Median Genes per Cell	3,954
Total Genes Detected	24,874

右侧折线图：

Y轴是每个细胞中UMI的值，X轴是单个细胞的按照UMI大小的排序（降序），所以这个图中的曲线是下降的趋势。蓝色的线是选取的细胞（和**2\ 细胞和基因数的统计**中的细胞数是一致的），灰色的线是背景。

正常的数据来说会有两个下降的趋势，第1个下降的趋势：区分完整细胞和背景物质（因为细胞和其他物质相比，真正细胞中会有更多的UMI，而其他物质可能没有或者由于一些污染能捕获到少量的转录本，所以会出现第一个下降的趋势）；第2个下降的趋势：区分细胞的质量，捕获率低或细胞破碎（这类细胞中基因数会很少，导致UMI数也少），而正常的细胞中UMI多且分布比较接近，所以质量好和不好的细胞在UMI上也会存在很大的差异，随后就出现了第2个下降趋势。

当数据出现了这两个下降趋势，且在蓝色区域的线条比较平稳时，也能说明我们的数据质量好

1. 单细胞数据定量之CellRanger



雲南農業大學
Yunnan Agricultural University

云南省生物大数据重点实验室



生信帮

Mapping ?

和参考基因组比对结果

Reads Mapped to Genome	93.4%
Reads Mapped Confidently to Genome	89.8%
Reads Mapped Confidently to Intergenic Regions	4.6%
Reads Mapped Confidently to Intronic Regions	15.0%
Reads Mapped Confidently to Exonic Regions	70.2%
Reads Mapped Confidently to Transcriptome	79.4%
Reads Mapped Antisense to Gene	5.4%

Sample

样本基本信息，使用的参考基因组，软件版本等

Sample ID	
Sample Description	
Chemistry	Single Cell 3' v3
Include introns	True
Reference Path	...singlecell/10X/refdata-gex-mm10-2020-A
Transcriptome	mm10-2020-A
Pipeline Version	cellranger-7.0.0

1. 单细胞数据定量之CellRanger



雲南農業大學
Yunnan Agricultural University



生信帮

云南省生物大数据重点实验室

Summary

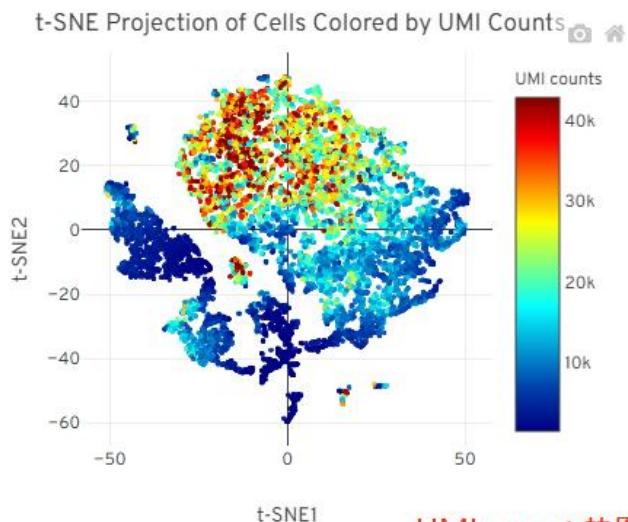
Gene Expression

可选择聚类的方法

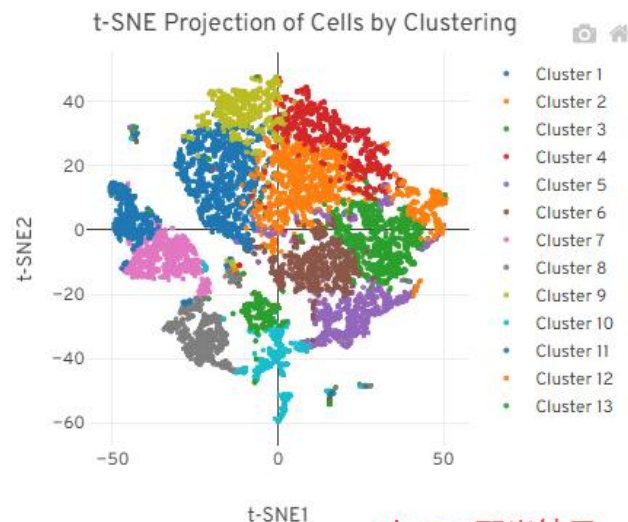
Clustering Type: Graph-based

Cellranger聚类的方式有两种，图（graph-base）聚类和K-means聚类，下方差异分析是每个cluster相对于其他所有cluster做差异分析的结果
UMAP:
TSNE:

t-SNE Projection



UMI_count 热图



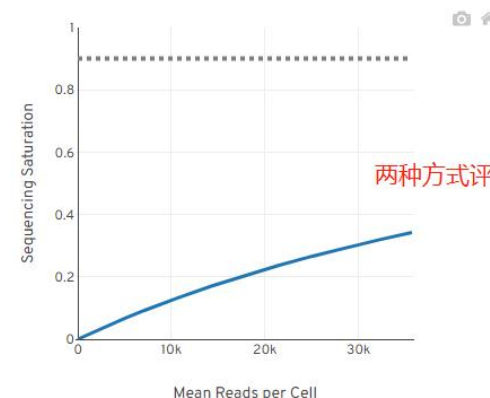
cluster聚类结果

Top Features by Cluster (Log2 fold-change, p-value)

Feature		Cluster 1		Cluster 2		Cluster 3		Cluster 4		Cluster 5		Cluster 6		L2
ID	Name	L2FC	p-value	L2FC	p-value	L2FC	p-value	L2FC	p-value	L2FC	p-value	L2FC	p-value	
ENSMUSG00000007...	Ighg1	3.50	3e-2	-2.73	1e+0	-3.73	1e+0	-2.43	1e+0	-1.97	9e-1	-2.09	9e-1	-
ENSMUSG0000000...	Neil3	2.84	3e-47	-2.39	2e-15	-3.56	1e-12	-2.98	3e-14	-4.01	3e-7	-4.02	1e-9	-
ENSMUSG0000000...	Aspm	2.76	7e-44	-2.79	8e-19	-3.56	1e-12	-2.98	3e-14	-4.01	3e-7	-4.02	1e-9	-
ENSMUSG0000000...	Kif4	2.68	8e-43	-2.35	3e-15	-3.56	1e-12	-2.98	3e-14	-4.01	3e-7	-4.02	1e-9	-
ENSMUSG0000000...	Cenpf	2.64	2e-41	-2.53	2e-17	-3.78	1e-14	-3.21	3e-16	-3.85	3e-7	-3.87	1e-9	-
ENSMUSG0000000...	Iqgap3	2.63	1e-40	-2.48	3e-16	-3.61	6e-13	-2.93	7e-14	-3.21	1e-5	-2.94	2e-6	-
ENSMUSG0000000...	Prc1	2.58	4e-40	-3.10	4e-23	-3.63	8e-14	-3.57	1e-18	-3.82	3e-7	-4.01	3e-10	-
ENSMUSG0000000...	Cit	2.55	5e-38	-2.11	1e-12	-3.19	6e-11	-2.25	2e-9	-3.52	3e-6	-3.75	8e-9	-
ENSMUSG0000000...	Kif11	2.55	1e-38	-2.04	8e-22	-3.19	6e-11	-2.25	2e-9	-3.52	3e-6	-3.75	8e-9	-

