第三课 单细胞转录组计算方法与数据分析

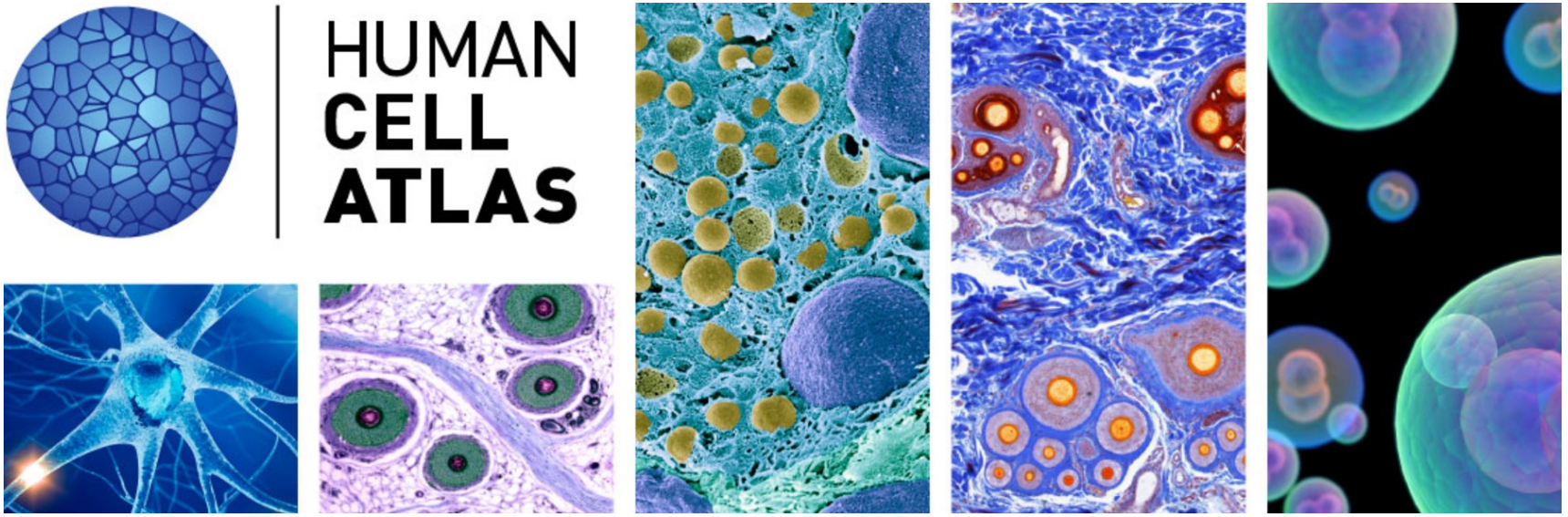
中国医学科学院基础医学研究所

陈阳

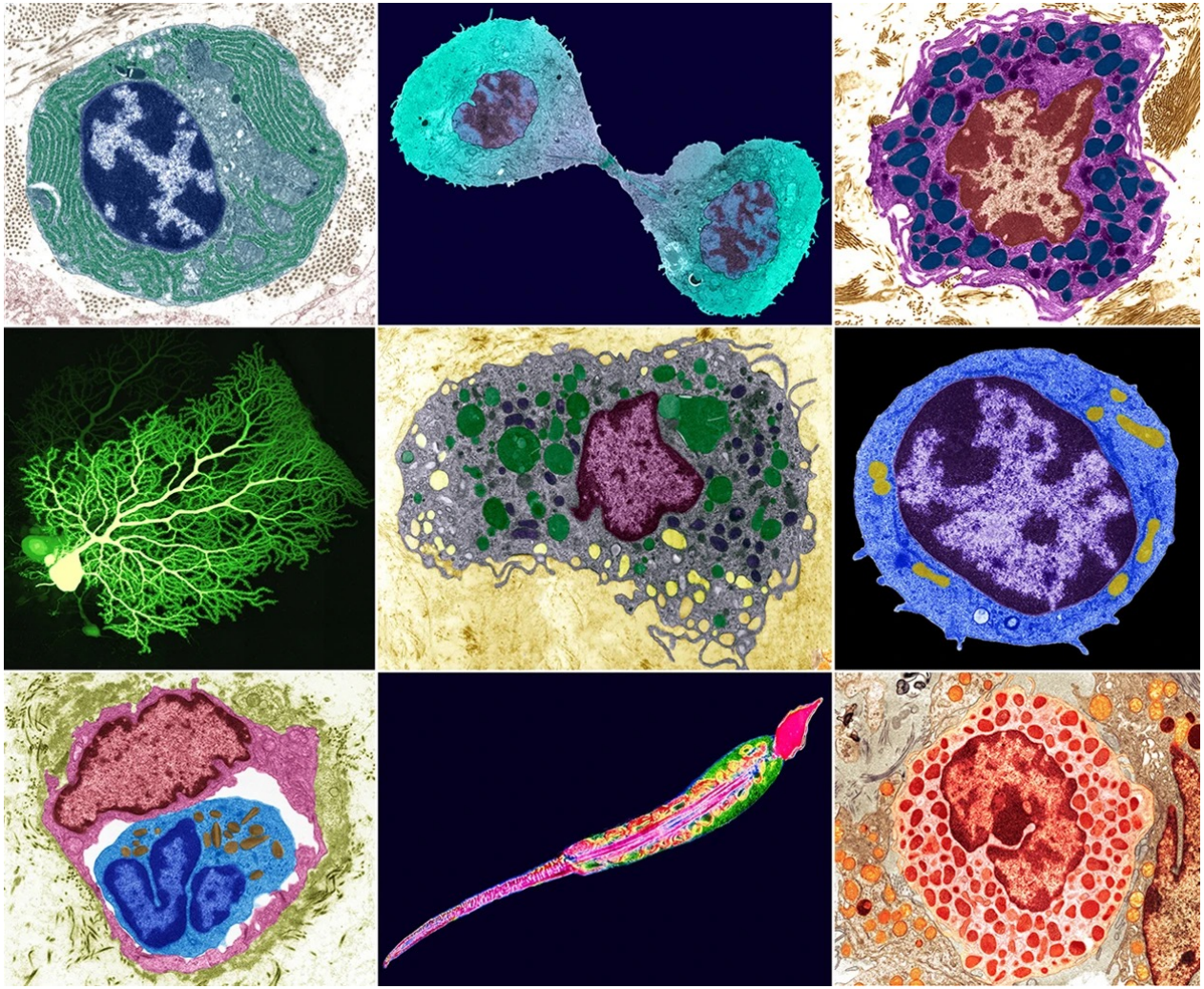
yc@ibms.pumc.edu.cn

20240402

平均而言每个人由40-60万亿个细胞组成

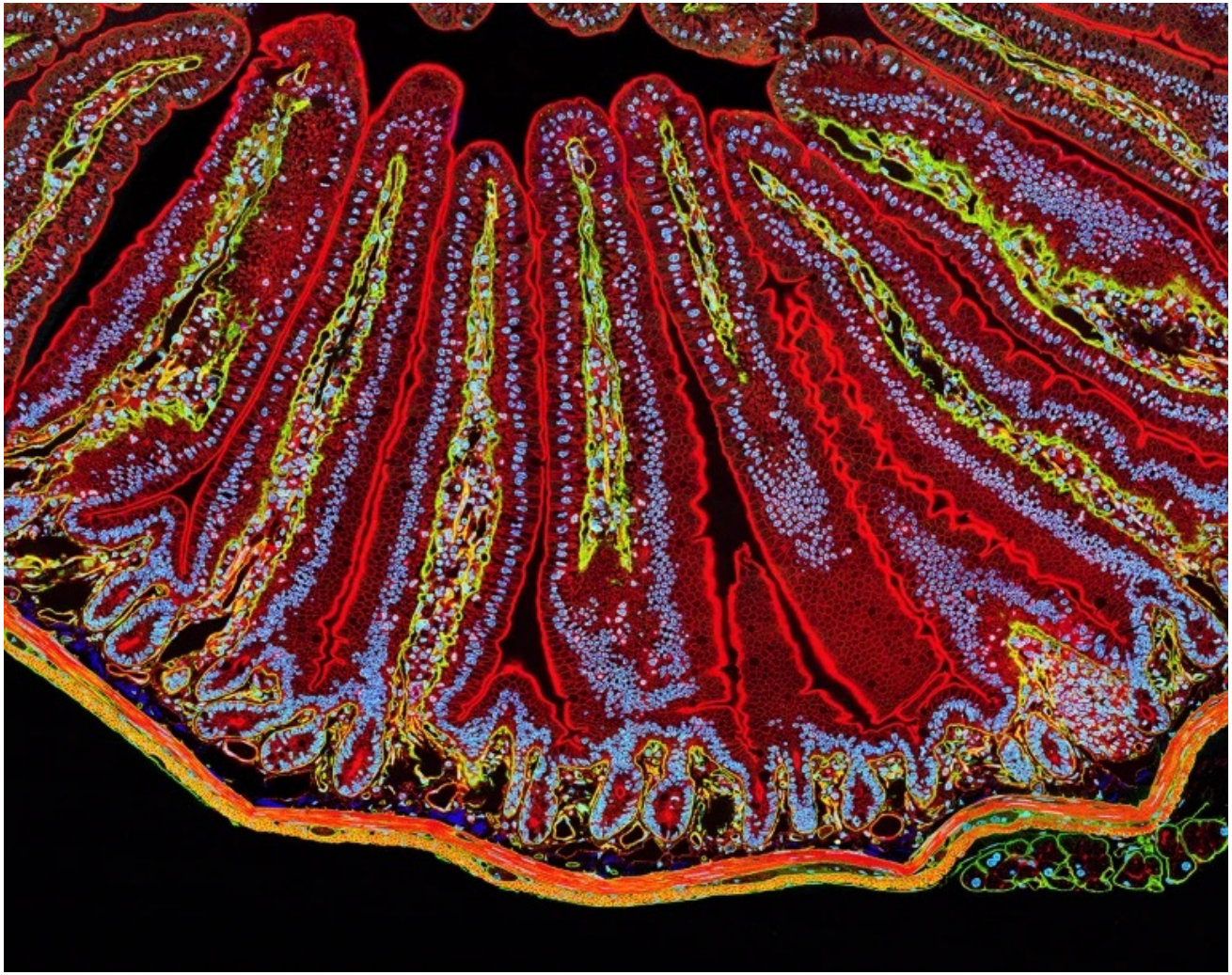


What is a cell type, really?



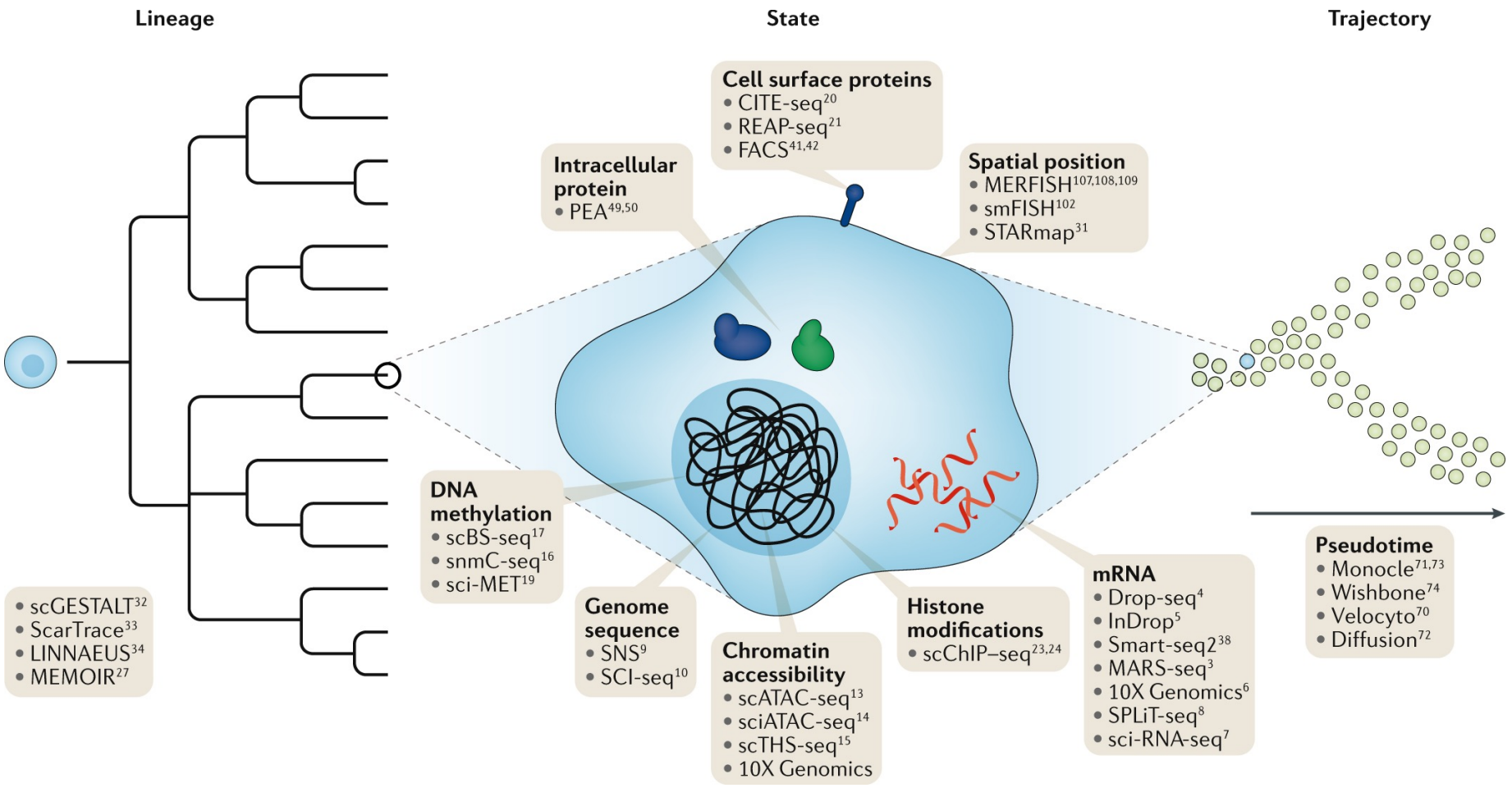
https://www.nature.com/articles/d41586-024-03073-2

Villi in the small intestine



https://www.nature.com/articles/d41586-024-03073-2

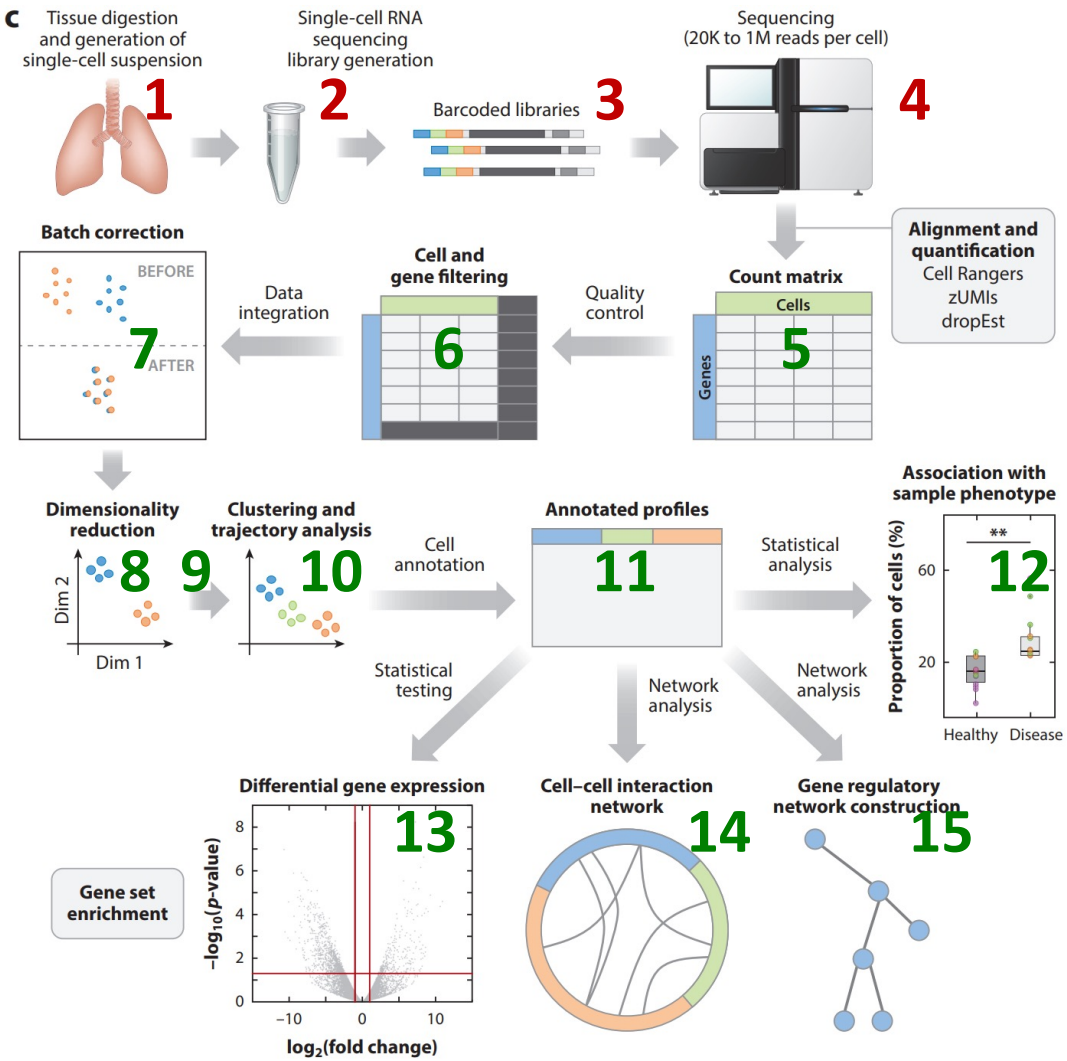
单细胞组简介



https://www.nature.com/articles/s41576-019-0093-7

单细胞转录组技术及分析基本流程

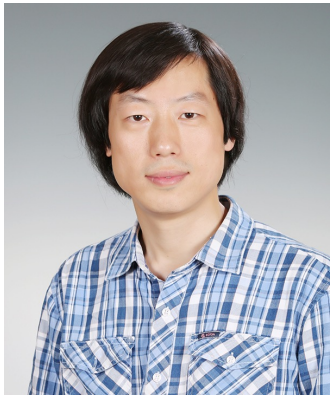
1-样本
2-文库
3-文库
4-测序
5-比对
6-过滤
7-批次
8-归一化
9-降维
10-聚类
11-标注
12-比例
13-差异
14-通讯
15-调控



https://www.annualreviews.org/doi/abs/10.1146/annurev-biodatasci-012220-100601

scRNA-seq

分子实验设计



汤富酬

nature methods

Explore content | About the journal | Publish with us

nature > nature methods > articles > article

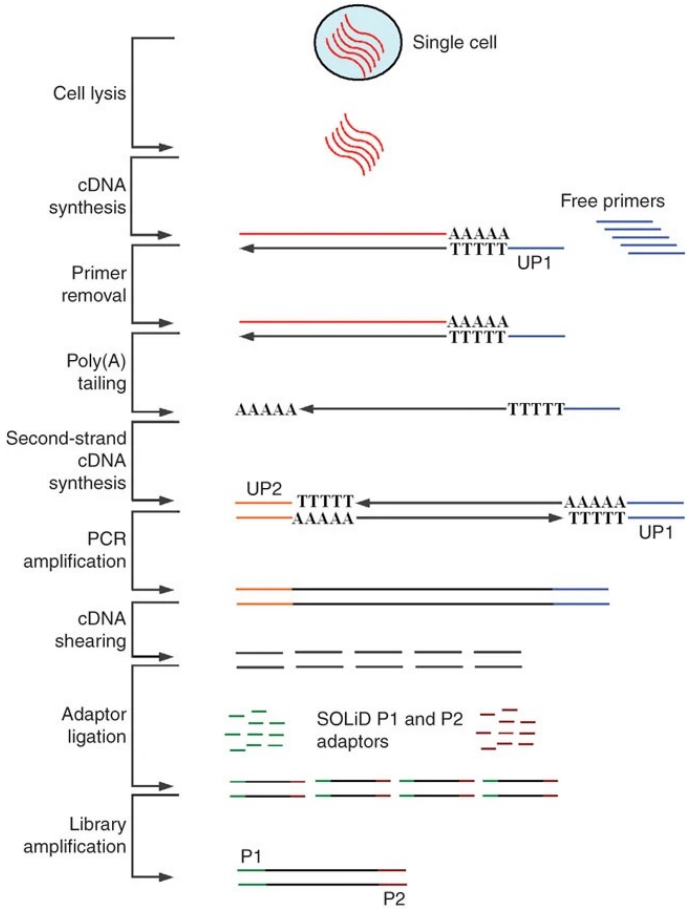
Published: 06 April 2009

mRNA-Seq whole-transcriptome analysis of a single cell

Fuchou Tang, Catalin Barbacioru, Yangzhou Wang, Ellen Nordman, Clarence Lee, Nanlan Xu, Xiaohui Wang, John Bodeau, Brian B Tuch, Asim Siddiqui, Kaiqin Lao & M Azim Surani

