



# 单细胞转录组学

## scRNA-seq 数据分析

伊现富 (Yi Xianfu)

2024 年 7 月 8 日

基础医学院  
生物信息学系

研究生暑期培训

# 章节概览

## ① 导言

- 常规转录组
- 细胞异质性
- 单细胞测序

## ② 单细胞转录组学

- 技术简介
- 实验原理
- 分析流程

## ■ 分析角度

## ③ 空间转录组学

- 技术简介
- 实验原理
- 分析策略

## ④ 单细胞多组学

- 其他单细胞组学
- 单细胞多组学

## ① 导言

- 常规转录组
- 细胞异质性

## ■ 单细胞测序

- ② 单细胞转录组学
- ③ 空间转录组学
- ④ 单细胞多组学

## ① 导言

- 常规转录组
- 细胞异质性

■ 单细胞测序

- ② 单细胞转录组学
- ③ 空间转录组学
- ④ 单细胞多组学

## 转录组

特定组织或细胞在某一发育阶段或功能状态下转录出来的所有 RNA 的集合，包括编码蛋白质的信使 RNA (mRNA) 和非编码 RNA (rRNA、tRNA 和其他 ncRNAs)。

转录组是连接基因组和蛋白质组遗传信息和生物功能的纽带。

## 转录组

特定组织或细胞在某一发育阶段或功能状态下转录出来的所有 RNA 的集合，包括编码蛋白质的信使 RNA (mRNA) 和非编码 RNA (rRNA、tRNA 和其他 ncRNAs)。

转录组是连接基因组和蛋白质组遗传信息和生物功能的纽带。

## 转录组学

从整体转录水平系统研究基因转录图谱并揭示复杂生物学通路和性状调控网络分子机制的学科。

## 转录组

特定组织或细胞在某一发育阶段或功能状态下转录出来的所有 RNA 的集合，包括编码蛋白质的信使 RNA (mRNA) 和非编码 RNA (rRNA、tRNA 和其他 ncRNAs)。

转录组是连接基因组和蛋白质组遗传信息和生物功能的纽带。

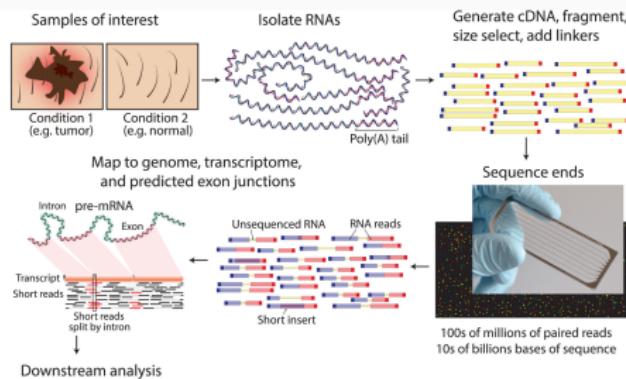
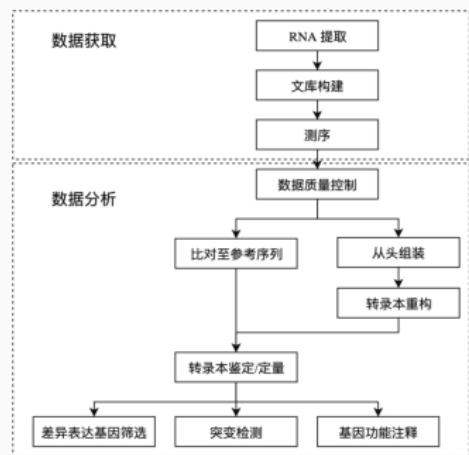
## 转录组学

从整体转录水平系统研究基因转录图谱并揭示复杂生物学通路和性状调控网络分子机制的学科。

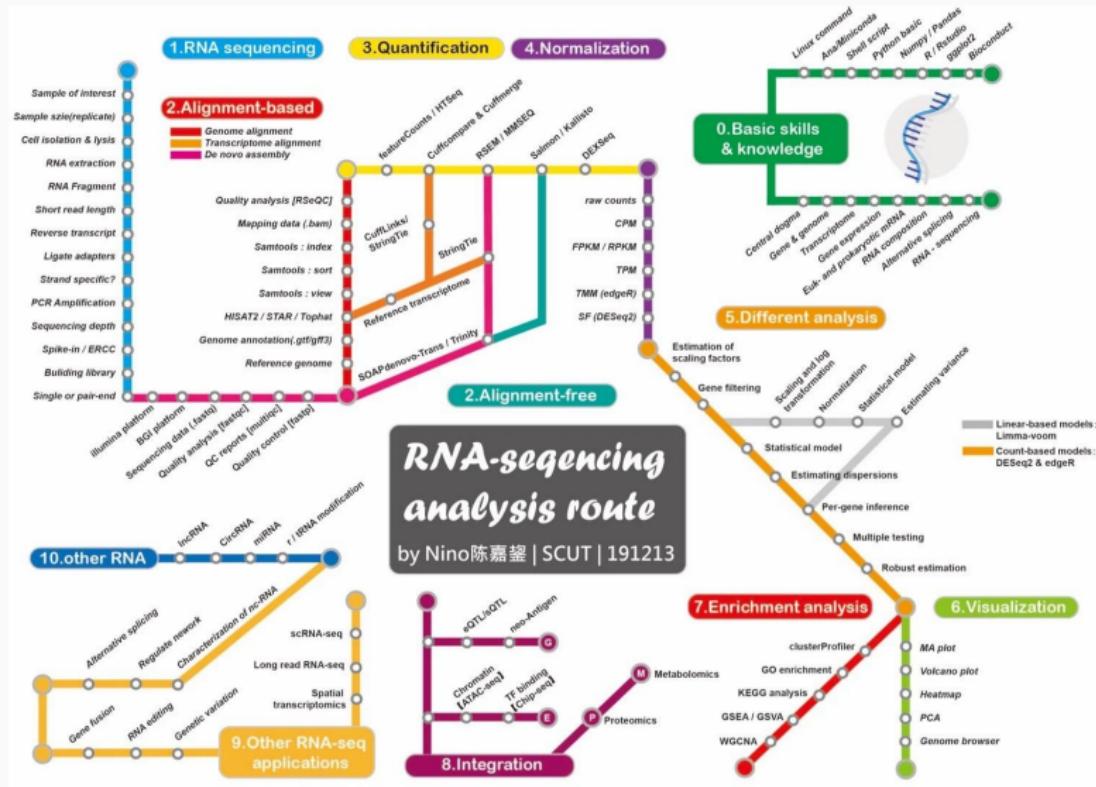
## 转录组测序 (RNA-seq, RNA sequencing, bulk sequencing)

利用高通量测序技术对组织或细胞中所有 RNA 反转录而成的 cDNA 文库进行测序，从整体水平研究基因的功能及其结构，揭示特定生物学过程以及疾病发生过程中的分子机制。

# 导言 | 转录组 | 流程

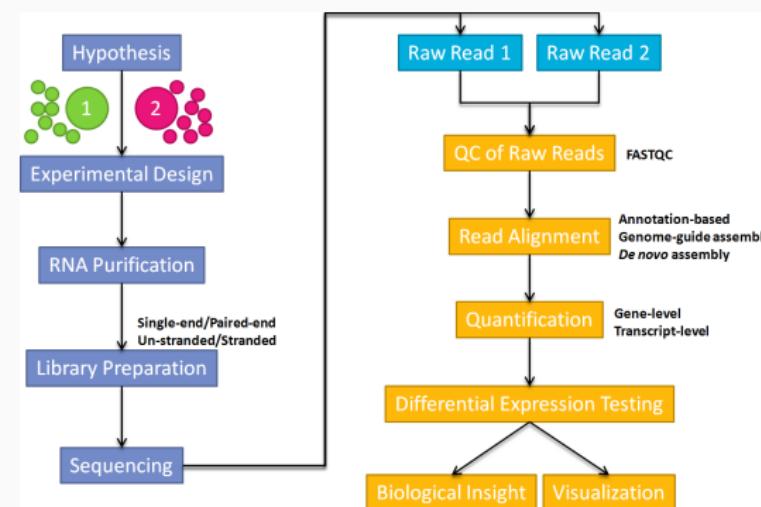


# 导言 | 转录组 | 流程



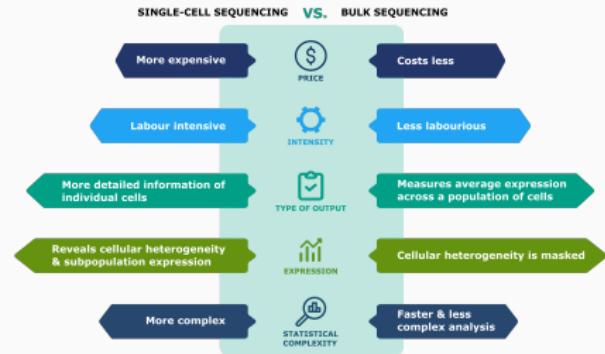
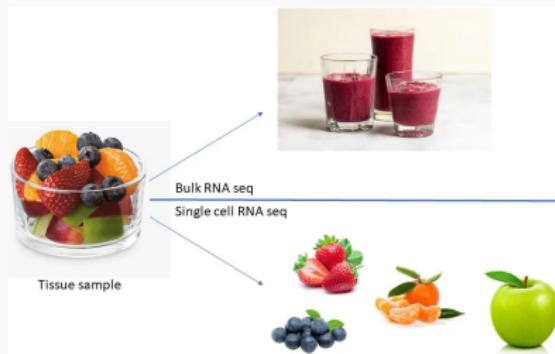
## RNA-seq 应用

- 基因表达定量
- 差异表达分析
- 新转录本预测
- 可变剪接研究
- 共表达网络分析
- 变异检测
- 融合基因识别
- ... ...



## RNA-seq 局限：异、时、空

- 组织中的细胞存在异质性  $\Rightarrow$  单细胞转录组 (scRNA-seq)
- 转录是实时变化的动态过程  $\Rightarrow$  活细胞转录组 (Live-seq)
- 组织中的细胞具有空间位置  $\Rightarrow$  空间转录组 (ST)



## ① 导言

- 常规转录组
- 细胞异质性

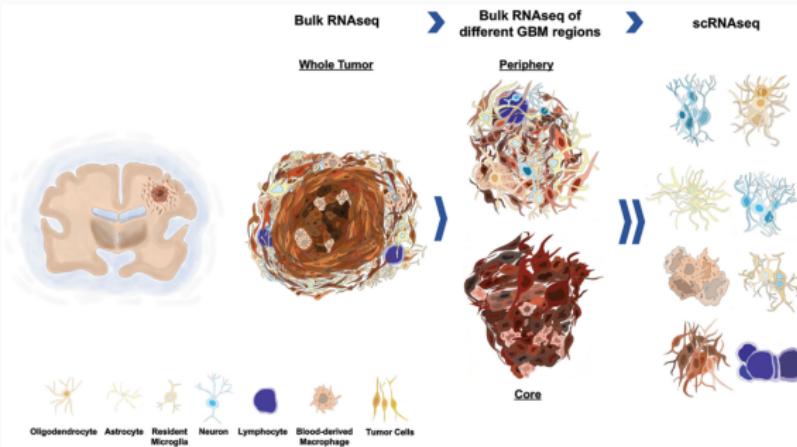
■ 单细胞测序

- ② 单细胞转录组学
- ③ 空间转录组学
- ④ 单细胞多组学

## 细胞异质性 (Heterogeneity)

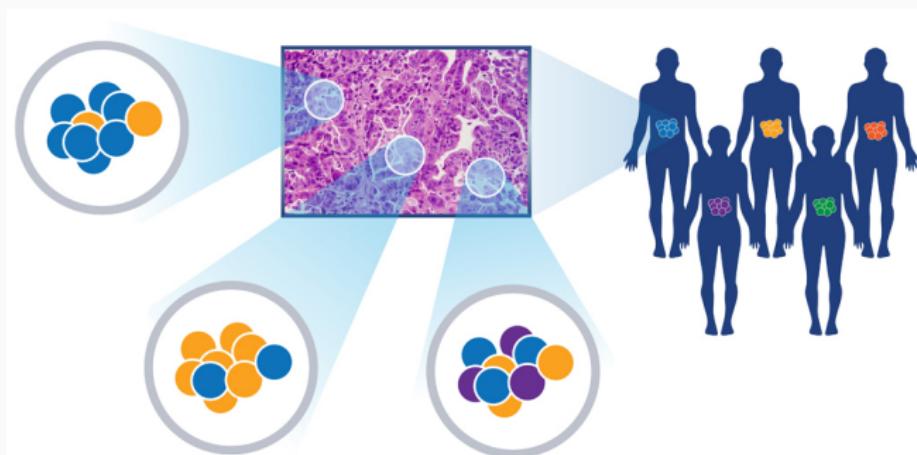
多细胞生物个体由多种形态功能不同的细胞组成；多种类型细胞有序地结合在一起，形成了组织和器官。细胞的异质性是一个普遍存在的生物学现象。

研究细胞异质性，是一个单细胞层面的范畴；单细胞间的异质性存在于 DNA、RNA、蛋白质等各个层面。



## 肿瘤异质性 (Tumor heterogeneity)

同一种恶性肿瘤在不同患者个体间（肿瘤间异质性，**Intertumor heterogeneity**）或者同一患者体内不同部位肿瘤细胞间（肿瘤内异质性，**Intratumor heterogeneity**）从基因型到表型上存在的差异。



## 肿瘤微环境 (Tumor microenvironment)

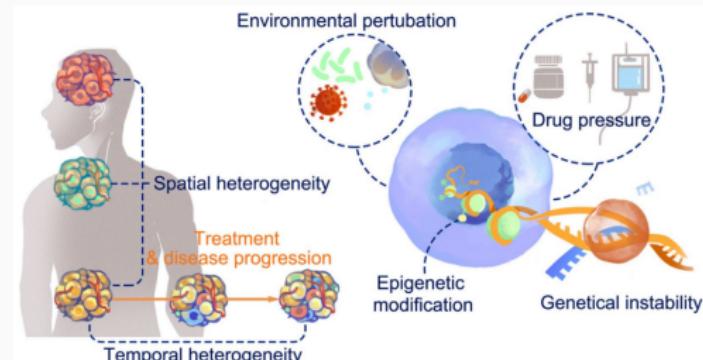
肿瘤细胞存在的周围微环境，包括周围的血管、免疫细胞、成纤维细胞、各种信号分子和细胞外基质（ECM）等。肿瘤微环境有助于肿瘤异质性的形成。

## 肿瘤微环境 (Tumor microenvironment)

肿瘤细胞存在的周围微环境，包括周围的血管、免疫细胞、成纤维细胞、各种信号分子和细胞外基质（ECM）等。肿瘤微环境有助于肿瘤异质性的形成。

## 肿瘤免疫微环境

肿瘤可能被多种免疫相关成分浸润，包括细胞因子/趋化因子、细胞毒性活性或免疫抑制因子。这种免疫异质性在几乎所有实体瘤中普遍存在，并且随着肿瘤的发展以及治疗干预而发生时空变化。





## ① 导言

- 常规转录组
- 细胞异质性

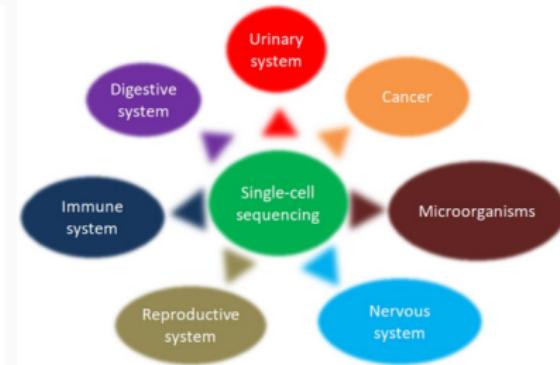
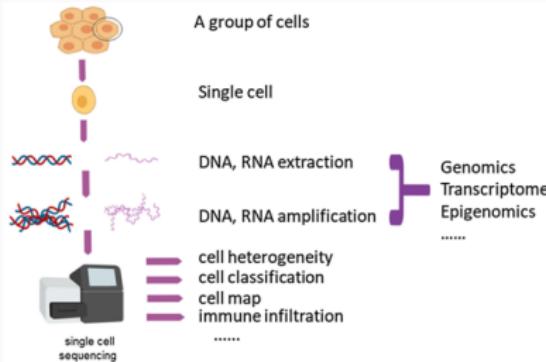
## ■ 单细胞测序

- ② 单细胞转录组学
- ③ 空间转录组学
- ④ 单细胞多组学

## 单细胞测序 (Single cell sequencing)

采取优化的 NGS 技术检测单细胞的序列，可以获得特定微环境下的细胞序列差异以方便研究其功能差异等。

- DNA 测序：了解例如在癌症中的小范围细胞的变异
- RNA 测序：了解和鉴别不同的细胞类型与其表达的基因



# 导言 | 单细胞测序 | 技术

**Table 1 Single-cell transcriptome sequencing.**

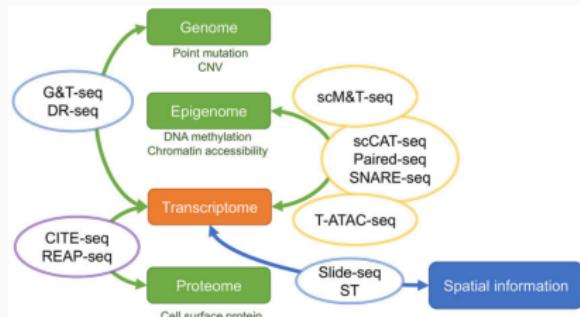
Method	Feature	References
Smart-seq	WTA method; template switching	<sup>3</sup>
CEL-seq	WTA method; in vitro transcription	<sup>6</sup>
Quartz-seq	WTA method; poly(A) tagging	<sup>5</sup>
C1-CAGE	5'-end RNA-seq	<sup>15</sup>
RamDa-seq	Total RNA-seq	<sup>7</sup>
Drop-seq	Microdroplet-based method	<sup>8</sup>
Microwell-seq	Microwell-based method	<sup>10</sup>

**Table 2 Single-cell genome sequencing.**

Method	Feature	References
MDA	WGA method; isothermal amplification	<sup>36</sup>
DOP-PCR	WGA method; PCR-based	<sup>38</sup>
MALBAC	WGA method; hybrid	<sup>37</sup>

**Table 3 Single-cell epigenome sequencing.**

Method	Target	Feature
scBS-seq	DNA methylation	Whole-genome BS-seq
scRRBS	DNA methylation	RRBS
scAba-seq	DNA methylation	ShmC sequencing
scATAC-seq	Chromatin accessibility	ATAC-seq
Drop-ChIP	Histone modification	ChIP-seq; microdroplet-based
scChIC-seq	Histone modification	Ab-Mnase
CUT&Tag	Histone modification	Ab + protein A-Tn5 transposase
Single-cell Hi-C	Chromatin structure	Hi-C