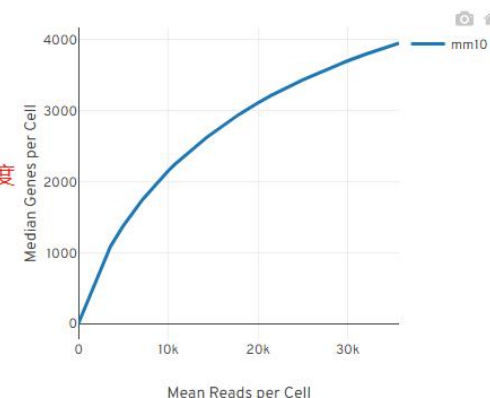
每个基因在所有cluster差异分析结果

Sequencing Saturation



两种方式评估饱和度

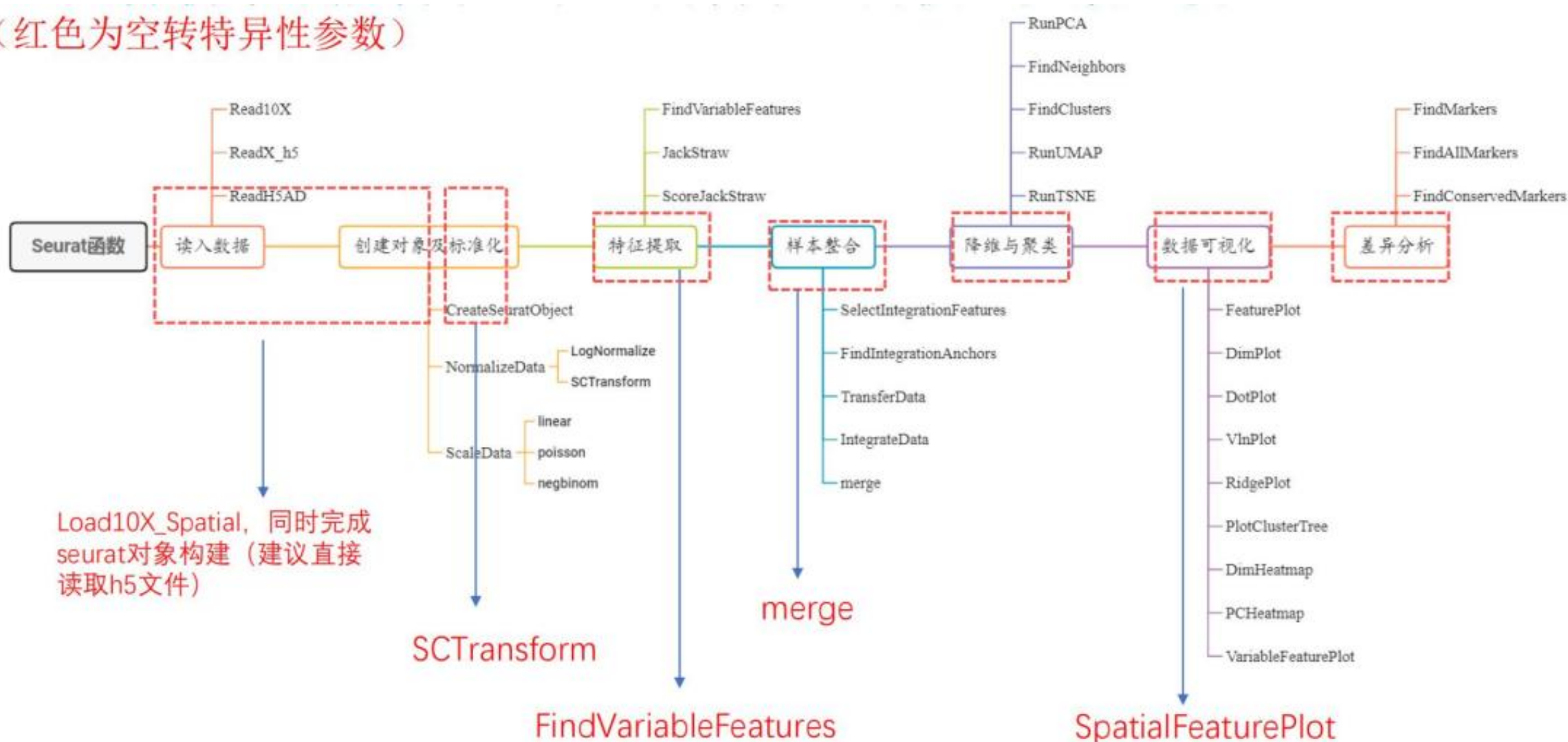
Median Genes per Cell



2. Seurat基础分析



(红色为空转特异性参数)



3. Seurat基础分析之数据过滤和标准化

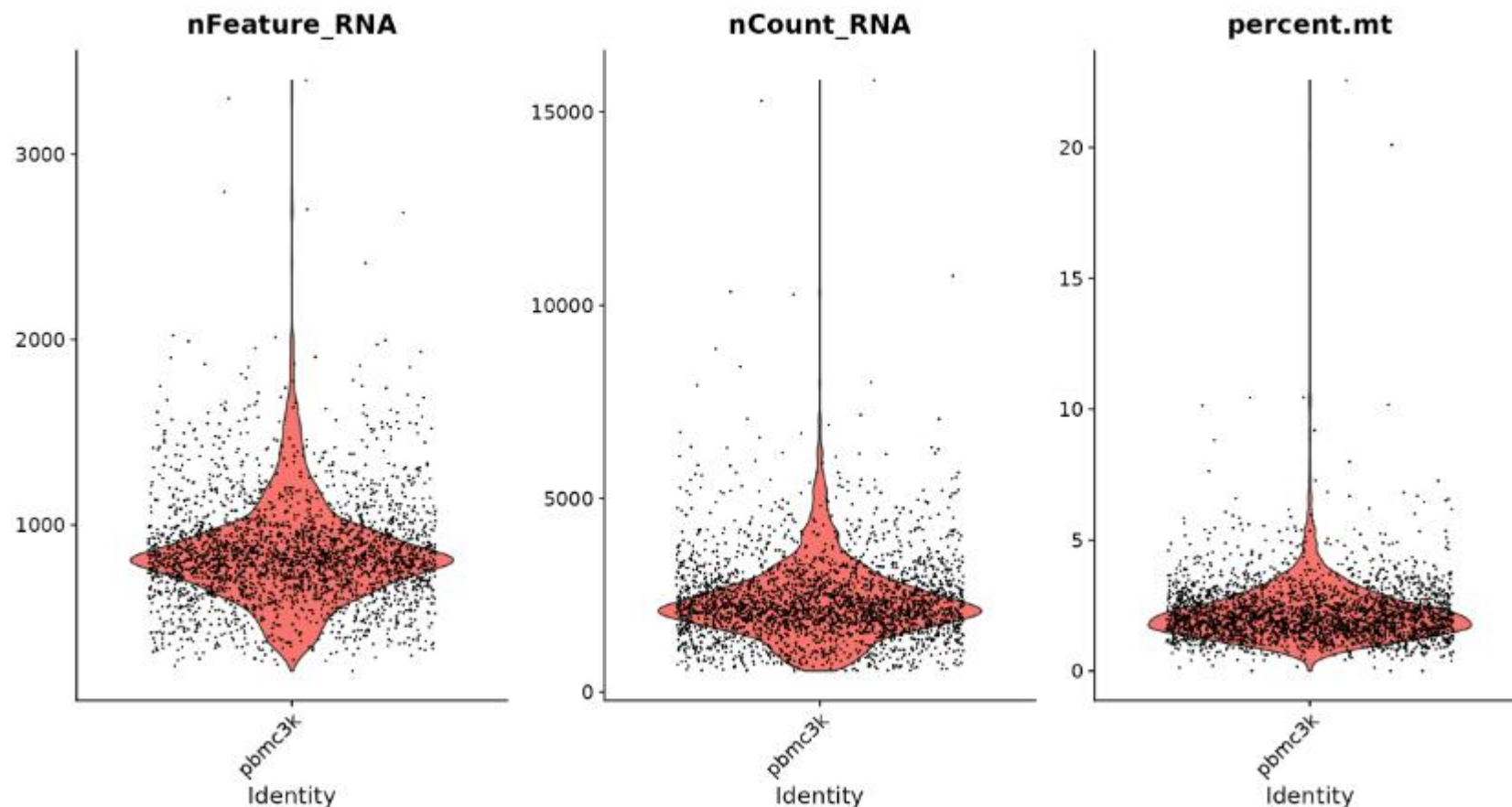


雲南農業大學
Yunnan Agricultural University

云南省生物大数据重点实验室



生信帮



使用R包Seurat分析单细胞数据

单细胞的数据过滤主要包含:

- 1.gene的过滤: 去除仅在个别细胞表达的基因
- 2.细胞的过滤: 去除死细胞, 双重细胞等

细胞的过滤需要考虑样本来源的差异

4. Seurat基础分析之筛选高变基因

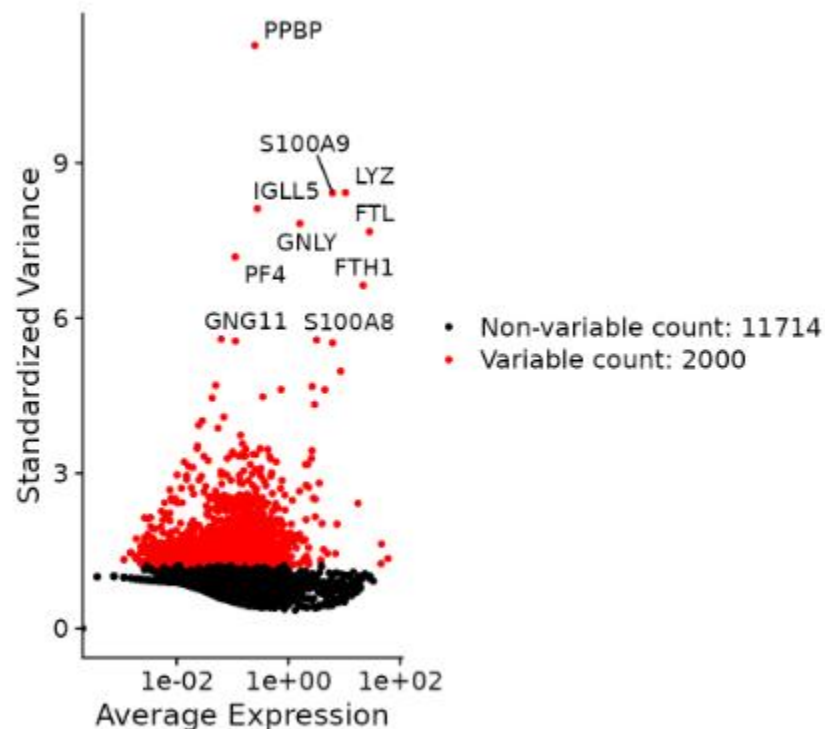
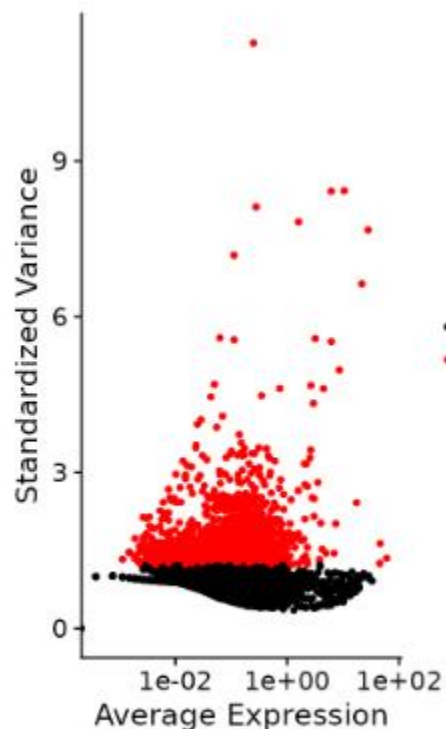