Nature Methods **6**, 377–382 (2009) | Cite this article

47k Accesses | **1847** Citations | **126** Altmetric | Metrics

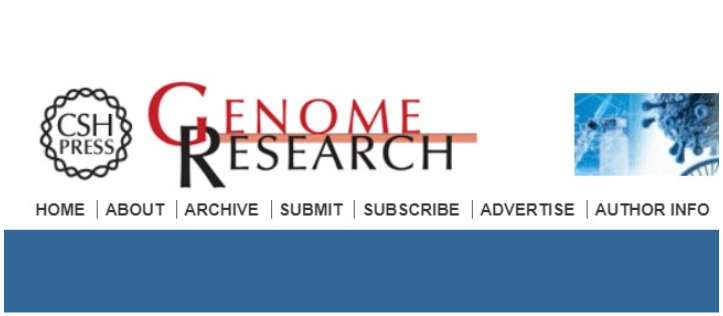


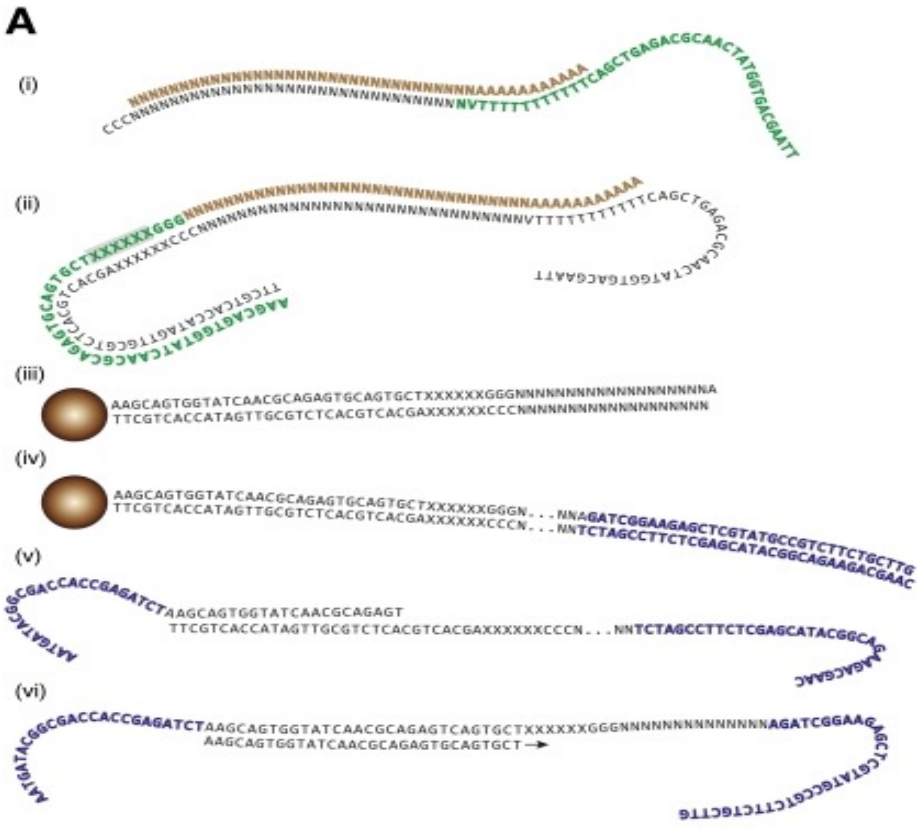
2009年，汤富酬进行第一个基于NGS的单细胞转录组测序

https://www.nature.com/articles/nmeth.1315

STRT-seq

分子实验设计





Sten Linnarsson
Karolinska Institute

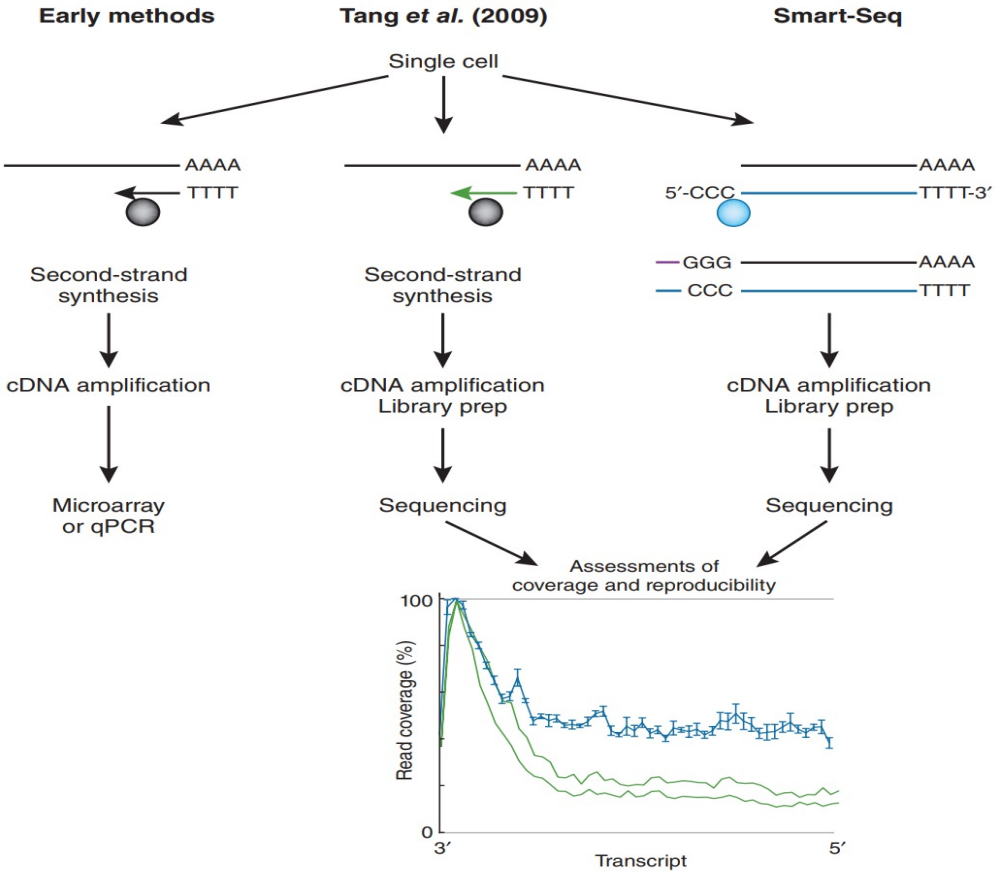
2011年，首次引入barcode

https://genome.cshlp.org/content/21/7/1160.short

SMART-seq

分子实验设计

Rickard Sandberg
Karolinska Institutet



nature biotechnology

Explore content | About the journal | Publish with us

nature › nature biotechnology › news & views › article

Published: 07 August 2012

Transcriptome sequencing of single cells with Smart-Seq

Jillian J Goetz & Jeffrey M Trimarchi

Nature Biotechnology **30**, 763–765 (2012) | Cite this article

10k Accesses | **55** Citations | **7** Altmetric | Metrics

2012年，单细胞全长转录组测序技术：SMART-seq

https://www.nature.com/articles/nbt.2325

STRT-seq

分子实验设计

nature methods

Explore content ⌄ About the journal ⌄ Publish with us ⌄

nature › nature methods › brief communications › article

Published: 22 December 2013

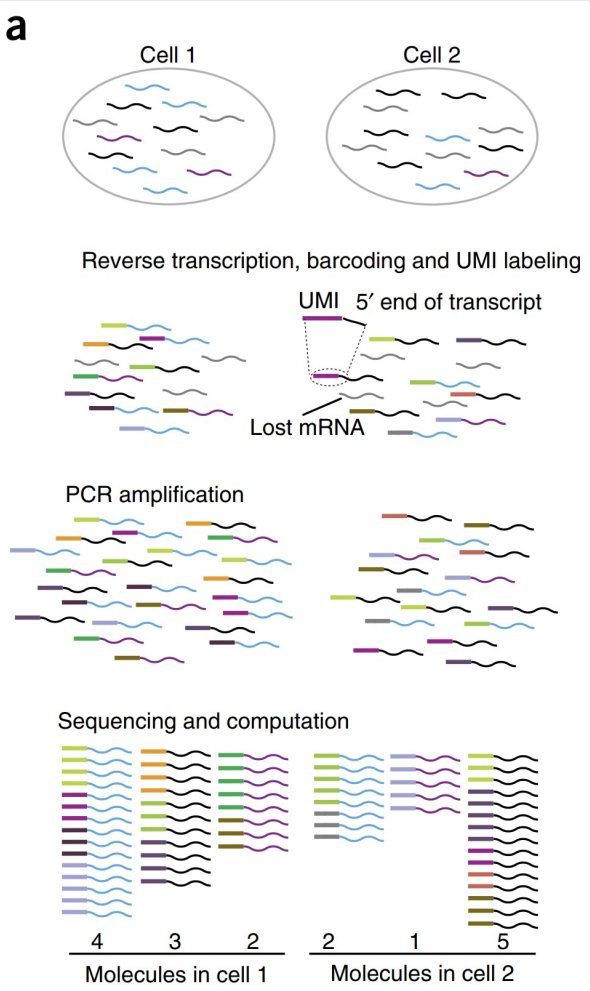
Quantitative single-cell RNA-seq with unique molecular identifiers

Saiful Islam, Amit Zeisel, Simon Joost, Gioele La Manno, Pawel Zajac, Maria Kasper, Peter Lönnerberg & Sten Linnarsson

Nature Methods **11**, 163–166 (2014) | Cite this article

64k Accesses | **735** Citations | **48** Altmetric | Metrics

Sten Linnarsson
Karolinska Institute



2013年，首次引入UMI

https://www.nature.com/articles/nmeth.2772

SMART-seq2

分子实验设计

nature protocols

Explore content ▾ About the journal ▾ Publish with us ▾

nature › nature protocols › protocols › article

Published: 02 January 2014

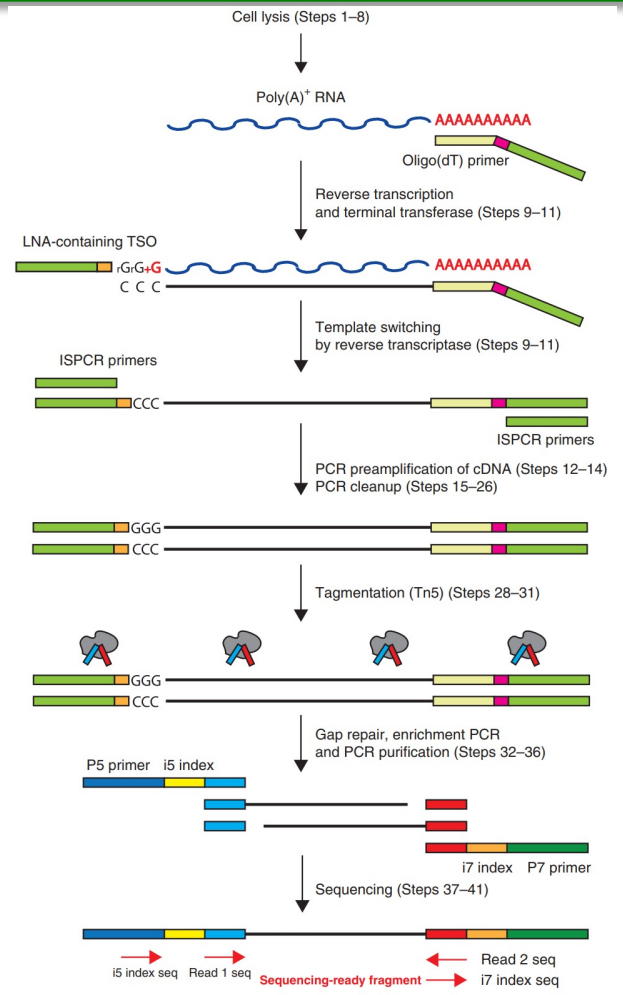
Full-length RNA-seq from single cells using Smart-seq2

Simone Picelli, Omid R Faridani, Åsa K Björklund, Gösta Winberg, Sven Sagasser & Rickard Sandberg

Nature Protocols **9**, 171–181 (2014) | Cite this article

143k Accesses | **1993** Citations | **75** Altmetric | Metrics

Rickard Sandberg
Karolinska Institute



2014年，SMART-seq2：TSO与模板结合热稳定性增强

https://www.nature.com/articles/nprot.2014.006

C1 microfluidics based method

细胞分离方式的改进

Stephen Quake

nature methods

Explore content | About the journal | Publish with us

nature › nature methods › analyses › article

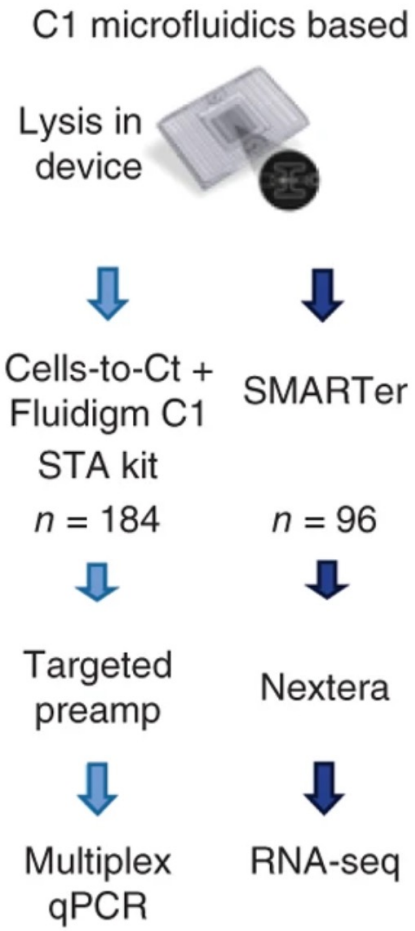
Published: 20 October 2013

Quantitative assessment of single-cell RNA-sequencing methods

Angela R Wu, Norma F Neff, Tomer Kalisky, Piero Dalerba, Barbara Treutlein, Michael E Rothenberg, Francis M Mburu, Gary L Mantalas, Sopheak Sim, Michael F Clarke & Stephen R Quake

Nature Methods **11**, 41–46 (2014) | Cite this article

61k Accesses | **505** Citations | **55** Altmetric | Metrics



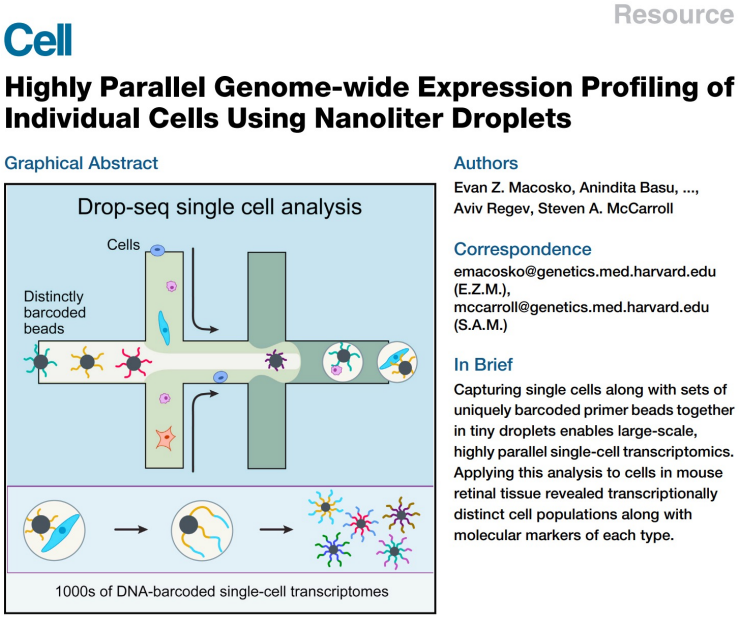
2014年，微流控分离细胞：C1

https://www.nature.com/articles/nmeth.2694

Drop-seq

细胞分离方式的改进

Steve McCarroll
Harvard University



Resource

Cell

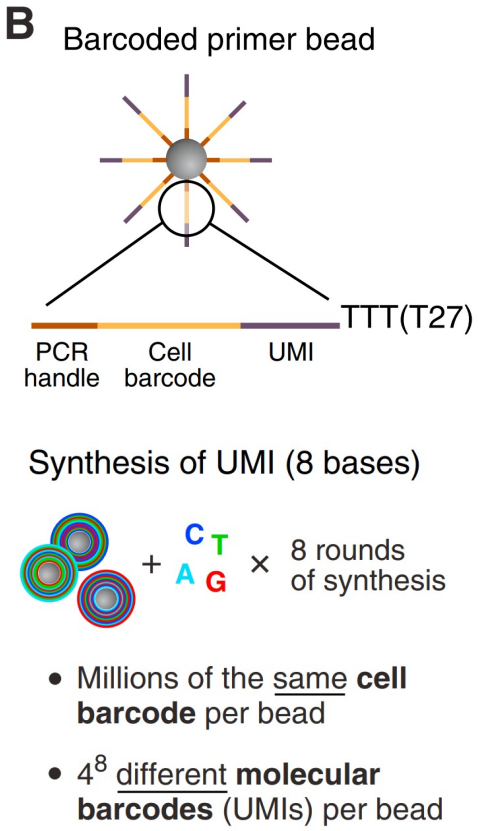
Highly Parallel Genome-wide Expression Profiling of Individual Cells Using Nanoliter Droplets

Graphical Abstract

Authors
Evan Z. Macosko, Anindita Basu, ..., Aviv Regev, Steven A. McCarroll

Correspondence
emacosko@genetics.med.harvard.edu (E.Z.M.), mccarroll@genetics.med.harvard.edu (S.A.M.)

In Brief
Capturing single cells along with sets of uniquely barcoded primer beads together in tiny droplets enables large-scale, highly parallel single-cell transcriptomics. Applying this analysis to cells in mouse retinal tissue revealed transcriptionally distinct cell populations along with molecular markers of each type.

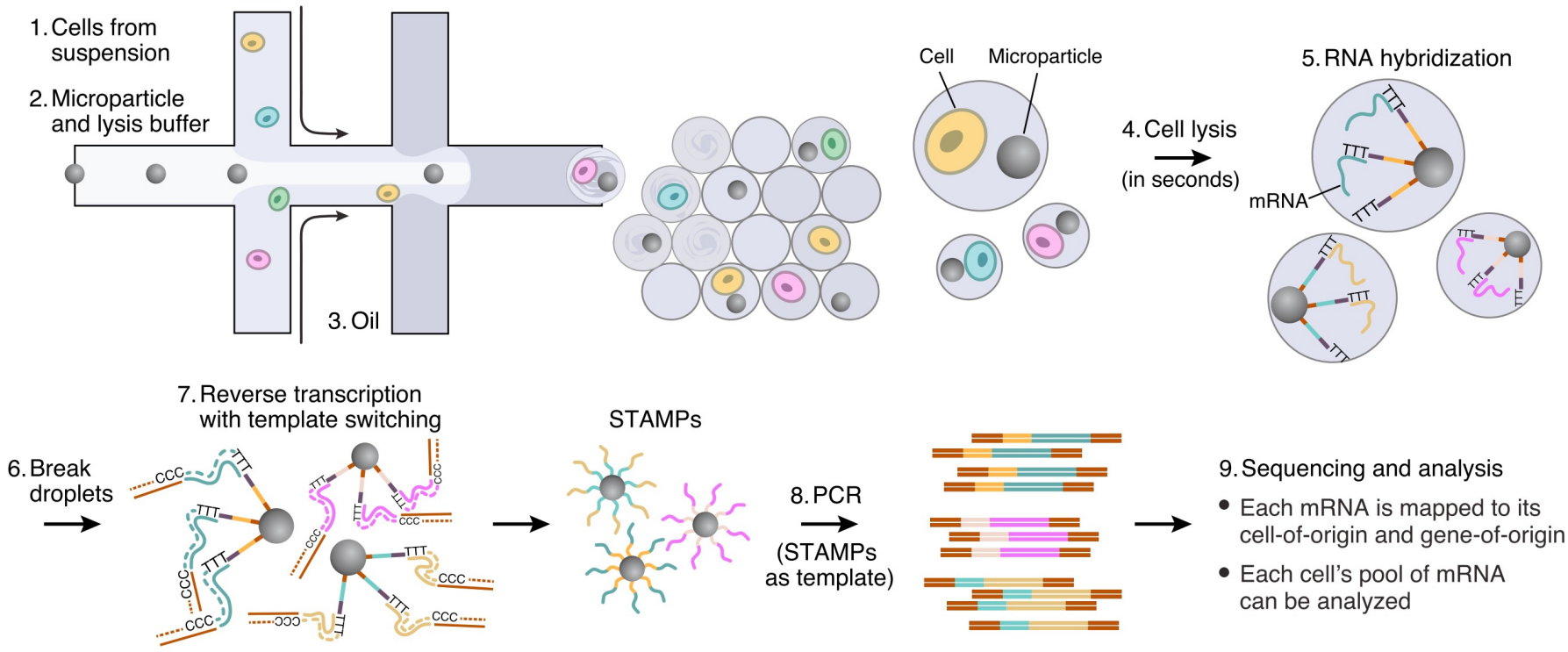


2015年，纳米液滴系统-油包水分离细胞：Drop-seq

https://www.cell.com/cell/fulltext/S0092-8674(15)00549-8

Drop-seq

细胞分离方式的改进



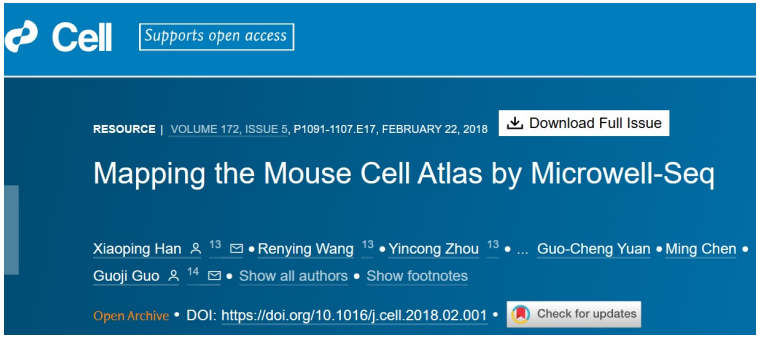
Drop-seq

https://www.cell.com/cell/fulltext/S0092-8674(15)00549-8

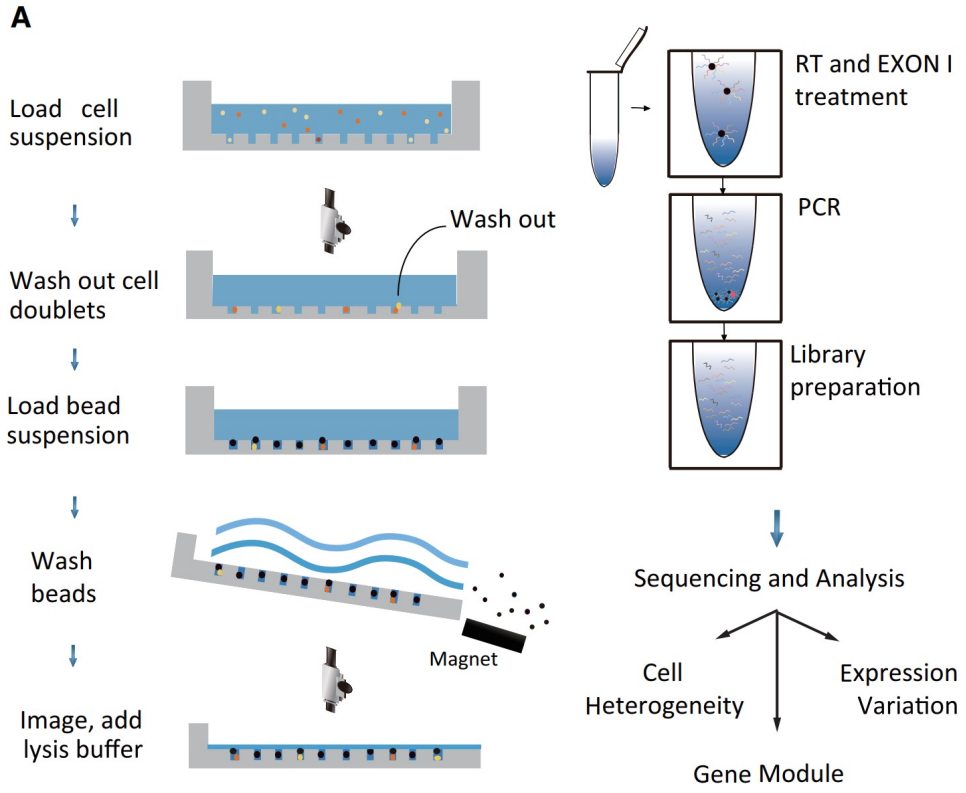
Microwell-seq

细胞分离方式的改进

400，000 single cells



郭国骥
浙江大学

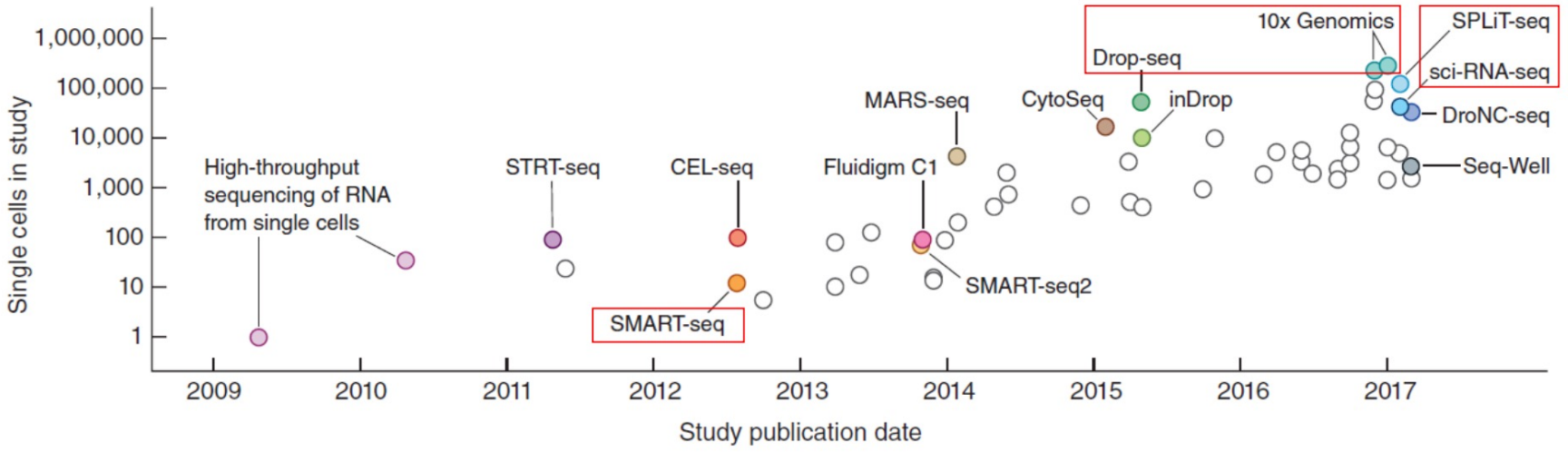


2018年，微孔系统：Microwell-Seq

https://www.cell.com/cell/fulltext/S0092-8674(18)30116-8

scRNA-seq技术发展时间轴

Cell numbers reported in representative publications by publication date. Key technologies are indicated.