# 章节概览



① 导言

■ 分析流程

② 单细胞转录组学

■ 分析角度

■ 技术简介

③ 空间转录组学

■ 实验原理

④ 单细胞多组学

# 教学提纲



① 导言

■ 分析流程

② 单细胞转录组学

■ 分析角度

■ 技术简介

③ 空间转录组学

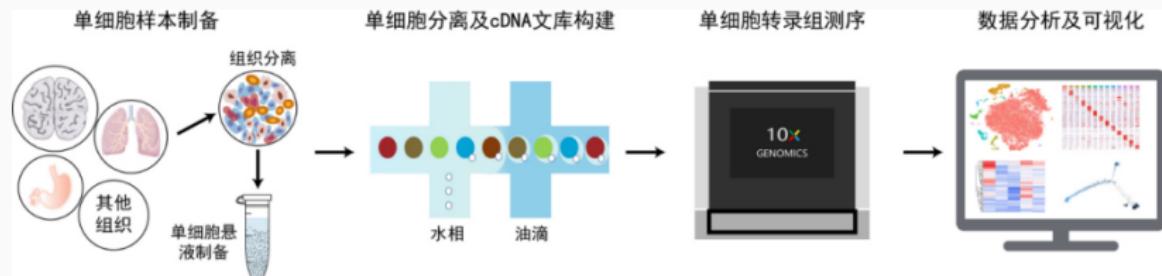
■ 实验原理

④ 单细胞多组学

## 单细胞 RNA 测序 (scRNA-seq, single-cell RNA sequencing)

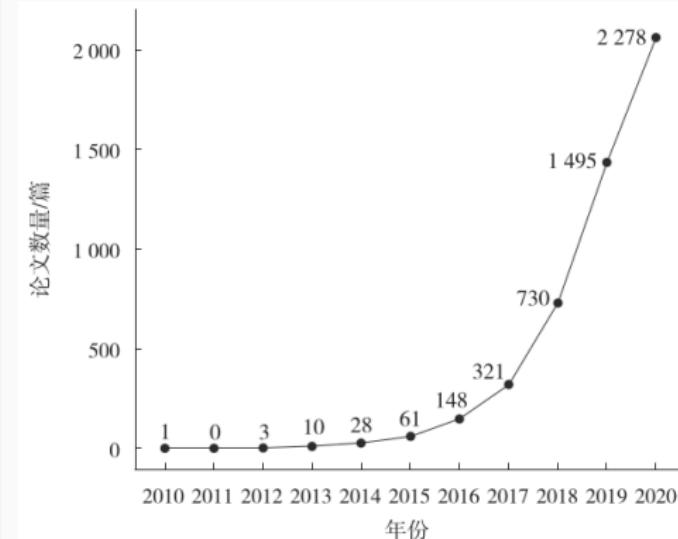
在**单细胞的分辨率水平**上进行 RNA 测序，检测细胞的基因表达水平。

与传统的转录组学测序相比，scRNA-seq 技术可以描绘组织块（或细胞悬液）中单个细胞独特的基因表达模式，反映群体的**细胞异质性**。



## 发展简史

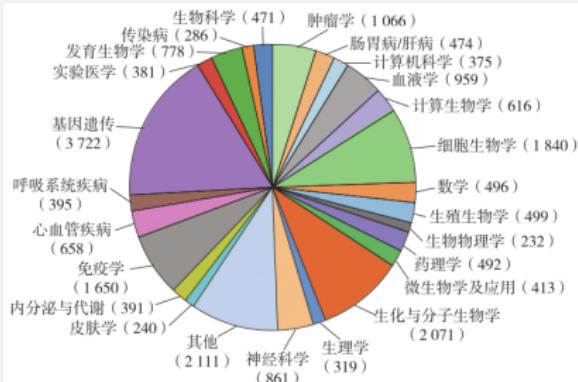
- scRNA-seq 技术由汤富酬等人在 2009 年首次报道。
- 随后，Smart-seq、Smart-seq2、Drop-seq 等不同平台的技术陆续被开发，该领域迅速繁荣起来。
- scRNA-seq 已经迅速成为当前生命科学领域最活跃和前沿的技术之一。

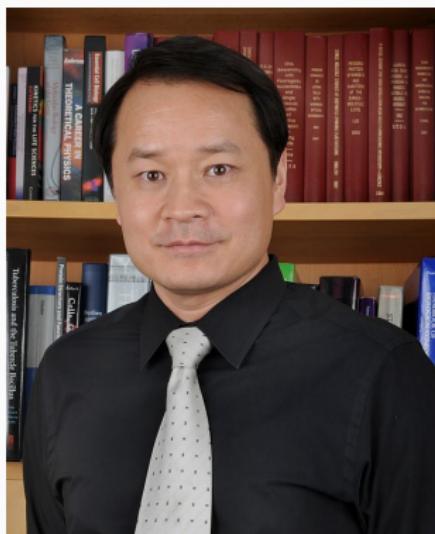


## 荣誉

- 2013 年, scRNA-seq 被 *Nature Methods* 杂志列为年度最主要的方法学进展
- 2019 年, 以 scRNA-seq 为核心代表的单细胞多组学方法再次被 *Nature Methods* 评选为年度方法

## 应用





谢晓亮  
(单细胞基因组学的  
开拓者)



汤富酬  
(开启了单细胞转录组  
测序时代)

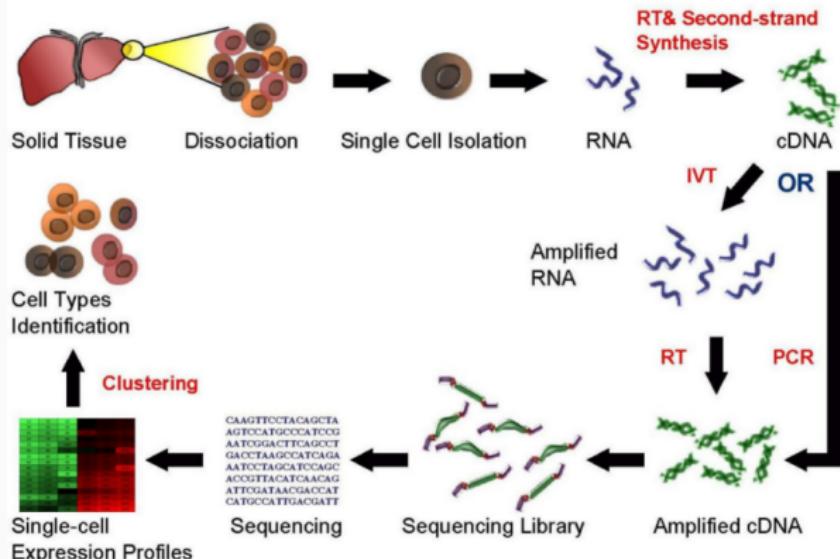


张泽民  
(计算癌症生物学的  
开拓者与引领者)

## scRNA-seq 流程

- ① 组织分离
- ② 单细胞捕获
- ③ 细胞裂解
- ④ 逆转录
- ⑤ 扩增
- ⑥ 文库构建
- ⑦ 测序
- ⑧ 数据分析

### Single Cell RNA Sequencing Workflow

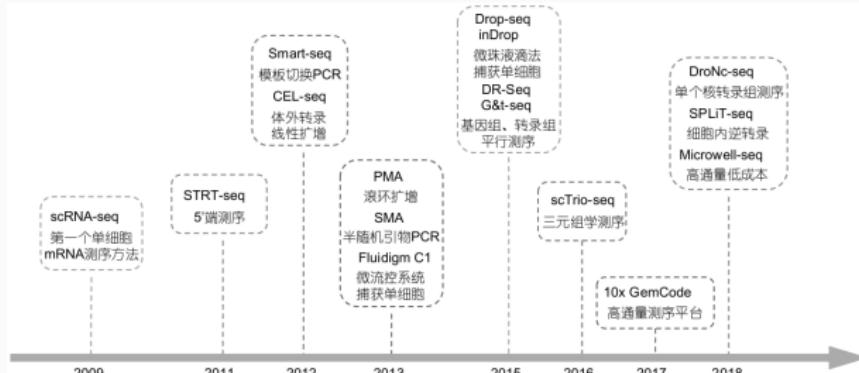


## 主流单细胞捕获技术

### 微流控平台：

- Drop-seq
- InDrop
- Chromium
- .....

技术名称	细胞捕获方法	扩增策略	目的mRNA	细胞总数(个)	技术特点
scRNA-seq	口吸管/FACS	末端加尾法PCR	全长	1~100	灵敏度高, 3'端偏倚较严重
STRT-seq	口吸管/FACS	模板切换法PCR	5'末端	1~100	避免3'端偏倚, 稳定高效, 成本较高
Smart-seq	口吸管/FACS	模板切换法PCR	全长	1~100	避免3'端偏倚, 稳定高效, 成本较高
CEL-seq	口吸管/FACS	体外转录	3'末端	1~100	避免片段长度偏倚, 反应效率较低
PMA	口吸管/FACS	滚环扩增	全长	1~100	避免片段长度偏倚, 反应效率较低
SMA	口吸管/FACS	半随机引物PCR	全长	1~100	避免片段长度偏倚, 反应效率较低
Fluidigm C1	微流控系统	模板切换法PCR	全长	100~1000	操作简便, 仪器芯片价格昂贵
MALBAC-RNA	口吸管/FACS	MALBAC	全长	1~100	避免片段长度偏倚, 扩增效率低
Drop-seq	微珠液滴系统	模板切换法PCR	3'末端	>1000	通量较高, 产物浓度高, 依赖微流控装置
inDrop	微珠液滴系统	体外转录	3'末端	>1000	通量较高, 产物浓度高, 依赖微流控装置
10x GemCode	微珠液滴系统	模板切换法PCR	3'末端	>10000	高通量, 操作简便, 测得基因数少
DroNc-seq	微珠液滴系统	模板切换法PCR	3'末端	>10000	适用于冻存样本, 损失胞质数据
Microwell-seq	琼脂糖微孔板	模板切换法PCR	全长	>10000	成本低, 操作复杂, 测得基因数少
SPLIT-seq	细胞内逆转录	模板切换法PCR	全长	>10000	成本低, 样本得率低, 操作复杂
NICHE-seq	TPLSM, FACS	体外转录	3'末端	>1000	原位单细胞测序, 依赖特定细胞及仪器



# 教学提纲



① 导言

■ 分析流程

② 单细胞转录组学

■ 分析角度

■ 技术简介

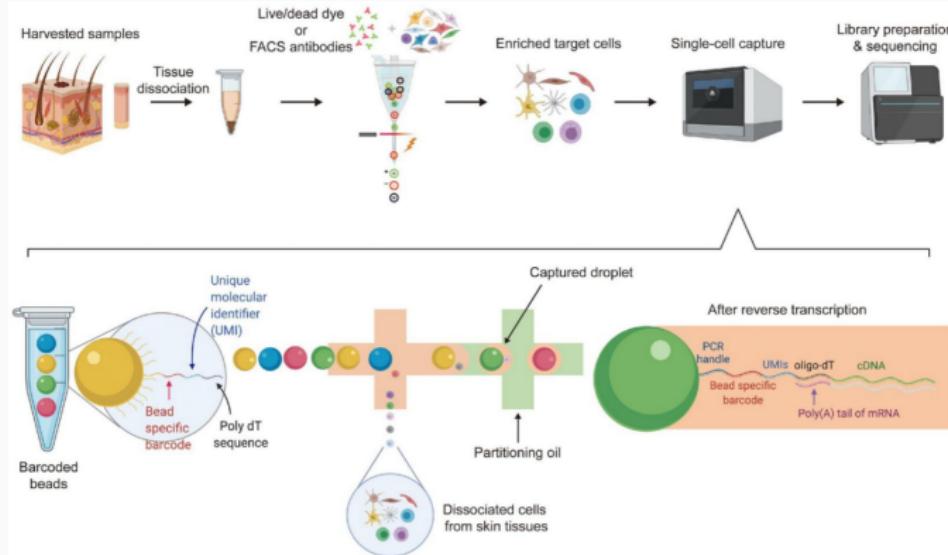
③ 空间转录组学

■ 实验原理

④ 单细胞多组学

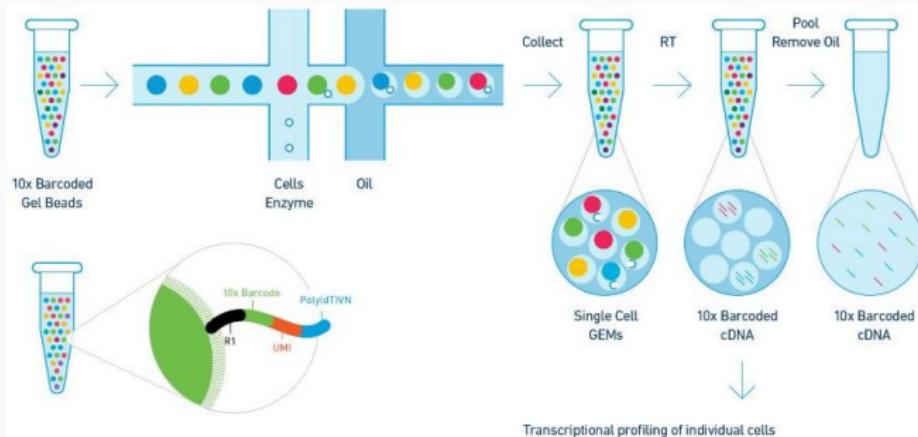
## 微流控平台

在油包水的环境下，单细胞与包含特异寡核苷酸标签（barcode）的凝胶珠结合，给每个细胞带上不同的特异性序列，之后将液滴混合、裂解细胞并进行逆转录等后续步骤。

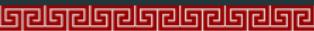


## 10x Genomics 的 Chromium 系统

利用 8 通道的微流体“双十字”交叉系统，将含 barcode 的凝胶珠 (Gel Beads)、细胞和酶的混合物、油三者混合，形成 GEMs (油包水的微体系)，GEMs 形成后，细胞裂解，凝胶珠自动溶解释放大量 barcode 序列。随后 mRNA 逆转录产生带有 10x barcode 和 UMI 信息的 cDNA，构建标准测序文库。



# 教学提纲



① 导言

② 单细胞转录组学

■ 技术简介

■ 实验原理

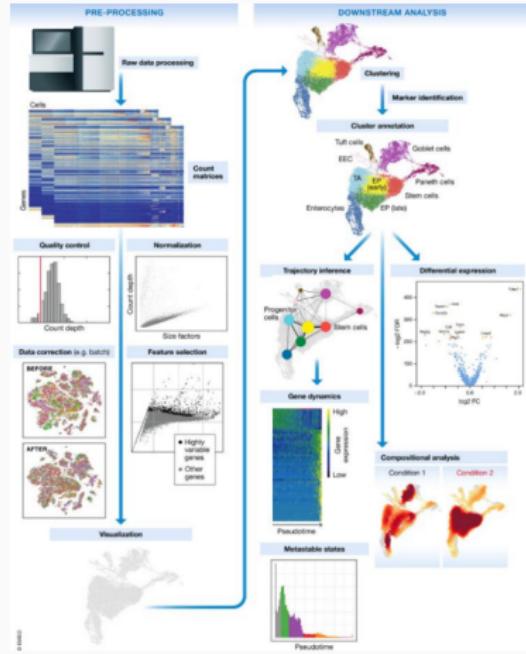
■ 分析流程

■ 分析角度

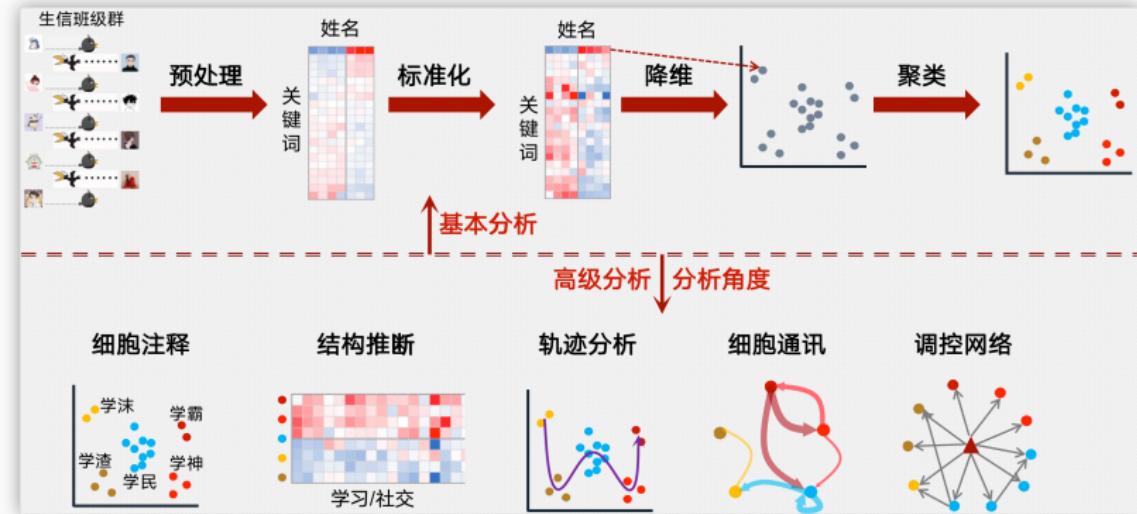
③ 空间转录组学

④ 单细胞多组学

# scRNA | 分析 | 概览



# scRNA | 分析 | 概览



## 对应关系

- 微信聊天记录  $\Rightarrow$  原始测序数据
- 学生  $\Rightarrow$  细胞；关键词  $\Rightarrow$  基因



测序后，得到每个单细胞的转录组表达谱，其中由 **barcode** 标记细胞、**UMI** 标记基因并记录表达量。

在对 barcode 进行分解（demultiplexing）和修剪单次测序所得到的碱基序列 reads 后，每个 barcode 分配给特定细胞中特定基因的 reads（或使用 UMI 的分子），再将 reads 比对到参考基因组，得到单细胞转录组测序的基因的表达矩阵。

每个 10x 样本经 **Cell Ranger** 处理后，将得到 **barcodes.tsv**、**genes.tsv**、**matrix.mtx** 3 个标准的输出文件。下游处理的时候，必须保证这 3 个文件同时存在，且在同一个文件夹下。