雲南農業大學
Yunnan Agricultural University

云南省生物大数据重点实验室



生信帮



highly variable features:

数据集中在细胞间表现出特异变化的特征基因子集（即，它们在一些细胞中高度表达，而在另一些细胞中低表达）

在下游的分群聚类中使用就是 highly variable features 这些基因代表了数据的主要特征

5. Seurat基础分析之PCA分析

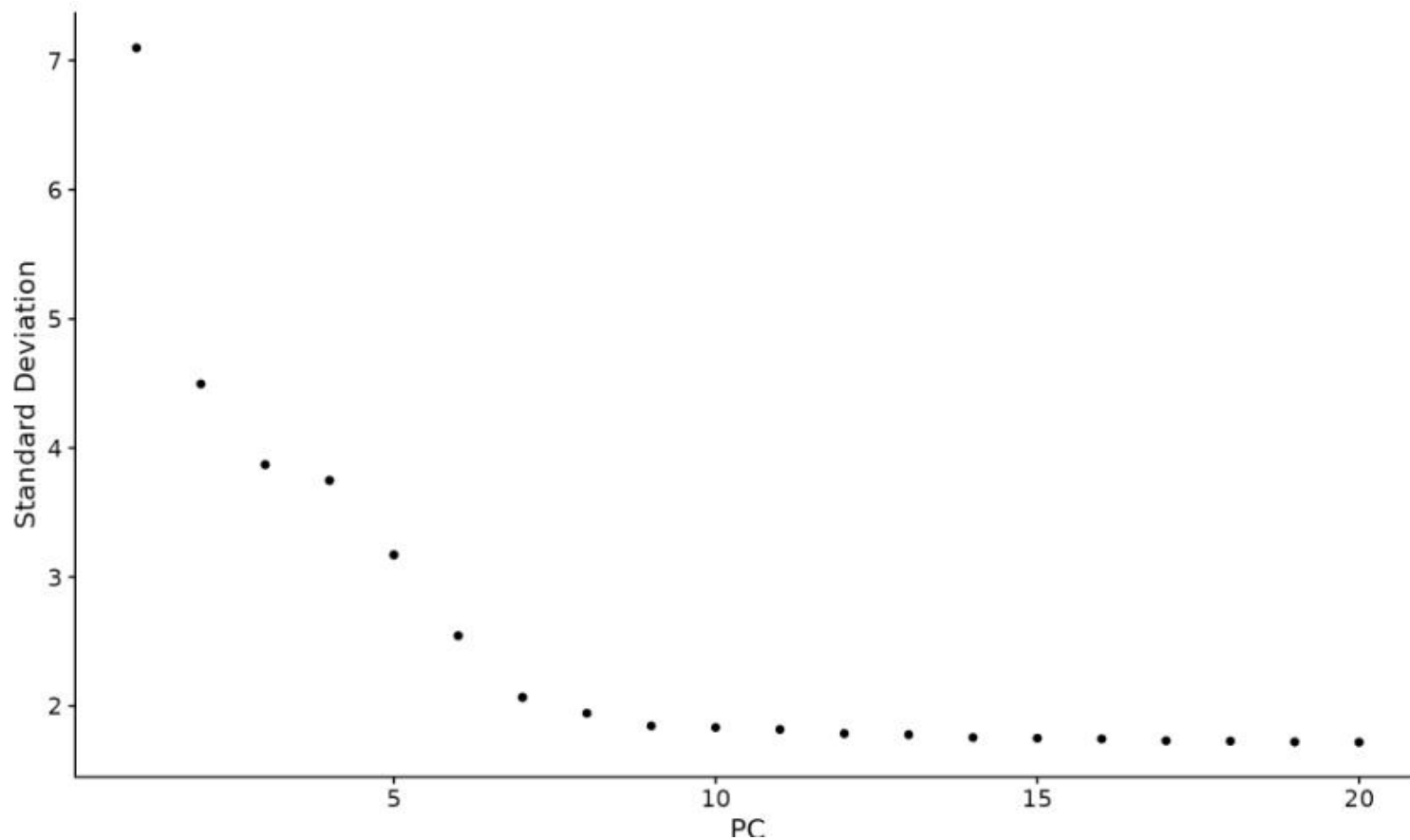


雲南農業大學
Yunnan Agricultural University

云南省生物大数据重点实验室



生信帮



对highly variable features 进行归一化，并对数据进行PCA分析，筛选合适的PC进行下游分析

6、Seurat基础分析之分群聚类



雲南農業大學
Yunnan Agricultural University



云南省生物大数据重点实验室

在单细胞分析中，UMAP (Uniform Manifold Approximation and Projection) 和 t-SNE (t-Distributed Stochastic Neighbor Embedding) 是常用的降维可视化方法。它们在单细胞分析中的区别如下：

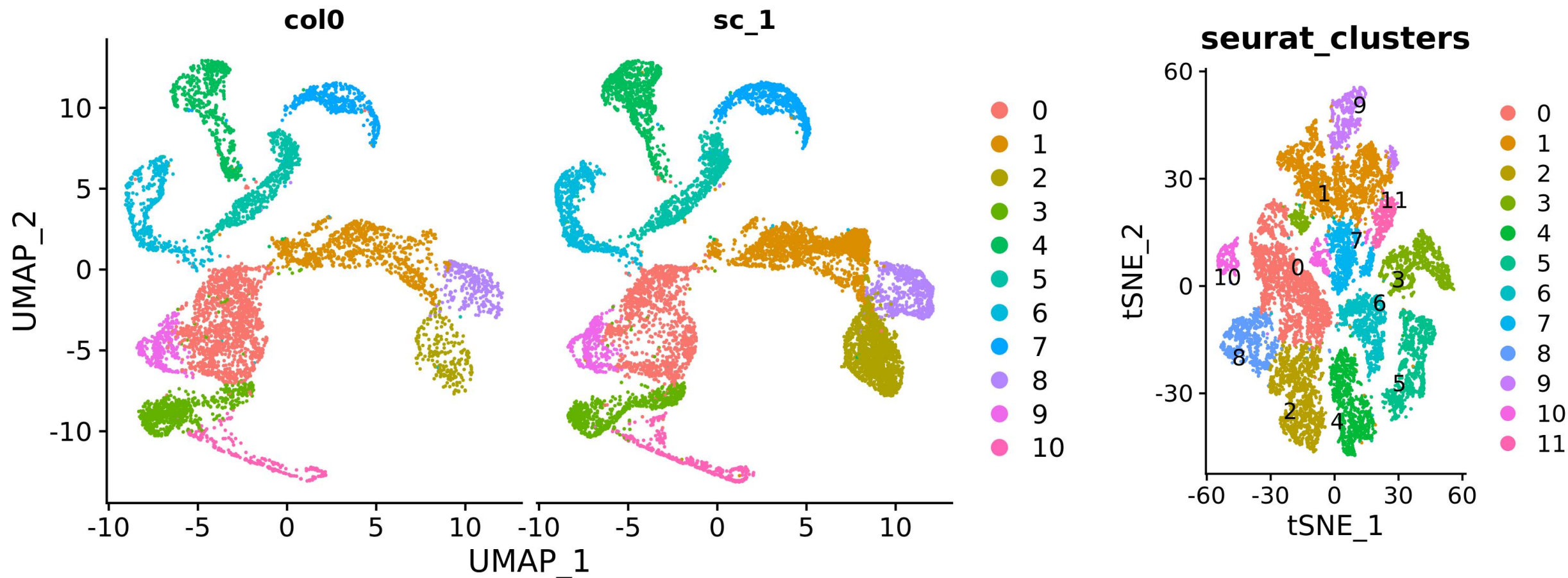
1. 处理大规模数据集的效率：UMAP通常比t-SNE更快且具有更好的可扩展性，在处理大规模单细胞数据集时表现更好。UMAP使用了一些近似计算技术和高效的数据结构，使得其能够处理包含成千上万个细胞的数据集。
2. 保留全局结构和局部结构：t-SNE倾向于保留数据集的全局结构，尽管这可能导致聚类和亚群之间的相对距离被放大。UMAP更注重保留局部结构，即更重视在降维后的空间中细胞之间的局部关系。这可能有助于更好地捕捉细胞类型之间的相似性和差异性。
3. 参数设置：UMAP和t-SNE都有一些参数需要进行调整。对于t-SNE，你需要选择困惑度(perplexity)和学习率(learning rate)等参数。对于UMAP，你需要选择邻域大小(n_neighbors)和最小距离(min_dist)等参数。正确设置这些参数可以影响降维结果的质量。
4. 可视化效果：UMAP通常能够更好地展示细胞之间的细微差异和聚类结构，尤其是在高维空间中具有复杂和混合细胞群的情况下。相比之下，t-SNE可能会产生较大的聚类和距离扩展，导致细胞类型之间的界限不够清晰。