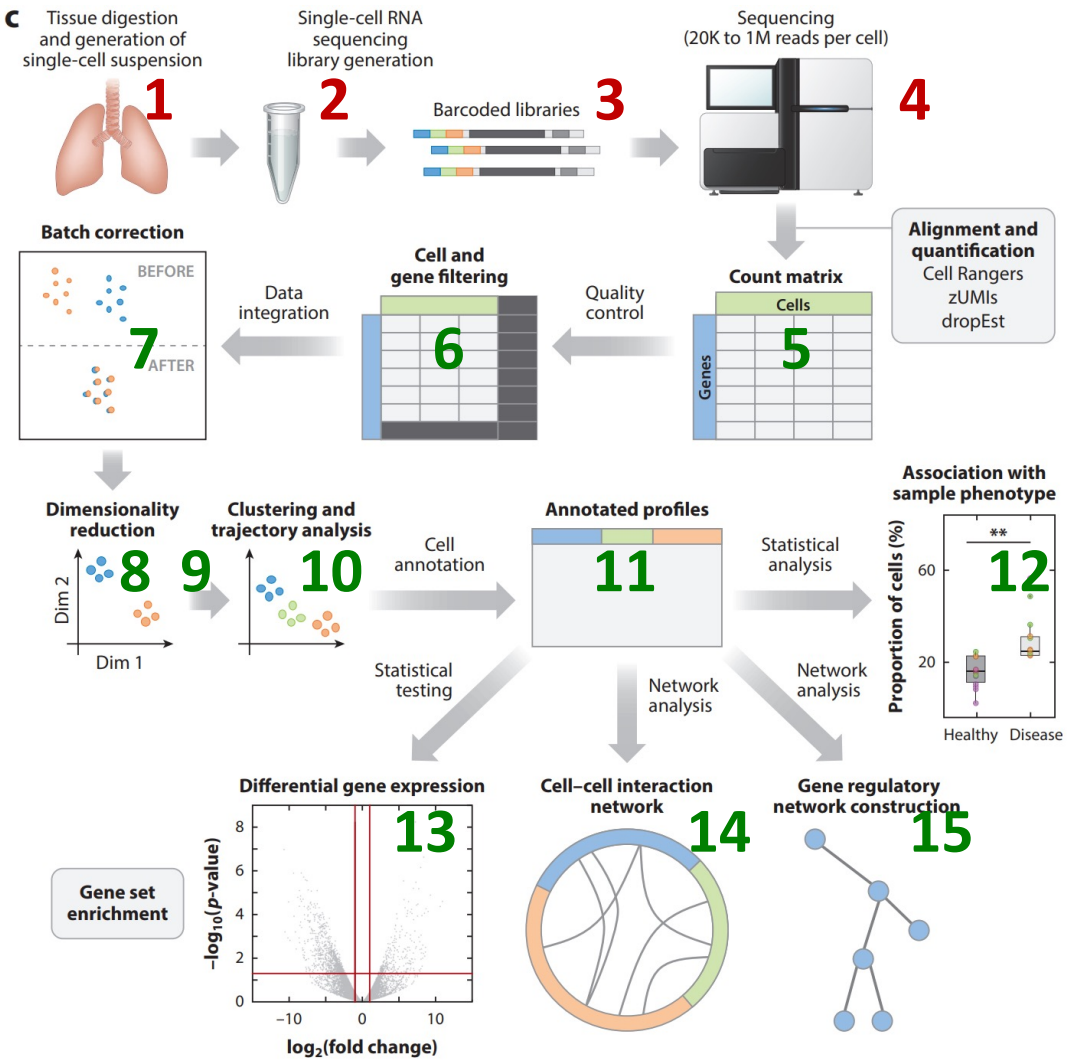


Svensson V, Vento-Tormo R, Teichmann SA. Nat Protoc. 2018 Apr;13(4):599-604.

https://www.nature.com/articles/nprot.2017.149

单细胞转录组技术及分析基本流程

1-样本
2-文库
3-文库
4-测序
5-比对
6-过滤
7-批次
8-归一化
9-降维
10-聚类
11-标注
12-比例
13-差异
14-通讯
15-调控



https://www.annualreviews.org/doi/abs/10.1146/annurev-biodatasci-012220-100601

5. 比对-Cellranger

1.提取Cell barcode和UMI

2.数据质控:cutadapt

3.数据比对:STAR

4.转录本定量:featureCounts

5.获取单细胞基因表达数矩阵:umi_toolscount

https://support.10xgenomics.com/single-cell-multiome-atac-gex/software/pipelines/latest/algorithms/overview

5. 比对-Cellranger

Estimated Number of Cells

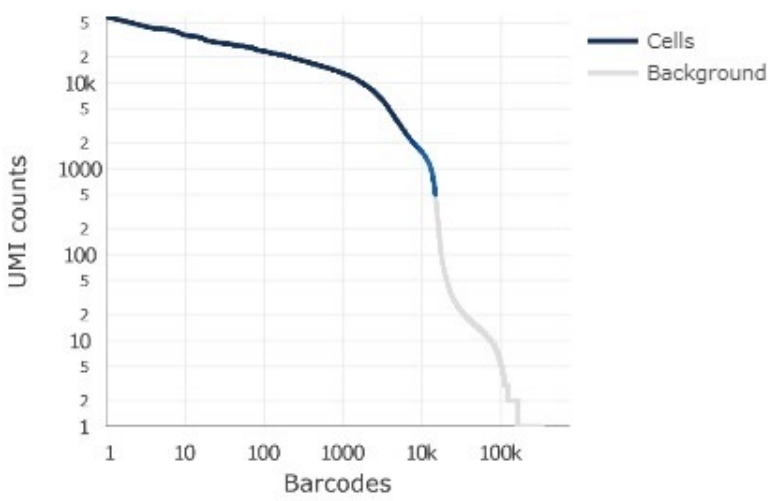
12,342

Mean Reads per Cell	Median Genes per Cell
28,019	1,082

Sequencing

Number of Reads	345,812,670
Valid Barcodes	94.0%
Sequencing Saturation	73.2%
Q30 Bases in Barcode	96.6%
Q30 Bases in RNA Read	94.5%
Q30 Bases in RNA Read 2	93.1%
Q30 Bases in Sample Index	94.1%
Q30 Bases in UMI	96.3%

Cells

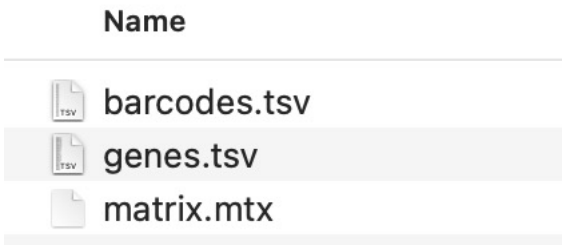


Estimated Number of Cells	12,342
Fraction Reads in Cells	92.3%
Mean Reads per Cell	28,019
Median Genes per Cell	1,082
Total Genes Detected	23,006
Median UMI Counts per Cell	2,594

https://support.10xgenomics.com/single-cell-multiome-atac-gex/software/pipelines/latest/algorithms/overview

Cellranger输入输出文件

- 输入文件：[SampleName]_S1_L00[LaneNumber]_[ReadType]_001.fastq.gz
- 输出文件：
  - 比对文件
  - 索引文件
  - 矩阵文件



6. 质控

QC and selecting cells for further analysis

- 细胞的过滤
  - 基因数:>200，1k， 2k
  - 线粒体比例:<5%
- 基因的过滤
  - 过滤不表达基因
  - 细胞数:>3
- 线粒体基因和红细胞基因比例的筛选
  - 线粒体基因含量高的细胞可能是凋亡的细胞
  - 很少有细胞表达红细胞基因

Seurat - Guided Clustering Tutorial

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

Setup the Seurat Object

For this tutorial, we will be analyzing the a dataset of Peripheral Blood Mononuclear Cells (PBMC) freely available from 10X Genomics. There are 2,700 single cells that were sequenced on the Illumina NextSeq 500. The raw data can be found here.

We start by reading in the data. The `Read10X()` function reads in the output of the cellranger pipeline from 10X, returning a unique molecular identified (UMI) count matrix. The values in this matrix represent the number of molecules for each feature (i.e. gene; row) that are detected in each cell (column). Note that more recent versions of cellranger now also output using the h5 file format, which can be read in using the `Read10X_h5()` function in Seurat.

We next use the count matrix to create a `Seurat` object. The object serves as a container that contains both data (like the count matrix) and analysis (like PCA, or clustering results) for a single-cell dataset. For more information, check out our [Seurat object interaction vignette], or our GitHub Wiki. For example, in Seurat v5, the count matrix is stored in `pbmc[["RNA"]]$counts`.

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

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

Seurat - Guided Clustering Tutorial

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

Setup the Seurat Object

For this tutorial, we will be analyzing the a dataset of Peripheral Blood Mononuclear Cells (PBMC) freely available from 10X Genomics. There are 2,700 single cells that were sequenced on the Illumina NextSeq 500. The raw data can be found here.

We start by reading in the data. The `Read10X()` function reads in the output of the cellranger pipeline from 10X, returning a unique molecular identified (UMI) count matrix. The values in this matrix represent the number of molecules for each feature (i.e. gene; row) that are detected in each cell (column). Note that more recent versions of cellranger now also output using the h5 file format, which can be read in using the `Read10X_h5()` function in Seurat.

We next use the count matrix to create a `Seurat` object. The object serves as a container that contains both data (like the count matrix) and analysis (like PCA, or clustering results) for a single-cell dataset. For more information, check out our [Seurat object interaction vignette], or our GitHub Wiki. For example, in Seurat v5, the count matrix is stored in `pbmc[["RNA"]]$counts`.

Copy to clipboard

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

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

QC and selecting cells for further analysis

QC and selecting cells for further analysis

Seurat allows you to easily explore QC metrics and filter cells based on any user-defined criteria. A few QC metrics commonly used by the community include

- The number of unique genes detected in each cell.
  - Low-quality cells or empty droplets will often have very few genes
  - Cell doublets or multiplets may exhibit an aberrantly high gene count
- Similarly, the total number of molecules detected within a cell (correlates strongly with unique genes)
- The percentage of reads that map to the mitochondrial genome
  - Low-quality / dying cells often exhibit extensive mitochondrial contamination
  - We calculate mitochondrial QC metrics with the `PercentageFeatureSet()` function, which calculates the percentage of counts originating from a set of features
  - We use the set of all genes starting with `MT-` as a set of mitochondrial genes

```
# The [[ operator can add columns to object metadata. This is a great place to stash QC stats
pbmc[["percent.mt"]] <- PercentageFeatureSet(pbmc, pattern = "^MT-")
```

▶ **Where are QC metrics stored in Seurat?**

QC and selecting cells for further analysis

```
# The [[ operator can add columns to object metadata. This is a great place to stash QC stats
pbmc[["percent.mt"]] <- PercentageFeatureSet(pbmc, pattern = "^MT-")
```

▼ Where are QC metrics stored in Seurat?

- The number of unique genes and total molecules are automatically calculated during `CreateSeuratObject()`
  - You can find them stored in the object meta data

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

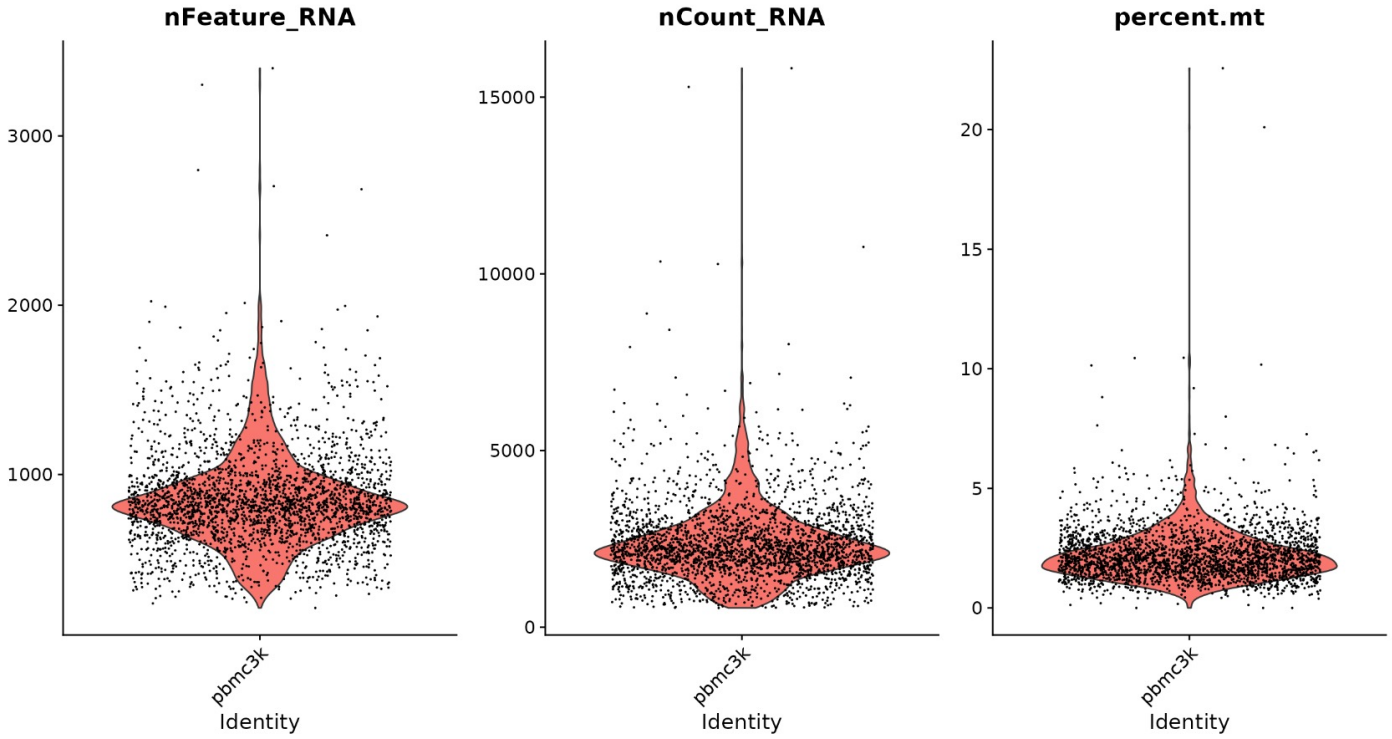
```
##                 orig.ident nCount_RNA nFeature_RNA percent.mt
## AAACATACAACCAC-1     pbmc3k       2419          779  3.0177759
## AAACATTGAGCTAC-1     pbmc3k       4903         1352  3.7935958
## AAACATTGATCAGC-1     pbmc3k       3147         1129  0.8897363
## AAACCGTGCTTCCG-1     pbmc3k   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

QC and selecting cells for further analysis

In the example below, we visualize QC metrics, and use these to filter cells.

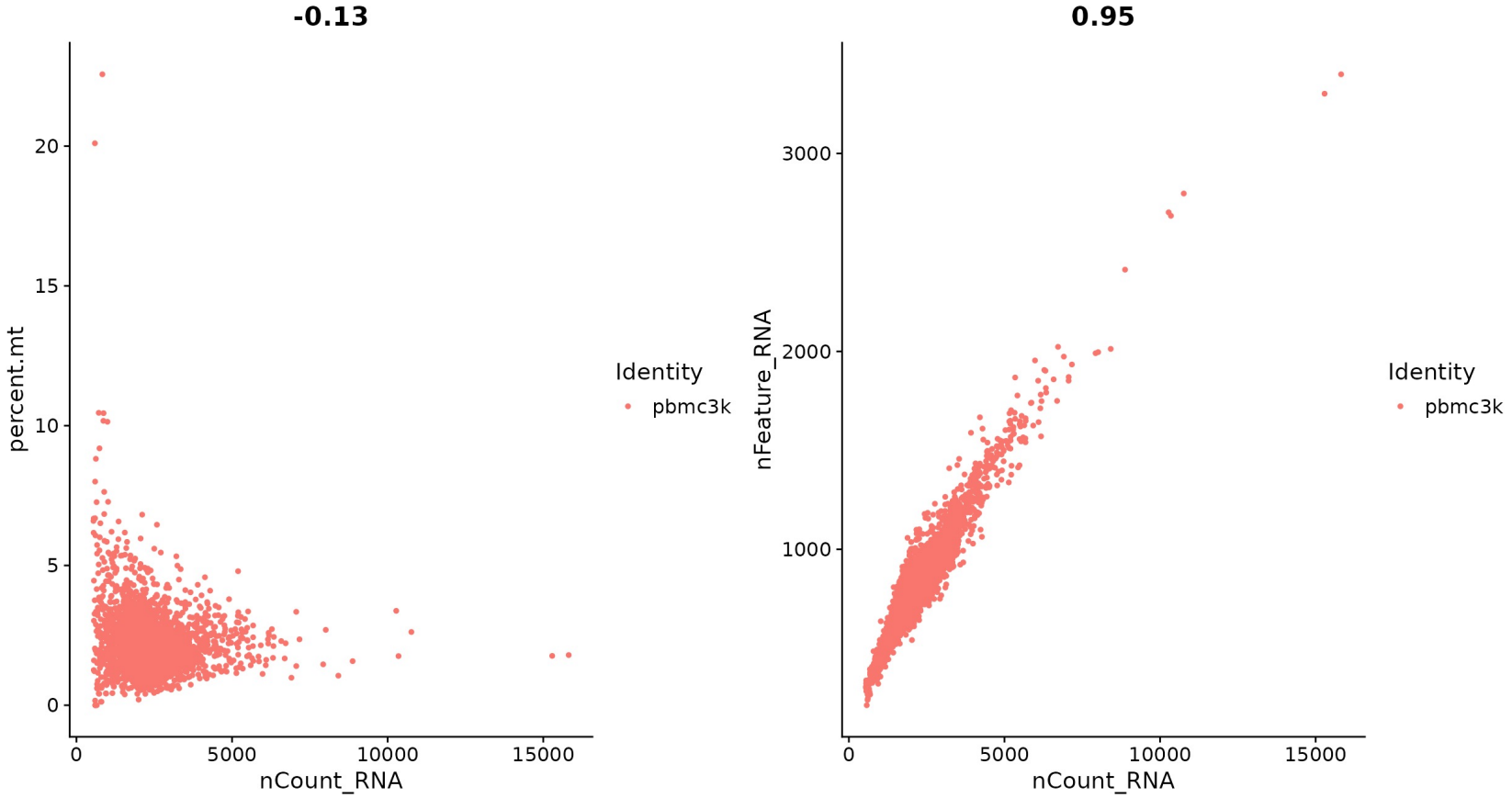
- We filter cells that have unique feature counts over 2,500 or less than 200
- We filter cells that have >5% mitochondrial counts

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



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

FeatureScatter is typically used to visualize feature-feature relationships



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

Normalizing the data（单细胞数据的标准化）

- **常见标准化方法——CPM**

➢ 从取样到建库上机每个操作步骤的可变性，即便同一个细胞测序两次获得的计数深度也可能会有所不同。因此，当基于原始计数数据比较细胞之间的基因表达时，得到的差异可能来自于技术原因。

➢ 使得**细胞**之间可以相互比较

➢ 测序深度（sequencing depth）标准化

- Downsampling：抽取等量reads
- **CPM (count per million)**

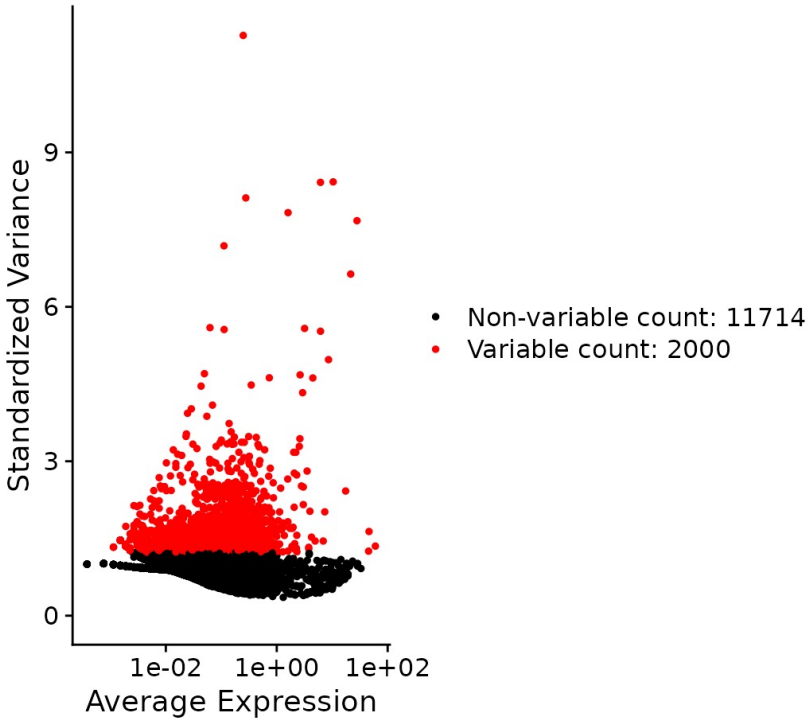
➢ $CPM=10^6*Cij/\sum_i Cij$

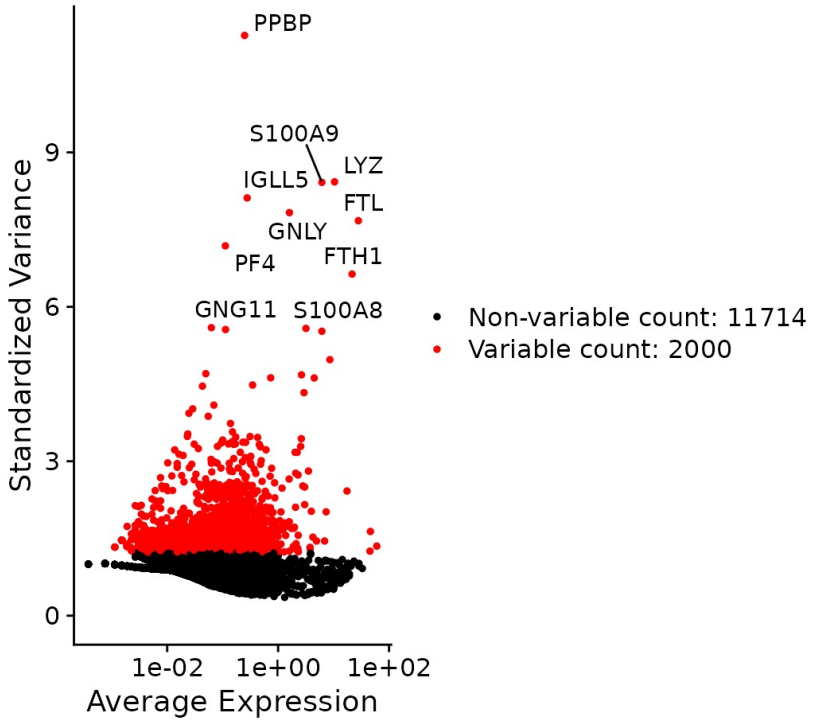
➢ data=log((CPM/10) + 1)

$$CPM = \frac{Number\ of\ reads\ mapped\ to\ gene \times 10^6}{Total\ number\ of\ mapped\ reads}$$

Seurat的默认**标准化方法**是:每个细胞的某一count先除以该细胞的总count，然后乘以scale因子-10000，再做个对数转换。

Identification of highly variable features (feature selection)





单细胞数据的归一化(scale/z-score)

- **基因归一化是指一个基因减去其在所有样品表达的均值然后除以其在所有样品表达值的标准差。**

$$x_i' = \frac{x_i - \bar{x}}{s}$$

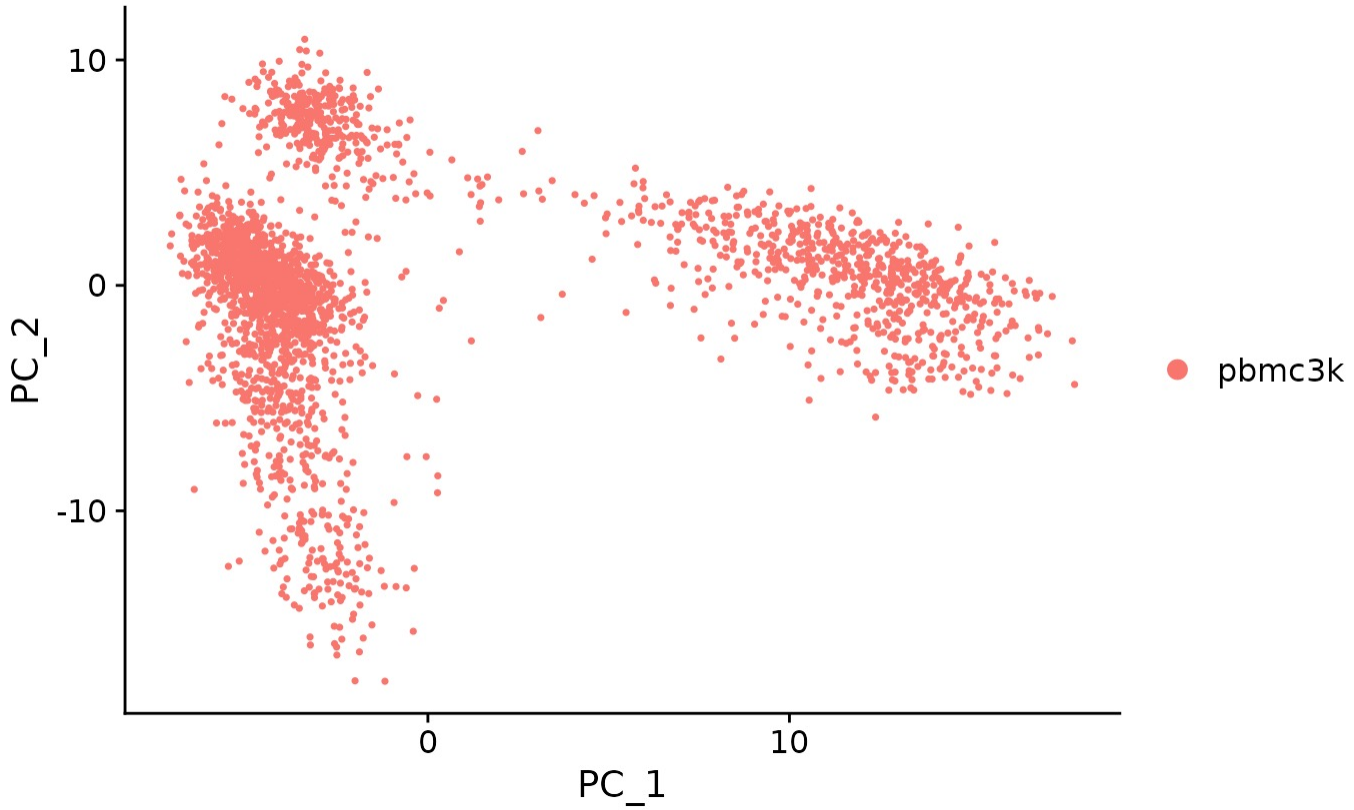
Scaling the data

Next, we apply a linear transformation ('scaling') that is a standard pre-processing step prior to dimensional reduction techniques like PCA. The `ScaleData()` function:

- Shifts the expression of each gene, so that the mean expression across cells is 0
- Scales the expression of each gene, so that the variance across cells is 1
  - This step gives equal weight in downstream analyses, so that highly-expressed genes do not dominate
- The results of this are stored in `pbmc[["RNA"]]@scale.data`

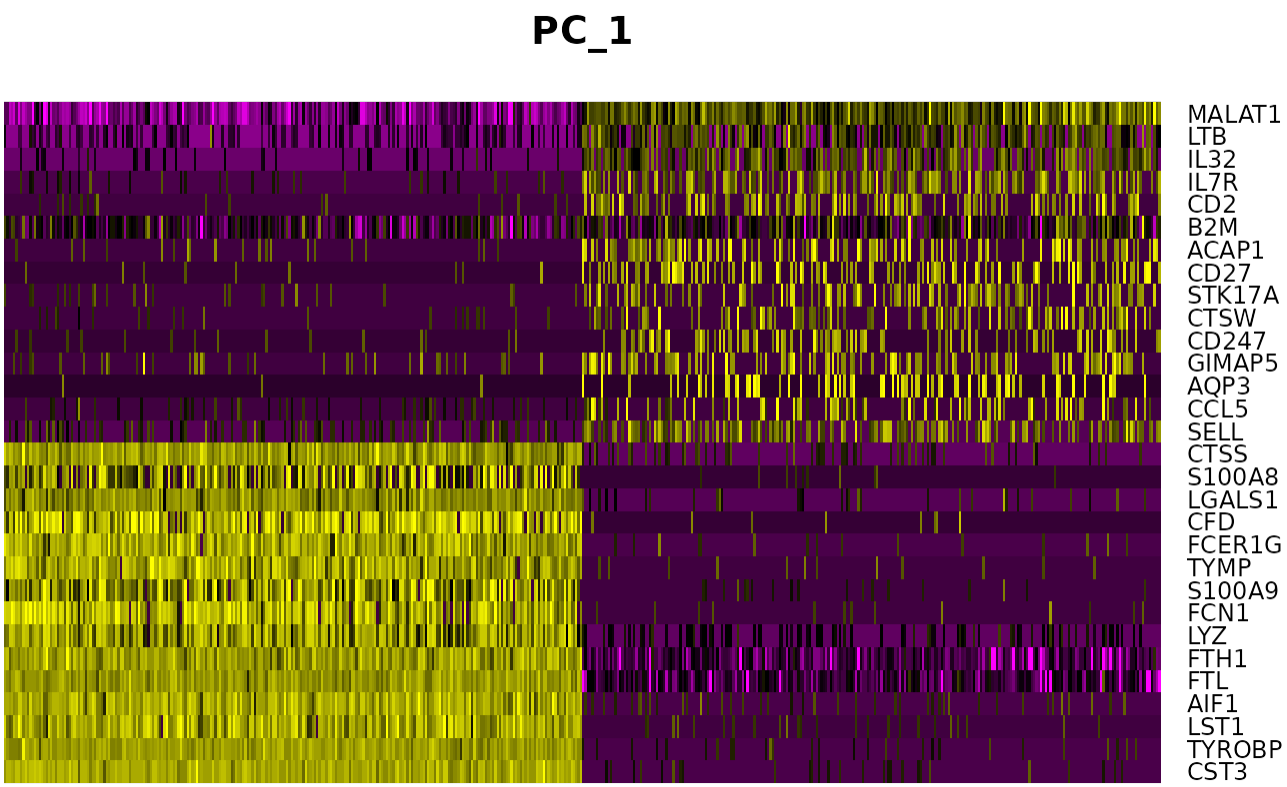
```
all.genes <- rownames(pbmc)
pbmc <- ScaleData(pbmc, features = all.genes)
```

Perform linear dimensional reduction



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

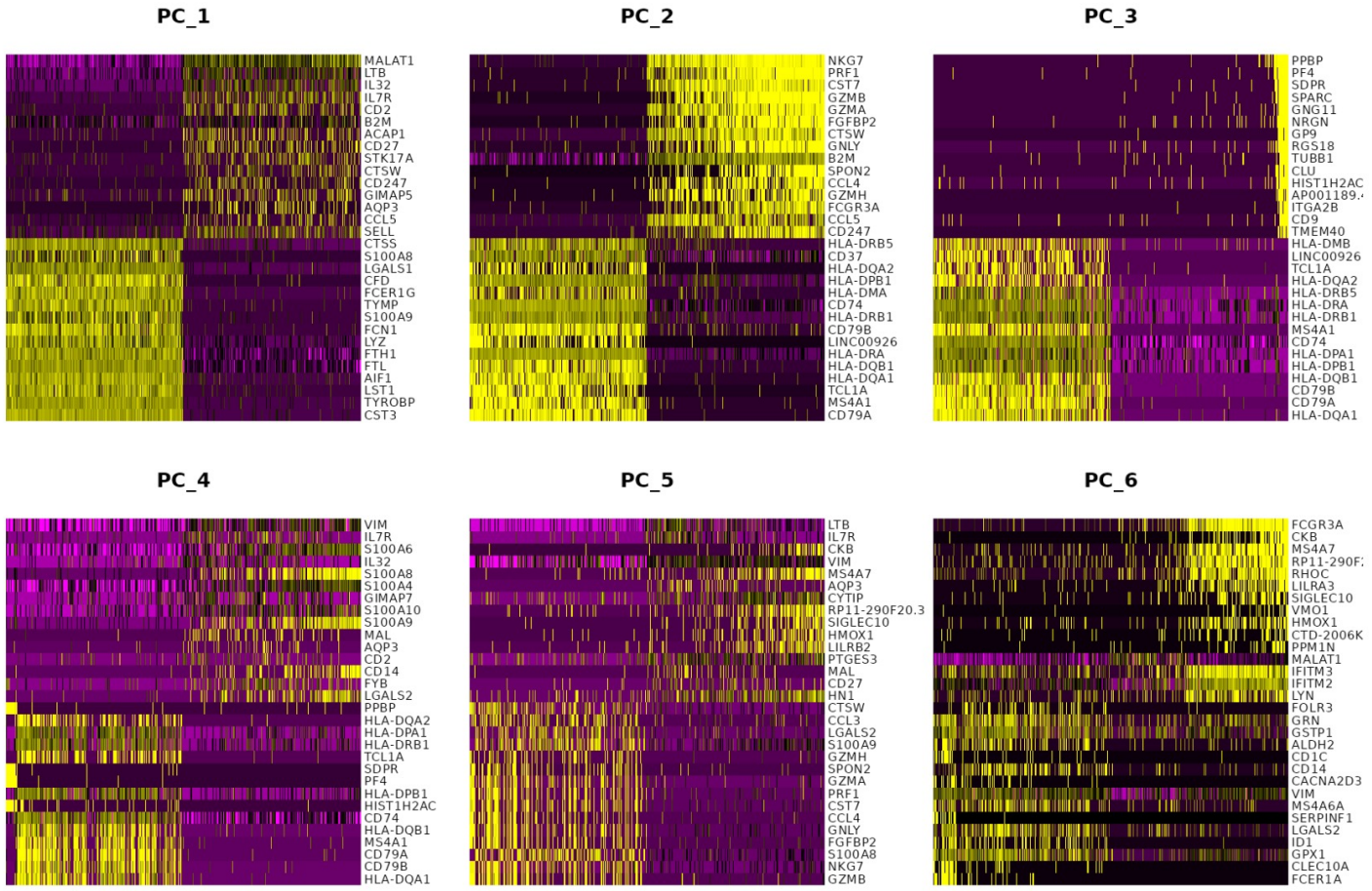
Perform linear dimensional reduction