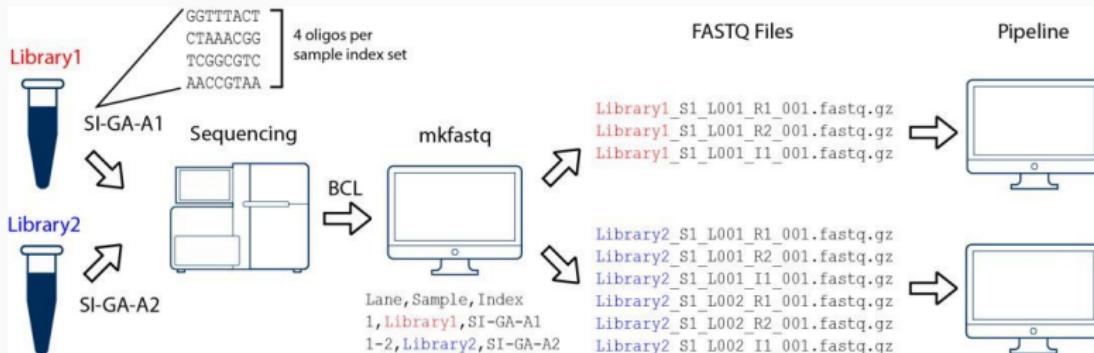


## Cell Ranger

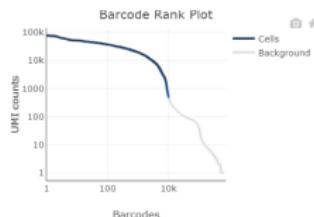
10x genomics 公司为单细胞 RNA 测序分析量身打造的数据分析软件，可以直接输入 Illumina 原始数据（raw base call, BCL）输出表达定量矩阵、降维（PCA）、聚类（Graph-based & K-Means）以及可视化（t-SNE）结果，结合配套的 Loupe Cell Browser 给予研究者更多探索单细胞数据的机会。

- 主要的流程：拆分原始数据 mkfastq、细胞表达定量 count、定量组合 aggr、调参 reanalyze
- 其他小工具：mkref、mkgtf、upload、sitecheck、mat2csv、vdj、mkvdjref、testrun 等

# scRNA | 分析 | 预处理



Cells ⓘ



## Sequencing ⓘ

Number of Reads	833,960,818
Number of Short Reads Skipped	0
Valid Barcodes	97.4%
Valid UMs	100.0%
Sequencing Saturation	81.0%
Q30 Bases in Barcode	97.4%
Q30 Bases in RNA Read	91.4%
Q30 Bases in UMI	97.4%

9,403

Estimated Number of Cells

88,691

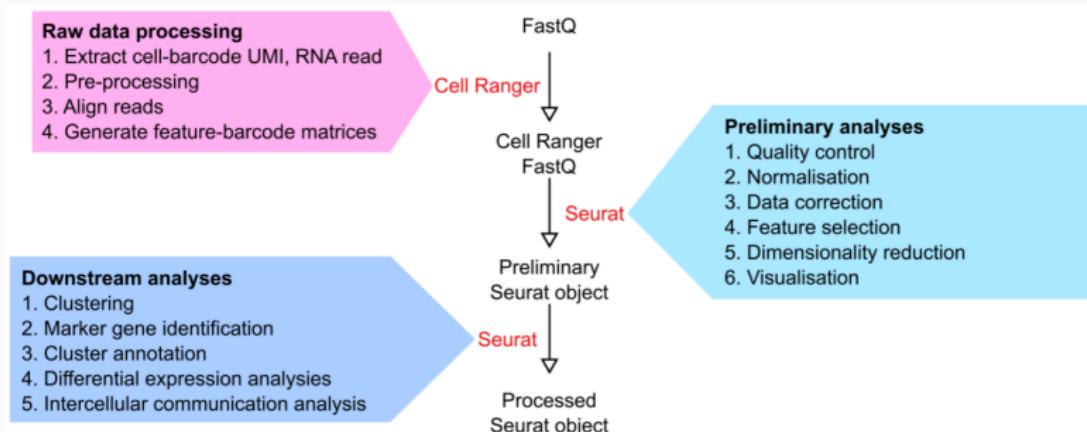
Mean Reads per Cell

1,780

Median Genes per Cell

## 基本分析的主要步骤

- ① 质量控制 (Quality control): 过滤细胞、去除双细胞
- ② 数据标准化 (Normalization): 纠正数据、去除批次效应
- ③ 降维 (Dimensionality reduction): 降低维度、捕获主要信息
- ④ 聚类 (Clustering): 分群细胞、识别相似群体





## 质控

数据识别和去除低质量细胞是 scRNA-seq 质量控制 (Quality control, QC) 的关键步骤。

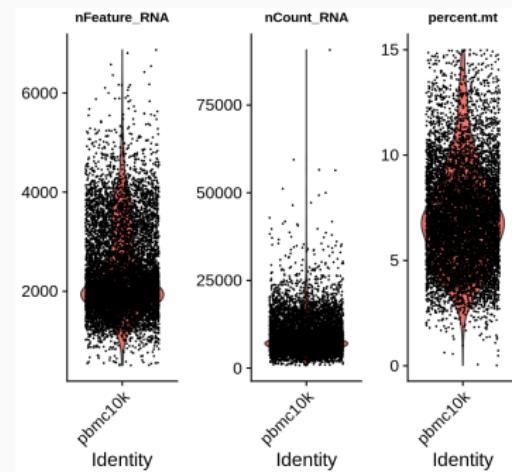
## 质控建议

在实际操作中，考虑到严格筛选可能删除具有实际意义的细胞，通常会采用较为**松弛的标准进行初筛**，在后续分析的过程中再进行二次筛选或人工识别。

## 质控策略

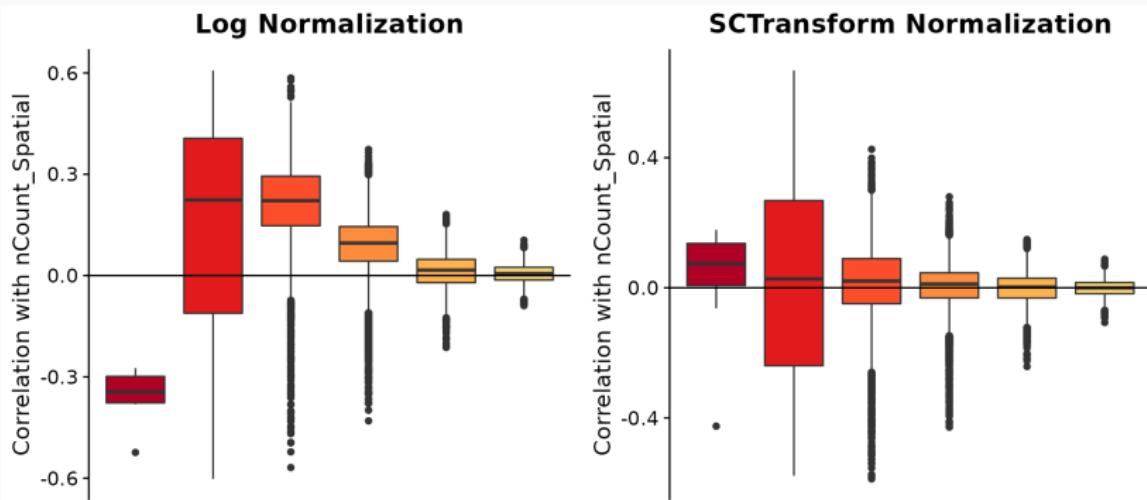
由于技术和环境等原因，细胞标签(Barcodes)可能会标记 doublets、死细胞或外膜破损的细胞。针对这些问题，目前有以下几种解决策略：

- 根据每个细胞中转录本的总量或库的大小进行合格细胞的筛选
- 依据线粒体基因读长 (Reads) 所占的百分比来筛选合格细胞
- 采用 Spikein 占基因总表达量的比例来判断细胞是否符合标准
- 根据每个基因在所有细胞中表达量的总和来筛选基因



## 数据标准化

单细胞 RNA 测序中，由于细胞之间的异质性及技术因素，各单细胞文库大小和测序深度会有不同，需要通过统计学方法消除这种差异，即数据标准化（Normalization）。



## CPM 标准化

以 CPM (Counts per million) 标准化为例，该方法主要基于如下假设：所有细胞中包含等量的 mRNA 分子，故所有的 Count 深度差异全部来自于抽样。CPM 标准化后的 Count 矩阵需要进行对数转换，便于后续的差异表达分析。需要注意的是，对数转换过程中通常会添加一个极小值（如 1），以避免对数底数为 0。

## SCT 标准化

SCTransform 函数可以代替三个函数 (NormalizeData, ScaleData, FindVariableFeatures) 的运行。且其对测序深度的校正效果要好于 log 标准化（10 万以内的细胞都建议使用 SCT 标准化）。

SCTransform 对测序深度的校正效果很好，也可用于矫正线粒体等因素的影响，但不能用于批次矫正。

## 原因

高维性是 scRNA-seq 数据的显著特点。

## 降维

降维 (Dimensionality reduction) 指将包含冗余信息的高维度的单细胞，在保留有用信息的情况下，降至三维或二维可视化，从而减少大部分的计算量。

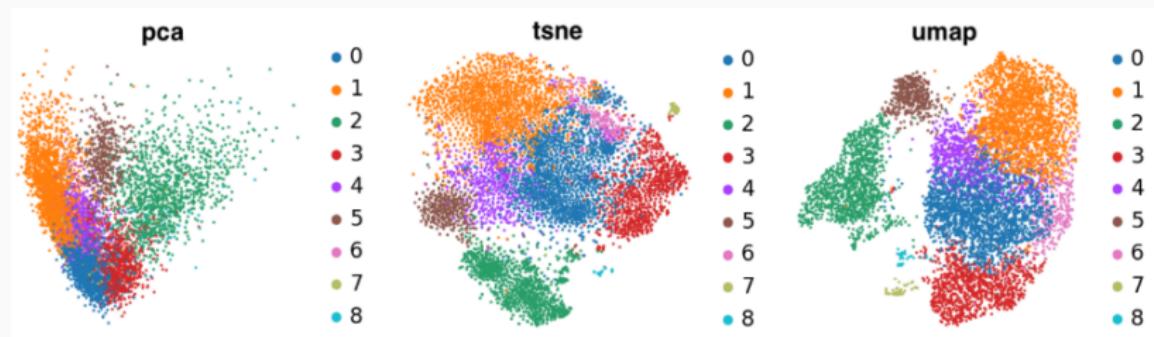
## 类型

- 线性降维：PCA (Principal component analysis, 主成分分析)
- 非线性降维
  - t-SNE (t-distributed stochastic neighbor embedding)
  - UMAP (Uniform manifold approximation and projection)

**PCA** 借助正交变换使线性维数减少，产生一组不相关的分量，通过最大化投影数据的方差，将高维数据投影到低维线性空间上。

**t-SNE** 通过捕获局部结构，将原始高维空间中不相似单元以大距离建模，而相似单元则以小距离建模，在不丢失数据点间相对距离的基础上，将高维数据嵌入到二维或三维空间中进行可视化。

**UMAP** 沿着分化轨迹排列簇并保留瞬时细胞的分化连续体，通过在二维或三维图上覆盖标记基因的表达或与生物过程有关的一组基因的活性，捕获 scRNA-seq 数据中局部和全局结构。



## 依据

相似的细胞具有相似的基因表达谱，因此根据每个细胞中的基因表达情况，可将相似类型的细胞聚集到一起，形成一个细胞簇。

## 聚类

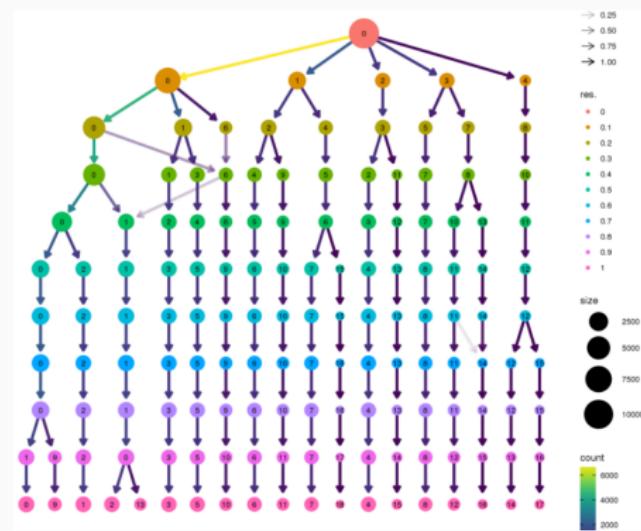
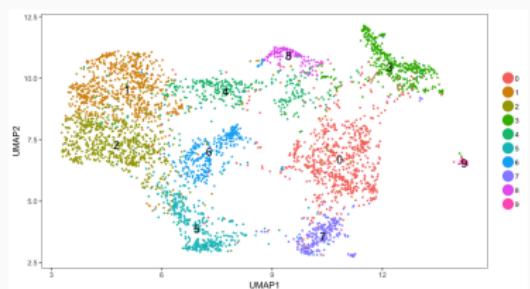
聚类（Clustering）主要是依据细胞-细胞距离矩阵将细胞归属到数目不等的类群中，使高度相似的细胞最大限度地聚为一个类群。

## 目标

聚类的目标是探究或鉴定组织样本中细胞类型或亚型，揭示组织的复杂结构和潜在功能。

## 注意事项

研究者事先并不知原始的所有细胞应该归属于几类以及这些细胞是否具有聚类的意义，因此聚类前用户需要初步判断数据集的聚类趋势。



## Seurat

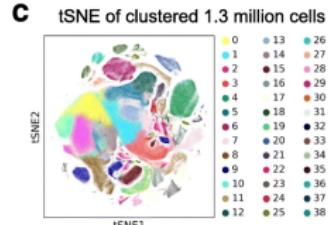
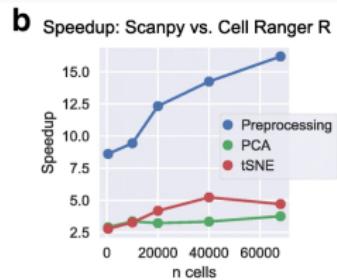
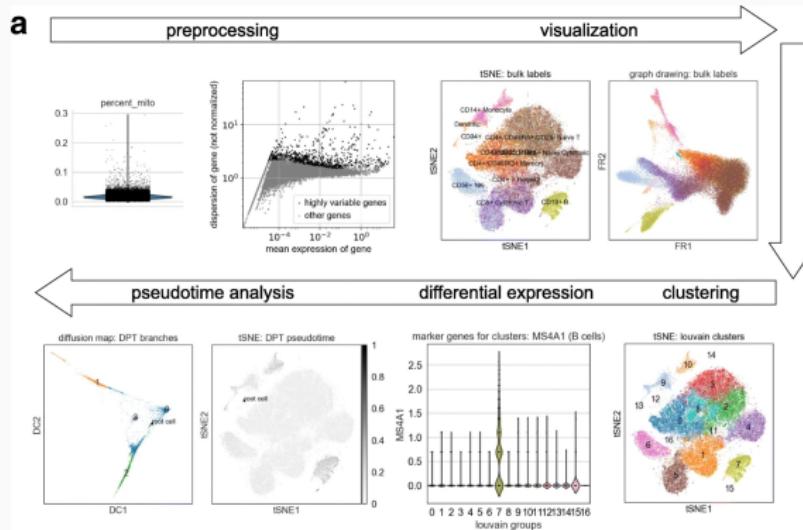
R 软件包，用于 QC、分析和探索单细胞 RNA-seq 数据。

- 基本分析：质控、细胞筛选、细胞类型鉴定、特征基因选择、差异表达分析、数据可视化，等
- 高级功能：时序单细胞数据分析、多组学单细胞数据整合分析，等



## Scanpy

基于 Python 分析单细胞数据的软件包，包括预处理、可视化、聚类、拟时序分析和差异表达分析等。



# scRNA | 分析 | 基本 | 工具

Step	Seurat	Scanpy	Python
Read the data from file	read.csv()*	scanpy.read_csv	pandas.read_csv()
Convert to special data format	CreateSeuratObject()	Already converted as AnnData	Keep as pd. DataFrame
Filter off outliers	Regular R functions	FilterCells(), FilterGenes()	Use general pandas functions for subsetting by threshold values
Normalize and log-transform	NormalizeData()	normalize_total()	normalize from Sklearn or self-made script
Remove invariant genes	FindVariableFeatures()	highly_variable_genes()	Use pandas DataFrame filter by <i>var</i> value. Use VarianceThreshold() from Sklearn
Scale gene expressions to 0-1 interval	ScaleData()	scale()	Normalize() in Sklearn
Run PCA, estimate significant components	RunPCA(), JackStraw()	pca()	Sklearn PCA()
Find or use predefined clusters	FindNeighbors(), FindClusters()	Import leiden, other options possible	Different options in Sklearn.cluster
Run tSNE, visualize clusters	RunTSNE(), TSNEplot()	Prefers UMAP (as imported package)	tSNE and other options in sklearn.manifold
Perform differential expression check	FindMarkers(), FindAllMarkers()	Build in options for Wilcoxon, t-test, logistic regression	t-test, oneway ANOVA, Wilcoxon, Kruskal-Wallis etc. in scipy.stats, RandomForest, ADAboost in sklearn

\*read.csv() in Seurat used for regular table read. Read10X() is for reading matrix data format.