综上所述，UMAP和t-SNE都是用于单细胞数据降维和可视化的强大工具，但它们在处理大规模数据集的效率、全局结构和局部结构的保留以及参数设置方面存在区别。选择适合你的数据和分析目标的方法非常重要。通常建议在进行单细胞分析时使用UMAP，并根据需要与t-SNE进行对比分析。

6. Seurat基础分析之分群聚类



根据PCA结果对单细胞数据进行UMAP/TSNE聚类，得到不同细胞类型的聚类结果



7. Seurat基础分析之细胞类型注释



雲南農業大學
Yunnan Agricultural University

云南省生物大数据重点实验室



生信帮

单细胞细胞类型注释需要丰富的生物学知识和经验，目前常用的细胞类型的注释策略：

- 1.根据已知标记基因确定细胞类型。
- 2.在不同物种中使用类似的标记基因。
- 3.在发育相似的组织区域中使用标记基因。
4. RNA测序产生的已知特定组织或细胞类型的差异表达基因(DEG)集合。
- 5.可以通过分析特定细胞簇中DEGs所编码的细胞功能来确定细胞类型。
- 6.可以通过相关性分析细胞簇间的关联关系来对类似细胞进行分类。
- 7.标记为未知细胞类型

方法1是最常用的方法，准确度高

方法2需要注意物种保守度

方法4类似于基因集打分的方法来注释

方法3、5、6需要相当的知识储备

方法7在研究新物种或确实无法识别细胞类型时可以标记为unknow

7. Seurat基础分析之细胞类型注释



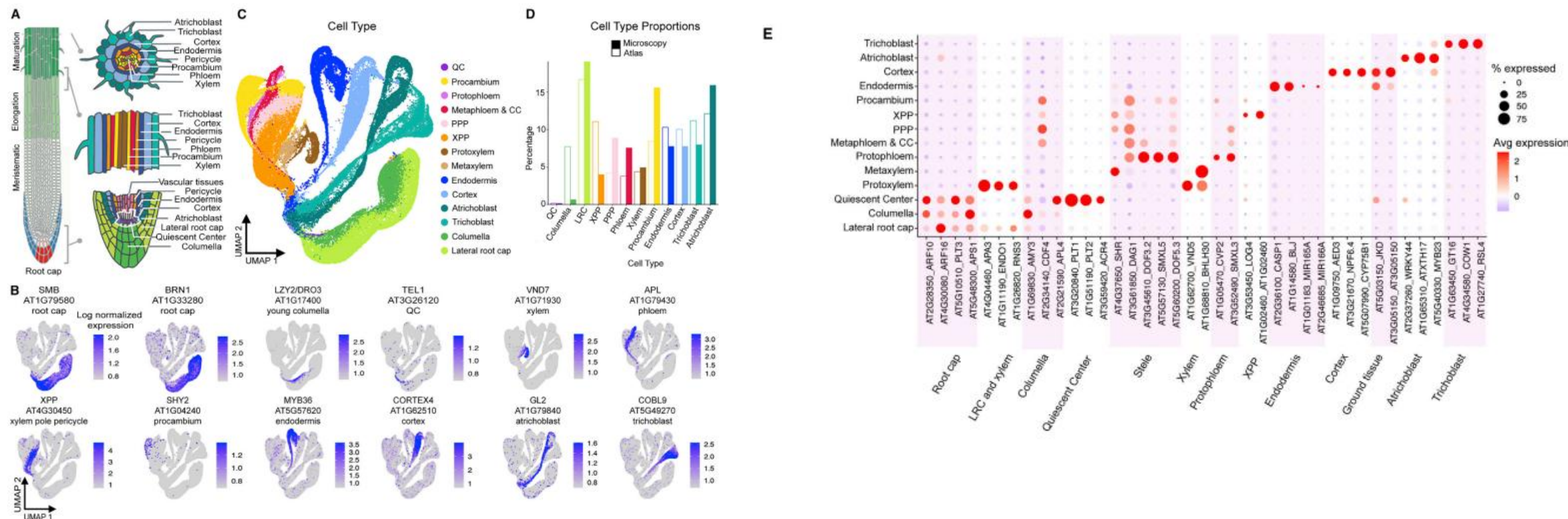
雲南農業大學
Yunnan Agricultural University



云南省生物大数据重点实验室

生信帮

marker基因展示方式



Shahan R,. A single-cell Arabidopsis root atlas reveals developmental trajectories in wild-type and cell identity mutants. Dev Cell. 2022 Feb 28;57(4):543-560.e9. doi: 10.1016/j.devcel.2022.01.008. Epub 2022 Feb 7. PMID: 35134336; PMCID: PMC9014886.

7. Seurat基础分析之细胞类型注释



雲南農業大學
Yunnan Agricultural University



生信帮

云南省生物大数据重点实验室

常用数据库

Database	Number of species (markers) collected	Release	Website				
<i>Comprehensive database</i>				Arabidopsis			
PlantscRNAdb	15 (112,666)	3.0	http://ibi.zju.edu.cn/plantscrnadb/	Root Cell Atlas	<i>A. thaliana</i>	1.0	http://wanglab.sippe.ac.cn/rootatlas/
PscctH	6 (98)	1.0	http://jinlab.hzau.edu.cn/PscctH/	PscB	<i>A. thaliana</i>	1.0	https://www.zmbp-resources.uni-tuebingen.de/timmermans/plant-single-cell-browser
PCMDB	6 (81,117)	1.1	http://www.tobacodb.org/pcmdb/	scRNAPal	<i>P. alba</i>	1.0	https://scu-populus.shinyapps.io/scRNAPal/
PlantPhoneDB	5 (*)	1.0	https://jasonxu.shinyapps.io/PlantPhoneDB/	The floral meristem 3D gene expression atlas	<i>A. thaliana</i>	1.0	http://threed-flower-meristem.herokuapp.com
<i>Cell atlas/Data resource</i>				Soybean nodules	<i>G. max</i>	1.0	http://159.138.151.218:3569/
SCEA	4 (*)	20	https://www.ebi.ac.uk/gxa/sc/home				
<i>Database of single species</i>							
BAR	<i>A. thaliana</i>	20-04	https://bar.utoronto.ca/efp/cgi-bin/efpWeb.cgi?dataSource=Single_Cell				

7. Seurat基础分析之细胞类型注释



雲南農業大學
Yunnan Agricultural University

云南省生物大数据重点实验室



生信帮

常用的植物细胞类型的marker基因常用数据库-PlantscRNAdb

 <i>Arabidopsis thaliana</i> 19 Tissues, 248 Cell Types, 20,862 Marker Genes	 <i>Oryza sativa</i> 7 Tissues, 71 Cell Types, 11,439 Marker Genes	364 cell types	248 <i>A. thaliana</i>	71 <i>O. sativa</i>	34 <i>S. lycopersicum</i>	75 <i>Z. mays</i>
 <i>Solanum lycopersicum</i> 7 Tissues, 34 Cell Types, 4,741 Marker Genes	 <i>Zea mays</i> 11 Tissues, 75 Cell Types, 11,599 Marker Genes		9 <i>F. vesca</i>	25 <i>Populus</i>	5 <i>N. attenuata</i>	12 <i>L. minuta</i>
 <i>Fragaria vesca</i> 1 Tissue, 9 Cell Types, 5,779 Marker Genes	 <i>Populus</i> 3 Tissues, 25 Cell Types, 8,406 Marker Genes	69,628 marker genes	20,862 <i>A. thaliana</i>	11,439 <i>O. sativa</i>	4,741 <i>S. lycopersicum</i>	11,599 <i>Z. mays</i>
 <i>Nicotiana attenuata</i> 1 Tissue, 5 Cell Types, 1,723 Marker Genes	 <i>Lemna minuta</i> 1 Tissue, 12 Cell Types, 4,238 Marker Genes		5,779 <i>F. vesca</i>	8,406 <i>Populus</i>	1,723 <i>N. attenuata</i>	4,238 <i>L. minuta</i>