```
DimHeatmap(pbmc, dims = 1, cells = 500, balanced = TRUE)
```

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

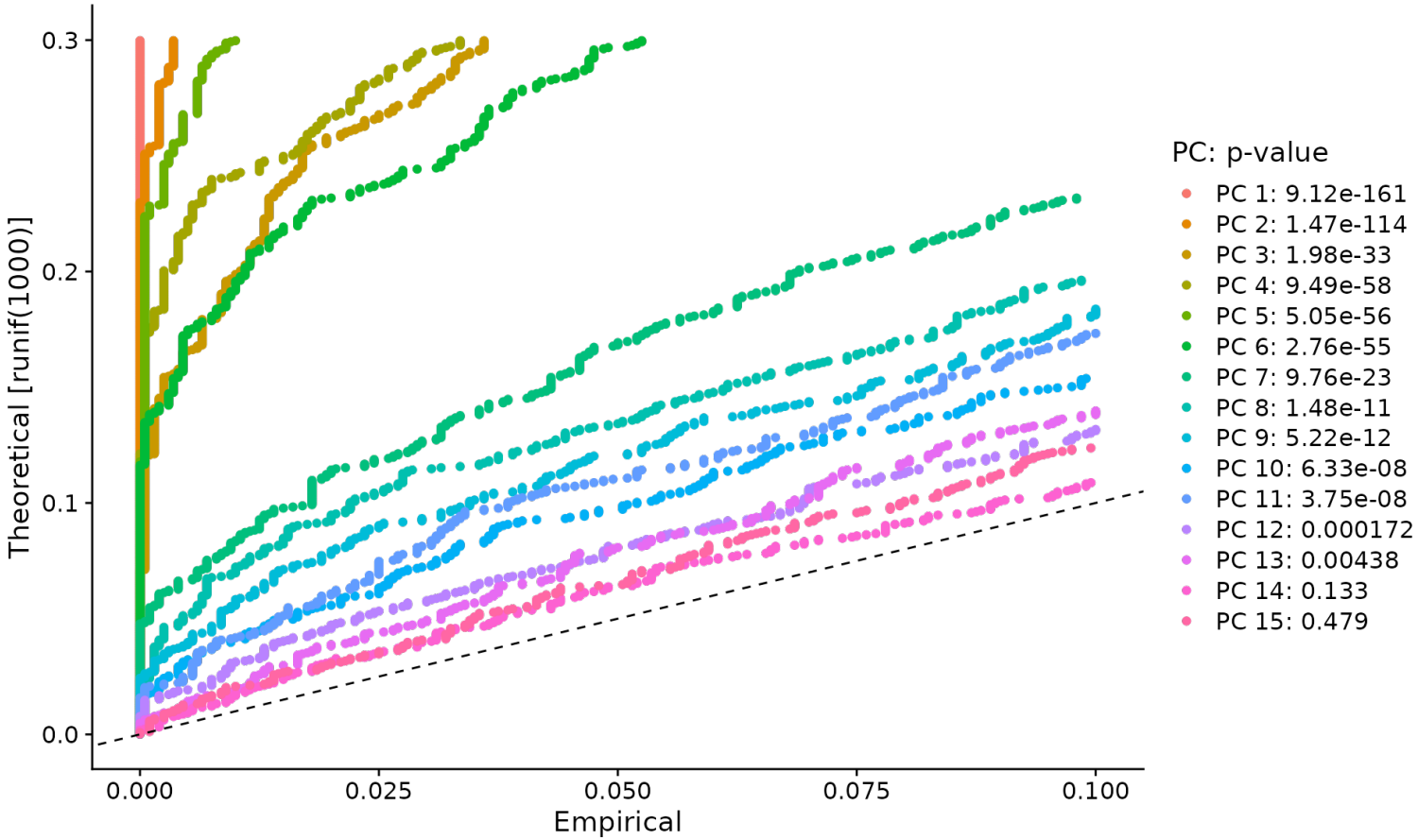
Perform linear dimensional reduction




```
DimHeatmap(pbmc, dims = 1, cells = 500, balanced = TRUE)
```

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

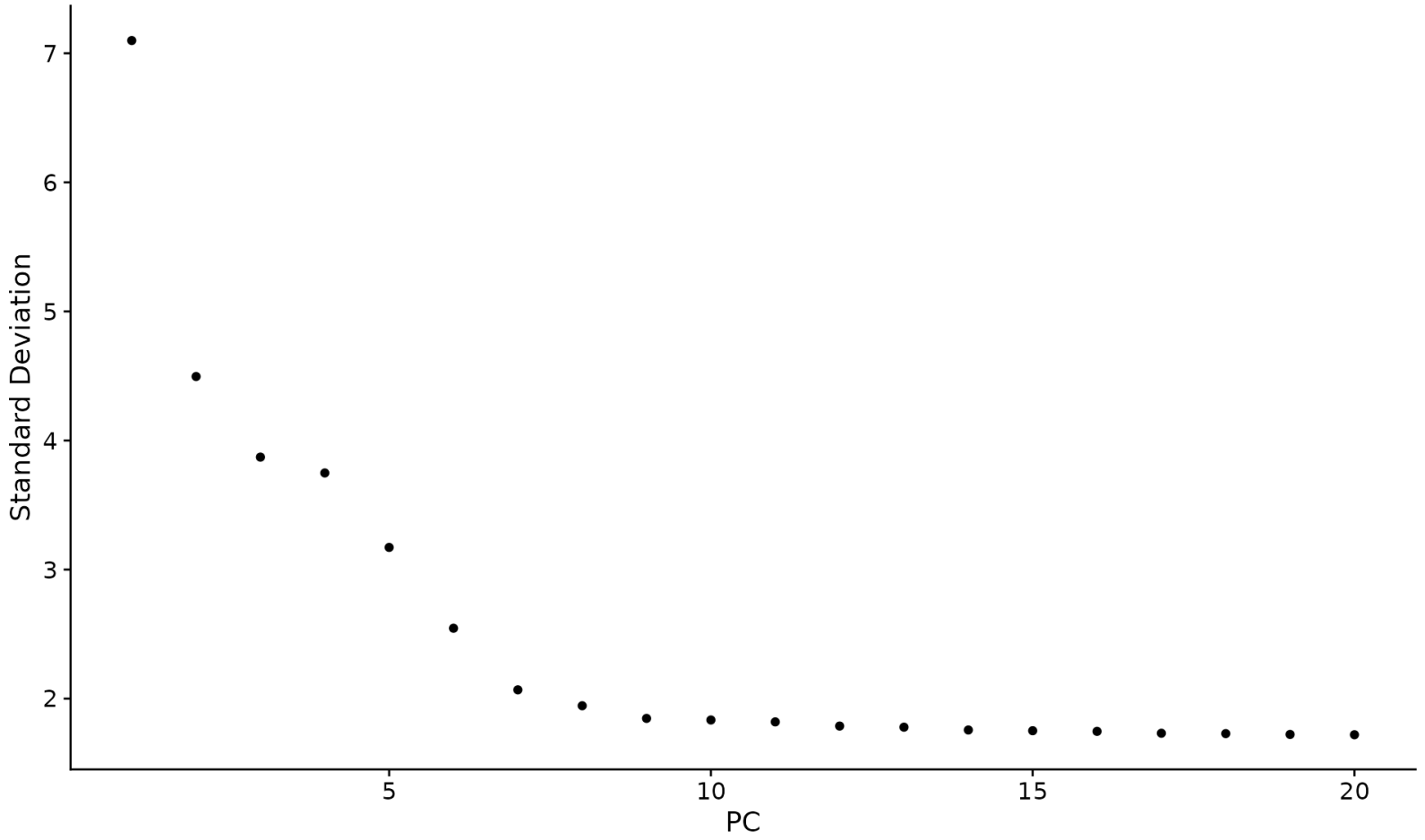
Determine the ‘dimensionality’ of the dataset




```
JackStrawPlot(pbmc, dims = 1:15)
```

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

Determine the ‘dimensionality’ of the dataset



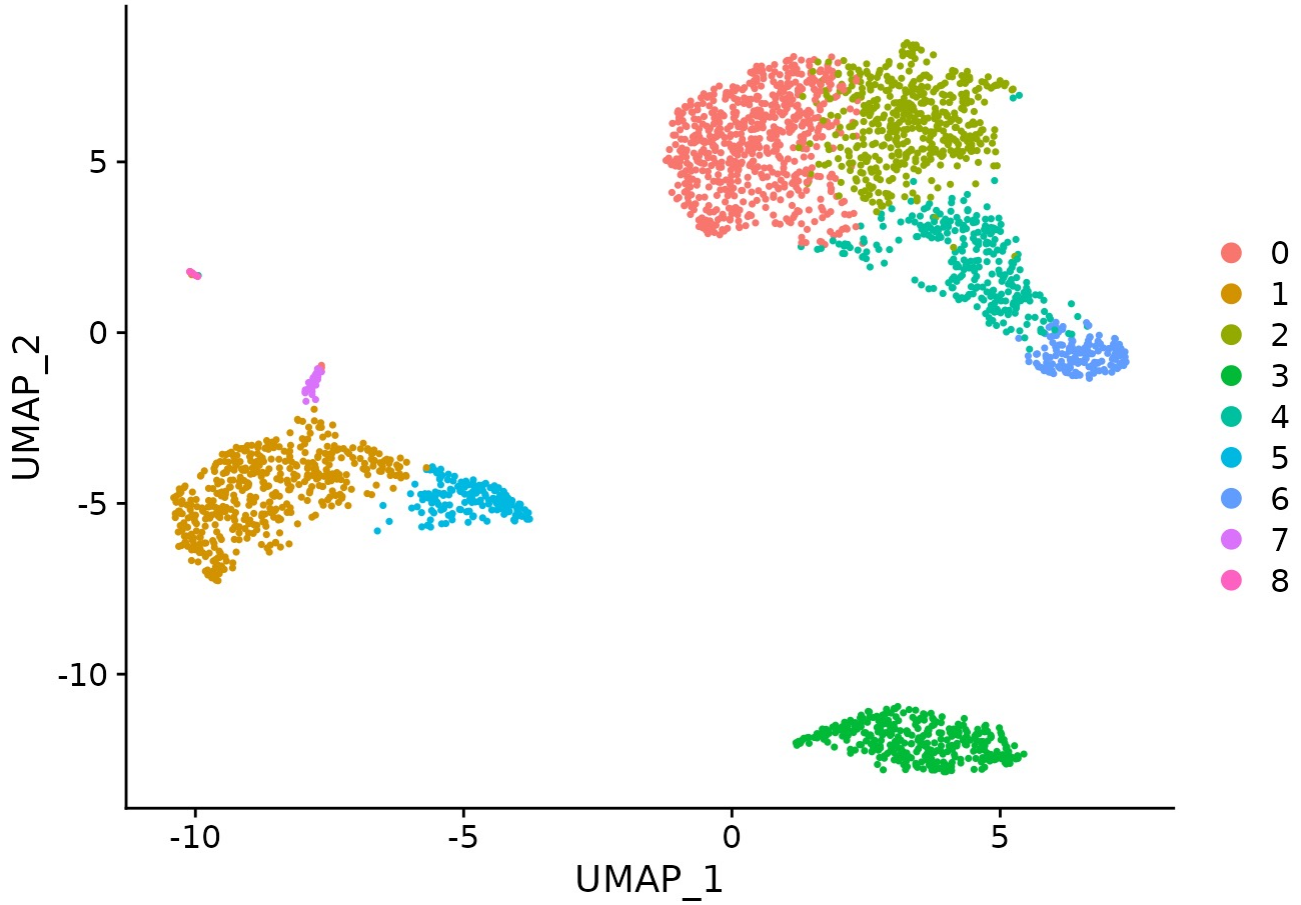
Cluster the cells

- Briefly, these methods embed cells in a graph structure - for example a K-nearest neighbor (KNN) graph, with edges drawn between cells with similar feature expression patterns, and then attempt to partition this graph into highly interconnected ‘quasi-cliques’ or ‘communities’.

- As in PhenoGraph, we first construct a KNN graph based on the euclidean distance in PCA space, and refine the edge weights between any two cells based on the shared overlap in their local neighborhoods (Jaccard similarity). This step is performed using the **FindNeighbors()** function, and takes as input the previously defined dimensionality of the dataset (first 10 PCs).

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

Run non-linear dimensional reduction (UMAP/tSNE)



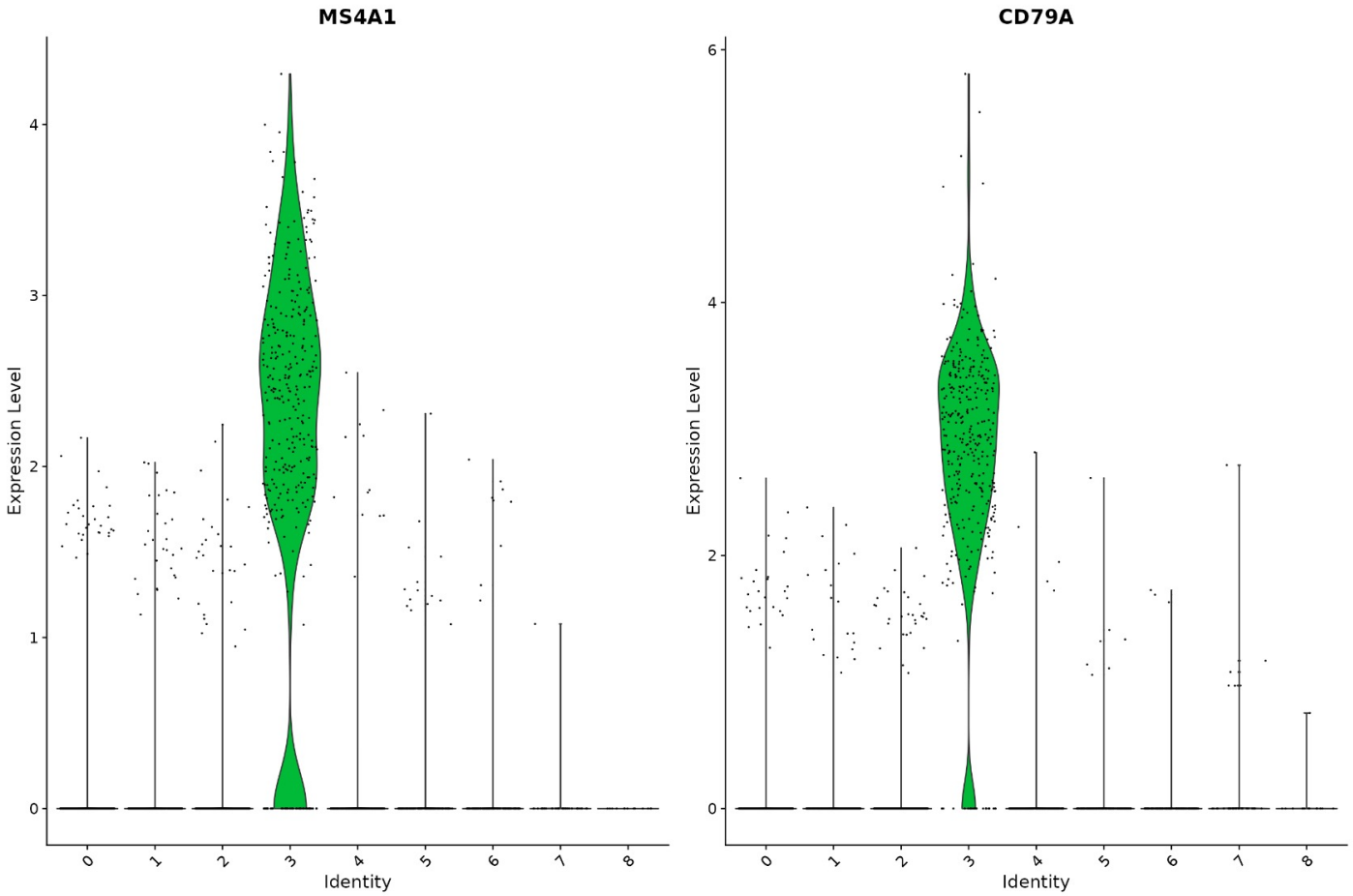
Finding differentially expressed features (cluster biomarkers)

```
# find all markers of cluster 2
cluster2.markers <- FindMarkers(pbmc, ident.1 = 2, min.pct = 0.25)
head(cluster2.markers, n = 5)
```