# 教学提纲



① 导言

■ 分析流程

■ 分析角度

② 单细胞转录组学

■ 技术简介

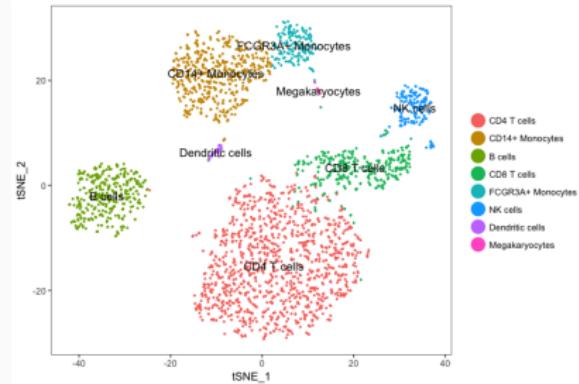
③ 空间转录组学

■ 实验原理

④ 单细胞多组学

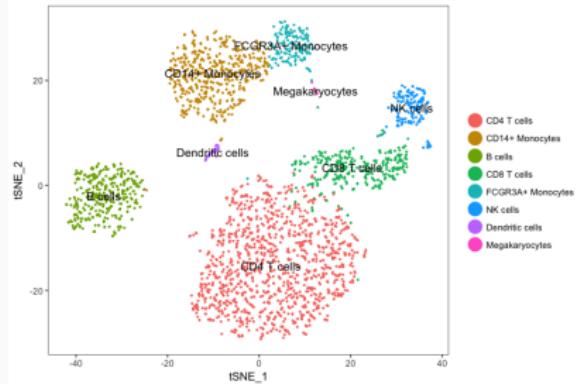
## 细胞类型注释

在 scRNA-seq 数据分析中，通过对单个细胞类群的 marker 基因进行鉴定，可赋予每个类一个有生物学意义的标签，该过程即细胞类型注释。



## 细胞类型注释

在 scRNA-seq 数据分析中，通过对单个细胞类群的 marker 基因进行鉴定，可赋予每个类一个有生物学意义的标签，该过程即细胞类型注释。



## 注释原理

- 鉴定和注释细胞类群主要依赖于外部参考数据库，如 Human Cell Atlas、CellMarker、CancerSEA、PanglaoDB 等。
- 在无相关参考库的情况下，可通过现有细胞的标志基因（Marker）和文献报道的特定类型细胞的标志基因进行匹配来鉴定未知细胞的类型。



### 注释策略：自动注释 vs. 人工注释

- 人工注释的方法在准确性上要优于自动注释
- 自动注释在注释效率和灵敏度上要优于手动注释
- 对于较大的数据集来说，现阶段最好的办法是同时进行软件或数据库自动注释及人工注释
- 对于细胞类型复杂度较低的数据集而言，人工注释更为经济有效



## SingleR

SingleR 是一个用于对 scRNA-seq 数据进行细胞类型自动注释的 R 包。



## 注释原理

- SingleR 通过给定的具有已知类型标签的细胞样本作为参考数据集，对测试数据集中与参考集相似的细胞进行标记注释。
- SingleR 自带 7 个参考数据集，其中 5 个是人类数据，2 个是小鼠的数据。

## 参考数据集

- 人类： BlueprintEncodeData；  
DatabaseImmuneCellExpressionData；  
HumanPrimaryCellAtlasData； MonacolImmuneData；  
NovershternHematopoieticData
- 小鼠： ImmGenData； MouseRNaseqData

## 注释过程

- ① 计算每个细胞的表达谱与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分布的固定分位数。
- ③ 对所有的标签重复此操作，然后将得分最高的标签作为此细胞的注释。

## InferCNV

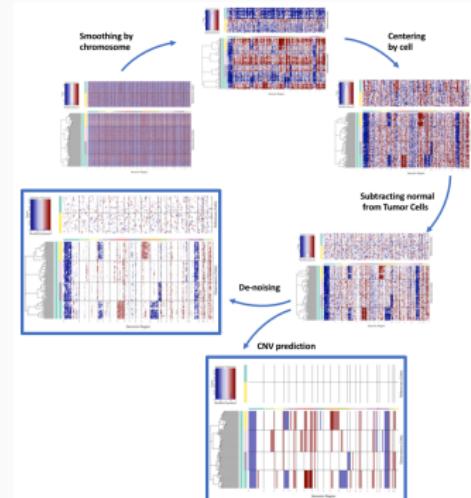
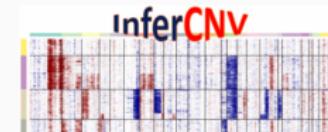
用于肿瘤单细胞 RNA-seq 数据中鉴定大规模染色体拷贝数变异 (copy number alterations, CNA)。

### 主要应用

判断肿瘤细胞；分析肿瘤异质性；探索克隆进化

### 基本原理

在整个基因组范围内，将每个肿瘤细胞基因表达与平均表达或“正常”参考细胞基因表达对比，确定其表达强度，分析肿瘤基因组上各个位置的基因表达量强度变化。



## 拟时序分析法

根据单个细胞的基因表达模式推断出细胞发育或分化的动态路径。注意分析结果不一定代表实际的细胞分化过程。

## 拟时序分析法

根据单个细胞的基因表达模式推断出细胞发育或分化的动态路径。注意分析结果不一定代表实际的细胞分化过程。

## 基本原理

通常 scRNA-seq 技术只能描绘在某一时刻细胞中的基因转录表达状态，只能得到一张细胞的“快照”。而“伪时间 (pseudotime)”可以近似作为衡量细胞分化发育的相对次序，该次序通过对细胞间表达谱的相似性计算和推测得到。随后，这些细胞将会按照这种相对次序被分配到一个一维空间中，代表着细胞发育分化进程中的一种独特状态。

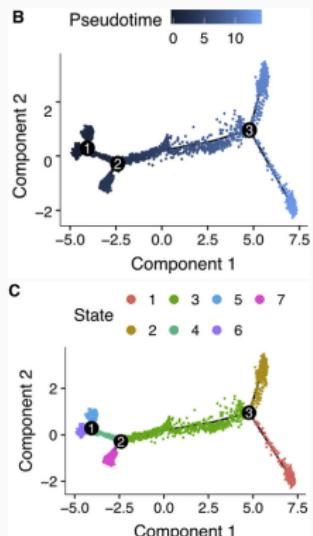


## Monocle

首个用于不同分化阶段对细胞排序且兼具鲁棒性和高效性的软件。

## 基本原理

在细胞分化发育的过程中都会执行一套特定的基因表达程序，而在整体的生物学过程中，各个细胞执行这套程序往往是不同步的，Monocle 正是利用了这一特点，将测序得到的细胞放置在计算得到的一条轨迹上，从而刻画出生物学过程（如发育分化等）中细胞的伪时序发展路径，并进一步提供聚类和差异表达分析等手段帮助我们更好地理解生物过程发展和调控的机制。



## RNA 速率 (RNA velocity)

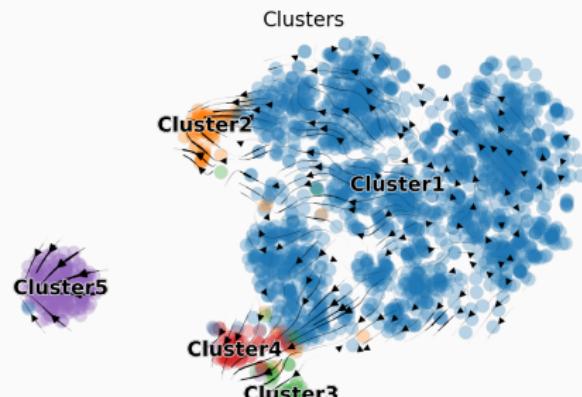
- 基因表达状态的时间导数
- 通过区分 scRNA-seq 中未剪接和剪接的 mRNA 来直接估计
- 可以在数小时的时间尺度上预测单个细胞的未来状态
- velocityo 是 RNA 速率分析的常用工具

## RNA 速率 (RNA velocity)

- 基因表达状态的时间导数
- 通过区分 scRNA-seq 中未剪接和剪接的 mRNA 来直接估计
- 可以在数小时的时间尺度上预测单个细胞的未来状态
- velocityo 是 RNA 速率分析的常用工具

## 基本原理

通过计算细胞内 mRNA 剪切前后的比例来估算 RNA 丰度随时间的变化，可以用于细胞分化、谱系发育、肿瘤微环境中细胞成分的动态轨迹变化等研究。

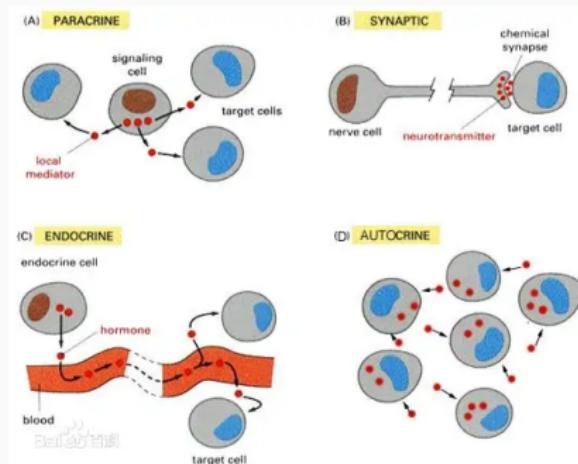


## 细胞通讯分析

根据不同细胞膜表面和游离蛋白之间的配体-受体关系，鉴定不同细胞之间可能的相互作用。

## 细胞-细胞互作 (Cell-cell interaction, CCI)

作为生命活动的基本单位，细胞与细胞之间能通过表面受体-配体蛋白识别结合，并以旁分泌和自分泌等方式传递信号，调控受体细胞的分化、凋亡以及有丝分裂等过程，因此 CCI 网络在整个生命活动中发挥着重要作用。



## scRNA-seq 细胞通讯分析

**主要目标** 比较不同样品组的细胞在各细胞类型之间的配体与受体基因表达差异。

**基本思路** 从单细胞基因表达矩阵出发，结合已有的配体-受体信息，量化配体-受体相互作用的强度，进而推测细胞间的互作关系。

**常用工具** CellPhoneDB、CellTalker、iTALK 和 CellChat 等。

### CellChat

开源 R 包，主要用于细胞间通讯的分析和可视化。它使用配体受体对应的基因表达对来量化细胞间的相互作用。



**CellphoneDB** 储存受体、配体以及两种相互作用的数据库，考虑了结构组成，能够描述异构复合物。(配体-受体 + 多聚体)

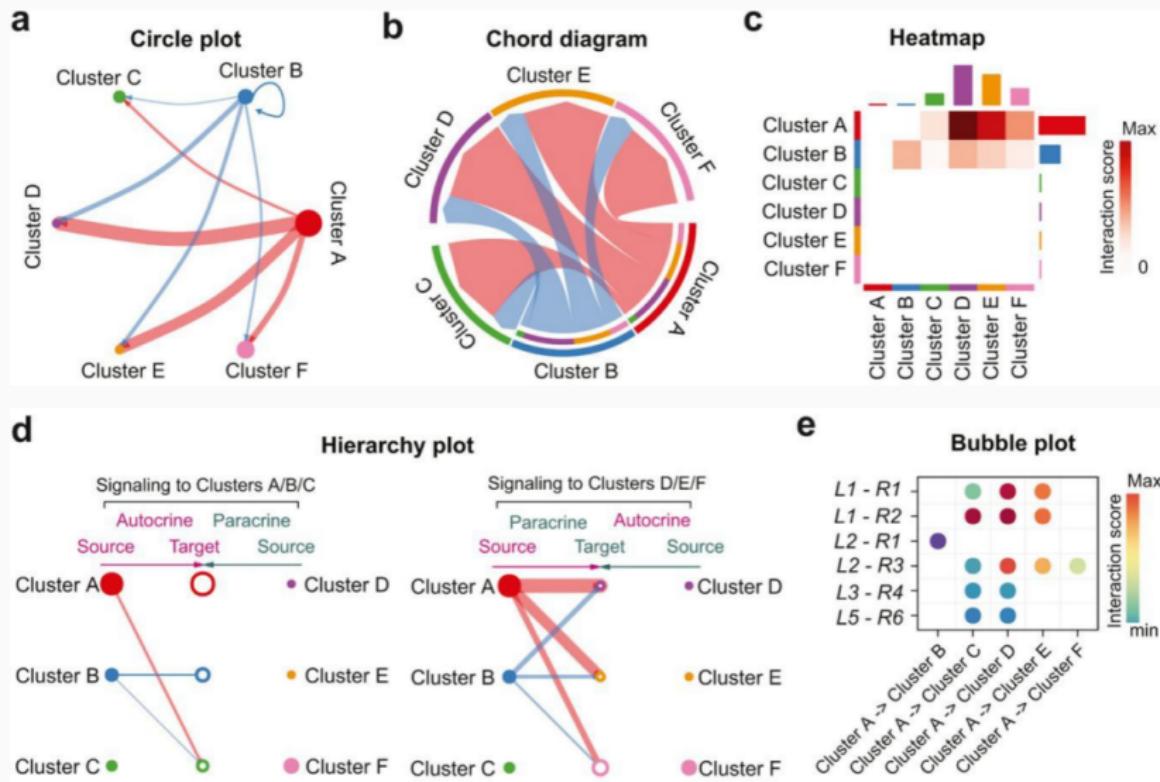
**iTALK** 通过平均表达量方式，筛选高表达的配体和受体，根据结果作圈图。(配体-受体)

**CellChat** 将基因表达数据作为输入，结合配体受体及其辅助因子的相互作用来模拟细胞间通讯。(配体-受体 + 多聚体 + 辅因子)

**NicheNet** 通过将相互作用细胞的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-靶标联系的方法。(配体-受体 + 信号通路)

**Celltalker** 通过寻找细胞群内和细胞群之间已知的配体和受体对的表达来评估细胞间的交流。(配体-受体)

# scRNA | 分析 | 角度 | 细胞通讯



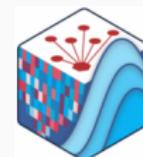
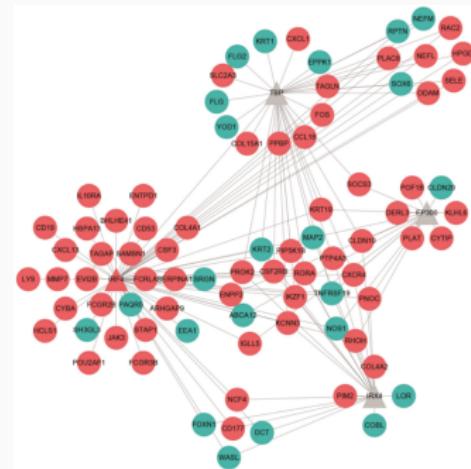
## 转录调控网络

基因转录调控网络描述转录因子及其调控的基因之间的关系。

阐明转录调控网络的结构和功能是很多研究的核心目标，该项研究面临的挑战主要是重建调控网络，因为基因或转录本代表的节点或边界之间都是相互作用。

## SCENIC

识别转录因子与潜在靶基因之间的共表达模块（regulon）。



SCENIC+  
Single-cell enhancer-gene  
regulatory networks

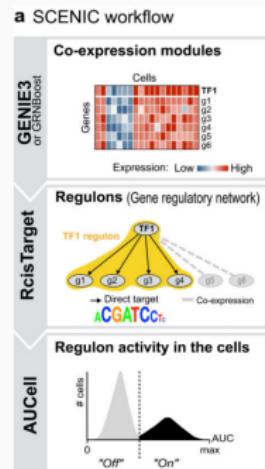
## SCENIC

SCENIC 可以得到细胞类型中 regulon 的活性强度，找出 regulon 与每种细胞类型之间特定的对应关系。同时可获得不同 regulon 之间的关联性，具有较高关联性的 regulon 可能共同调控下游基因，并共同负责细胞功能。

## SCENIC

SCENIC 可以得到细胞类型中 regulon 的活性强度，找出 regulon 与每种细胞类型之间特定的对应关系。同时可获得不同 regulon 之间的关联性，具有较高关联性的 regulon 可能共同调控下游基因，并共同负责细胞功能。

- ① GENIE3 运用随机森林推断潜在的转录因子靶标。
- ② RcisTarget 分析每个 regulon 的基因，以鉴定富集的 motif。每个转录因子及其潜在的直接靶标被称为一个 regulon。
- ③ AUCell 使用 AUC 来计算输入基因集的关键子集是否在每个细胞的表达基因中富集，为细胞间的 regulon 打分赋值。



## scRNA-seq 分析资料与工具合辑

- scRNA-seq data analysis tools and papers: Single-cell RNA-seq related tools and genomics data analysis resources.
- awesome-single-cell: List of software packages (and the people developing these methods) for single-cell data analysis, including RNA-seq, ATAC-seq, etc.
- scRNA-tools: A database of software tools for the analysis of single-cell RNA-seq data.



# 章节概览

- ① 导言
- ② 单细胞转录组学
- ③ 空间转录组学

- 技术简介
  - 实验原理
  - 分析策略
- ④ 单细胞多组学

# 教学提纲

① 导言

② 单细胞转录组学

③ 空间转录组学

## ■ 技术简介

■ 实验原理

■ 分析策略

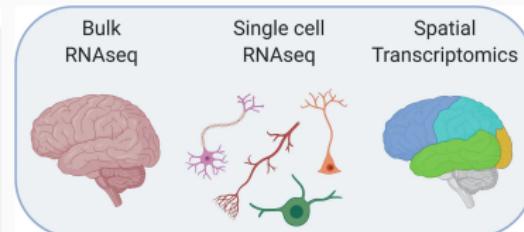
④ 单细胞多组学

## 基因表达：动态 + 时空异质性

**传统的转录组测序** 将 RNA 的表达量平均化，忽略了细胞群体内不同细胞之间基因表达的异质性。

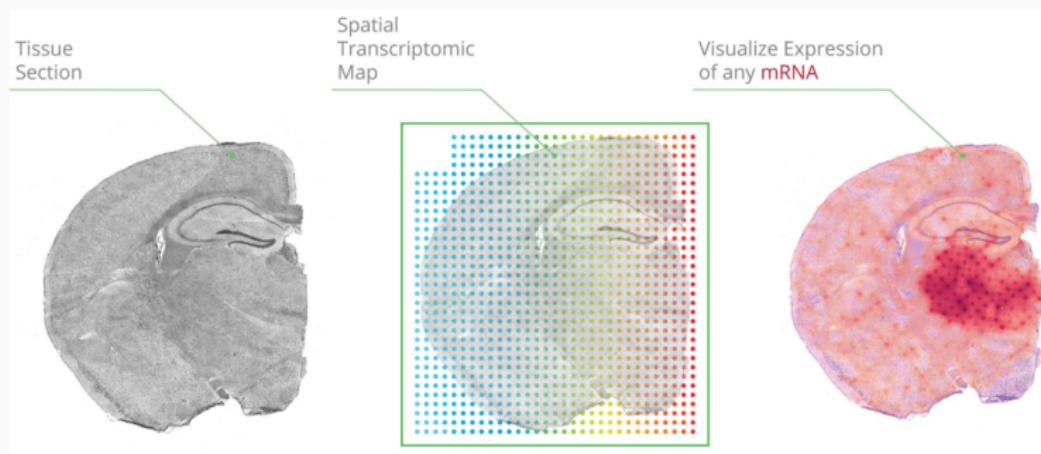
**单细胞 RNA 测序** 能够独立地提供每个细胞的 RNA 表达谱，区分出细胞之间的基因表达差异，并鉴定出异质细胞群中的稀有细胞；需要将细胞从组织中解离，从而导致细胞空间位置信息的丢失。

**空间转录组** 记录所检测 RNA 分子的空间位置信息。



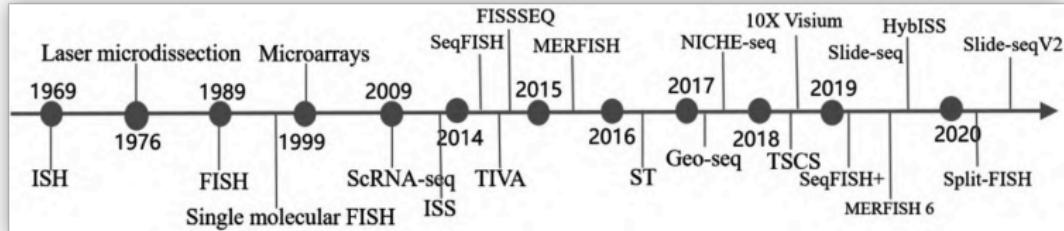
## 空间转录组 (Spatial transcriptomics, ST)

结合成像、生物标记、测序及生物信息学等工具对组织切片的基因表达进行空间定位的一项技术，可揭示各细胞类型在组织中的空间分布、各细胞群体间的相互作用以及绘制不同组织区域的基因表达图谱，对于理解疾病和癌症的发生机制具有深远的应用价值。



## 技术发展

- 一系列能够进行高通量原位 RNA 检测分析的技术都被归为空间转录组学技术的范畴
- 空间转录组的发展可以追溯到 1969 年的原位杂交技术 (*in situ* hybridization, ISH) 的应用
- 2020 年, “空间转录组技术”被 *Nature Methods* 评为年度技术方法
- 2022 年, 空间多组学被 *Nature* 评为值得关注的七大技术之一