<http://ibi.zju.edu.cn/plantscrnadb/index.php>

7. Seurat基础分析之细胞类型注释



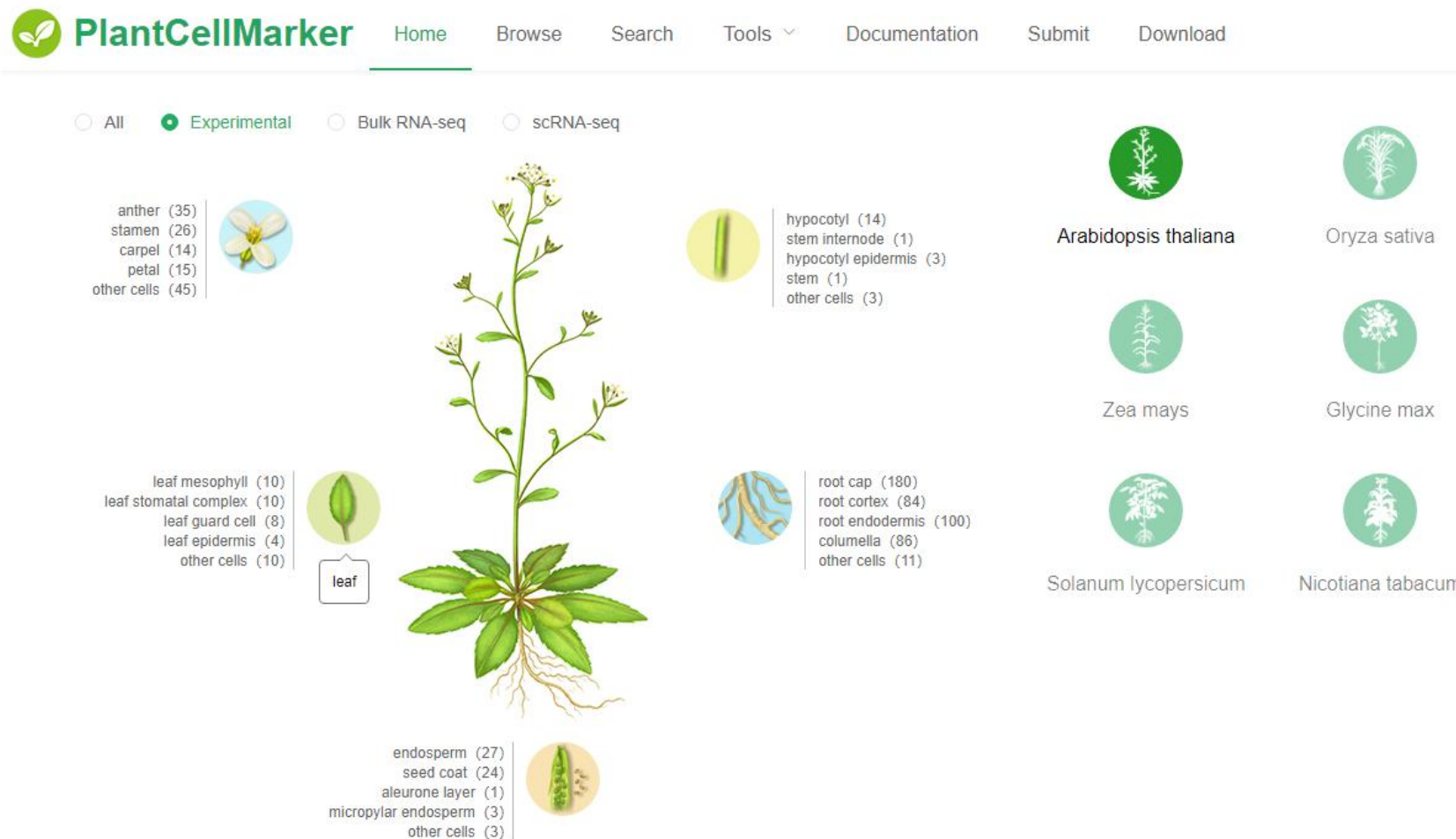
雲南農業大學
Yunnan Agricultural University



生信帮

云南省生物大数据重点实验室

常用的植物细胞类型的marker基因常用数据库



8. Seurat基础分析之差异分析

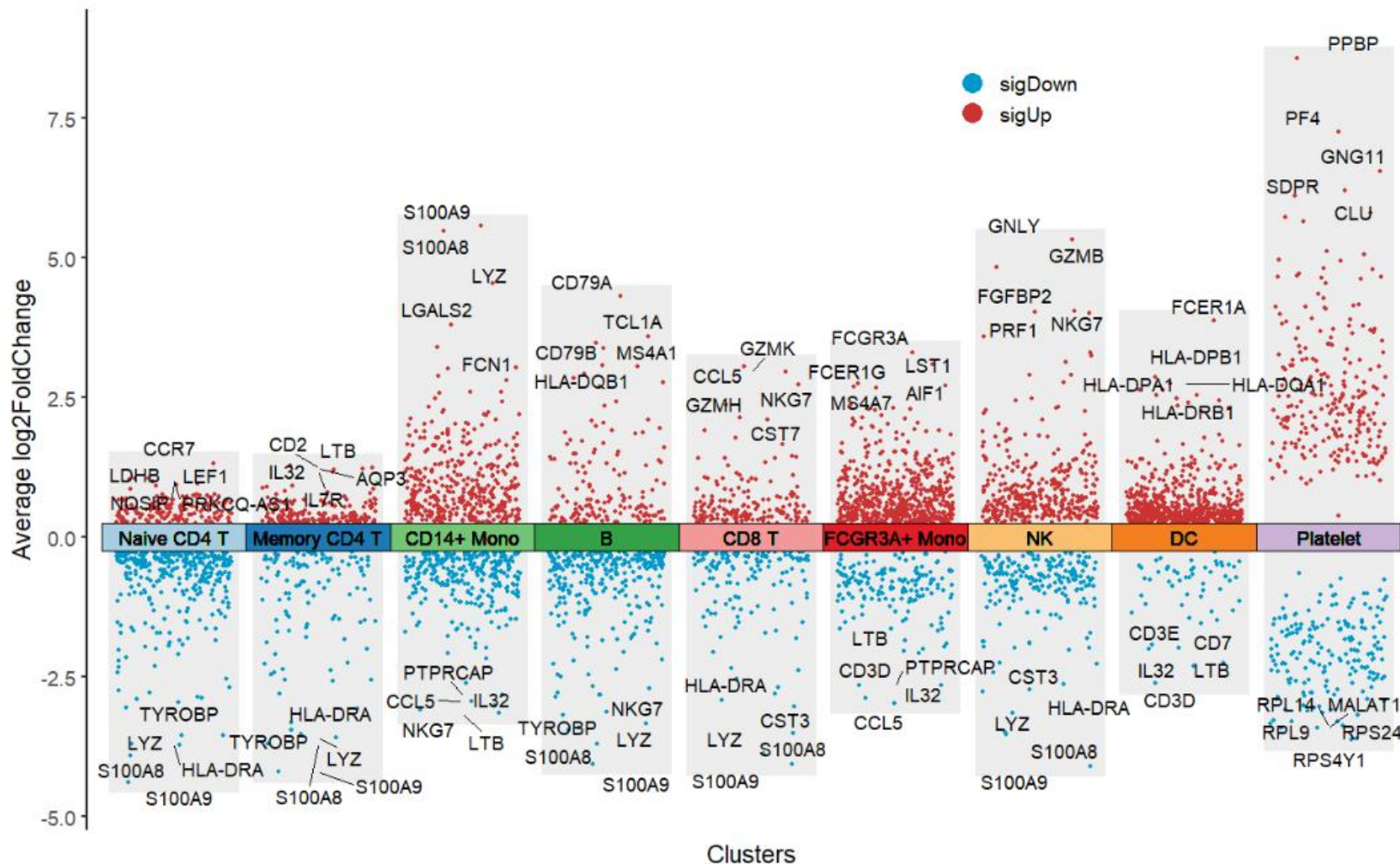


云南农业大学
Yunnan Agricultural University



云南省生物大数据重点实验室

生信帮



差异分析: Seurat包含多种差异分析方法, 可以比较不同细胞类型或样本分组间的差异

可选差异分析方法:

wilcox (default)

bimod

roc

t-test

negbinom

poisson

LR

MAST

DESeq2

这些差异基因同时可以作为细胞类型的marker基因

<https://junjunlab.github.io/scRNAtoolVis-manual/>



雲南農業大學
Yunnan Agricultural University

云南省生物大数据重点实验室



生信帮

第三章 单细胞高级分析

3.1 单细胞数据整合和批次校正



单细胞批次效应产生的原因：

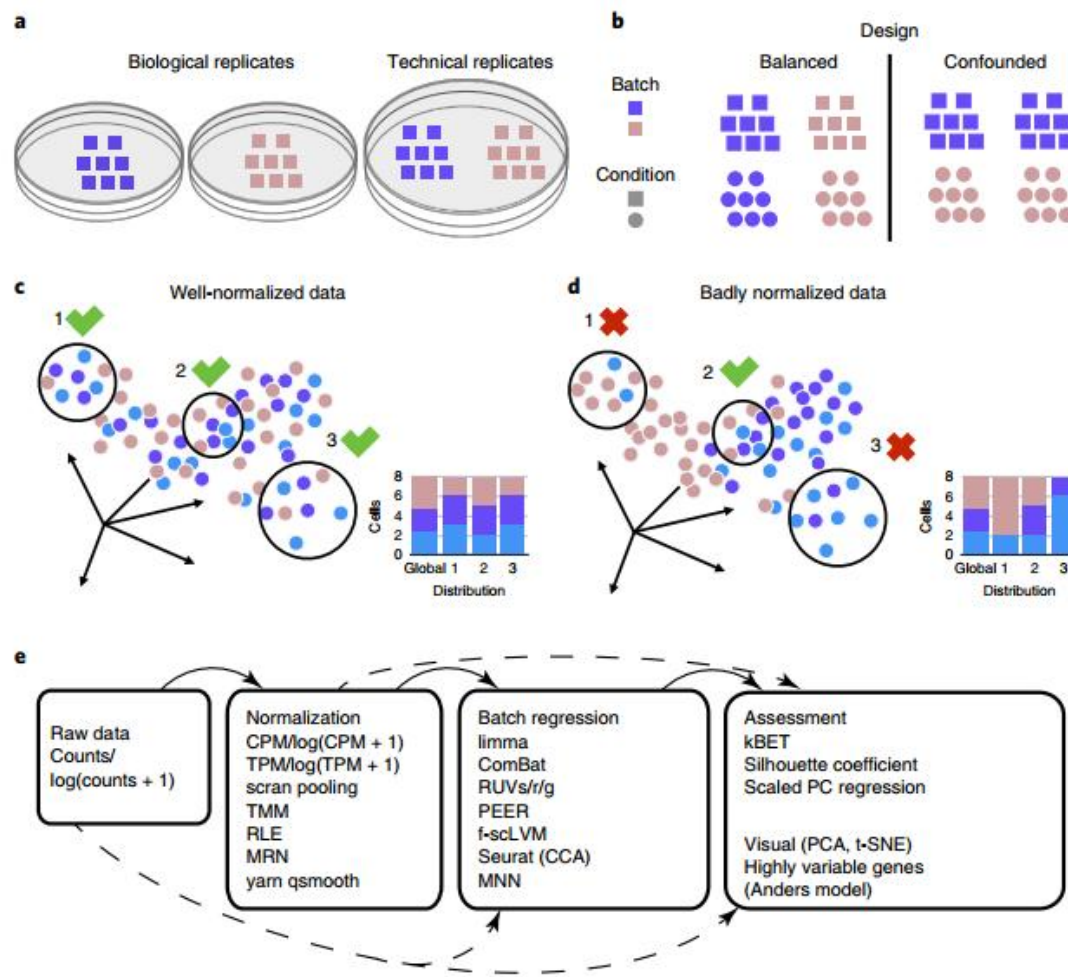
1.实验技术差异：不同批次的实验条件、设备、试剂盒、测序平台等会引入技术上的变异。例如，不同批次的RNA测序可能使用不同的流程、引物、文库制备方法等，这些因素可能造成样本间的技术差异。