```
##              p_val avg_log2FC pct.1 pct.2    p_val_adj
## IL32 2.892340e-90  1.2013522 0.947 0.465 3.966555e-86
## LTB  1.060121e-86  1.2695776 0.981 0.643 1.453850e-82
## CD3D 8.794641e-71  0.9389621 0.922 0.432 1.206097e-66
## IL7R 3.516098e-68  1.1873213 0.750 0.326 4.821977e-64
## LDHB 1.642480e-67  0.8969774 0.954 0.614 2.252497e-63
```

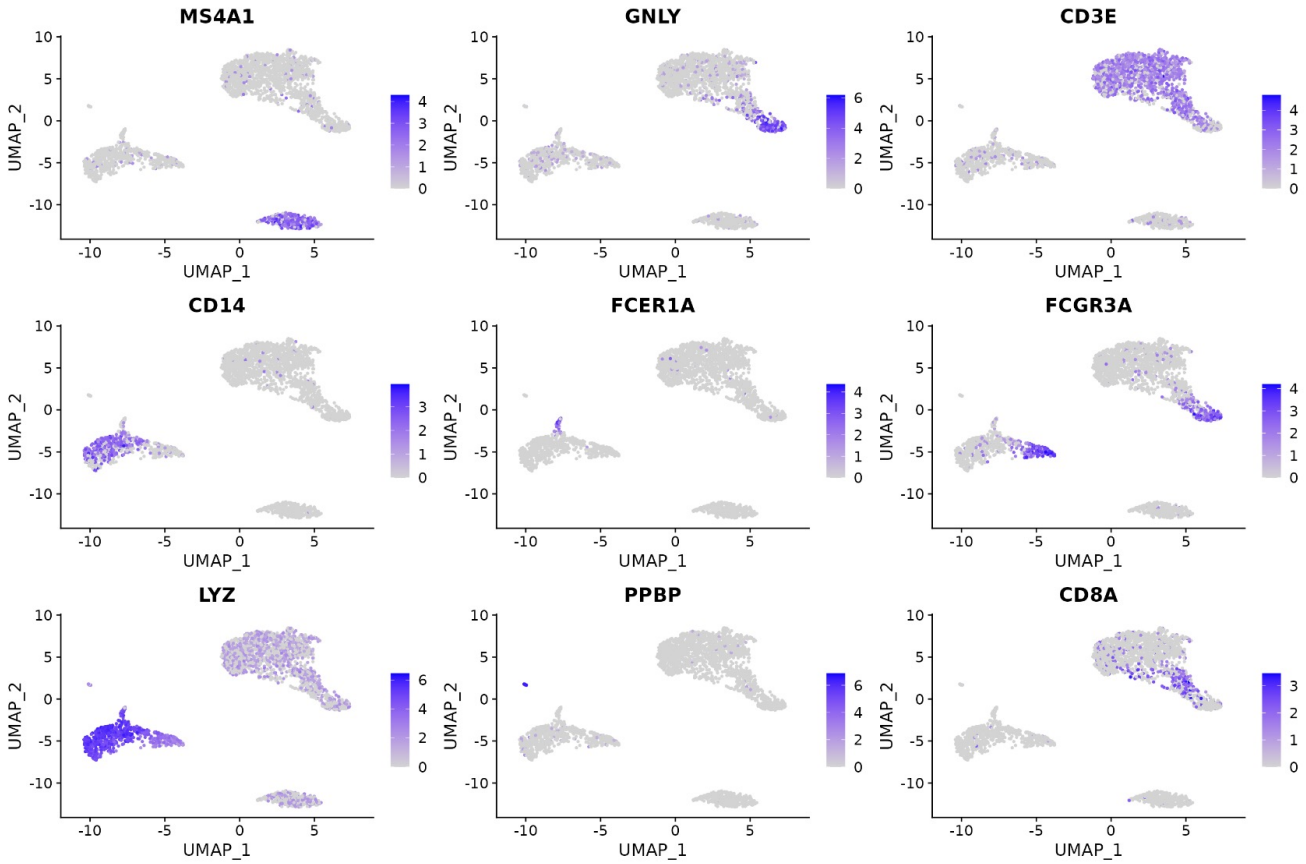
visualizing marker expression

```
VlnPlot(pbmc, features = c("MS4A1", "CD79A"))
```

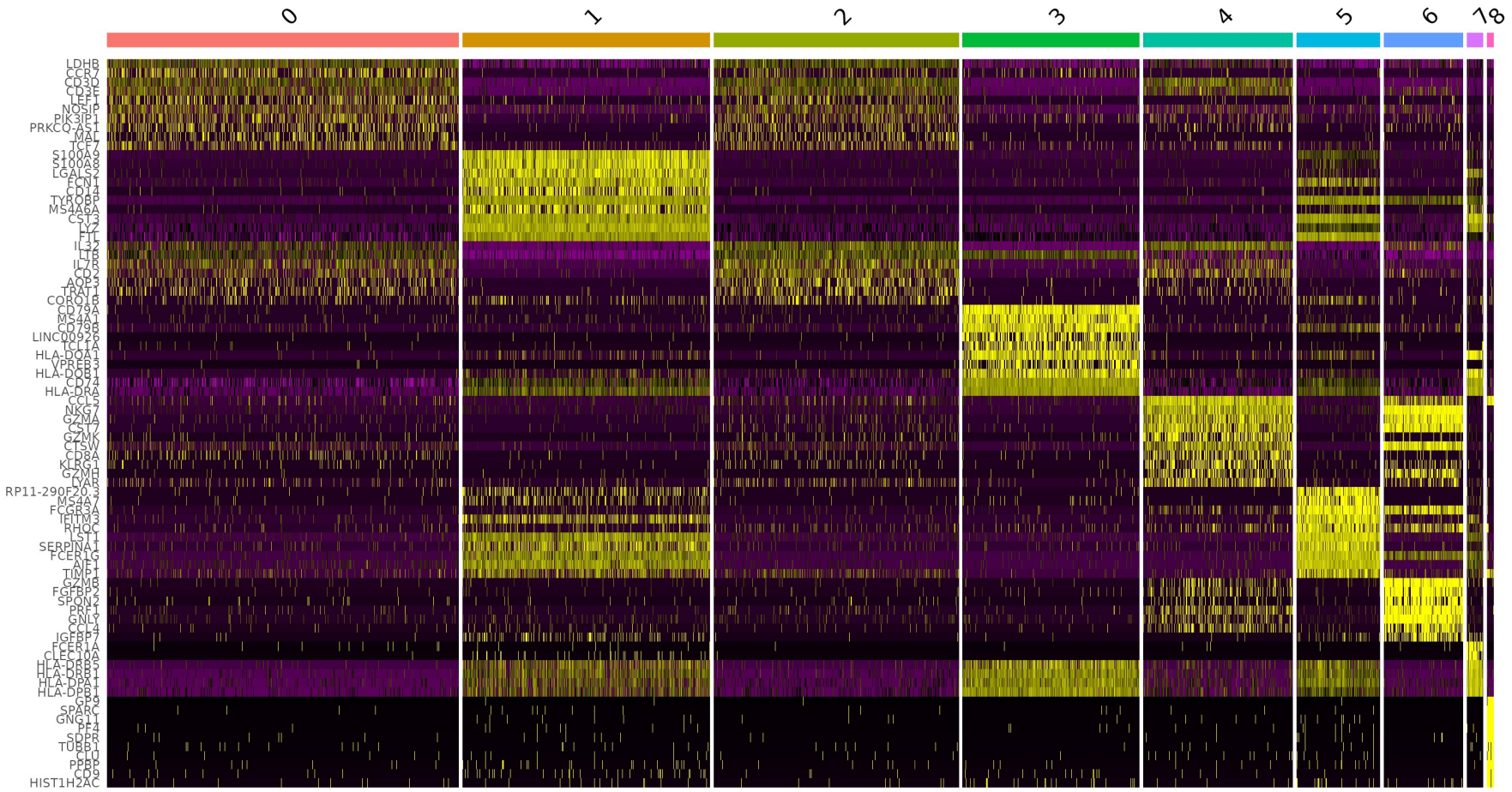


visualizing marker expression