## 技术概述

根据空间转录组获取空间信息的原理不同，分为四类：

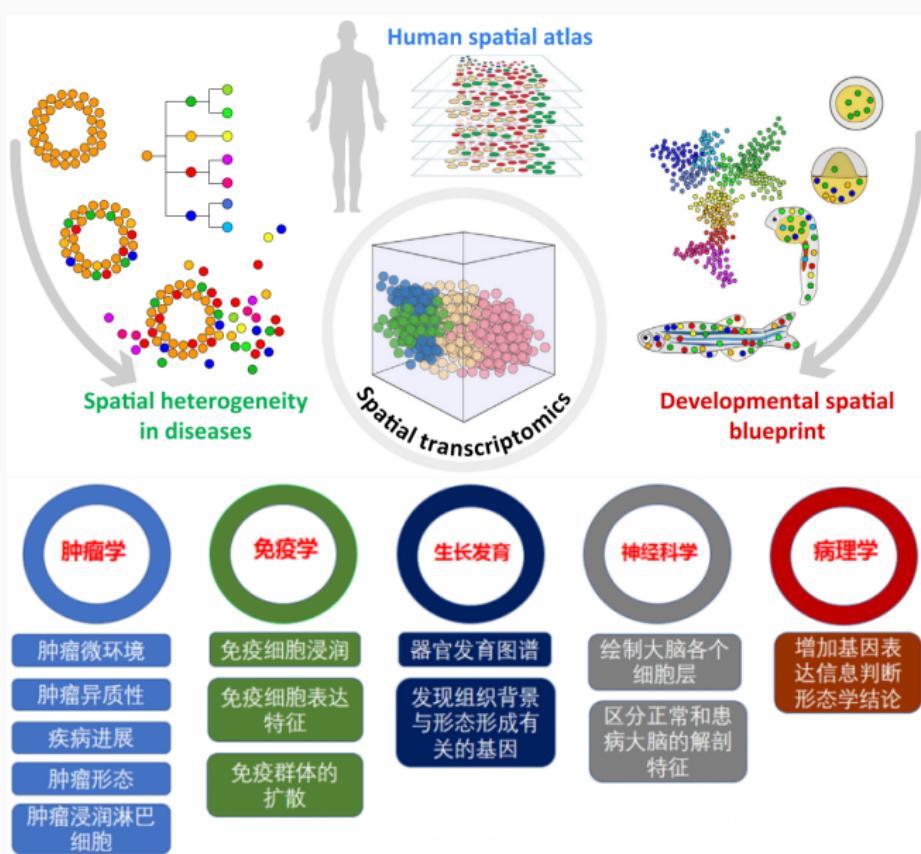
**显微切割技术** 在显微镜下通过手工操作或仪器采样的方法从组织切片或细胞图片上将所研究的目标细胞从中分离出来。

**原位杂交技术** 使用同位素标记或荧光标记的探针与预定的靶 RNA/DNA 杂交，确定组织或细胞中的 RNA/DNA 丰度，逐渐发展至单分子分辨水平。

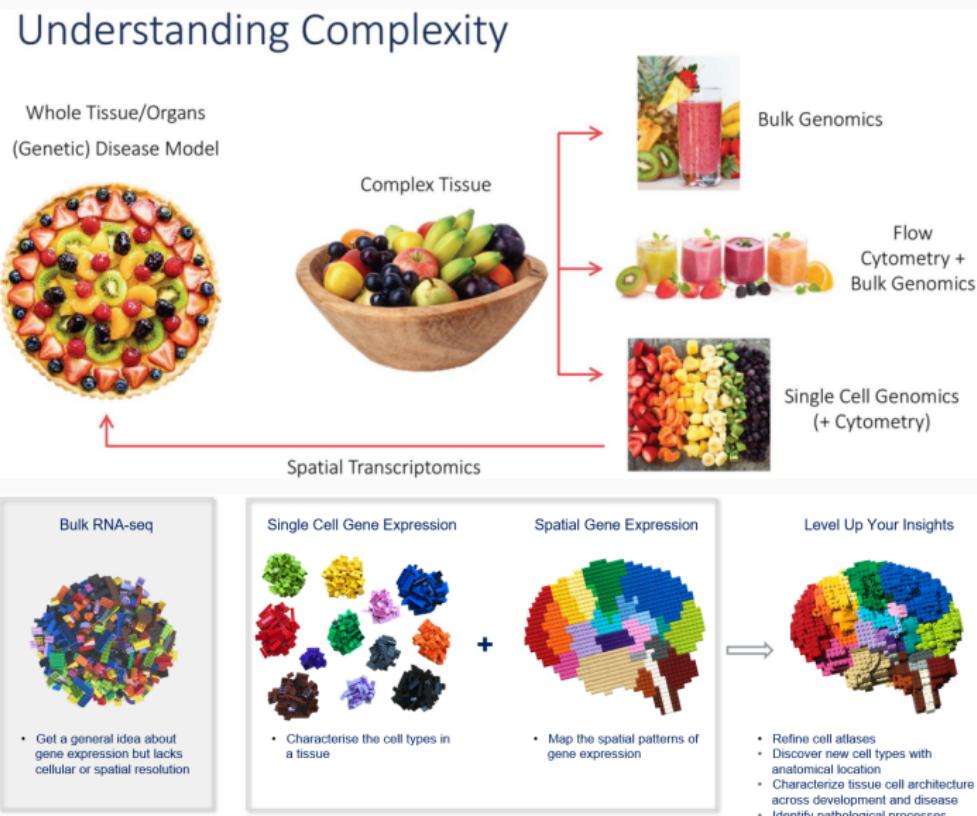
**原位测序技术** 利用单碱基特异性荧光杂交和 DNA 连接反应对组织中已知或未知转录本的序列进行原位测定。

**原位捕获技术** 利用具有空间位置信息的 DNA 标签引物捕获并标记转录本的空间位置。

技术分类	样本类型	方法	空间分辨率	优势	缺点
<b>显微切割技术</b>					
Geo-seq	新鲜冰冻组织	全转录本	10个细胞	比LCM更敏感	低通量
TIVA	完整活细胞或组织	全转录组	细胞水平	能够应用于活细胞	不能应用于人类样本研究
NICHE-seq	活细胞	全转录组	细胞水平	高通量	不能应用于人类标本
TSCS	新鲜冰冻组织	全转录组	细胞水平	可分析10 000个基因	仅应用于100个细胞
<b>原位杂交技术</b>					
smFISH	冰冻组织或石蜡切片组织	靶向	亚细胞	高敏感性	低通量
seqFISH	新鲜冰冻组织	靶向	亚细胞	信噪比smFISH提高20倍	需要专用设备，实验成本高
MERFISH	新鲜冰冻组织	靶向	细胞水平	实验效率高，成像时间短	实验成本高，杂交荧光背景强
seqFISH+	新鲜冰冻组织	靶向	细胞水平	可分析10 000个基因	实验成本高
split-FISH	新鲜冰冻组织	靶向	细胞水平	降低了假阳性率	实验成本高
<b>原位测序技术</b>					
BaristaSeq	培养的细胞	靶向	亚细胞水平	探针缺口内读取长度可读取15个碱基	不能应用于实验组织中
FISSEQ	冰冻组织或石蜡切片组织	全转录组	亚细胞水平	不需要设定靶基因	转录本检测灵敏度相对较低
<b>原位捕获技术</b>					
ST	新鲜冰冻组织	全转录组	精确到100 μm	可检测完整组织切片中总mRNA	条码区域覆盖多个细胞，未达到单细胞
Slide-seq	新鲜冰冻组织	全转录组	精确到10 μm	高通量，可分析10 000个基因	转录本检测灵敏度相对较低
HDST	新鲜冰冻组织	全转录组	精确到2 μm	高通量，检测100 000个细胞	转录本检测灵敏度相对较低



## Understanding Complexity



# 教学提纲

① 导言

② 单细胞转录组学

③ 空间转录组学

■ 技术简介

■ 实验原理

■ 分析策略

④ 单细胞多组学

## 原位捕获技术

原位捕获技术的核心是利用具有空间位置信息的 DNA 标签引物捕获并标记转录本的空间位置。

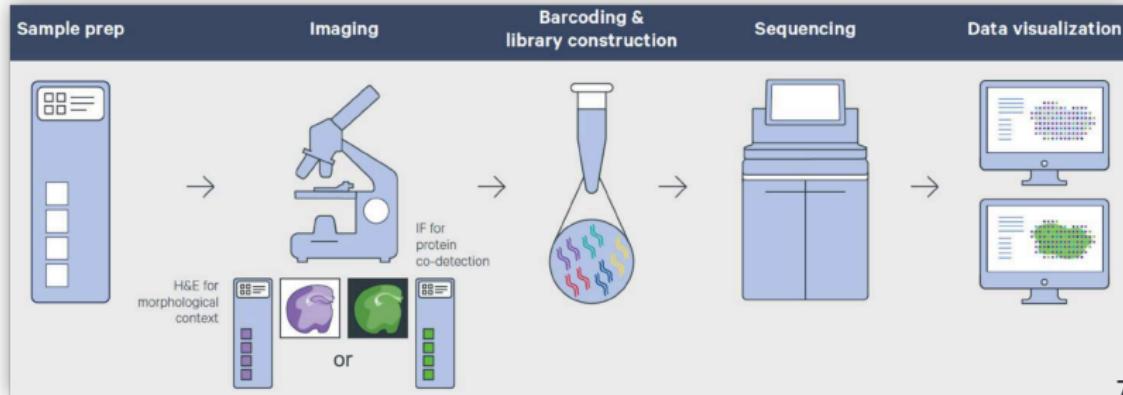
近几年原位捕获技术发展迅猛，并朝着提高空间分辨率及空间多组学联合分析两个方向发展。

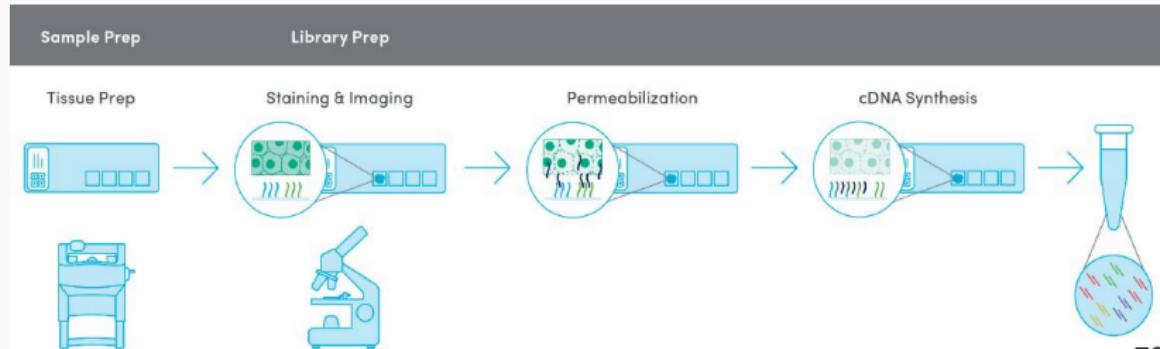
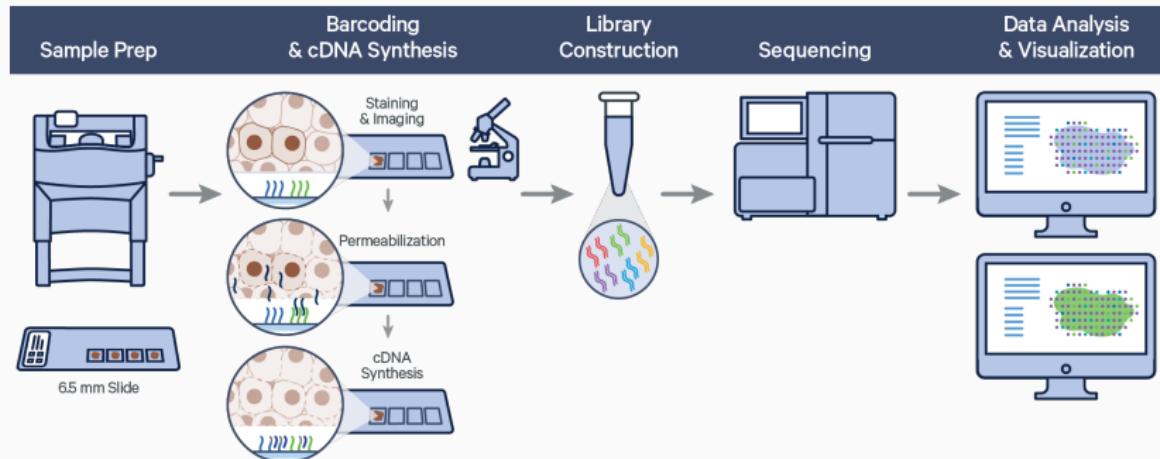
## 空间转录组 (Spatial transcriptomics, ST)

- 2016 年首次提出
- 能够在单个组织切片上提供 mRNA 分布的可视化和基因表达数据
- 2018 年被 10x Genomics 收购并更名为 10x Visium

## 基本原理

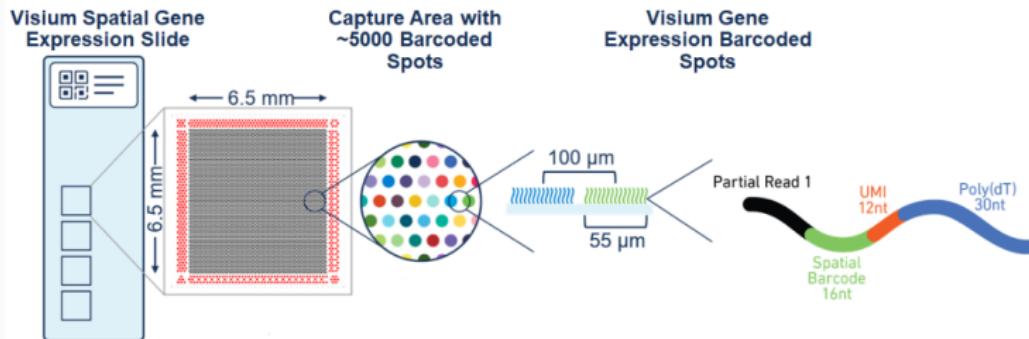
将含有空间编码、分子编码的 mRNA 特异捕获引物簇固定于载玻片表面以制备空间编码玻片；接着将组织切片贴于空间编码玻片，并对其进行固定、染色与成像；利用玻片上的引物捕获经透化后的组织释放的 mRNA，通过逆转录过程生成含有空间编码的 cDNA，经转录组文库制备、高通量测序与分析，将转录组序列映射至原空间位置，实现高覆盖度的组织空间转录组测量分析。





## 技术要点

- 一个载玻片可容纳 4 个切片（捕获区域），制备 4 个文库
- 每个捕获区域大小为 6.5mm × 6.5mm
- 每个捕获区域有 5000 个 barcode 标记的 ST 位点
- 每个 ST 位点直径 55 μm，点与点之间相距 100 μm
- 每个 ST 位点覆盖 1-10 个细胞 (**非单细胞水平**)
- 每个 ST 位点有百万个 UMI 探针，检测数千个基因



# 教学提纲

① 导言

■ 技术简介

■ 实验原理

■ 分析策略

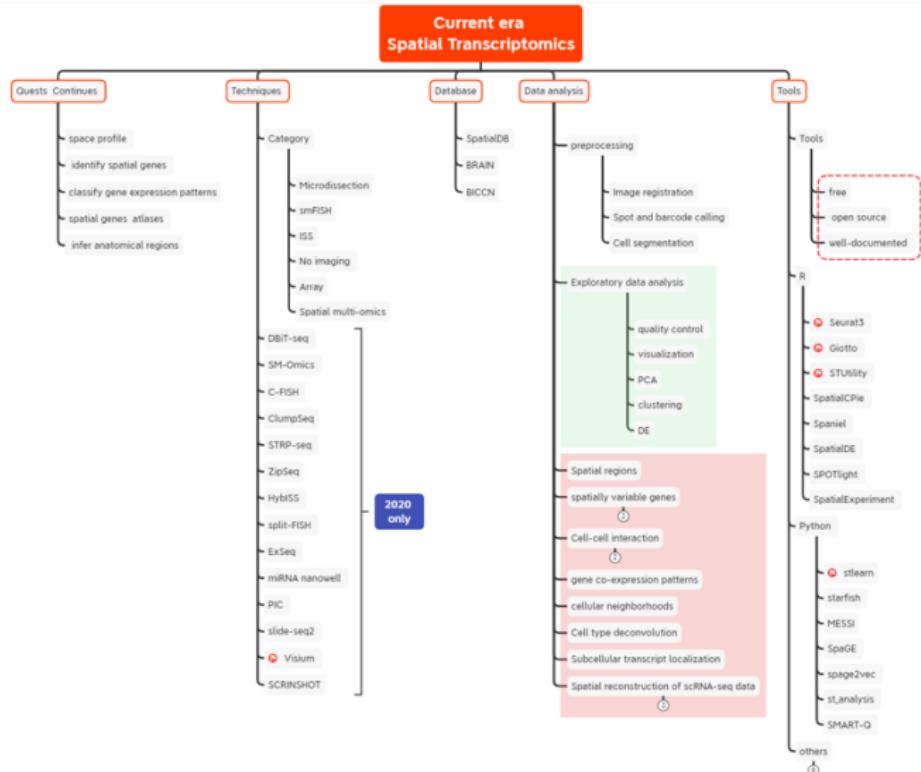
② 单细胞转录组学

④ 单细胞多组学

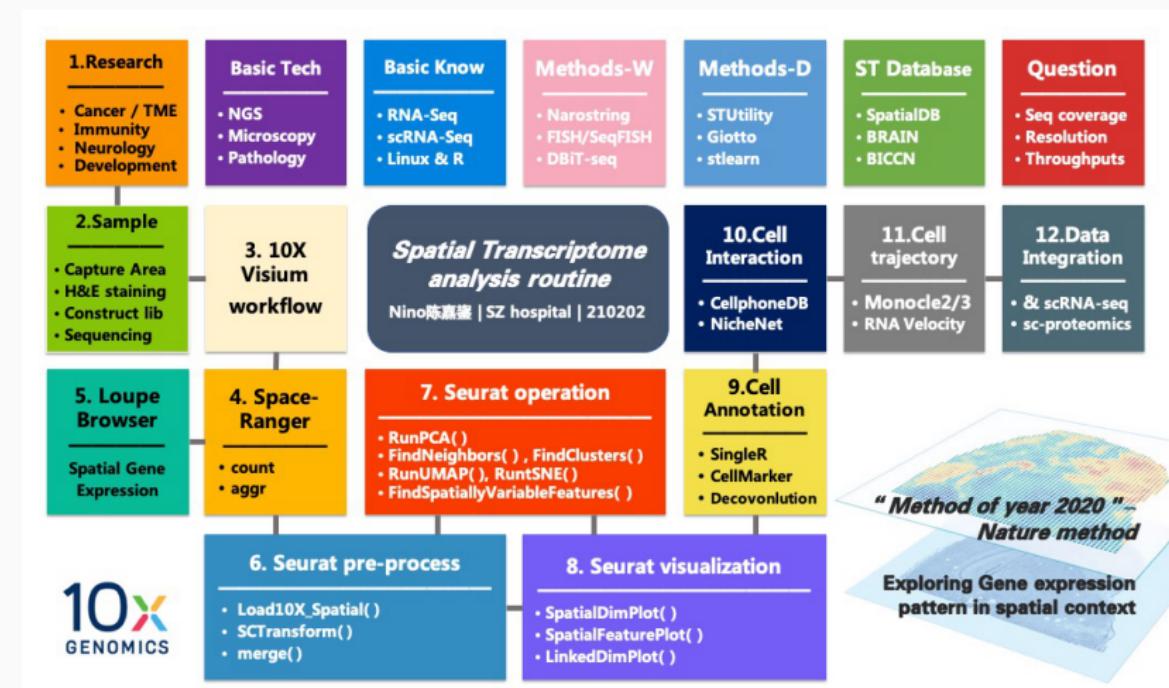
③ 空间转录组学

# ST | 分析 | 概览

高通量空间转录组时代的数据分析方法与工具



# ST | 分析 | 概览



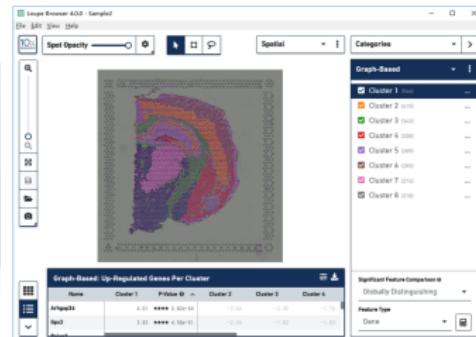
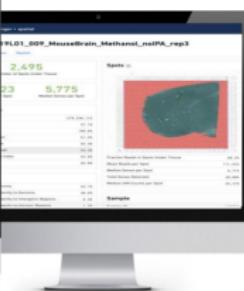
10X  
GENOMICS