2.样本处理差异：样本采集、处理和分离的时间点也可能存在差异。这些差异可能导致细胞状态、RNA质量、细胞组成等方面的变化，进而影响到测序结果。

3.实验条件差异：实验条件的变化，如培养基的配方、培养温度、染色剂的使用等，也可能对细胞产生影响，从而引入批次效应。

4.样本来源差异：如果细胞样本来源于不同的组织、生理条件或疾病状态，那么这些生物学差异也可能在单细胞测序数据中表现为批次效应。

Ps：注意区分生物学差异和批次效应，并不是所有的差异都是由批次造成的



3.1 单细胞数据整合和批次校正



雲南農業大學
Yunnan Agricultural University

云南省生物大数据重点实验室



生信帮

不同单细胞数据整合分析方法的比较

a

	Considerations	scANVI	Scanorama embed	scVI	FastMNN embed	scGen	Harmony	FastMNN gene	Seurat v3 RPCA	BBKNN	Scanorama gene	ComBat	MNN	Seurat v3 CCA	trVAE	Conos	DESC	LIGER	SAUCIE embed	SAUCIE gene
Input	Programming language																			
	Method runs without additional information	✗				✗														
Scib results	Consistent top performer	✓	✓	✓		✓														
	Top method on small/simple tasks		✓		✓	✓	✓													
	Top method on large/complex tasks	✓	✓	✓		✓														
	Top method on ATAC data	—		—			✓											✓		
Task details	Integrates strong batch effects	✓	—	—		✓			—	—				—						
	Top method for recovery cell states or modules	✓	✓								✓	✓	✓							
	Confounding of bio and batch variance	✓	—			✓														
	Top method for trajectories	—	✓	—	✓	✓														
	Method deals with varying compositions											✗								
Speed	Fast method for quick results									✓		✓								
	Scales well to large datasets on CPU	✓	—	✓						✓	—								✓	✓
	Method has GPU support	✓		✓		✓									✓		✓		✓	✓
	Scales well to feature spaces beyond genes														✓	✓				
Output	Method shows corrected expression					✓		✓	✓		✓	✓	✓	✓						✓
	Method gives relative cell embeddings									✗						✗				

3.1 单细胞数据整合和批次校正



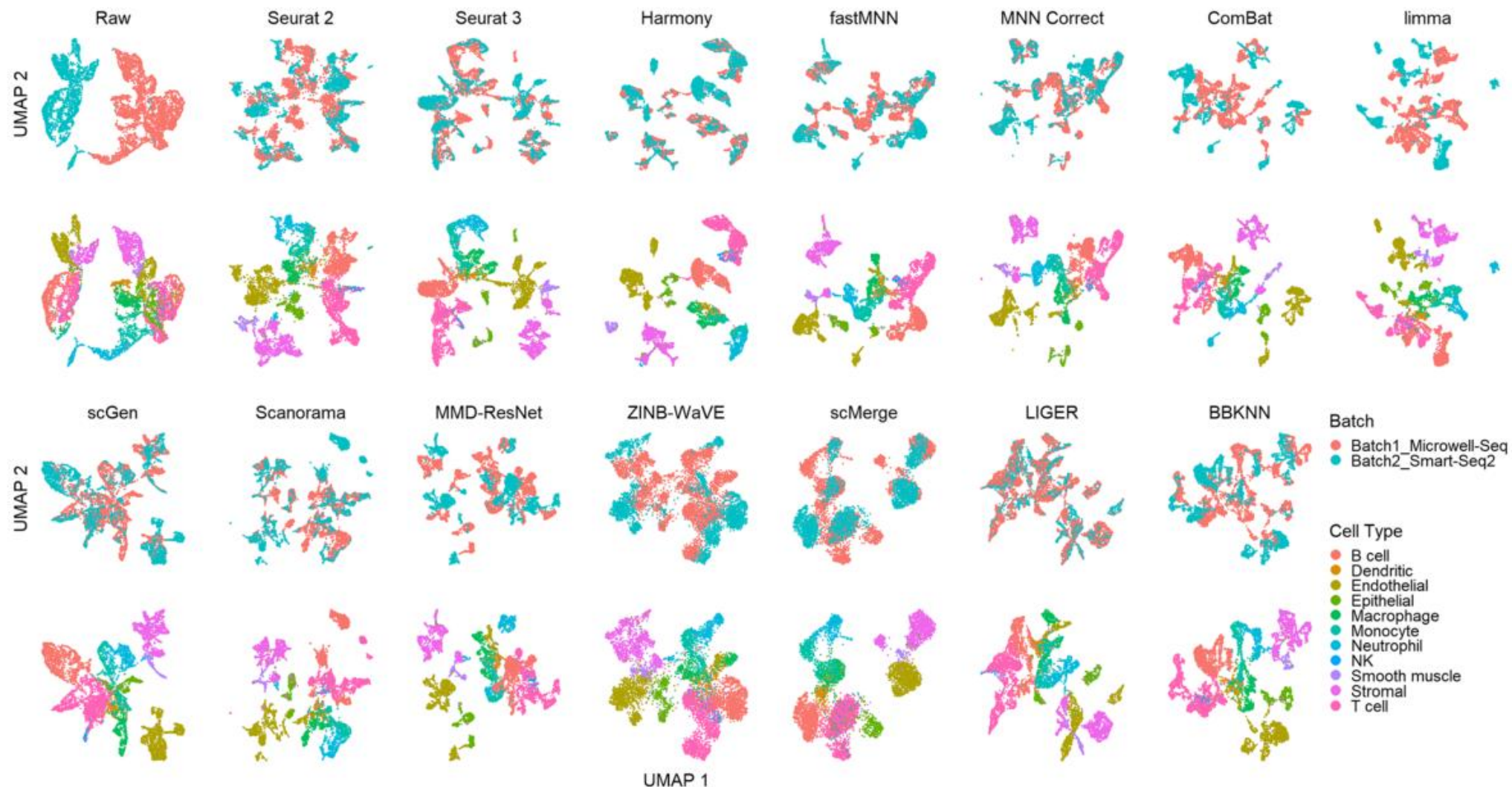
雲南農業大學
Yunnan Agricultural University



生信帮

云南省生物大数据重点实验室

不同单细胞数据整合分析方法的比较



3.1 单细胞数据整合和批次校正



雲南農業大學
Yunnan Agricultural University

云南省生物大数据重点实验室



生信帮

数据整合推荐方法

1. merge 。 简单的数据累加，不对数据做任何处理和校正，在样本聚类结果没有明显批次时，尽量不要对数据做其他任何处理，可能会影响到样本本身的差异
2. SCT 。 Seurat包自带，负二项回归方法对测序深度等进行校正，替代数据标准化，高变基因筛选，归一化，会重新生成一个新的表达矩阵，可能会对下游分析造成影响，不能校正批次效应
3. Harmony。 可结合Seurat，应用主成分分析将转录组表达谱嵌入到低维空间中，用迭代过程去除数据集特有影响，使用简单，速度快
4. CCA。 Seurat包自带，通过识别数据集之间共享的变异源，CCA 非常适合在细胞类型保守的情况下识别锚点，但不同实验之间的基因表达存在很大差异，可能存在过矫正的情况
5. RPCA 。 Seurat包自带，和CCA运行方式相同，更加‘保守’，建议在综合分析过程中使用 RPCA，即：
 - * 一个数据集中的大部分细胞在另一个数据集中没有匹配的类型
 - * 数据集源自同一平台（即 10x 基因组的多个泳道）
 - * 有大量数据集

Ps:

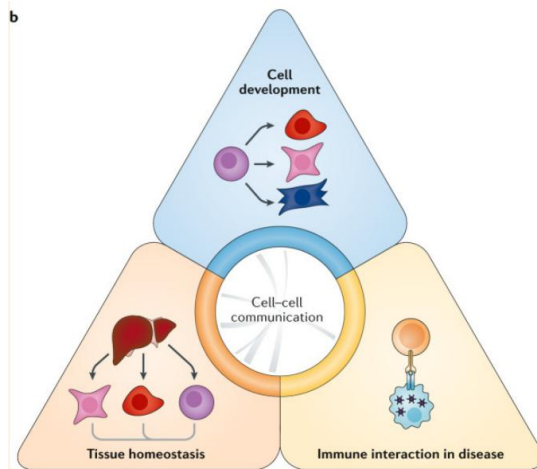
- 初步聚类结果批次不明显的情况下，不建议校正批次
- 不同的软件在不同的数据集会有不同的效果，建议多试几个软件，看聚合效果最好的软件。
- 初学者推荐使用Harmony方法校正批次

3.2 细胞通讯分析

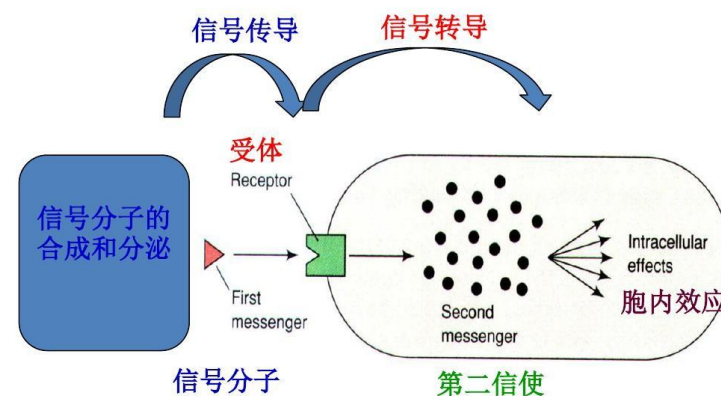


细胞通讯的原理:

- 1、多细胞生物是由不同类型的细胞组成的整体，细胞间存在频繁的相互作用。如生物体的生长发育、分化各种组织器官的形成、组织的维持以及各种生理活动的协调，都需要高度精确和高效的细胞间的通讯机制。
- 2、细胞之间的相互通讯是一个极其复杂的过程，通常指一个细胞发出的信息通过介质传递到另一个细胞产生相应的反应。
- 3、细胞通讯中有两个基本概念: 细胞信号传导和信号转导，前者强调的产生与细胞间的传送，而信号转导是接受与接收后信号转换的方式途径和结果。
- 4、细胞有三种通讯方式: 第一种通过化学信号分子，这是动物和植物最普遍采用的通讯方式: 第二种通过相邻细胞表面分子的粘着: 第三种通过细胞与细胞外基质的粘着
- 5、细胞通讯基本过程:
 - 信号分子的合成
 - 信号分子从信号生成细胞释放到周围环境中
 - 信号分子向靶细胞运输
 - 靶细胞对信号分子的识别和检测
 - 细胞内信号作用于效应分子



细胞通讯的过程



3.2植物单细胞通讯



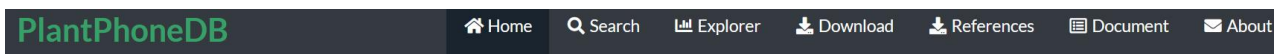
雲南農業大學
Yunnan Agricultural University



生信帮

云南省生物大数据重点实验室

植物的细胞通讯数据库，包含拟南芥，水稻等模式植物的通讯信息，可以使用对应的R包*PlantPhoneDB* 来对植物做单细胞细胞通讯分析



Welcome to PlantPhoneDB! [Cite our paper](#)

A manually curated pan-plant database of ligand-receptor pairs infers cell-cell communication

Arabidopsis thaliana



3,514 Ligand-Receptor pairs

Oryza sativa



3,762 Ligand-Receptor pairs

Solanum lycopersicum



1,751 Ligand-Receptor pairs

Zea mays



2,823 Ligand-Receptor pairs

Populus alba x Populus glandulosa



3,110 Ligand-Receptor pairs

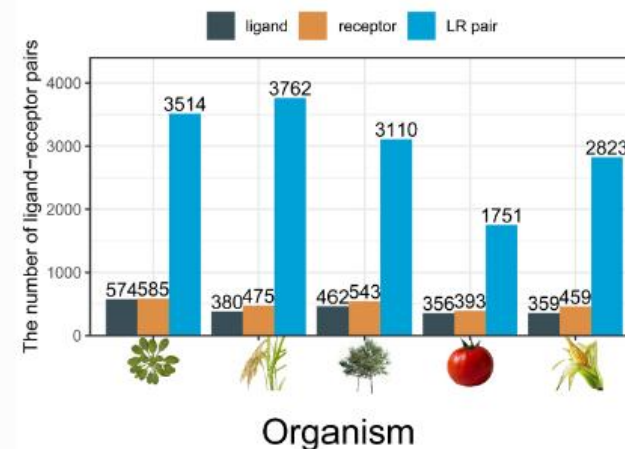
Arachis hypogaea L.



pending

scRNA-seq datasets statistics

Datasets (29)
Species (5)
Tissues (15)
Cells (~560,000)
Technology (6)



By assigning orthologs of ligand-receptor pairs between *Arabidopsis thaliana* and other four plant species proteomes using the InParanoid algorithm, the number of ligand-receptor pairs identified ranged from 1,751 (*Solanum lycopersicum*) to 3,762 (*Oryza sativa*).

Organism	Ligand	Receptor	LR pair
Arabidopsis thaliana	574	585	3514
Oryza sativa	380	475	3762
Populus alba x Populus glandulosa	462	543	3110
Solanum lycopersicum	356	393	1751
Zea mays	359	459	2823

3.3 拟时序分析之monocle



雲南農業大學
Yunnan Agricultural University

云南省生物大数据重点实验室



生信帮

- 在发育、刺激应答和全生命周期过程中，细胞从一种功能状态转变为另一种。不同状态的细胞表达不同的基因集，进而产生动态的表达谱和代谢谱。通过基因表达信息的动态变化来重构细胞的演变（发育、衰老、疾病进展）轨迹，揭示细胞的演变状态和规律，分析相关的驱动基因表达特征。
- 拟时序（Pseudotime），是测量细胞在生物学过程（如细胞分化）中**进展程度**的一种方法。它对一个生物学过程进行抽象，并描述了细胞当前状态与时间轨迹起始状态之间最短路径的距离。拟时分析适用于发育生物学中**发育轨迹**的研究，以及肿瘤微环境中免疫细胞状态变化的研究等。
- 拟时序分析是基于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。**如果是两个完全独立的cluster是不适合做时序分析的**。如果数据存在持续的分化，那么单个时间点的数据也是可以的。我们在做分析时候，最重要的还是带着某些生物学问题和假设去做轨迹分析。

3.3 拟时序分析之monocle



雲南農業大學
Yunnan Agricultural University



生信帮

云南省生物大数据重点实验室

Monocle 是一种广泛使用的单细胞RNA测序 (scRNA-seq) 数据分析工具，用于探索和理解细胞发展和分化过程。以下是使用 Monocle 进行分析的基本步骤：

1. 数据准备：将单细胞RNA测序数据导入R环境，并使用适当的包（如`Seurat`或`SingleCellExperiment`）对数据进行预处理和归一化。这通常包括质量控制、基因过滤、批次校正和归一化等步骤。
2. 构建单细胞数据对象：使用Monocle 提供的函数，将预处理后的数据转换为Monocle 特定的单细胞数据对象（`CellDataSet`或`CellDataSetSeq`）。这些对象包含了细胞的表达矩阵、样本信息和其他必要的元数据。
3. 细胞排序：使用Monocle 内置的非线性降维方法（例如DDRTree）对细胞进行排序，以捕获其在发育过程中的顺序。这将根据单细胞之间的相似性将其映射到一个连续的发育轨迹上。
4. 构建拓扑图：通过计算和连接具有相似表达模式的细胞，使用Monocle 的函数构建细胞发育的拓扑图。这将显示出细胞之间的关系和发育路径的分支等信息。
5. 发现差异基因：使用Monocle 提供的功能，对不同发育状态或细胞类型之间的差异进行统计分析，以发现在细胞发育过程中表达模式发生变化的基因。这有助于理解细胞的分化和功能。
6. 建模和可视化：使用Monocle 的函数和工具，可以根据发育轨迹的形状和细胞类型的特点，设计和构建模型以描述细胞发育的动态过程。然后，可以使用内置的可视化工具（如细胞发育轨迹图、基因表达图等）对结果进行展示和解释。