```
FeaturePlot(pbmc, features = c("MS4A1", "GNLY", "CD3E", "CD14", "FCER1A", "FCGR3A", "LYZ", "PPBP",
    "CD8A"))
```



visualizing marker expression

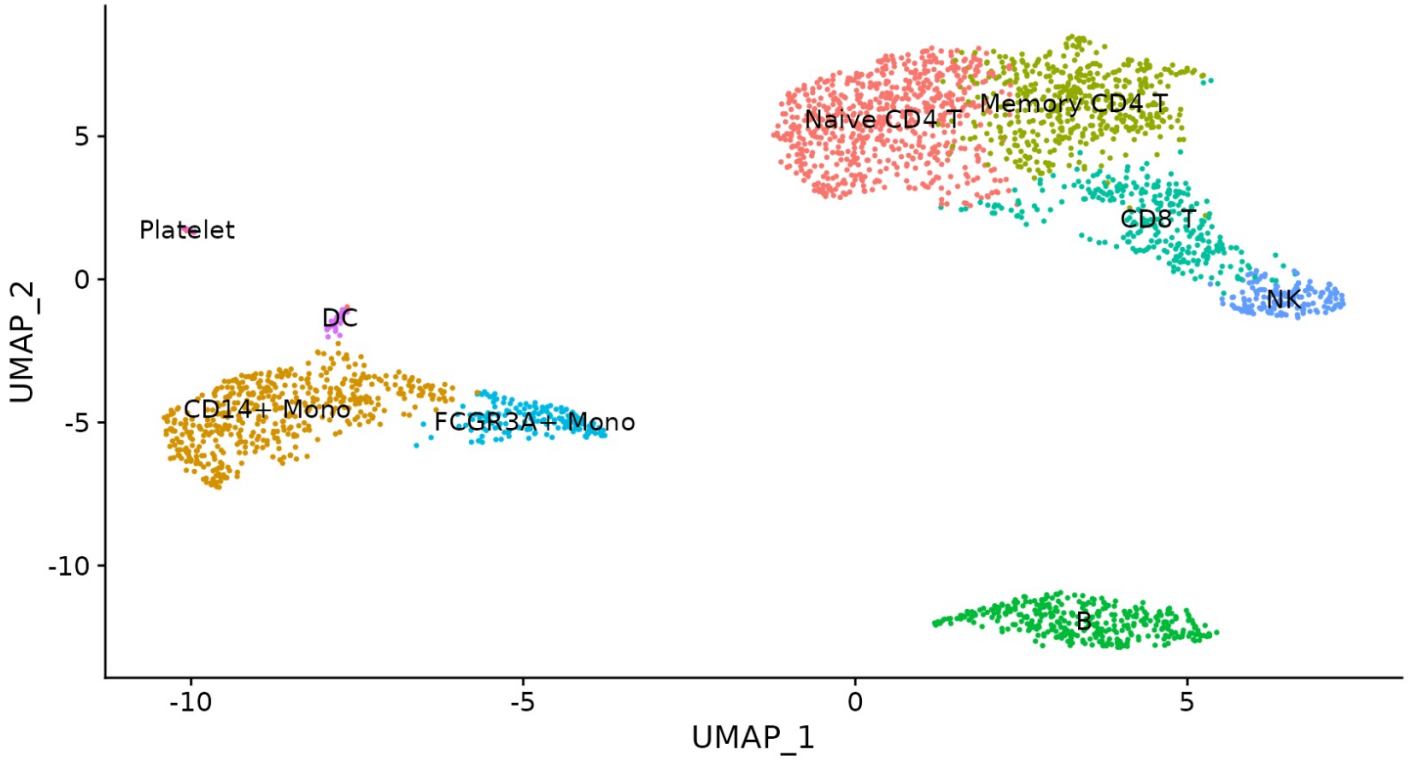


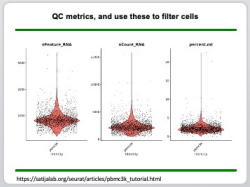
Assigning cell type identity to clusters

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

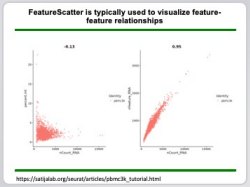
Assigning cell type identity to clusters

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

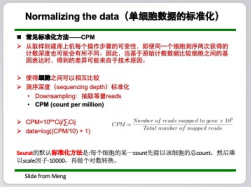




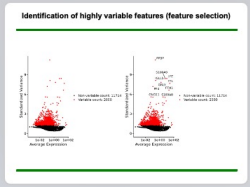
19



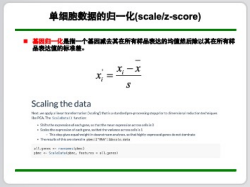
20



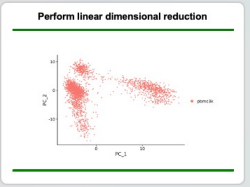
21



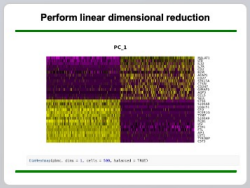
22



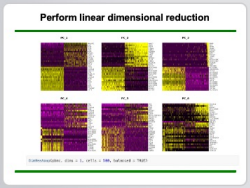
23



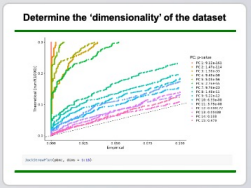
24



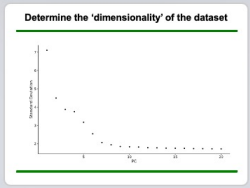
25



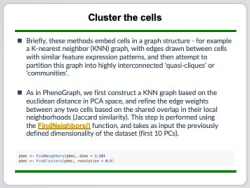
26



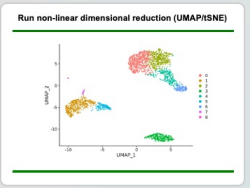
27



28

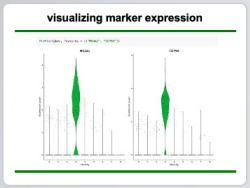


29

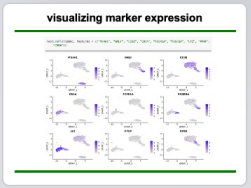


30

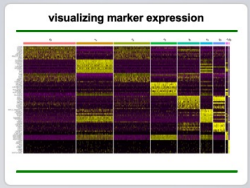
31



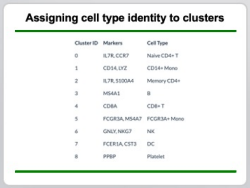
32



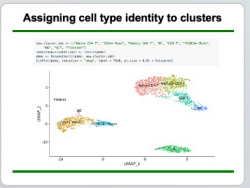
33



34



35



36

单细胞数据的降维

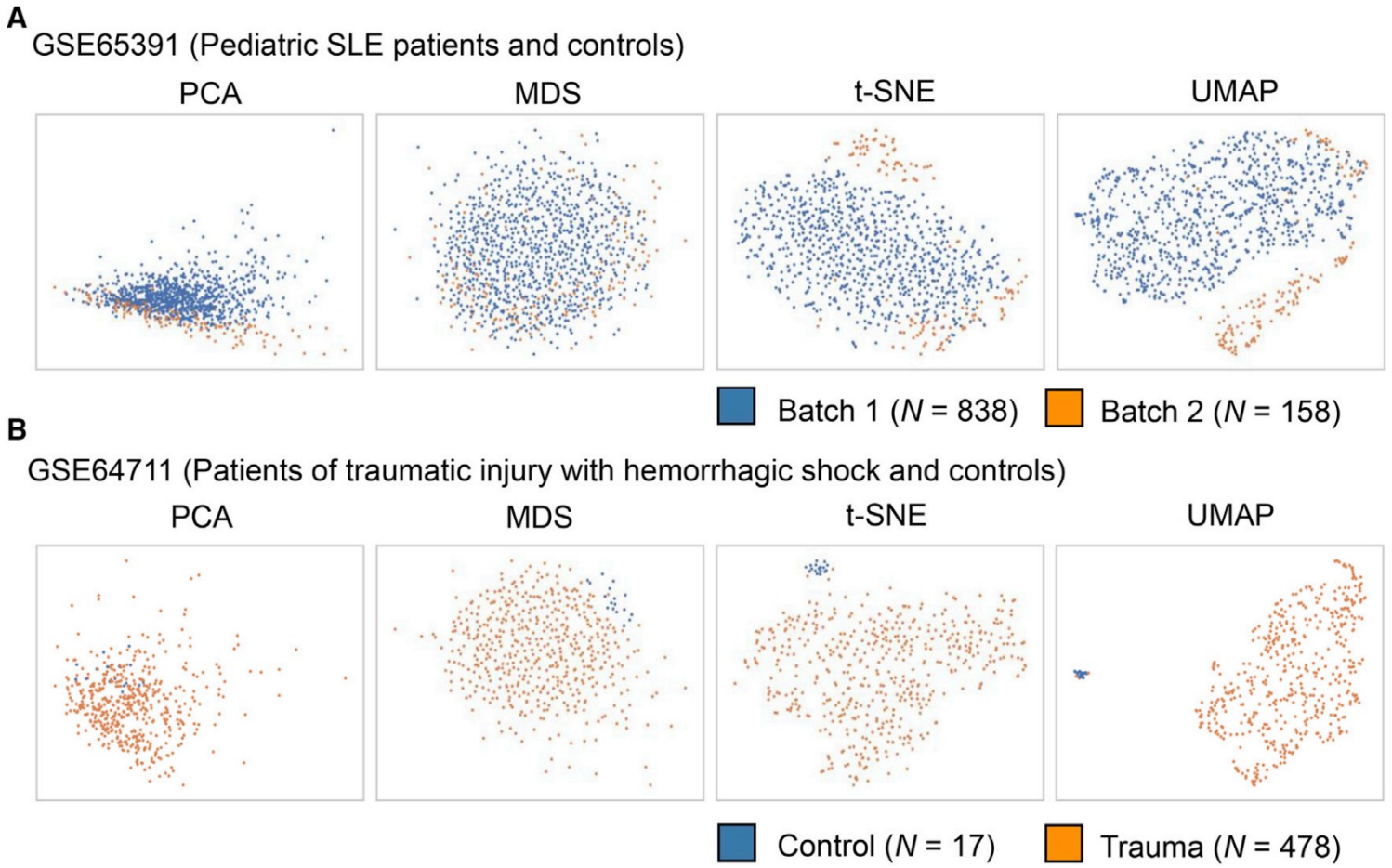


Figure 3. Biological explanation of clustering by batch effects and biological group

https://www.cell.com/cell-reports/fulltext/S2211-1247(21)00859-7

手动注释1：利用差异基因注释

- 差异基因标记细胞类型
  - • 转录因子:影响谱系lineage分化，遗传操作可控制该细胞类型
  - • 表面蛋白:抗体富集该种细胞类型
  - • 特异表达基因或基因集组合
- 如何找标志基因
  - 看文献收集:review文章;
  - 相关领域单细胞文章。
  - • 数据库
  - • 实验室传承
- 红细胞:
  - c('HBG1','HBG2','HBQ1','HBA1','HBA2','HBE1','HBD','HBM','HBZ','HBB')

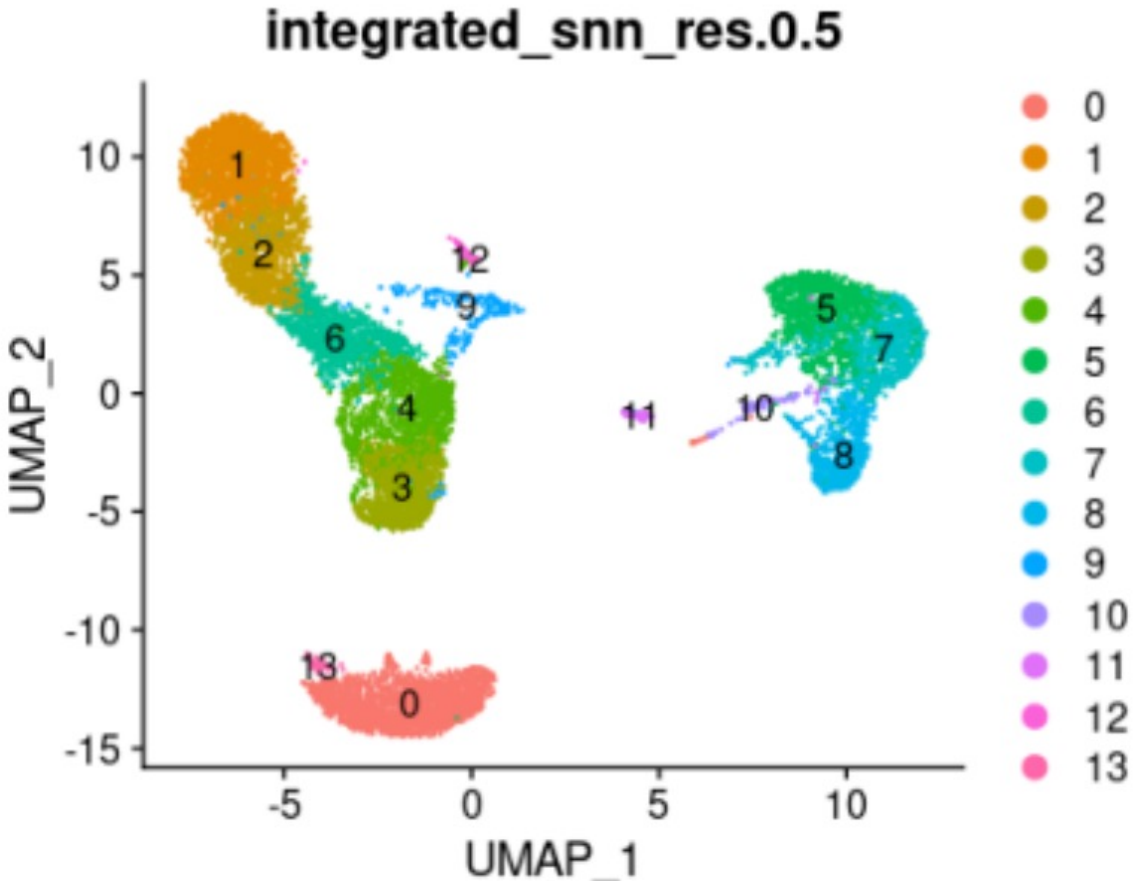
手动注释1：利用差异基因注释-数据库

- CellMarker http://xteam.xbio.top/CellMarker/

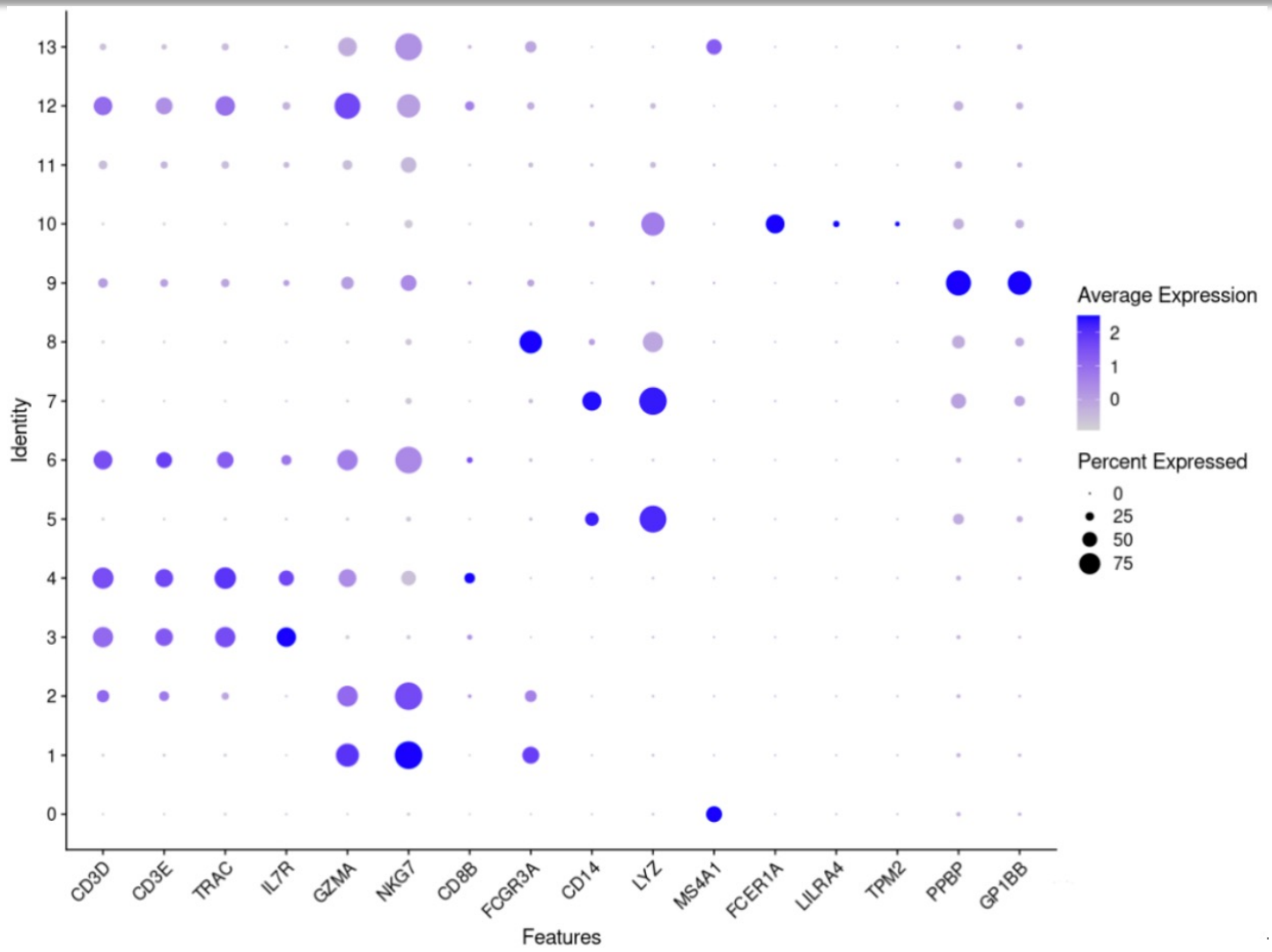
- Cell Taxonomy
  https://ngdc.cncb.ac.cn/celltaxonomy/

手动注释1：利用差异基因注释

- 比对网站中的基因
- 手动标记每个细胞簇对应的细胞类型

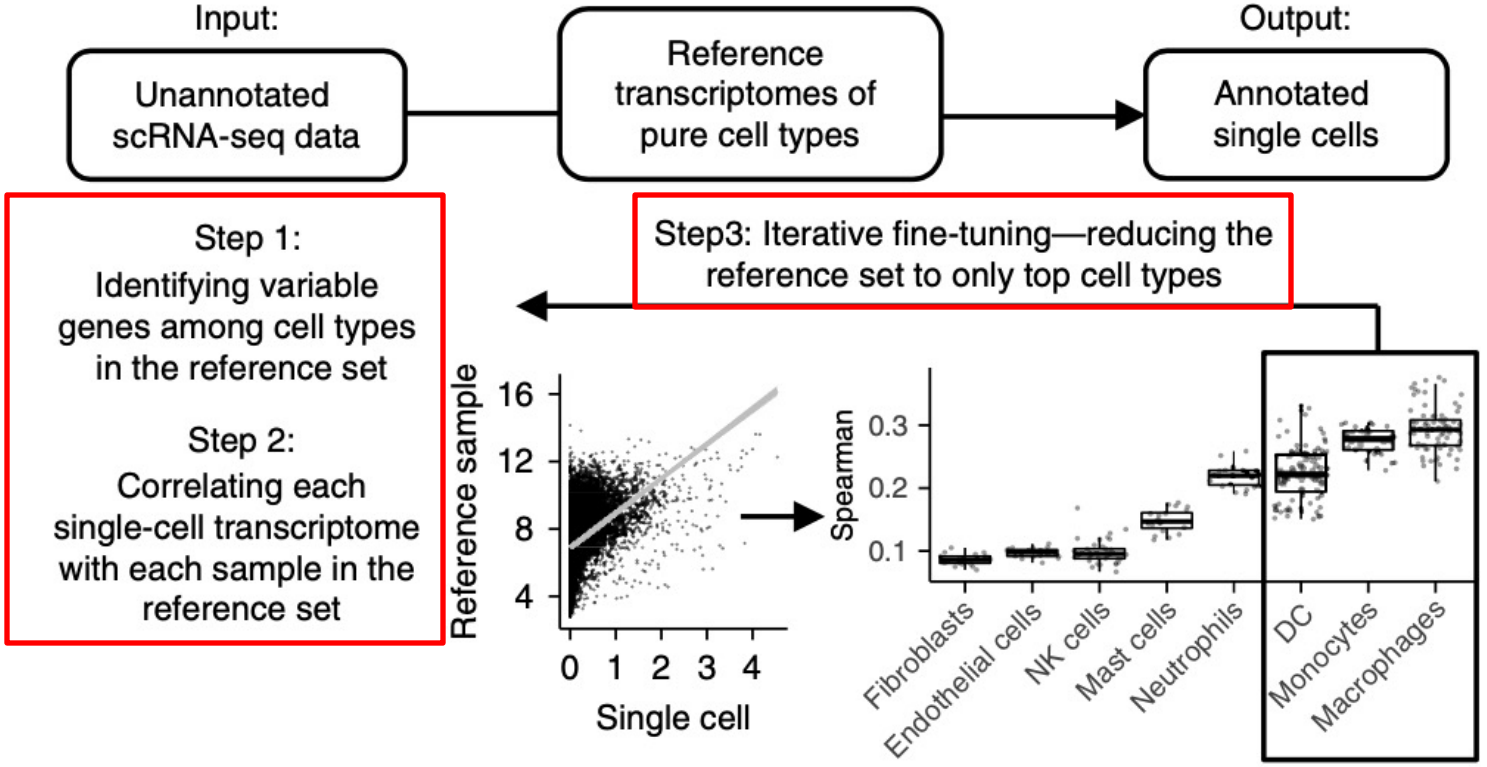


手动注释2：绘制气泡图进行注释



自动注释-SingleR

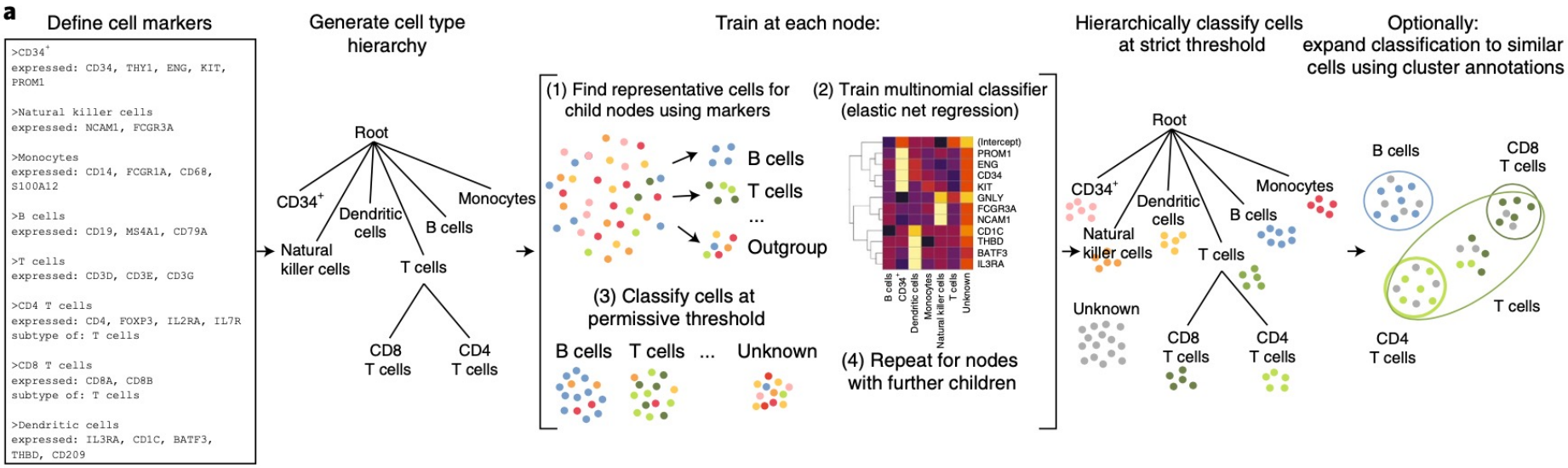
- 对于每个测试单元：
  1. 计算其表达谱与每个参考样品的表达谱之间的Spearman相关性。
  2. 将每个标签的分数定义为相关性分布的fixed quantile（默认为0.8)。
  3. 对所有标签重复此操作，然后将得分最高的标签作为此细胞的注释。



https://www.nature.com/articles/s41590-018-0276-y

自动注释-Garnett

基于细胞类型特异性基因的可解释的分层标记语言，用于在单细胞转录分析和单细胞染色质可及性数据集中快速注释细胞类型

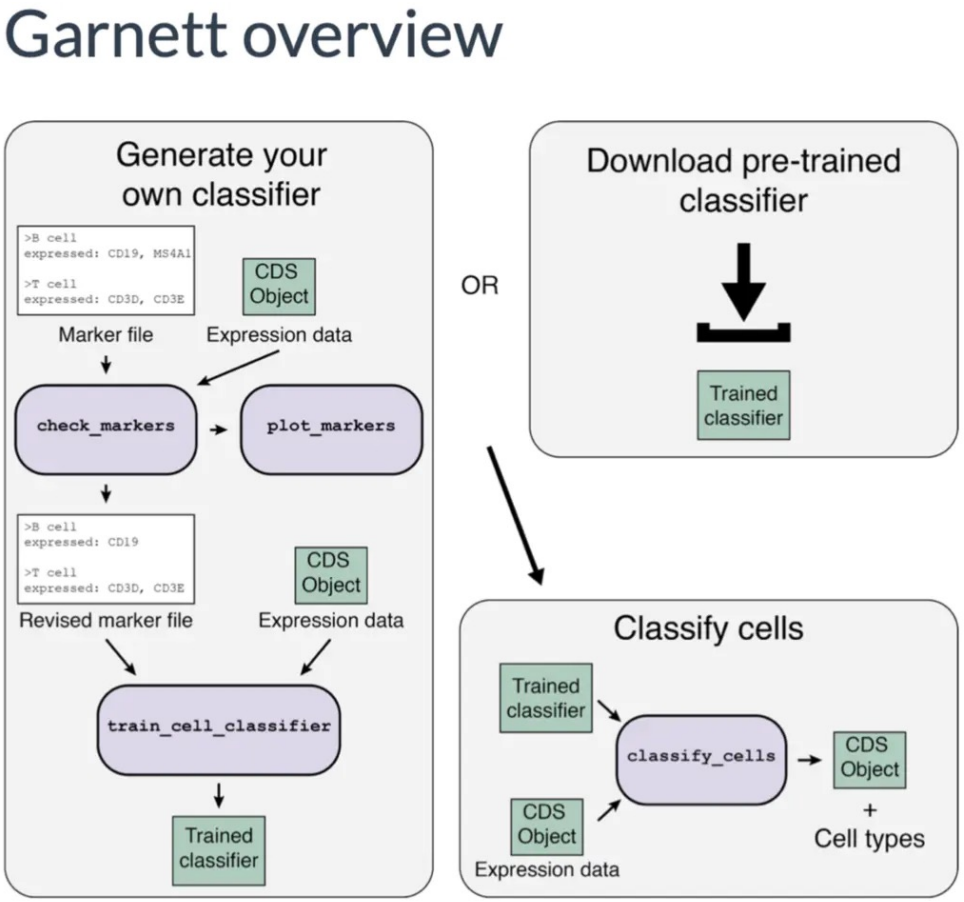


构建细胞类型标记文件 → 生成细胞类型树 → 训练分类器 → 完成细胞分类

https://www.nature.com/articles/s41592-019-0535-3

自动注释-Garnett

1. 制作符合garnett格式要求的定义细胞类型的marker基因文本文件(marker file)。
2. 使用单细胞数据创建monocle3的CDS数据对象(cds object)。
3. 将marker file和cds object输入garnett，对marker基因进行打分，然后根据评分结果优化marker file。
4. 将优化后的marker file和cds object输入garnett训练分类器。使用分类器对新的cds数据进行细胞分类。



https://cole-trapnell-lab.github.io/garnett/

拟时序分析

Published: 23 March 2014

The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells

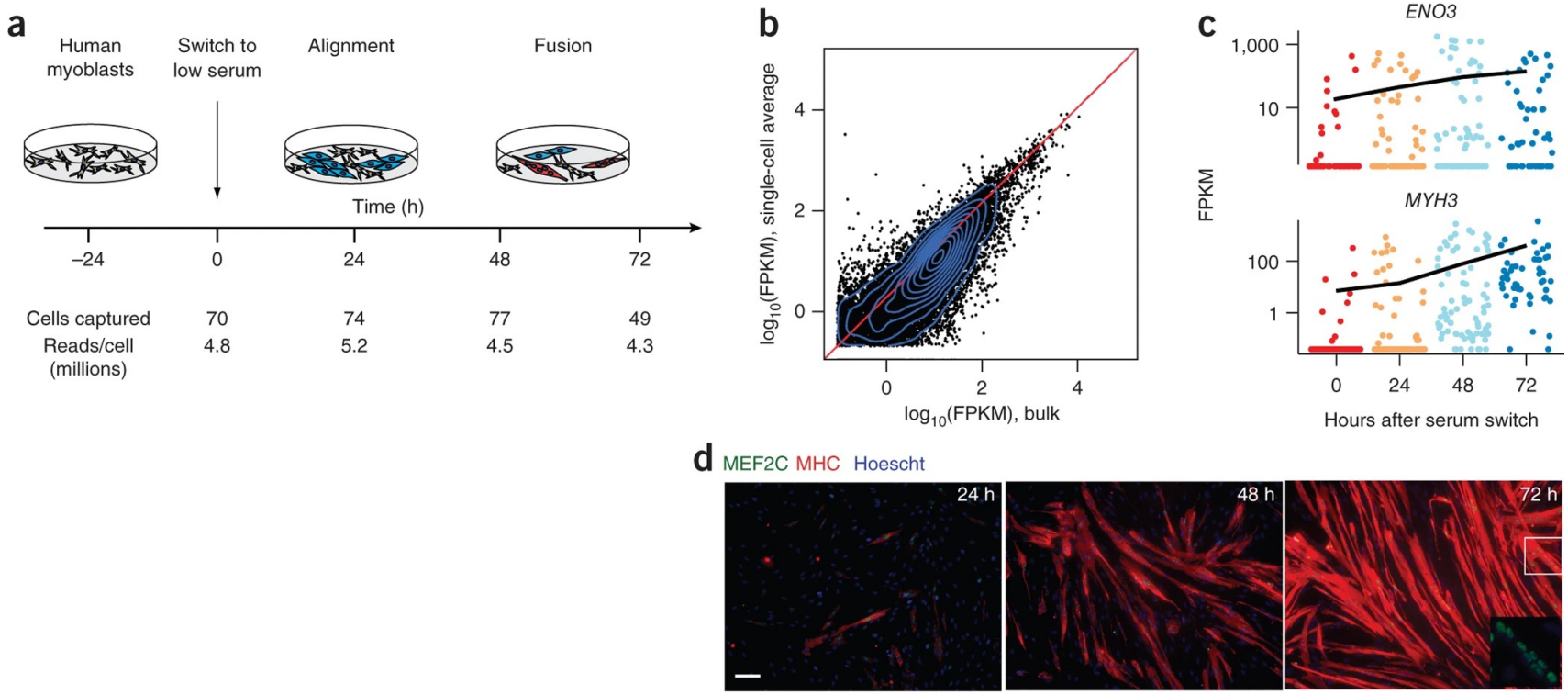
Cole Trapnell, Davide Cacchiarelli, Jonna Grimsby, Prapti Pokharel, Shuqiang Li, Michael Morse, Niall J Lennon, Kenneth J Livak, Tarjei S Mikkelsen & John L Rinn

Nature Biotechnology **32**, 381–386 (2014) | Cite this article

https://www.nature.com/articles/nbt.2859

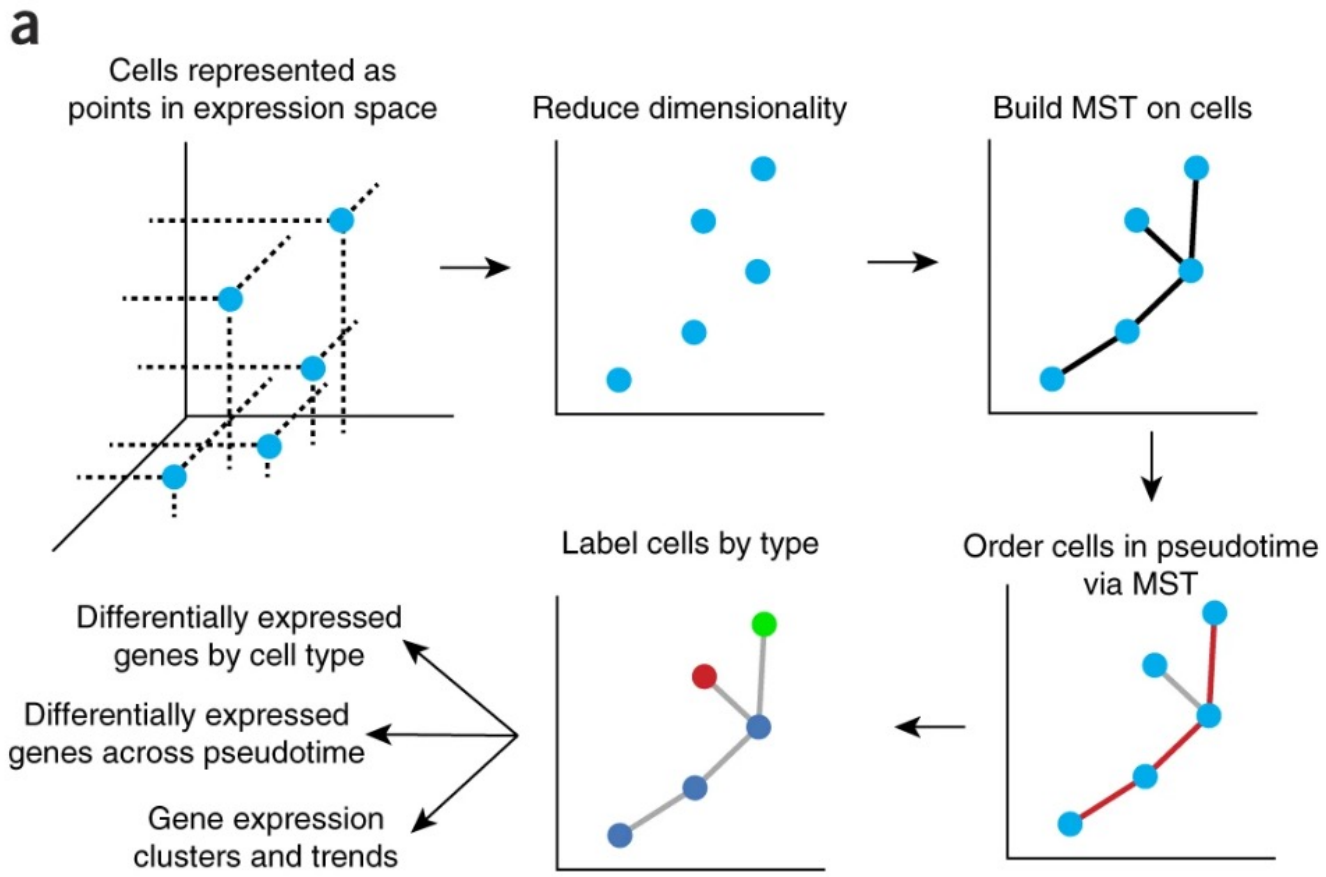
拟时序分析

single-cell RNA-Seq data of differentiating myoblasts



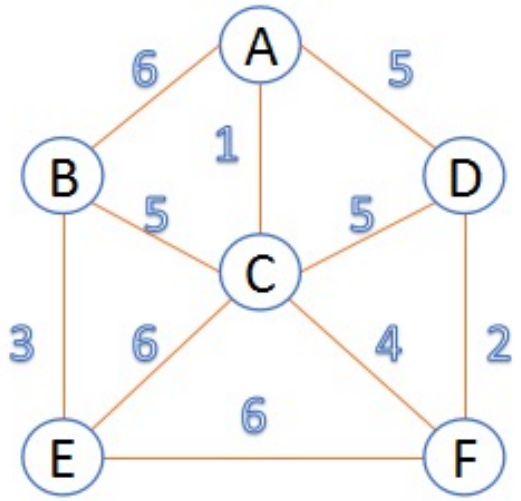
https://www.nature.com/articles/nbt.2859

拟时序分析

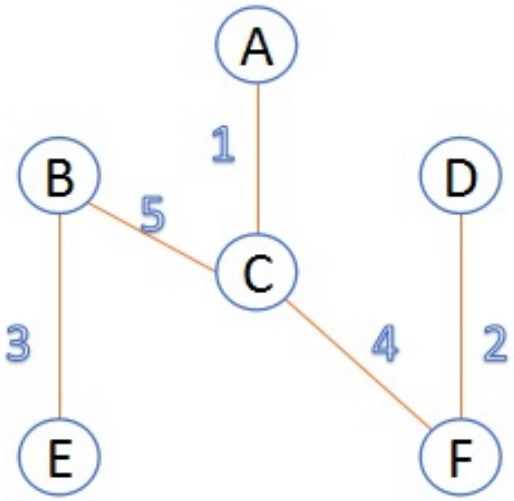