## Space Ranger

利用 10x Genomics 官方分析流程及软件（Space Ranger & Loupe Browser），可将细胞分群信息与其空间定位相结合，深入探索不同空间定位下特定细胞亚群的生物学功能。



SpaceRanger



Loupe Browser

spaceranger count 主要的功能是spotXgene定量，将spatial barcode在切片组织中空间位置进行可视化，并基于表达矩阵进行一个粗略的降维和聚类。如果需要进一步分析官方推荐使用 R 包Seurat。

# ST | 分析 | Space Ranger

**Sample2**

Summary    Analysis

**Spots** ②

2,698 Number of Spots Under Tissue

115,740 Mean Reads per Spot

5,861 Median Genes per Spot

**Sequencing** ②

Number of Reads	312,269,984
Valid Barcodes	97.4%
Valid UMs	100.0%
Sequencing Saturation	62.9%
Q30 Bases in Barcode	97.2%
Q30 Bases in RNA Read	95.3%
Q30 Bases in UMI	97.1%

Fraction Reads in Spots Under Tissue 92.2%

Mean Reads per Spot 115,740

Median Genes per Spot 5,861

Total Genes Detected 21,118

Median UMI Counts per Spot 25,979

**outs**

- web\_summary.html Metrics for quality assessment
- metrics\_summary.csv Loupe Browser file for data visualization and analysis
- cloupe.cloupe Secondary analysis results (clustering, DGE, Moran's I)
- analysis/ spatial/spatial\_enrichment.csv Fiducial alignment, low and high res images and spot alignment results
- spatial molecule\_info.h5 Molecular level info used in additional pipelines (e.g., targeted-compare)
- BAM (optional) possorted\_genome\_bam.bam possorted\_genome\_bambai Read alignment files
- Filtered GEX Matrix MEX: filtered\_feature\_bc\_matrix/HD5F: filtered\_feature\_bc\_matrix.h5 Secondary analysis in R/Python
- Raw GEX Matrix MEX: raw\_feature\_bc\_matrix/HD5F: raw\_feature\_bc\_matrix.h5

**Cluster**

- 1 Lamp5
- 2 Igfbp6
- 3 Tbx16
- 4 Ephb4
- 5 Lmp4
- 6 Vip
- 7 Satb1
- 8 Egr1
- 9 Epop
- 10 Nr5

**Cluster 4 mRNA**

78 / 113

## Seurat

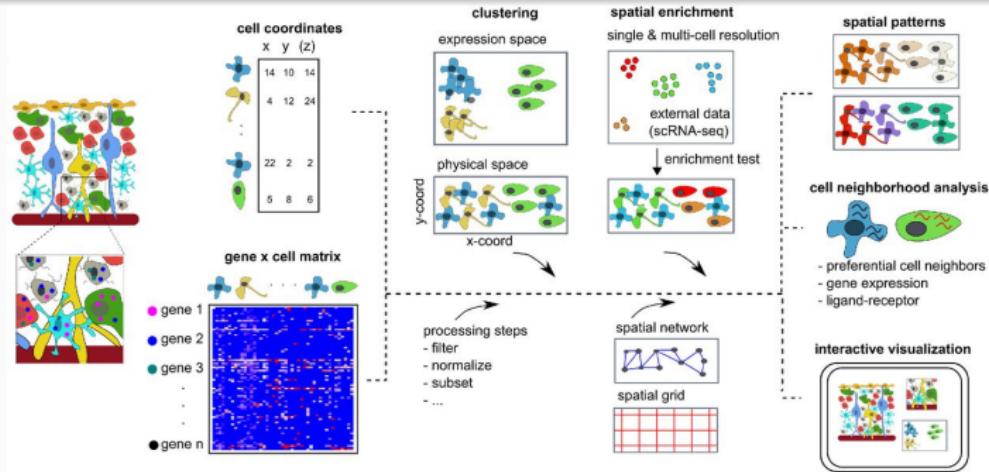
应用 Seurat 分析空间转录组的流程类似于 scRNA-seq 分析，但引入了更新的交互和可视化工具，特别强调了空间和分子信息的整合。

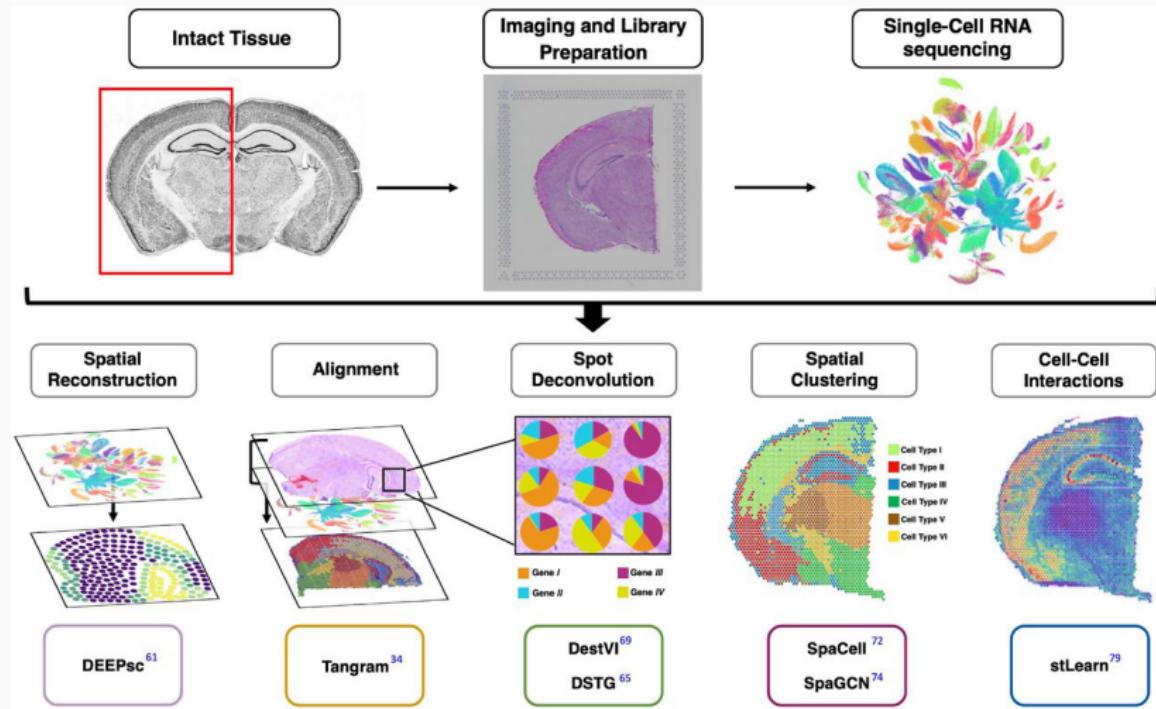
- 标准化、降维和聚类
- 检测空间高变基因；处理多个切片；交互式可视化
- 与单细胞 RNA-seq 数据整合

分析类型	空间转录组	单细胞转录组
聚类	不同的功能区域	细胞类型
降维	空间位置（代表了形态学特征）	TSNE UMAP
差异分析	针对部位的差异分析	针对细胞类型的差异分析
富集分析	功能“块”的功能差异	细胞类型的生物学功能
细胞类型	相互糅合在一起，强调统一体	批次分开，强调细胞类型的精确
肿瘤细胞区域	图片可以直接识别	算法推断
细胞类型分布	具有区域性	一个cluster或几个

## Giotto

Giotto 提供了一个全面的空间分析工具箱，包含两个独立但完全集成的模块。第一个模块（Giotto Analyzer）提供有关分析空间单细胞表达数据的不同步骤的分析说明，第二个模块（Giotto Viewer）在用户本地计算机上提供此类数据的响应式和交互式查看器。





# 章节概览

① 导言

② 单细胞转录组学

③ 空间转录组学

④ 单细胞多组学

■ 其他单细胞组学

■ 单细胞多组学

① 导言

② 单细胞转录组学

③ 空间转录组学

## ④ 单细胞多组学

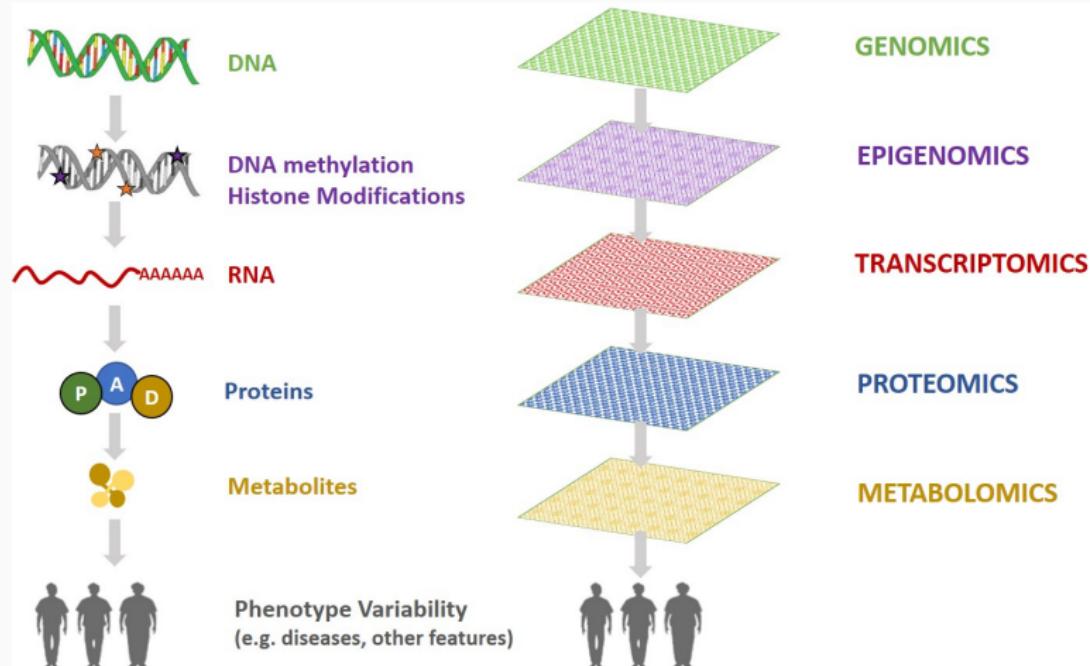
■ 其他单细胞组学

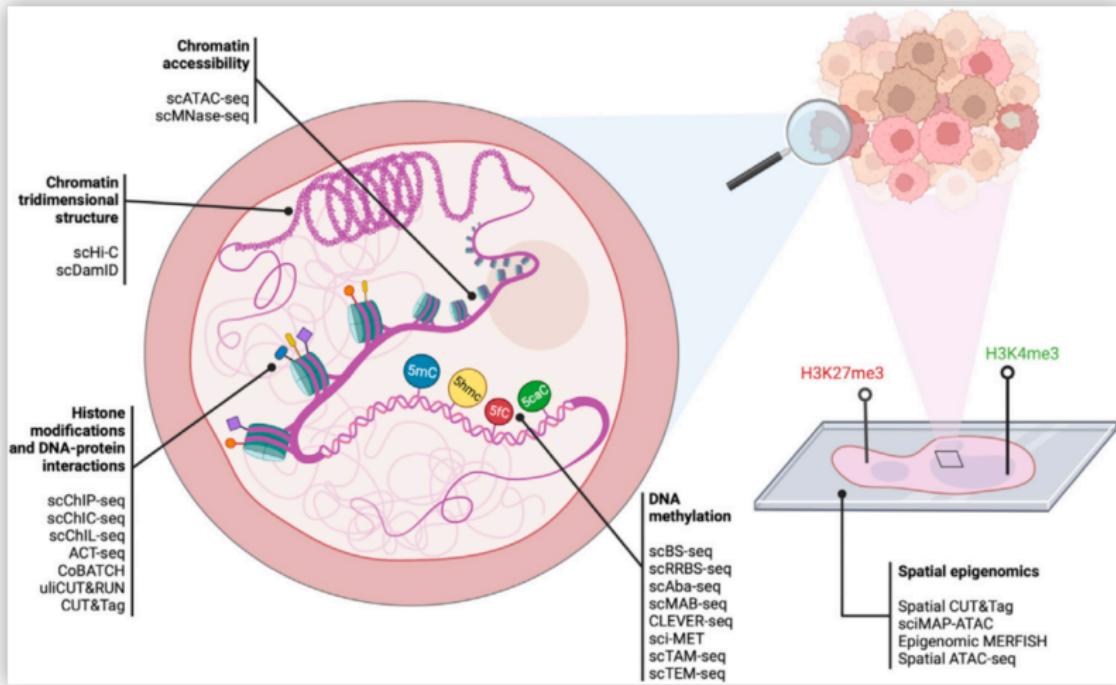
■ 单细胞多组学

## 组学 (Omics)

通常指生物学中对各类研究对象（一般为生物分子）的集合所进行的系统性研究，例如基因组学、蛋白质组学和代谢物组学等，而这些研究对象的集合被称为组。

方法	定义	常用技巧和技术
基因组学	<ul style="list-style-type: none"><li>重点关注生物体 DNA（基因组）中编码信息的结构、功能、演化、定位和编辑。</li></ul>	<ul style="list-style-type: none"><li>NGS: 全基因组测序, 外显子组测序, 靶向测序</li><li>芯片</li></ul>
表观遗传学	<ul style="list-style-type: none"><li>研究细胞如何通过 DNA 甲基化和组蛋白修饰等非遗传性修饰控制基因活性。</li></ul>	<ul style="list-style-type: none"><li>NGS: 甲基化测序, ChIP-Seq, ATAC-seq, HiC, 3C</li><li>芯片: 甲基化芯片</li></ul>
转录组学	<ul style="list-style-type: none"><li>研究转录组（由基因组产生的一整套 RNA 转录本）以及它们在调控过程、剪接、疾病或其他现象过程中的变化。</li></ul>	<ul style="list-style-type: none"><li>NGS: mRNA-seq, 全转录组以及靶向 RNA 测序</li></ul>
蛋白质组学	<ul style="list-style-type: none"><li>表征及识别响应特定刺激或随后的基因组或转录组变化的蛋白质表达模式。</li></ul>	<ul style="list-style-type: none"><li>质谱</li><li>质谱流式细胞技术</li><li>基于 NGS 的蛋白质检测（例如：CITE-seq、Olink、Ab-seq、BEN-Seq）</li></ul>