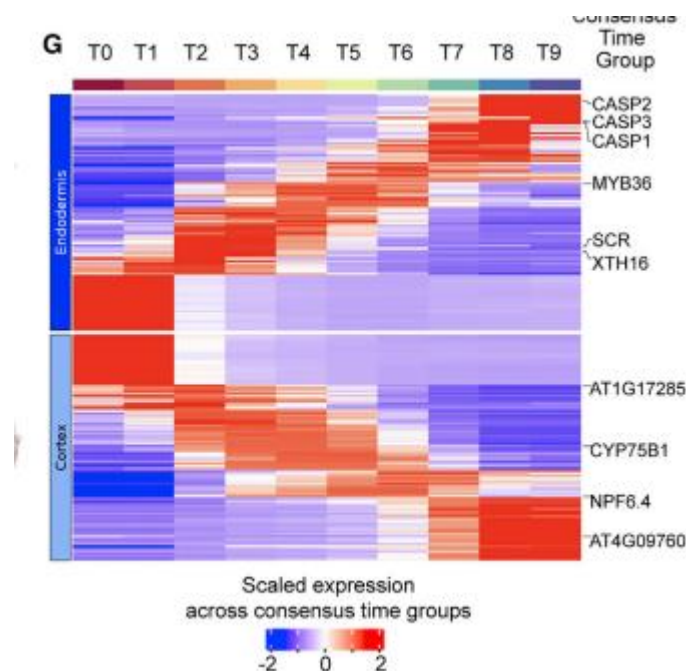
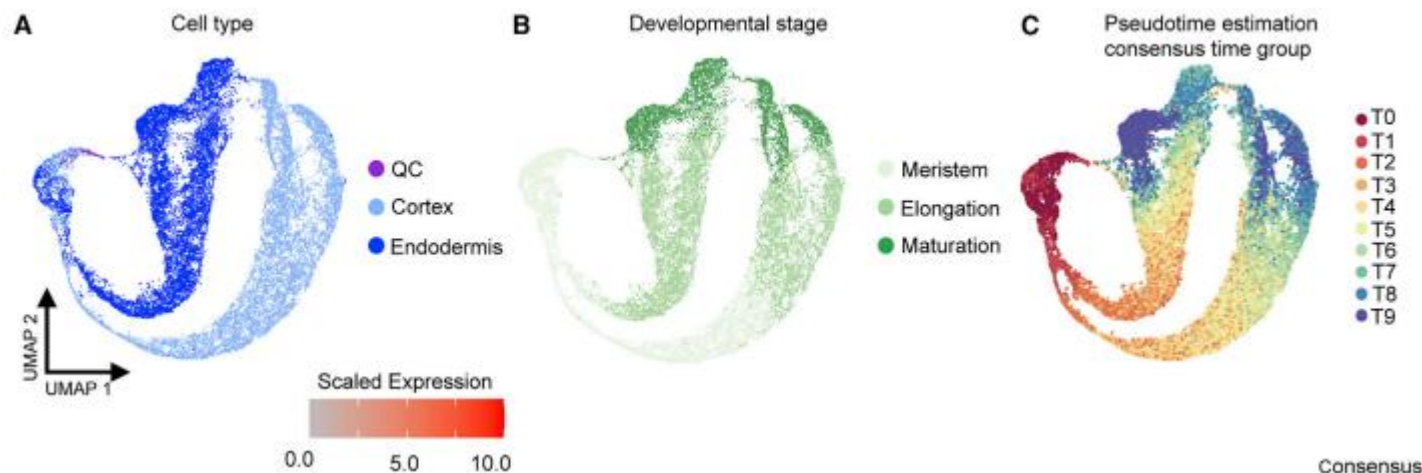
需要注意的是，这只是Monocle 分析的基本步骤，根据具体的研究目标和数据特点，可能需要进行其他额外的分析和解释步骤。

3.3 拟时序分析之monocle



基因集选择的几种常见策略:

- 1) 基于Seurat 鉴定的marker genes
 - 2) 基于monocle2差异分析鉴定的差异基因: 根据细胞群、实验条件等
 - 3) 基于高变基因
 - 4) 其他自定义的基因集
 - 5) 基因集的选择不宜过多, 也不宜过少 (几十到几百个基因)
- 基因的选择比较灵活, 但是要有生物学或者数据的依据(拟时分析有点玄学)



3.4 拷贝数变异 (CNV)



雲南農業大學
Yunnan Agricultural University

云南省生物大数据重点实验室



生信帮

CNV分析原理:

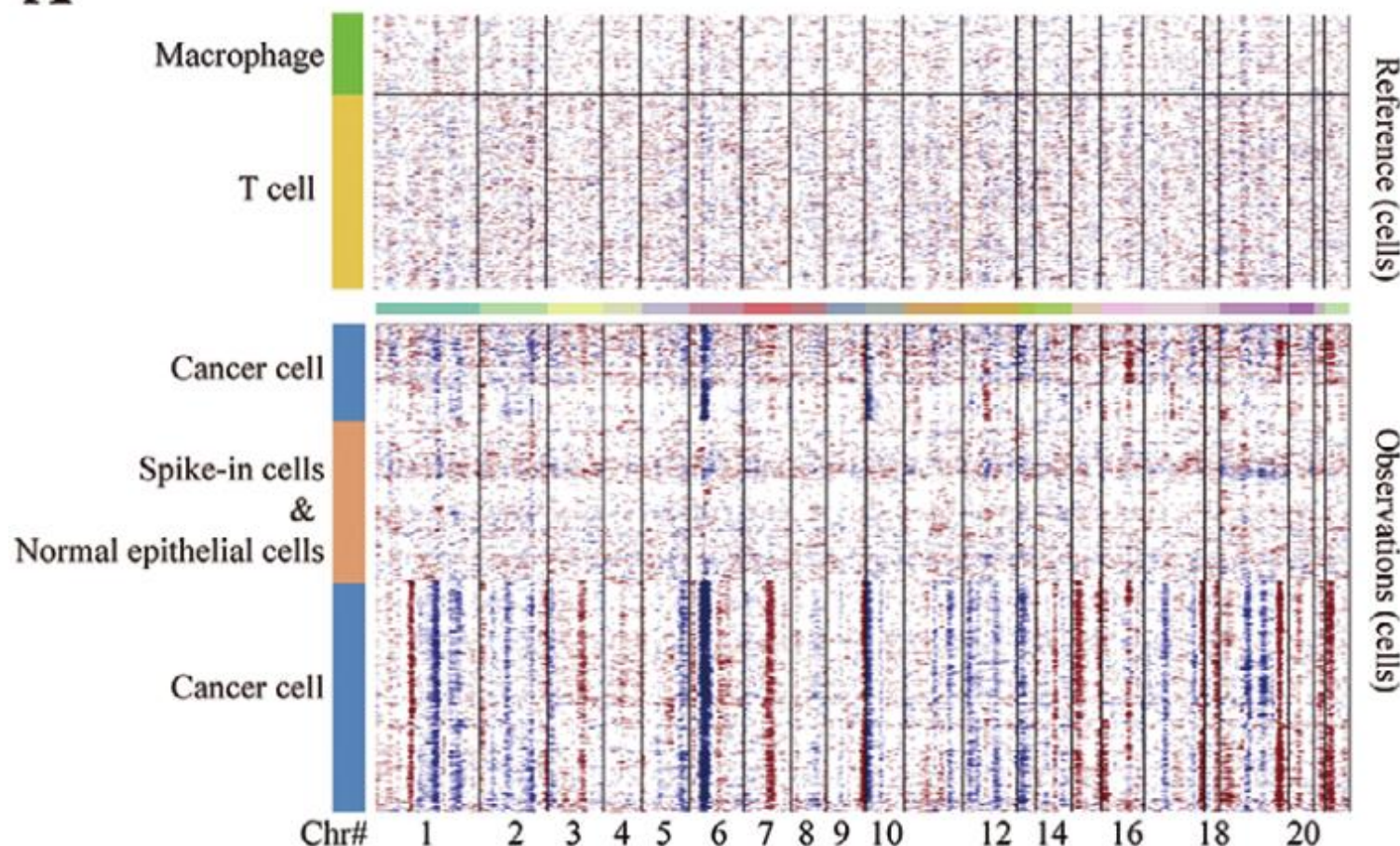
- 1.染色体结构: 人类细胞核内包含有23对染色体, 其中22对是自动体染色体, 1对是性染色体。每对染色体都有两个同源染色体, 即父母遗传的一份来自父亲, 一份来自母亲。正常情况下, 每个染色体区域都应该有两份基因拷贝, 即拷贝数为2。
- 2.拷贝数变异: 但在某些情况下, 由于遗传突变、基因重复、重组等因素, 某个或某些染色体区域的拷贝数可能发生增加或减少。这种拷贝数变异就是CNV。CNV可以导致基因拷贝数的改变, 进而影响基因表达水平和细胞的功能。
- 3.单细胞CNV预测分析可以推断肿瘤细胞, 解析肿瘤细胞异质性, 为探索肿瘤克隆进化、挖掘肿瘤发生机制提供参考。

常用的分析软件: inferCNV、copyKat

3.4 拷贝数变异 (CNV)



A



InferCNV是TrinityCTAT工具包的一个组成部分，可以用于肿瘤scRNA-seq数据中鉴定大规模染色体拷贝数变异 (copy number variation, CNV)

它的原理是：用一组“正常”细胞（如癌旁组织来源的细胞）作为对照，分析肿瘤基因组上各个位置的基因表达量强度变化，通过热图的形式展示每条染色体上的基因的相对表达量，那么相较于正常细胞，肿瘤基因组会表现出过表达或者低表达。

Ge Guangzhe, Han Yang, Zhang Jianye et al. Single-Cell RNA-seq Reveals a Developmental Hierarchy Super-Imposed Over Subclonal Evolution in the Cellular Ecosystem of Prostate Cancer.[J] .Adv Sci (Weinh), 2022, 9: e2105530.



雲南農業大學
Yunnan Agricultural University

云南省生物大数据重点实验室

生信帮，答你所问