https://www.nature.com/articles/nbt.2859

最小生成树



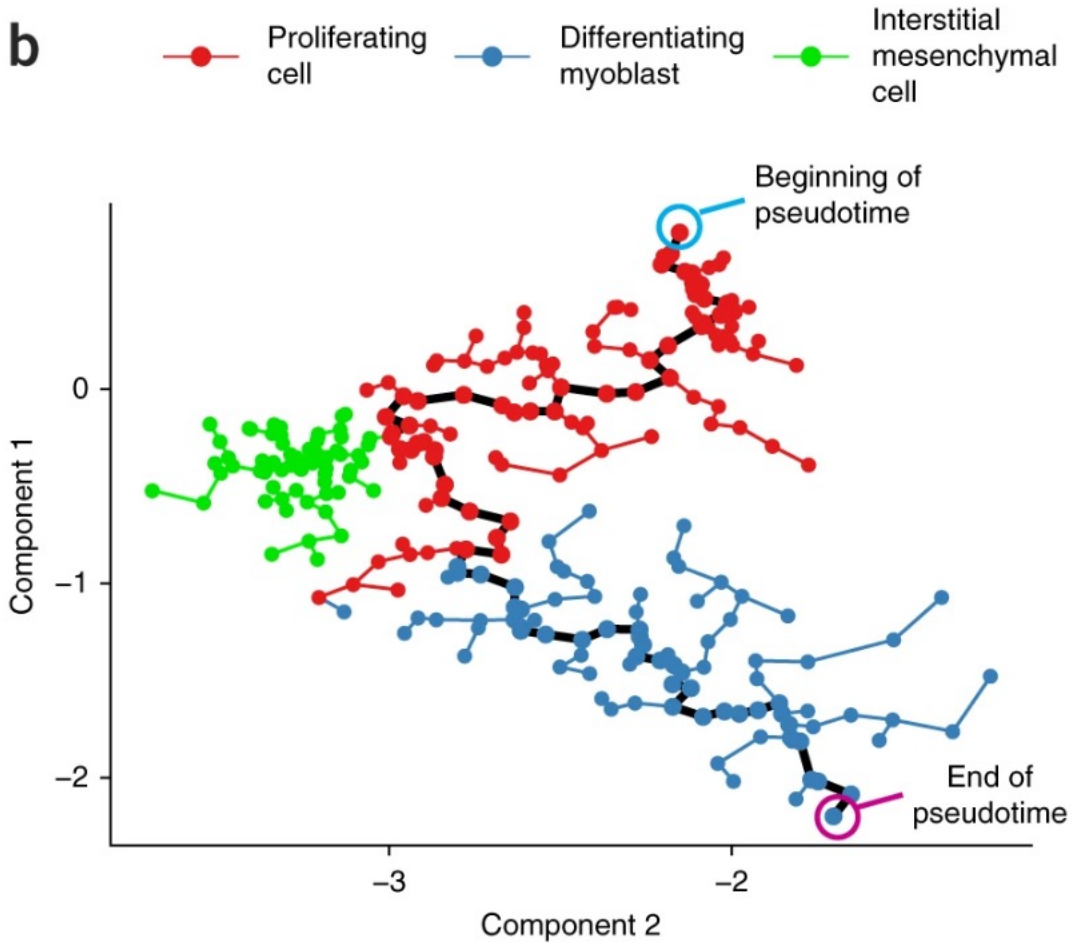
连通网G



最小生成树

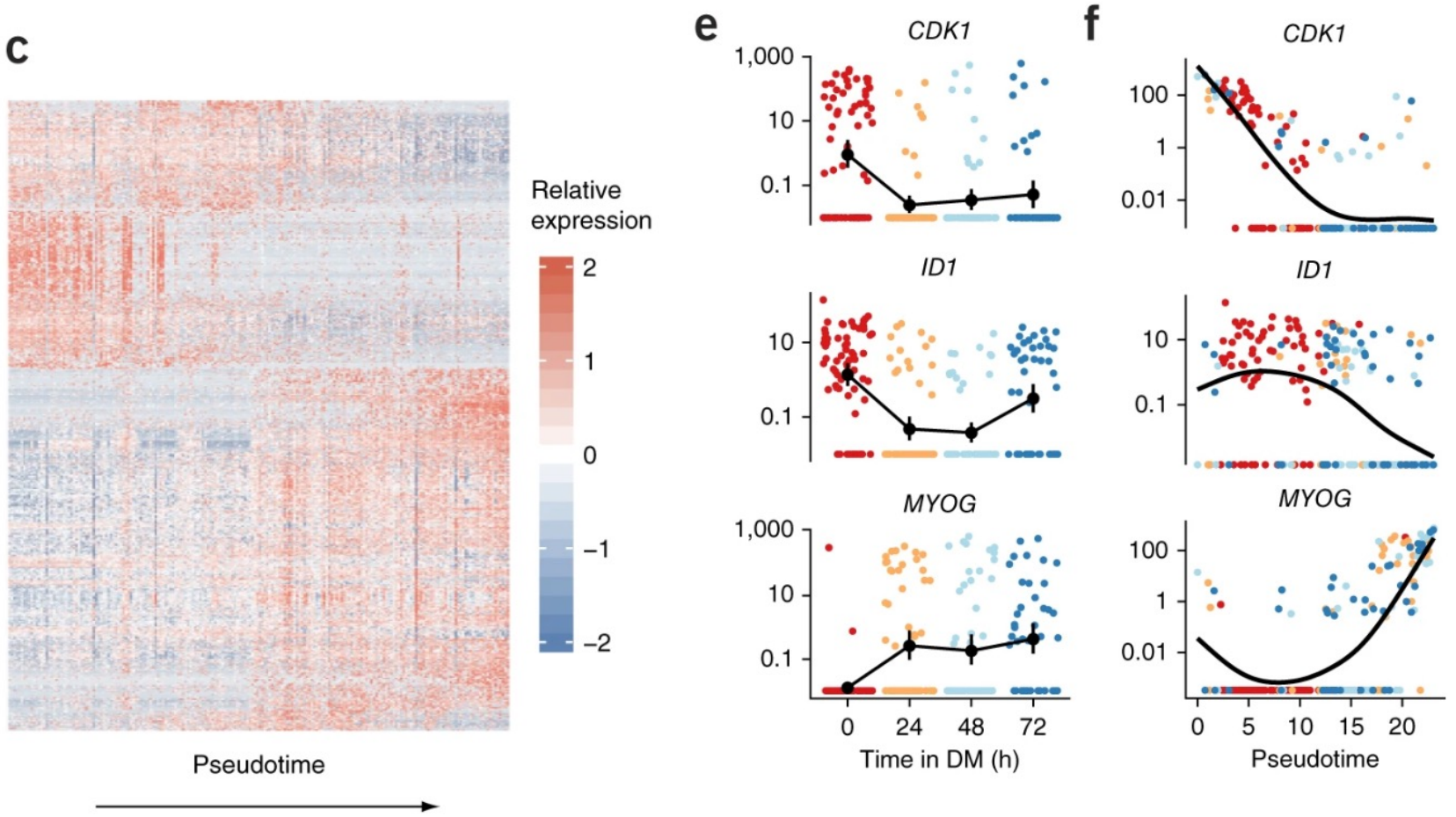
https://blog.csdn.net/luoshixian099/article/details/51908175

拟时序分析



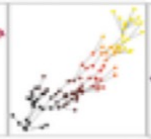
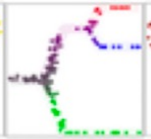
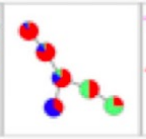
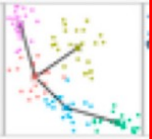
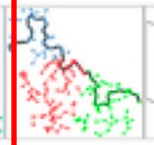
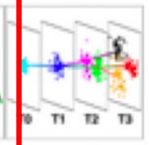
https://www.nature.com/articles/nbt.2859

拟时序分析



https://www.nature.com/articles/nbt.2859

拟时序分析系列方法

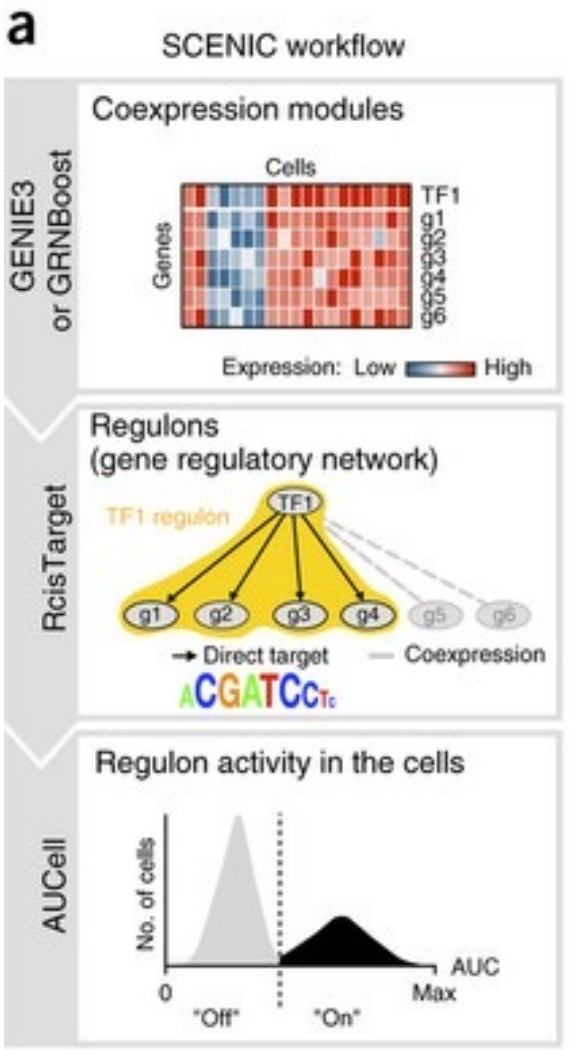
Method	SCUBA pseudotime	Wanderlust	Wishbone	SLICER	SCOUP	Waterfall	Mpath	TSCAN	Monocle	SCUBA
Visual abstract										
Structure	Linear	Linear	Single bifurcation	Branching	Branching	Linear	Branching	Linear	Branching	Branching
Robustness strategy	Principal curves	Ensemble, starting cell	Ensemble, starting cell	Starting cell	Starting population	Clustering of cells	Clustering of cells using external labelling	Clustering of cells	Differential expression	Simple model
Extra input requirements	None	Starting cell	Starting cell	Starting cell	Starting population	None	Time points	None	Time points	Time points
Unbiased	+	±	±	±	±	+	−	+	−	−
Scalability w.r.t. cells	−	−	±	±	−	±	+	+	−	±
Scalability w.r.t. genes	+	+	+	+	−	+	±	±	±	+
Code and documentation	−	±	+	±	+	±	+	+	+	±
Parameter ease-of-use	+	+	+	+	−	±	−	+	+	+

First Author	Marco	Bendall	Setty	Welch	Matsumoto	Shin	Chen	Ji	Trapnell	Marco
Last Author	GC Yuan	Dana Pe'er	Dana Pe'er	Hartemink, Prins	Kiryu	Hongjun Song	Poidinger	Ji	Rinn	GC Yuan
Journal	PNAS	Cell	Nature Biotechnology	Genome Biology	BMC Bioinformatics	Cell Stem Cell	Nature Communications	NAR	Nature Biotechnology	PNAS
Year	2014	2014	2016	2016	2016	2015	2016	2016	2014	2014

Slide from MIT

SCENIC的分析流程

- ■ 共表达模块的构建
  - ➢ 与TF共表达的基因归为一个共表达模块
  - ➢ 使用GENIE3或GRNBoost算法
- ■ 转录因子调控网络的构建
  - ➢ 根据TF的motif筛选共表达模块中的基因，形成最终的网络
  - ➢ 网络中每个TF及其靶基因构成一个regulon
  - ➢ 使用Rcistarget包
- ■ 转录因子活性评分
  - ➢ 使用AUCell对每个regulon进行评分



Regulatory information from single-cell data

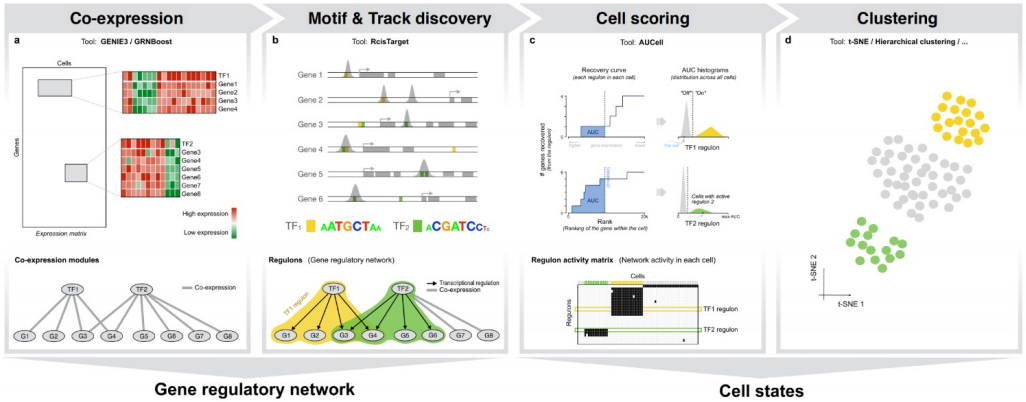
SCENIC Suite: Unveil regulatory information from single-cell data

SCENIC Suite is a set of tools to study and decipher gene regulation. Its core is based on SCENIC (Single-Cell rEgulatory Network Inference and Clustering) which enables you to infer transcription factors, gene regulatory networks and cell types from single-cell RNA-seq data (using SCENIC) or the combination of single-cell RNA-seq and single-cell ATAC-seq data (using SCENIC+).

Read the original SCENIC paper (Aibar et al., 2017)

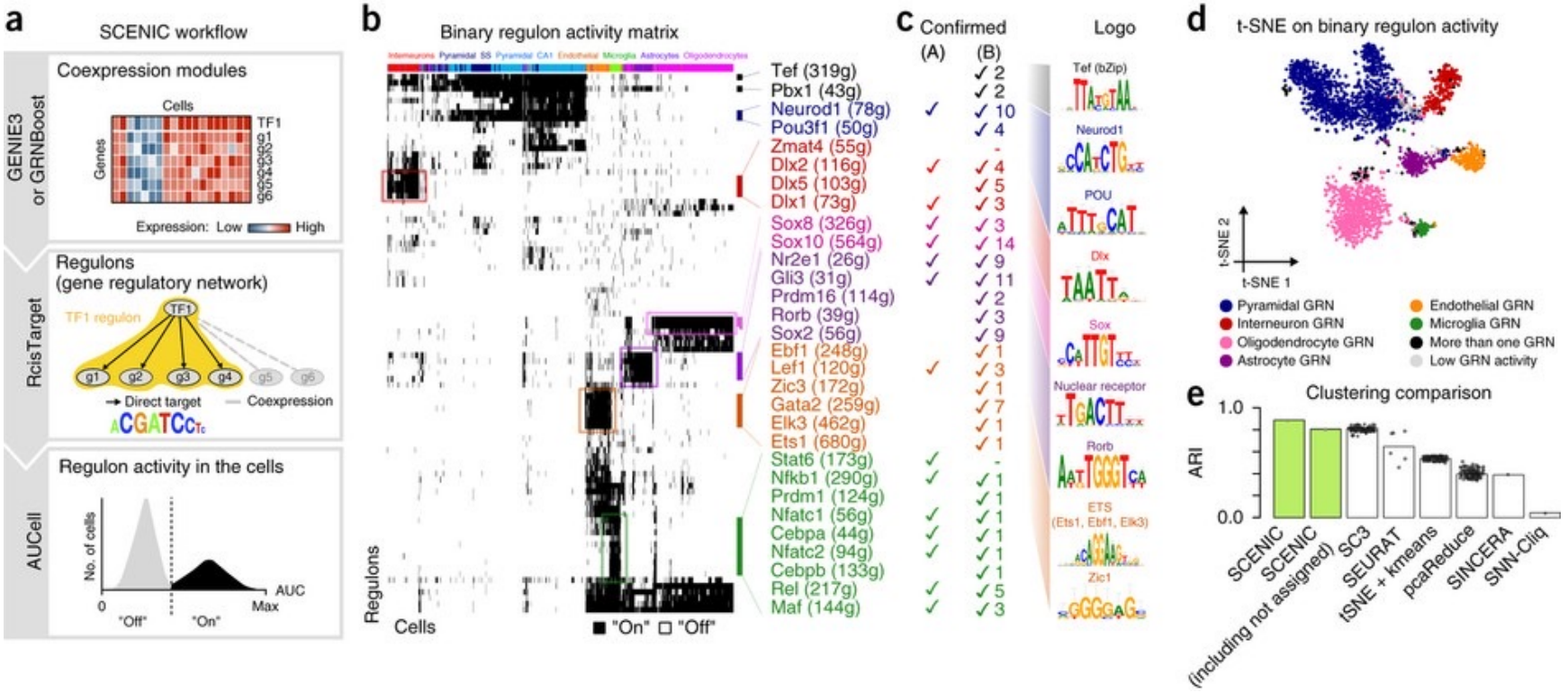
Read the SCENIC protocols paper (Van de Sande et al., 2020)

Read the SCENIC+ preprint (Bravo and De Winter et al., 2022)



https://scenic.aertslab.org/

The SCENIC workflow and its application to the mouse brain



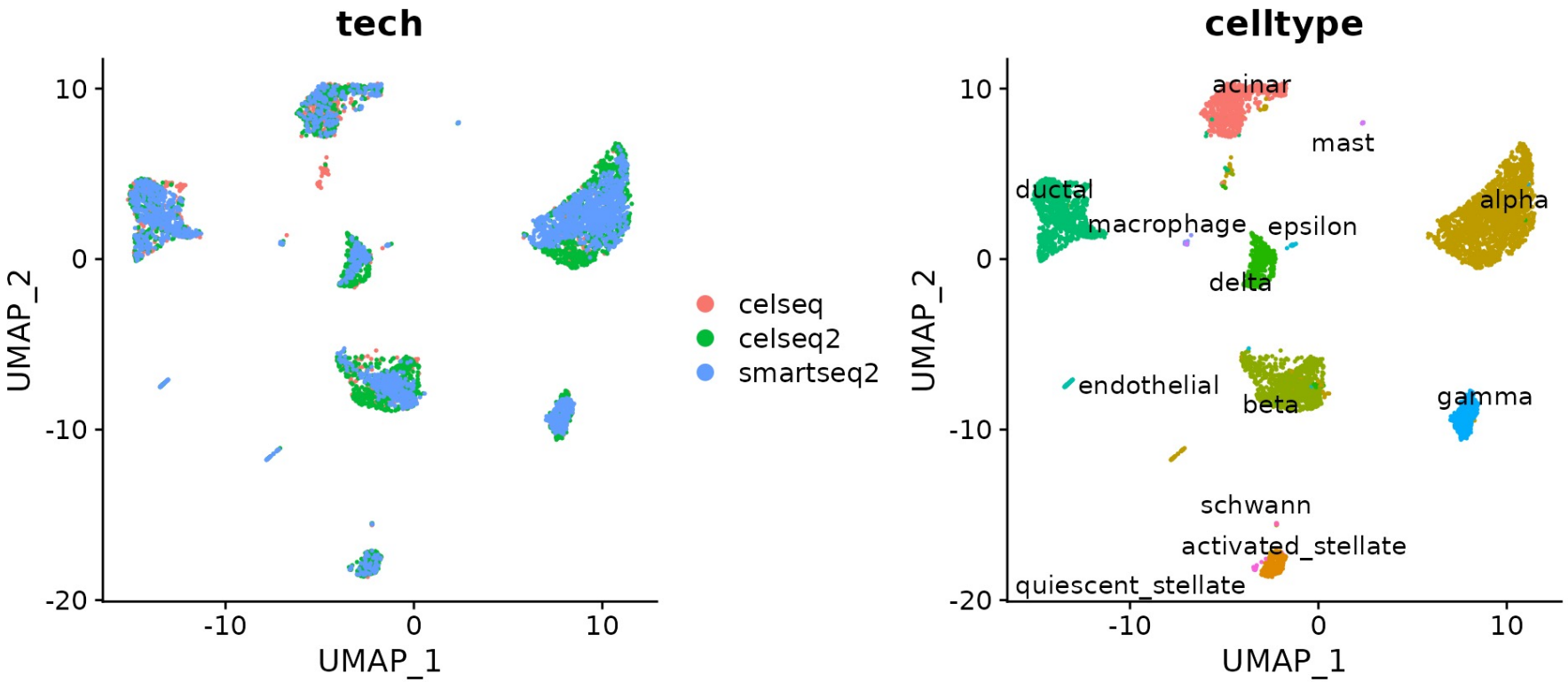
https://www.nature.com/articles/nmeth.4463

多样本数据整合-批次效应

- 单细胞测序产生批次效应原因有：
  - （1）不同建库策略，如10X平台，Drop-seq，SMART2-seq等
  - （2）不同测序平台
  - （3）不同公司的试剂或不同批次的试剂
  - （4）不同测序批次
  - （5）不同实验操作者……

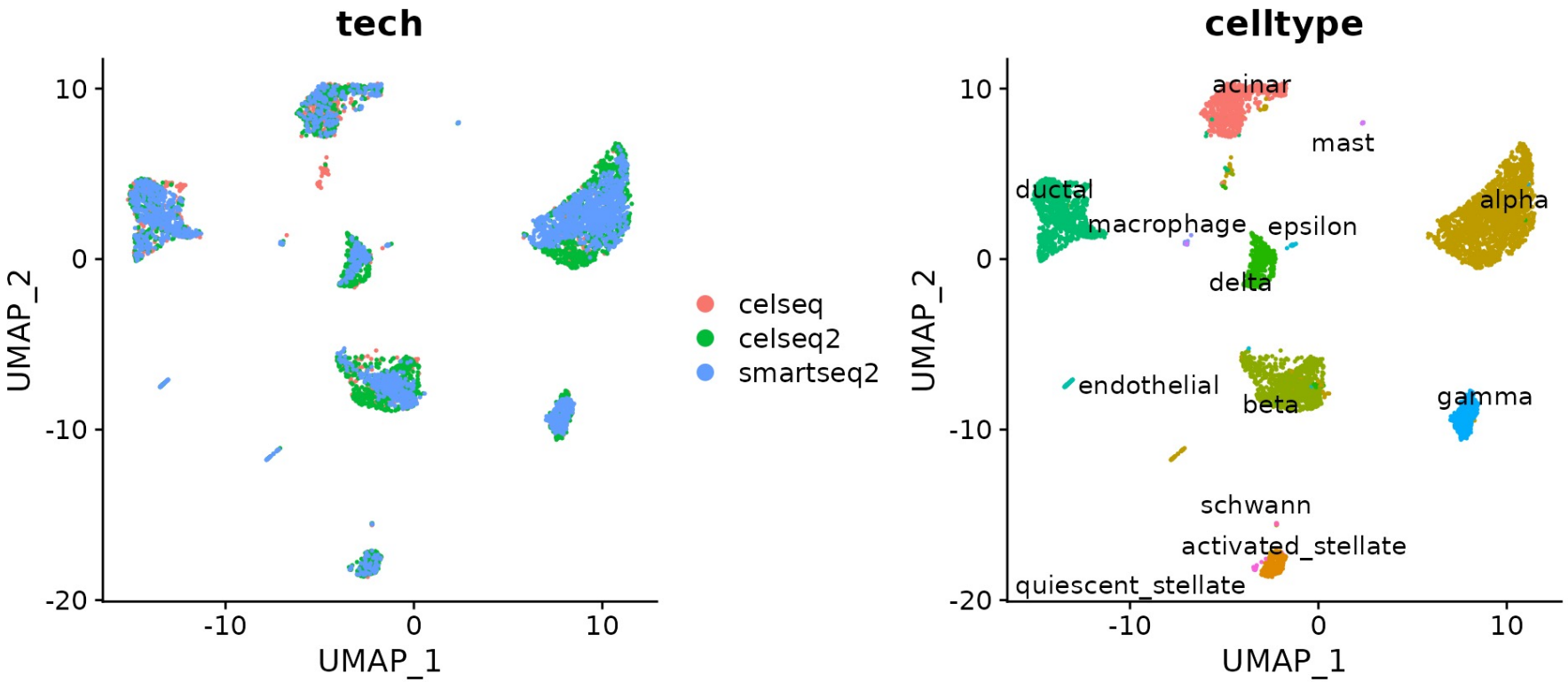
- 单细胞数据中最显著的技术变量是计数深度和批次

Integration of 3 pancreatic islet cell datasets



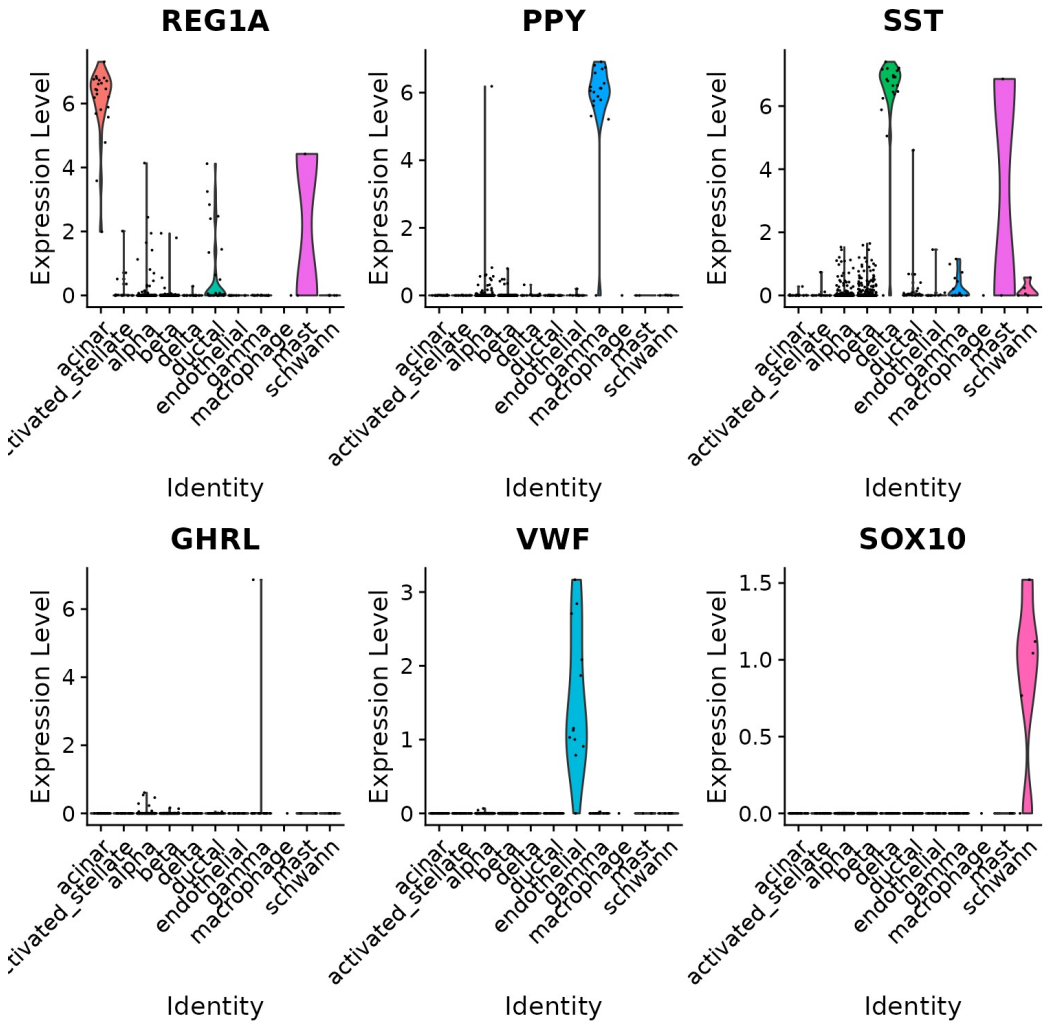
https://satijalab.org/seurat/articles/integration_mapping.html