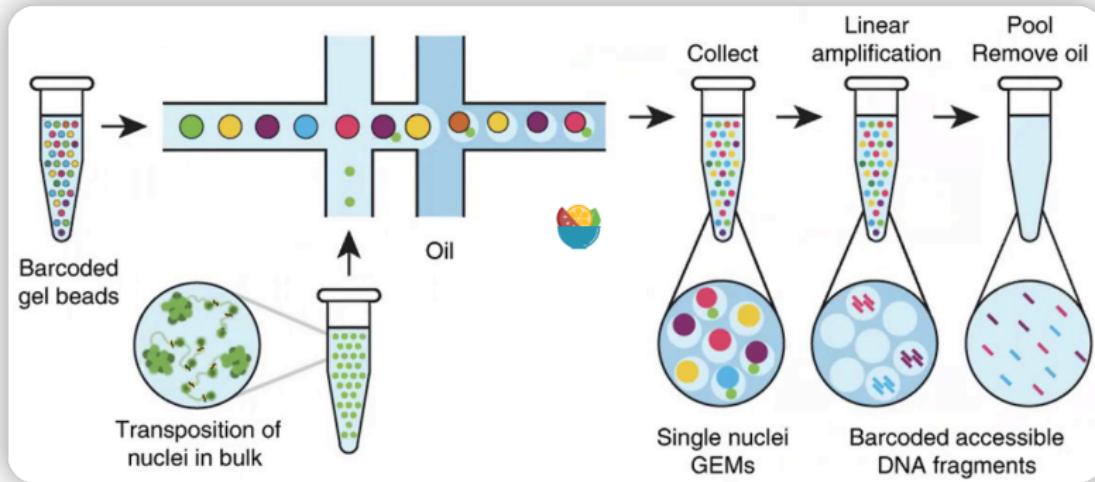




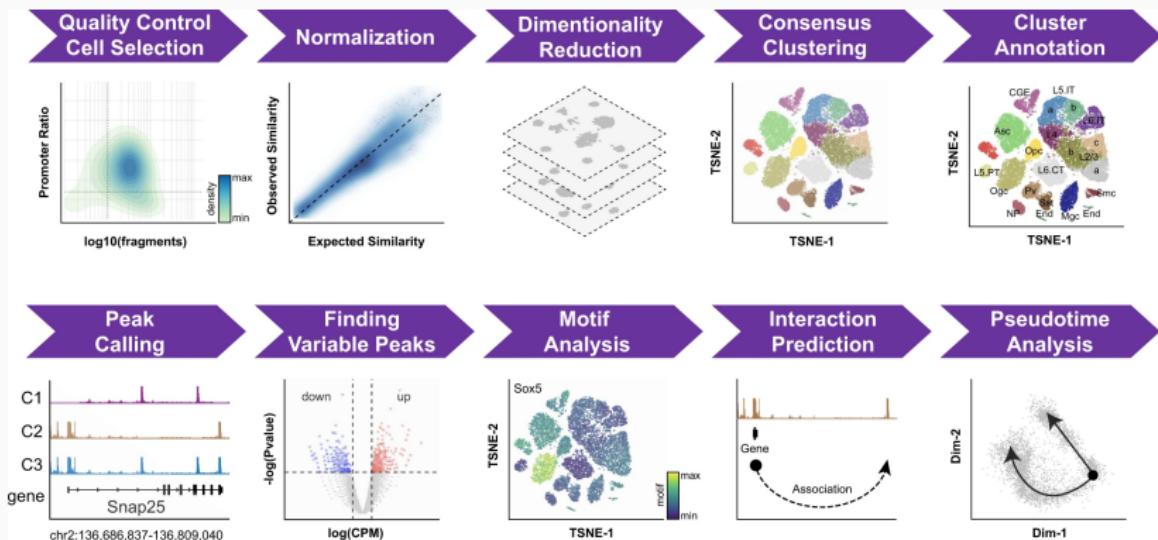
## ATAC-seq

ATAC-seq 是用于检测基因组染色质可及性的分子生物学手段。

ATAC-seq 分析可以用于研究许多染色质可及性特征。最常见的用途是核小体定位，也可用于定位转录因子结合位点。

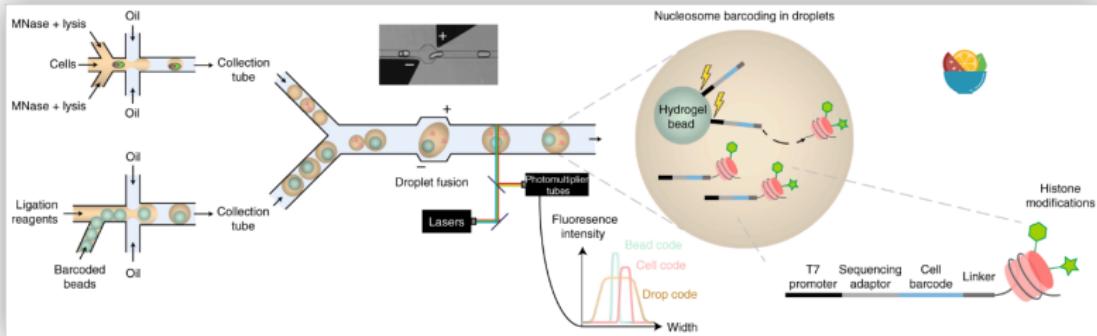


# 其他 | scATAC-seq



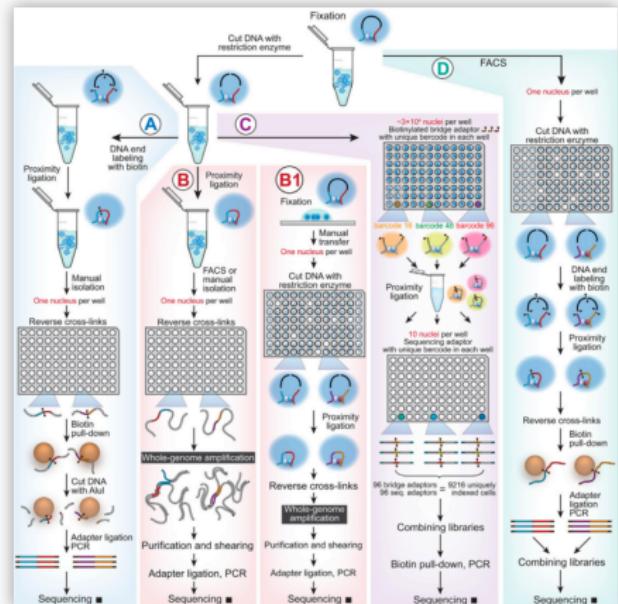
## ChIP-seq

ChIP-seq 被用于分析蛋白质与 DNA 的交互作用。该技术将染色质免疫沉淀 (ChIP) 与大规模并行 DNA 测序结合起来以鉴定与 DNA 相关蛋白的结合部位。可被用于精确绘制任意目的蛋白在全基因组上的结合位点，以确定转录因子和其他染色质相关蛋白如何影响表型影响机制。



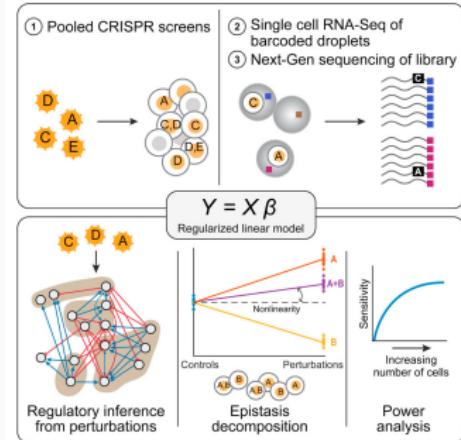
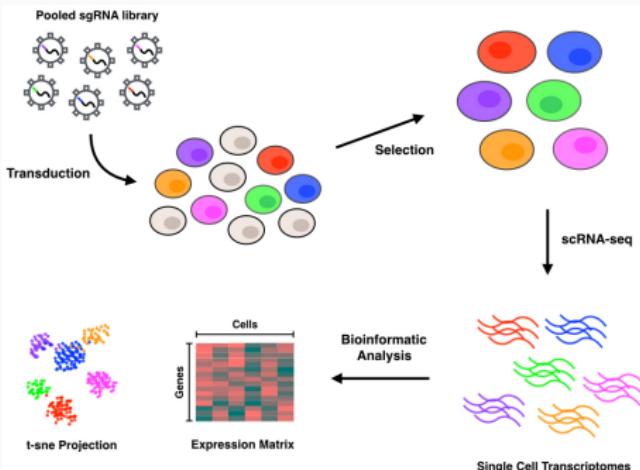
## Hi-C

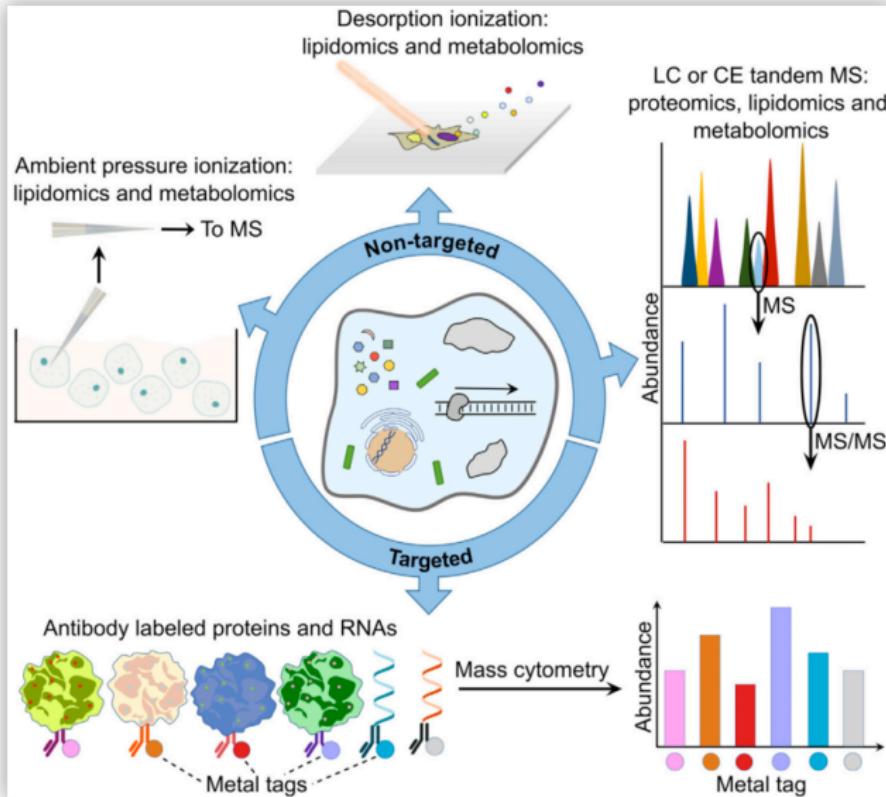
Hi-C 技术源于染色体构象捕获(3C)技术，利用高通量测序技术，结合生物信息分析方法，研究全基因组范围内整个染色质DNA在空间位置上的关系，获得高分辨率的染色质三维结构信息。最初用于捕获全基因组范围内所有的染色质内和染色质间的空间互作信息，目前已应用于基因表达的空间调控机制研究、构建染色体水平参考基因组、构建单体型图谱等。

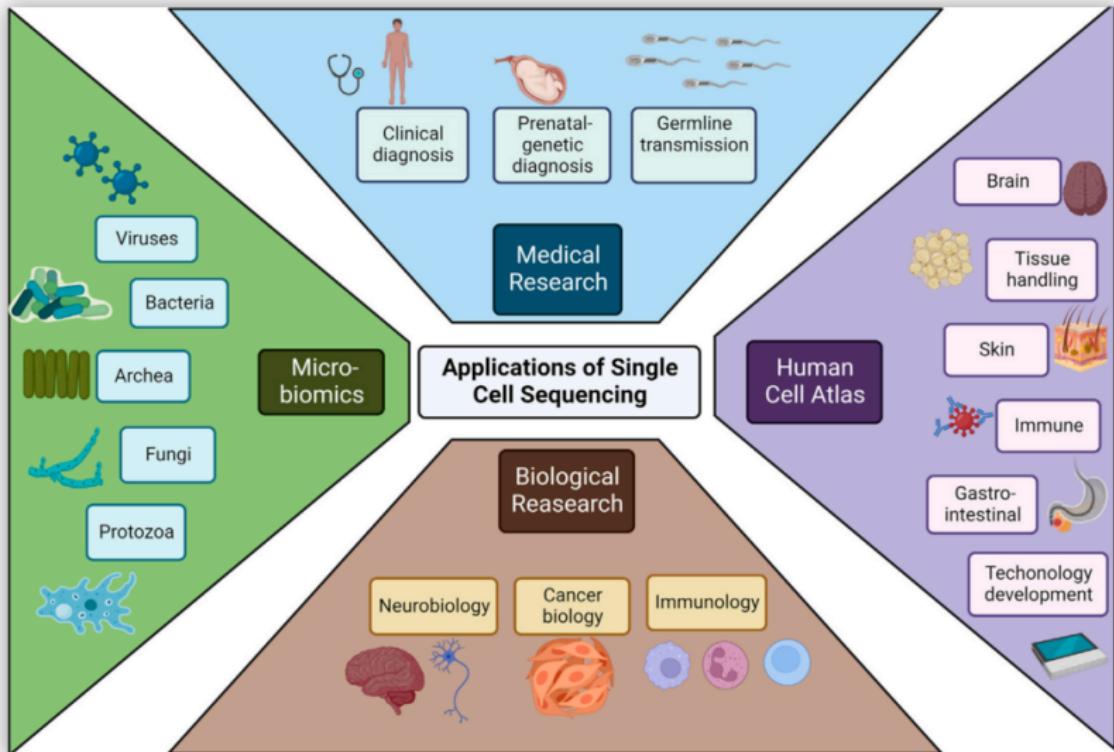


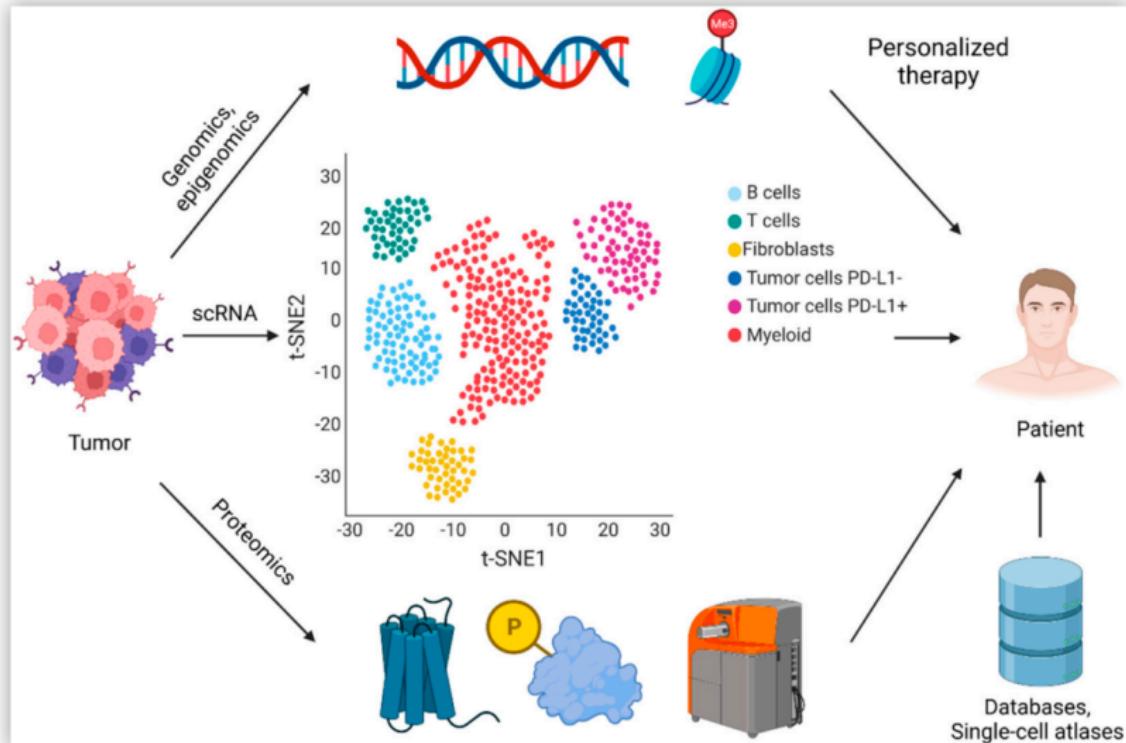
## Perturb-seq

利用 CRISPR-Cas9 技术将基因变化引入细胞内，然后使用单细胞转录组测序捕获特定基因变化导致的转录组信息变化，能够研究给定细胞类型的全面遗传扰动影响，可以以前所未有的深度跟踪打开或关闭基因的影响。









① 导言

② 单细胞转录组学

③ 空间转录组学

## ④ 单细胞多组学

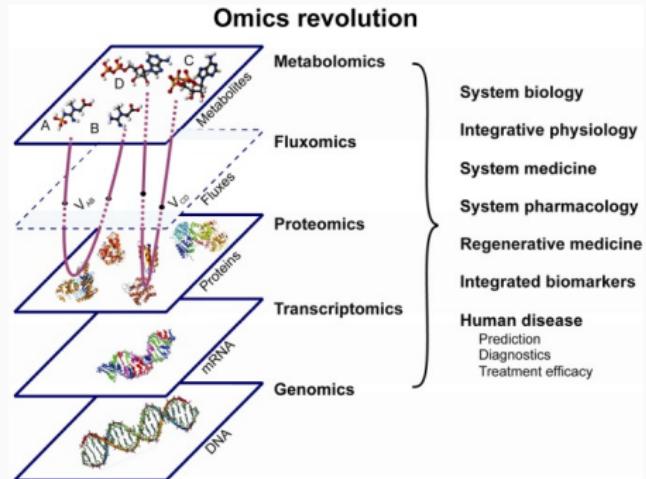
■ 其他单细胞组学

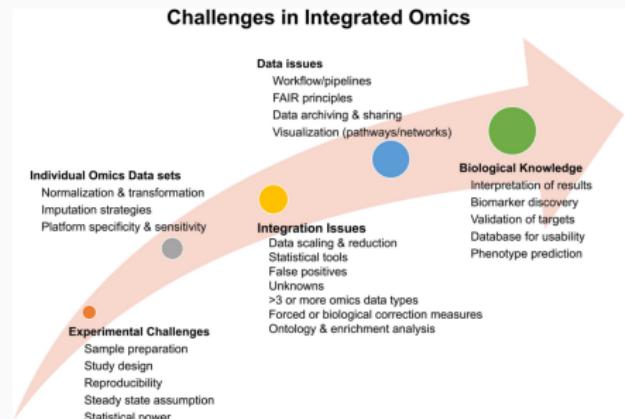
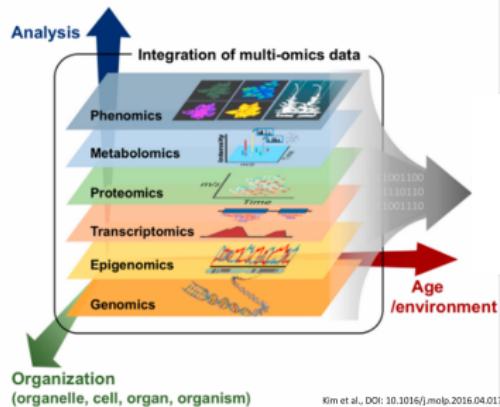
■ 单细胞多组学

## 多组学（Multiple omics）

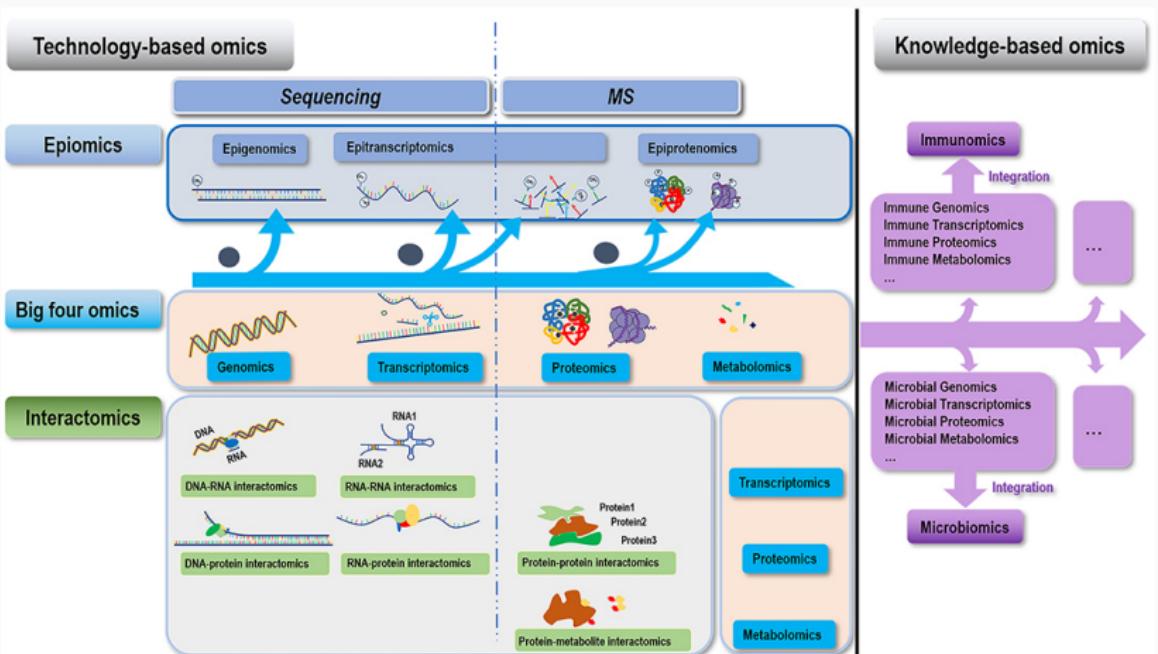
多组学提供了一个综合视角，这种分析方法将基因组数据与转录组学、表观遗传学、蛋白质组学其他模式的数据相结合，可测量基因表达、基因激活和蛋白质水平。