Integration of 3 pancreatic islet cell datasets



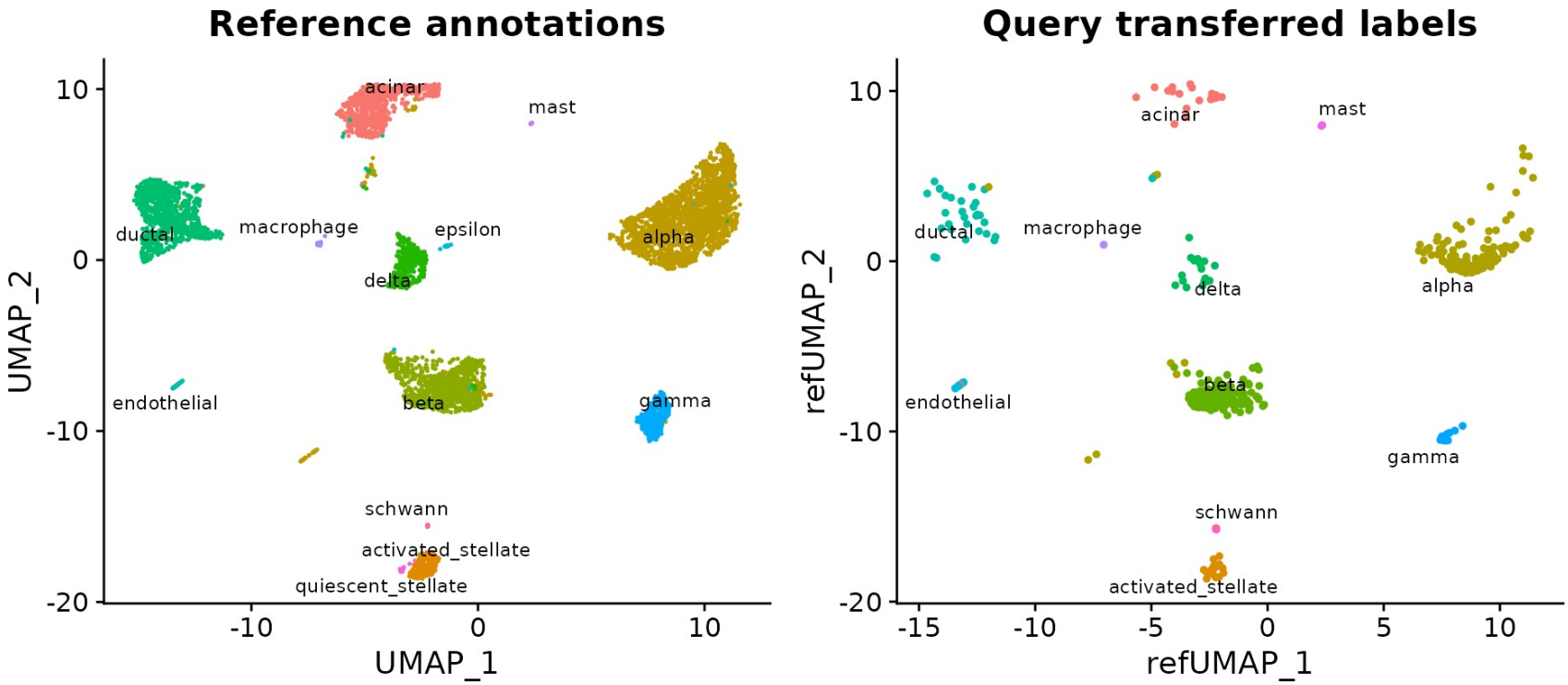
https://satijalab.org/seurat/articles/integration_mapping.html

Cell type classification using an integrated reference



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

What is MapQuery doing?



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

多样本数据整合-常用算法（Harmony）

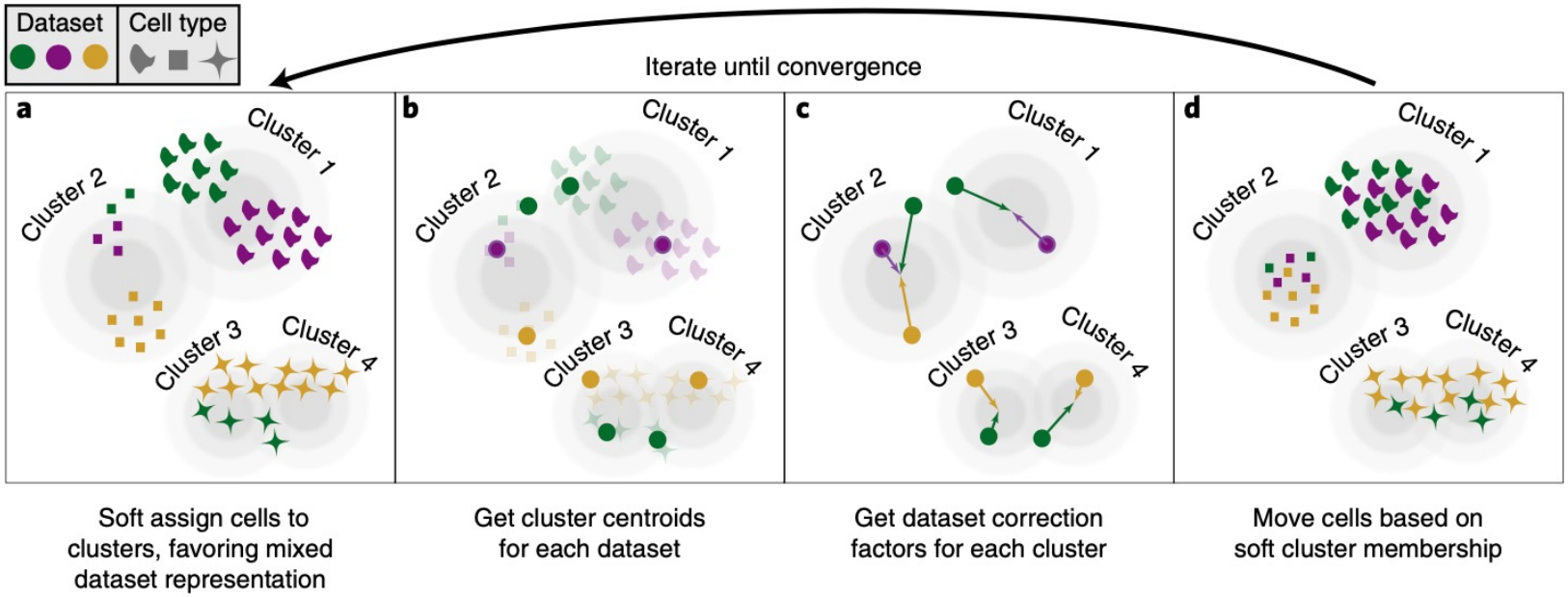
- Harmony方法在2019年提出
- 优势：
  - 整合数据的同时对稀有细胞的敏感性依然很好
  - 省内存
  - 适合于更复杂的单细胞分析实验设计，可以比较来自不同供体、组织和技术平台的细胞

Fast, sensitive and accurate integration of single-cell data with Harmony

Ilya Korsunsky[1,2,3,4], Nghia Millard[1,2,3,4], Jean Fan[5], Kamil Slowikowski[1,2,3,4], Fan Zhang[1,2,3,4], Kevin Wei[2], Yuriy Baglaenko[1,2,3,4], Michael Brenner[2], Po-ru Loh[1,3,4] and Soumya Raychaudhuri[1,2,3,4,6*]

https://www.nature.com/articles/s41592-019-0619-0

多样本数据整合-常用算法（Harmony）



•A) 使用模糊聚类将每个单元分配到多个聚类，而惩罚项确保每个聚类内数据集的多样性最大化。

•B) 计算每个集群的全局质心，以及每个集群的数据集特定质心。

•C) 在每个集群内，Harmony根据质心计算每个数据集的校正因子。

•D) 使用特定于细胞的因子校正每个细胞：由在步骤 A 中进行的软集群分配加权的数据集校正因子的线性组合。

https://www.nature.com/articles/s41592-019-0619-0

Article | Published: 20 February 2019

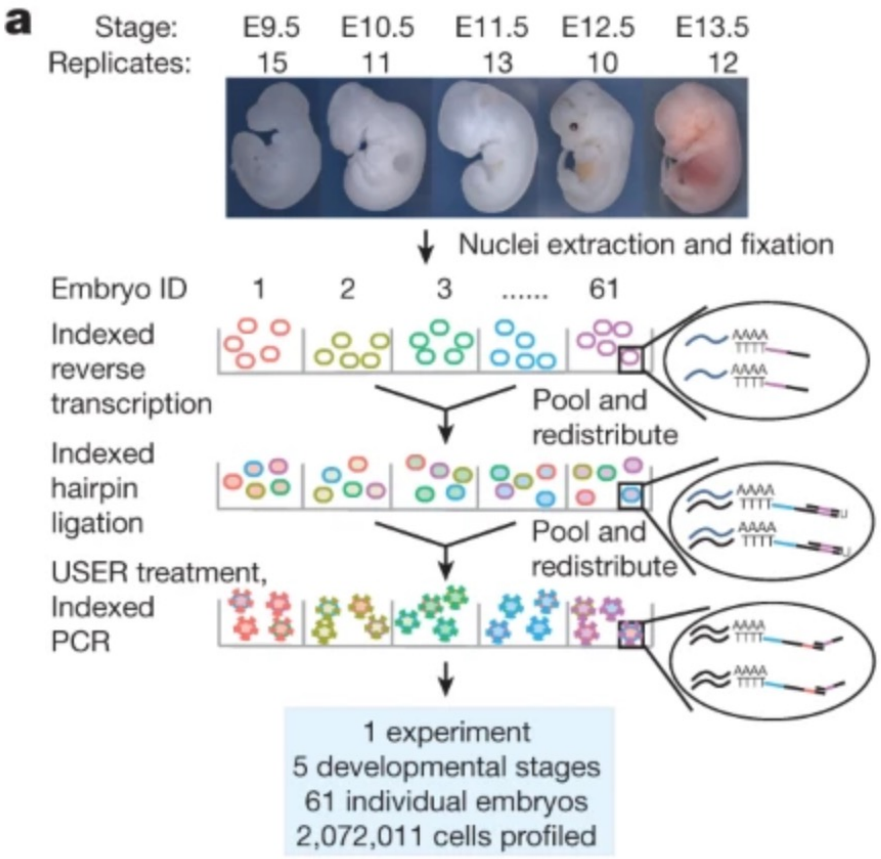
The single-cell transcriptional landscape of mammalian organogenesis

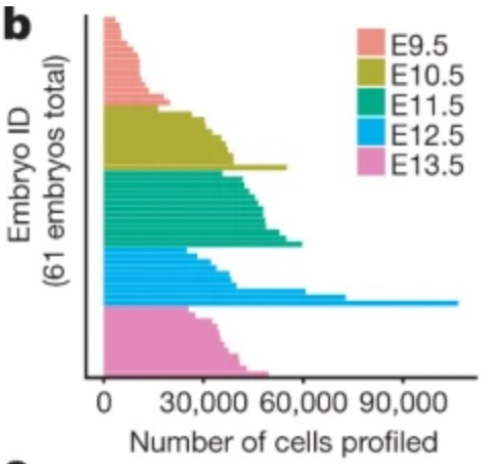
Junyue Cao, Malte Spielmann, Xiaojie Qiu, Xingfan Huang, Daniel M. Ibrahim, Andrew J. Hill, Fan Zhang, Stefan Mundlos, Lena Christiansen, Frank J. Steemers, Cole Trapnell ✉ & Jay Shendure ✉

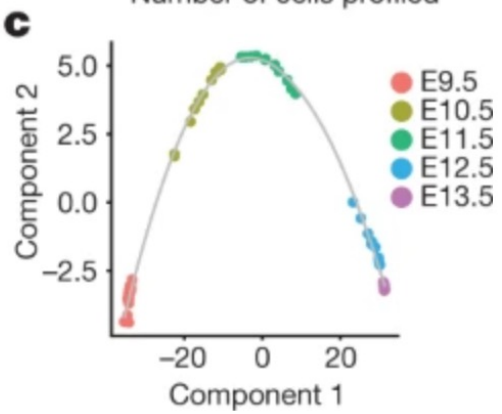
Nature **566**, 496–502 (2019) | Cite this article

https://www.nature.com/articles/s41586-019-0969-x

Fig. 1: sci-RNA-seq3 enables profiling of 2,072,011 cells from 61 mouse embryos across 5 developmental stages in a single experiment.

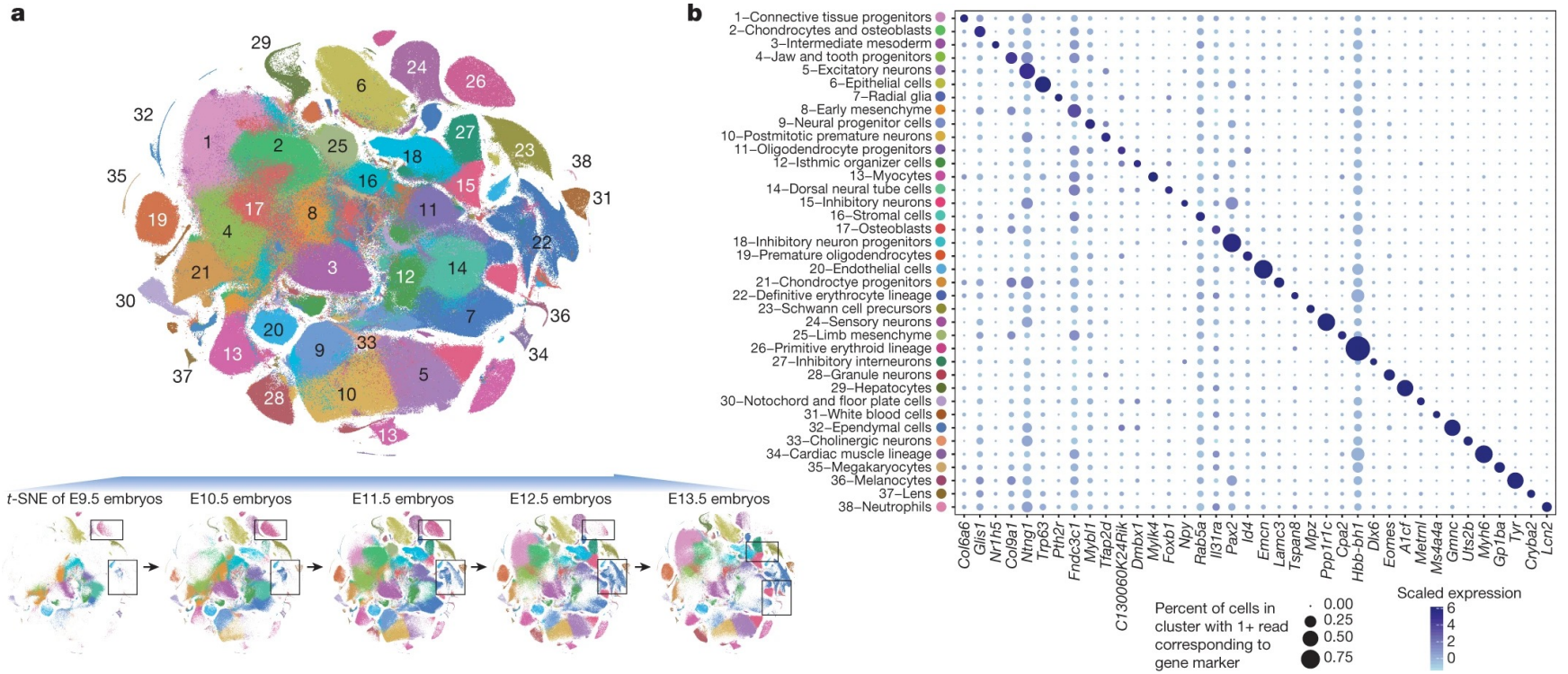






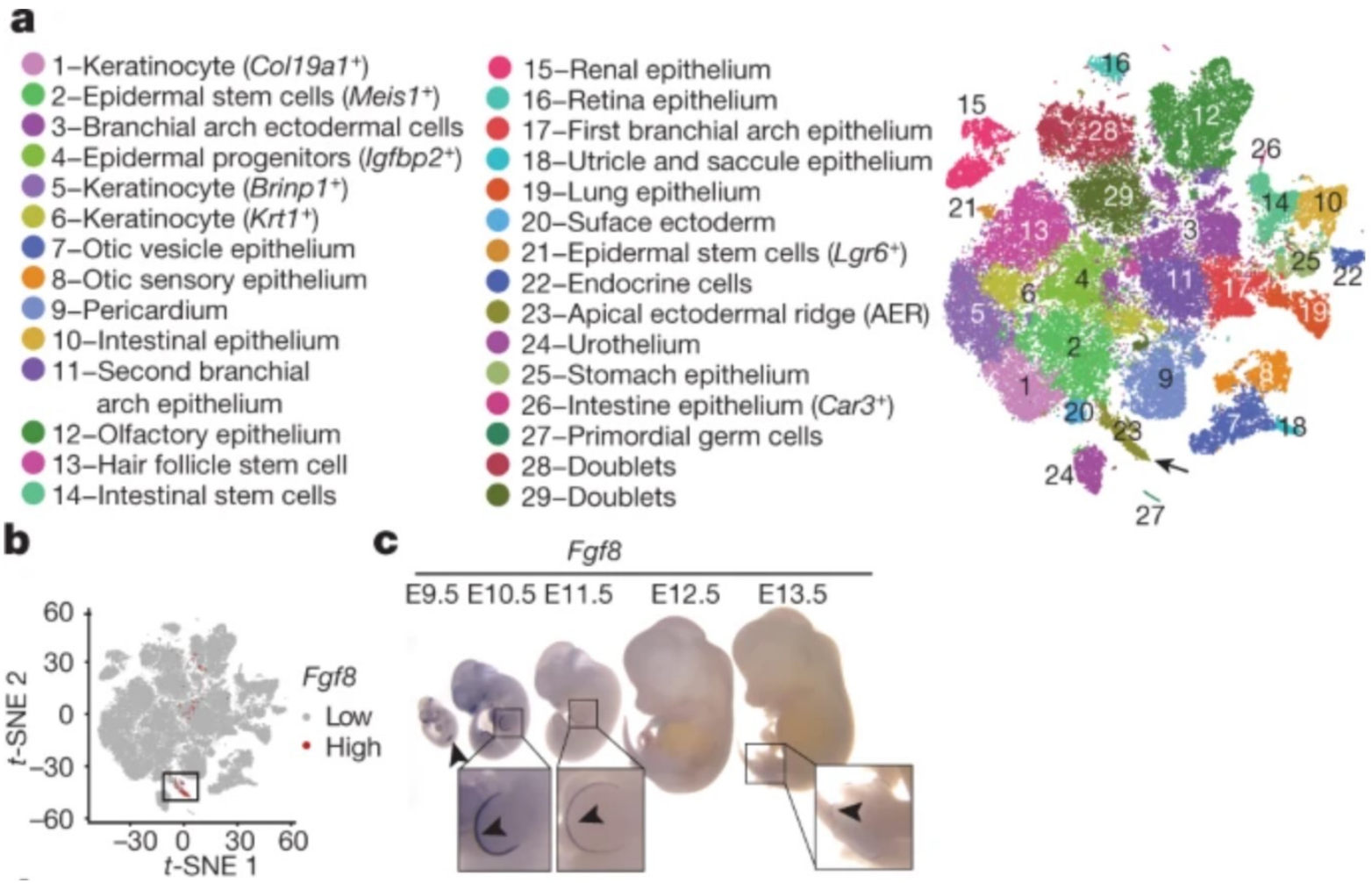
https://www.nature.com/articles/s41586-019-0969-x

Identifying the major cell types of mouse organogenesis



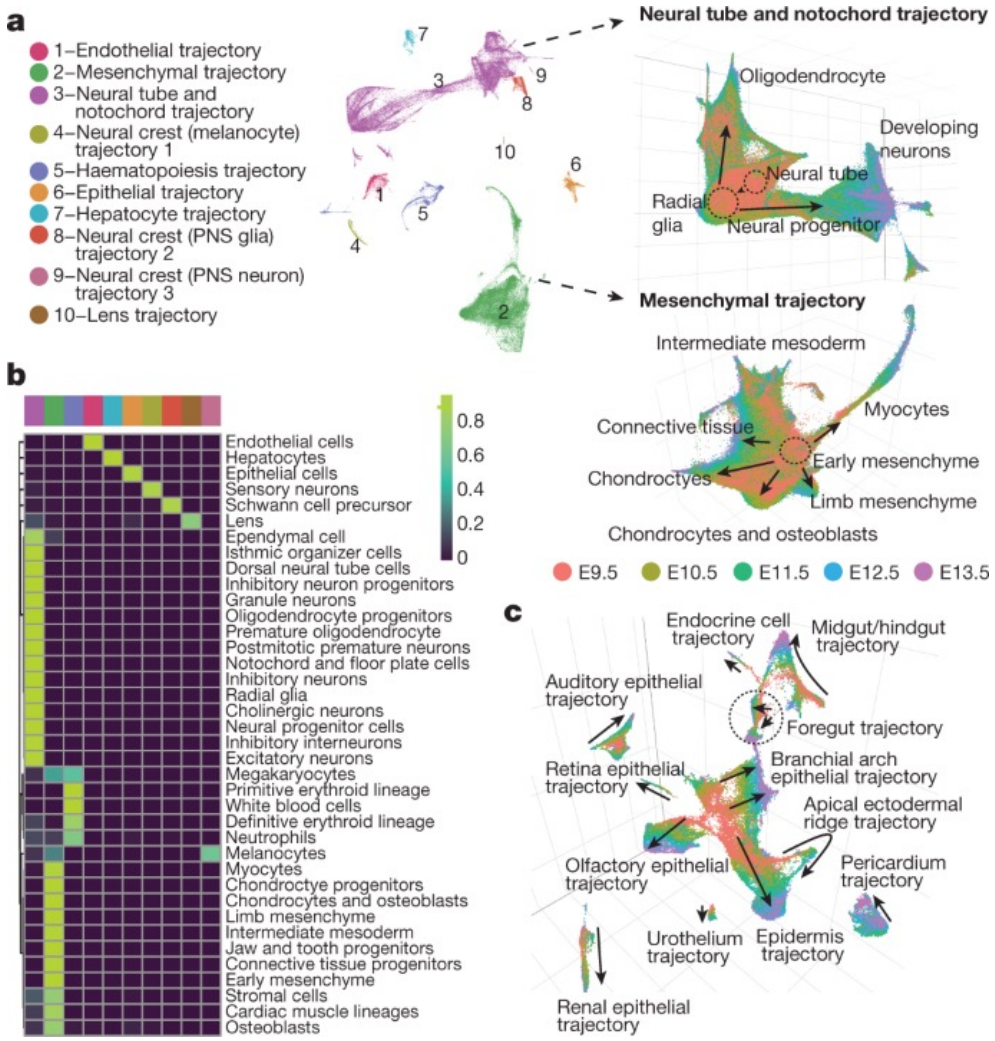
https://www.nature.com/articles/s41586-019-0969-x

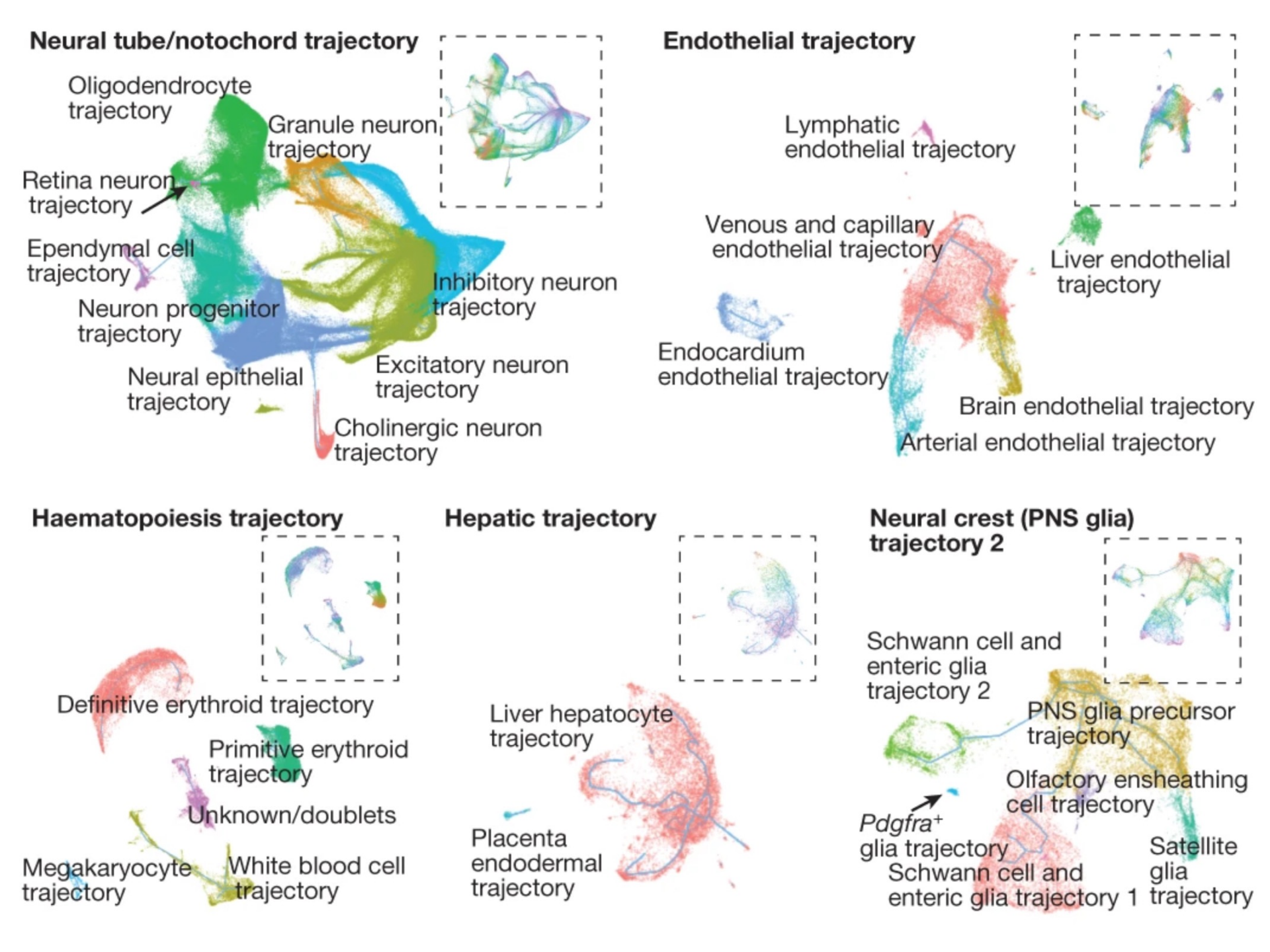
Identification and characterization of epithelial cell subtypes



https://www.nature.com/articles/s41586-019-0969-x

Characterization of ten major developmental trajectories present during mouse organogenesis




Neural tube/notochord trajectory
Oligodendrocyte trajectory
Granule neuron trajectory
Retina neuron trajectory
Ependymal cell trajectory
Neuron progenitor trajectory
Inhibitory neuron trajectory
Neural epithelial trajectory
Excitatory neuron trajectory
Cholinergic neuron trajectory
Endothelial trajectory
Lymphatic endothelial trajectory
Venous and capillary endothelial trajectory
Liver endothelial trajectory
Endocardium endothelial trajectory
Brain endothelial trajectory
Arterial endothelial trajectory
Haematopoiesis trajectory
Definitive erythroid trajectory
Primitive erythroid trajectory
Unknown/doublets
Megakaryocyte trajectory
White blood cell trajectory
Hepatic trajectory
Liver hepatocyte trajectory
Placenta endodermal trajectory
Neural crest (PNS glia) trajectory 2
Schwann cell and enteric glia trajectory 2
PNS glia precursor trajectory
Olfactory ensheathing cell trajectory
Pdgfra+ glia trajectory
Satellite glia trajectory
Schwann cell and enteric glia trajectory 1

Thanks for attention!