多组学分析研究能够帮助我们更全面地了解分子变化对正常发育、细胞响应和疾病的影响。研究人员可以借助多组学技术更好地将基因型与表型联系起来，不断探索、发现新药物靶点和生物标志物。



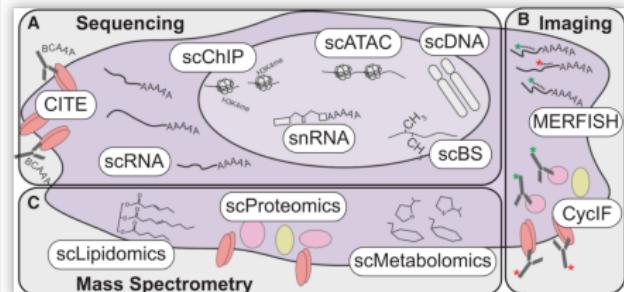


# 多组学 | 简介



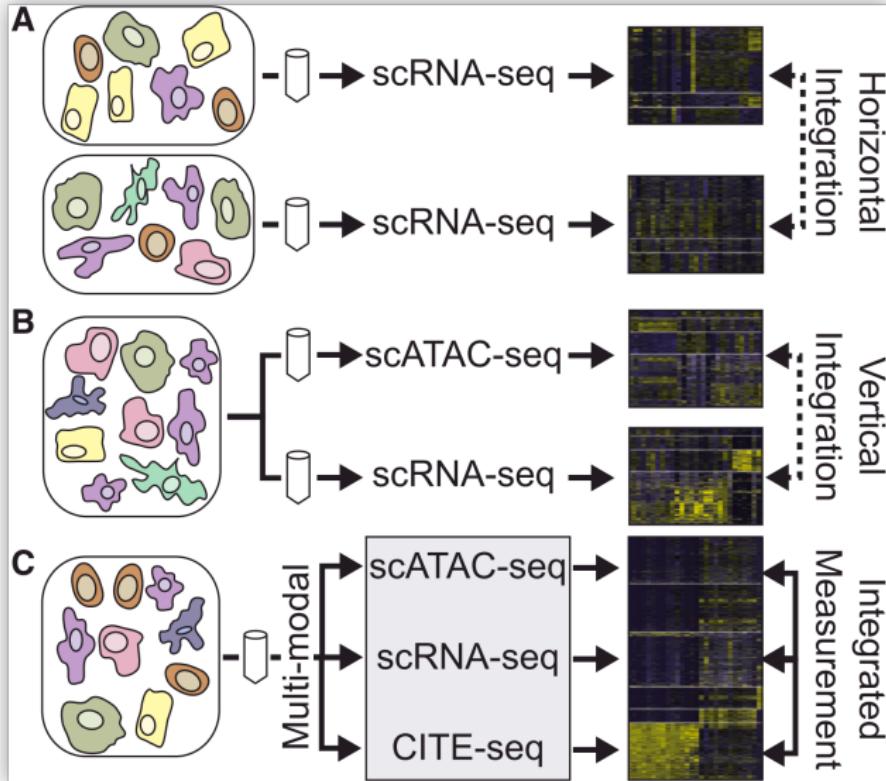
## 单细胞多组学

单细胞多组学技术通过测量来自同一单个细胞的多种分子，包括DNA、RNA、染色质和蛋白质分子等，整合来自多个组学水平的信息，可以为同一细胞构建多组学概况。



## 优势

- 单细胞多组学技术是单细胞分析技术发展的前沿，也是单组学技术发展的必然趋势。
- 与细胞的单组学相比，多组学更侧重对细胞完整信息的采集以及对时间空间的关注，因此可以更好和更全面地反映细胞特征，可提供更准确的生物学见解。



# 单细胞多组学 | 多模态



## 单细胞转录组和基因组联合分析

- 基因组是决定细胞内遗传变异分子机制的源代码，而转录组决定了细胞的特定功能。
- 基因组和转录组的联合分析通过探究 DNA 拷贝数、DNA 编码区与非编码区如何决定转录组表达水平，从而建立基因组和转录组的相关性，并阐述选择性表达分子作用机制，如 RNA 编辑、调控变异和等位基因特异性表达。
- 对不同细胞的基因组和转录组进行联合分析，可进一步**区分遗传亚克隆**，同时进行谱系示踪。

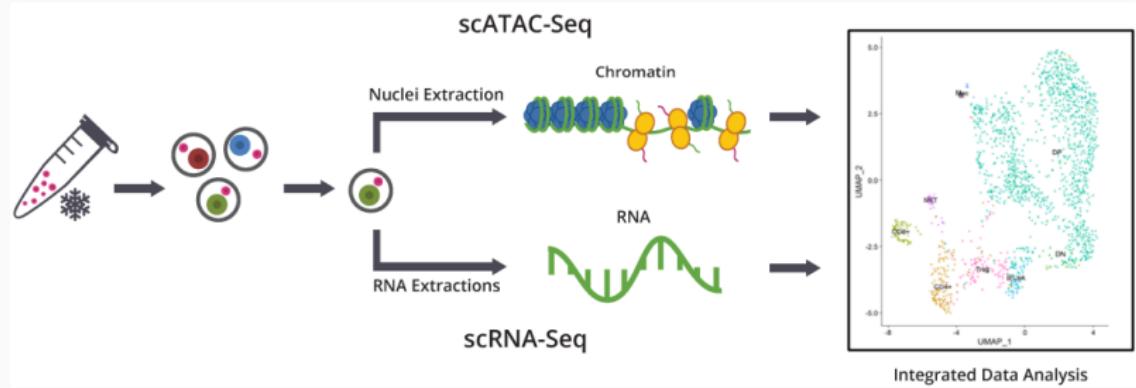
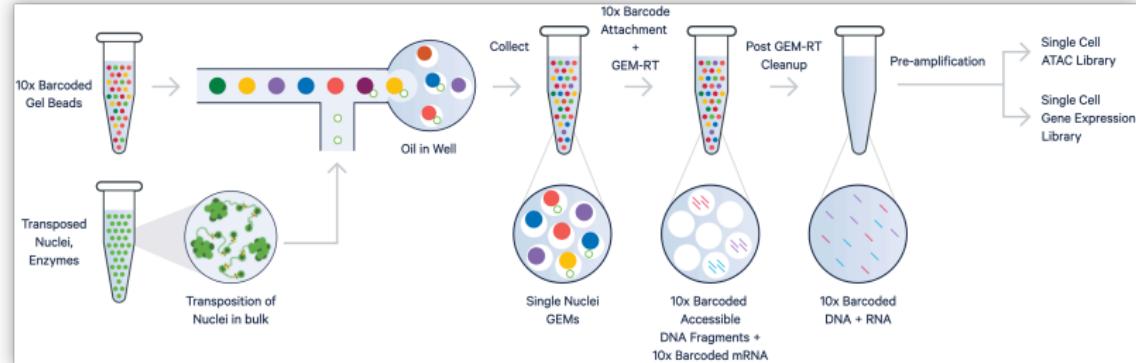
## 单细胞转录组和蛋白质组联合分析

- 蛋白质作为表型的直接输出者，直接反映细胞当下的生理状态和细胞功能。
- 转录组和蛋白质组的联合测量有助于解释转录异质性如何转化为功能表型多样性，深入理解转录/翻译后修饰过程。
- 有助于揭示 RNA 层面无法区分的表型差异，并探讨特定细胞亚型的基因表达网络和功能调控类型。

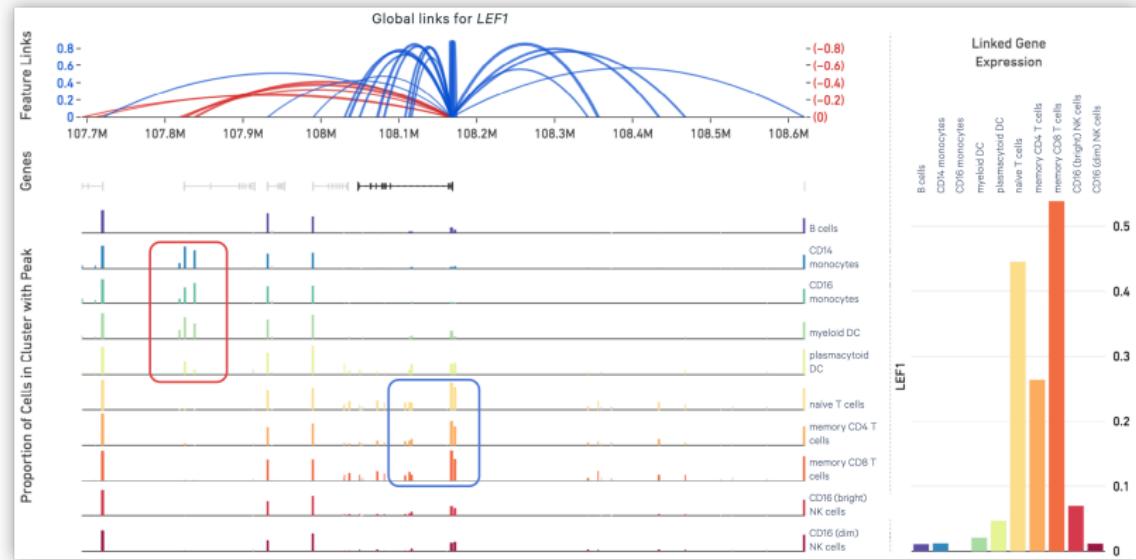
## 单细胞转录组和表观遗传组联合分析

- 细胞内表观遗传标记变化广泛，如 DNA 甲基化修饰、组蛋白修饰、转录因子结合、染色质可及性等，其通过调节染色质结构与构象促进/抑制转录组表达。
- 对转录组和表观基因组进行联合分析可以解析染色质结构与构象变化对细胞基因表达的可塑性，并构建细胞表观遗传调控网络，揭示基因表达调控机制。二者的同时分析可帮助鉴别细胞发育和疾病发展过程中的特殊标记并实现遗传谱系示踪。
- 染色质可及性和转录组的联合检测为揭示单个细胞的基因调控机制提供了强有力的工具。
- 单细胞 DNA 甲基化组和转录组的联合分析可鉴别突变产生的遗传谱系，全面揭示细胞间的异质性，同时探讨 DNA 甲基化修饰对基因表达的调控规律。

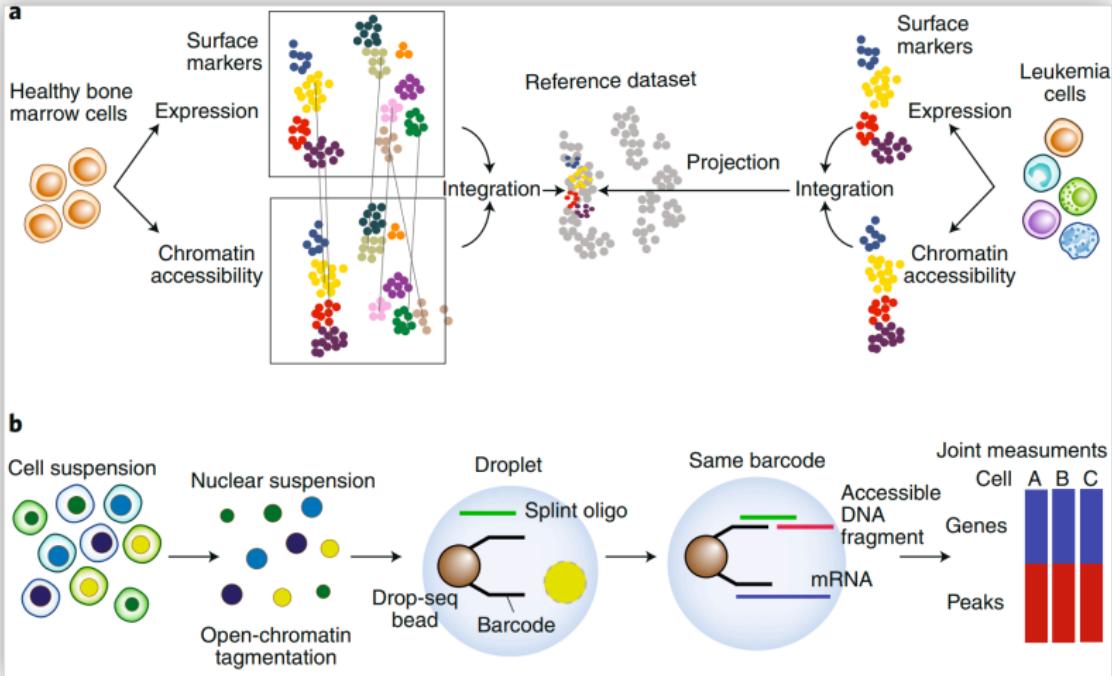
# 单细胞多组学 | 双组学



# 单细胞多组学 | 双组学



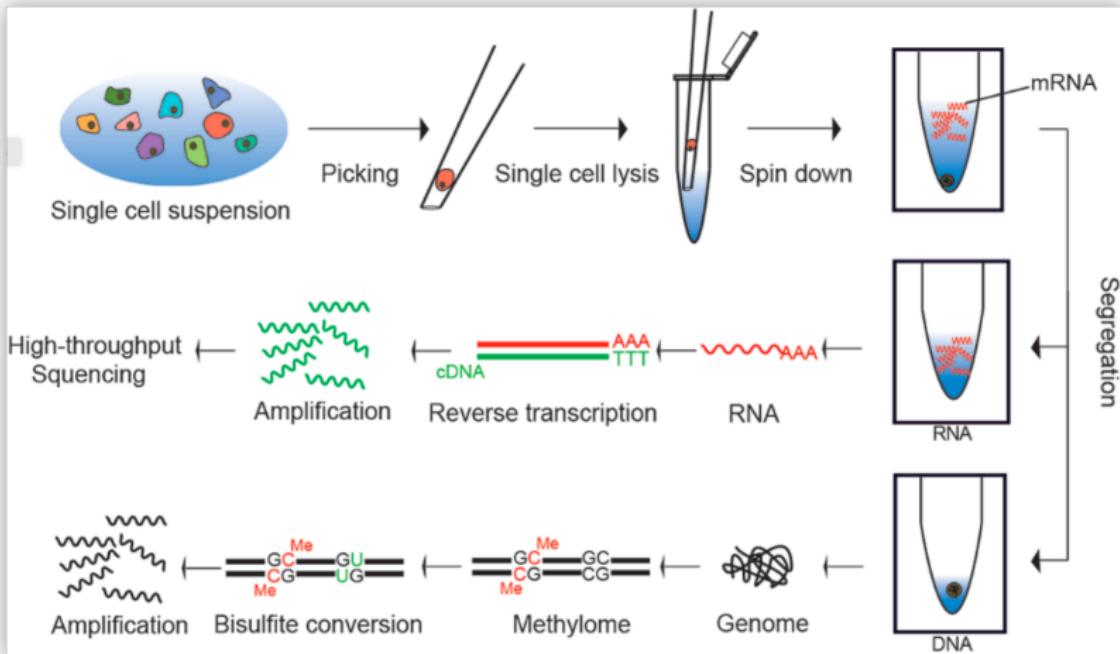
# 单细胞多组学 | 双组学 | 两种策略



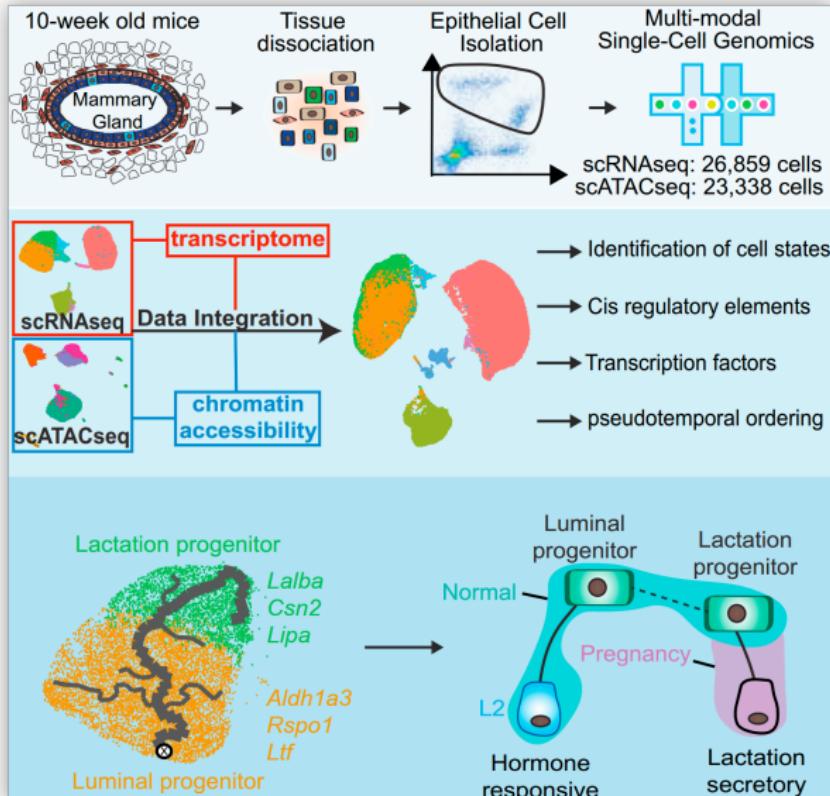
## scTrio-seq

- 首次同时分析了单个哺乳动物细胞中基因组、DNA 甲基化和转录组之间的关系，可以获取单个细胞的拷贝数变异、DNA 甲基化和转录组信息。
- 以 10 Mb 分辨率准确推导 CNV 模式，获得了 150 万个 CpG 位点的甲基化模式，并检测了单个哺乳动物细胞中平均 6179 个基因的表达水平。
- 分析了来自人类肝细胞癌组织样品的 25 个单细胞，发现 2 个细胞亚群的 DNA 拷贝数、DNA 甲基化或 RNA 表达水平不同；通过比较 2 个 HCC 亚群之间的多基因组差异，发现某一亚群具有更多的拷贝数变异，表达了更多的侵袭性细胞标志物，且更有可能逃避免疫监视。

# 单细胞多组学 | 三组学



# 单细胞多组学 | 应用 | 乳腺上皮细胞



## Single-cell transcription and chromatin landscapes of human cortical development

