关于单细胞测序技术的介绍

汪翔

2024/12/2

本文主要针对单细胞测序（Single-cell Sequencing, SCS）技术的背景知识，主要介绍不同组学的 SCS 技术，同时介绍单细胞单组学，单细胞多组学，空间转录组学及空间多组学四类研究对象间的区别及联系。针对不同组学的研究对象，以 RNA 转录组数据为例，介绍针对研究对象的不同的数据分析任务，主要介绍分析任务的内容及其难点[1][2]

1 单细胞单组学

单细胞测序技术允许研究者在单个细胞层面理解细胞间异质性，同时帮助研究者从不同层面深入理解细胞功能及基因调控关系。随着分子生物学、微流体技术和纳米技术的最新进展催生了大量 SCS 技术，极大的促进了细胞各个组学研究的发展。细胞生物组学根据研究对象的不同主要分为基因组学、转录组学、蛋白质组学和表观基因组学四大类，现有的 SCS 技术通常关注于细胞单个模态（组学）数据的测量，如 DNA 序列，RNA 表达及染色质可及性等。

[1]本文中关于测序技术的参考文献可参考 https://doi.org/10.1038/s41576-019-0093-7.

[2]本文及对应 slides 均以归档至 https://github.com/hello102/notes.

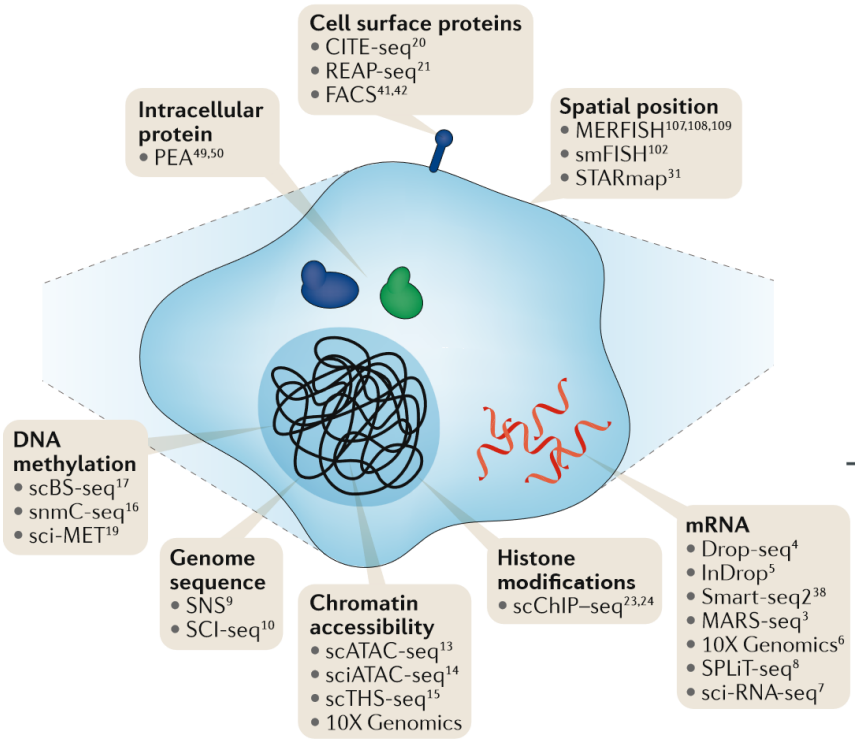

图 1: SCS 不同模态数据展示图。

如图1所示，针对单细胞不同组学数据的测量，已有多种不同的测序方案被提出：

- 单细胞基因组学。单细胞基因组学主要研究单个细胞的 DNA 序列和基因组特性，测序技术包括单核测序（Single-nucleus Sequencing, SNS）及单细胞组合条形码索引测序（Single-cell Combinatorial Indexed Sequencing, SCI-seq）等技术。

- 单细胞转录组学。单细胞转录组学主要研究细胞中的 RNA 表达，主要通过 scRNA-seq 技术进行测量，根据测序技术的不同，可分为 Drop-seq，InDrop 等测序技术。

- 单细胞表观组学。单细胞表观组学研究单细胞中由 DNA 序列变化以外的其他机制引起的基因活动的变化，包括 DNA 甲基化，染色质可及性及组蛋白修饰等。其中，单细胞转座酶可及染色质测序（single-cell Assay for Transposase-accessible Chromatin with Sequencing, scATAC-seq）技术可用于测量单细胞中染色质可及性，单细胞亚硫酸氢盐测序（Single-cell Bisulfite Sequencing, scBS-seq）用于测量单细胞 DNA 甲基化水平，单细胞染色质免疫沉淀测序

(Single-cell Chromatin Immunoprecipitation Followed by Sequencing, scChIP-seq) 能够在单细胞水平测量组蛋白修饰。

- 单细胞蛋白质组学。单细胞蛋白质组学主要评估单细胞中的蛋白质表达。单细胞转录组与表面蛋白质组索引测序 (Cellular Indexing of Transcriptomes and Epitopes by Sequencing, CITE-seq) 能够测量单细胞的表面蛋白表达，此外，CITE-seq 还可以同时测量 mRNA 的表达。

其中，不同的组学数据具有不同的数据特点。例如 scRNA-seq 数据通常可以使用负二项 (Negative Binomial, NB) 分布或零膨胀负二项 (Zero-inflated Negative Binomial, ZINB) 分布进行建模 (详见之前的介绍文章, `https://github.com/hello102/notes`)，而 scATAC-seq 数据主要测量染色质可及性，每个区域只有开放和不开放两种状态，通常服从 Bernoulli 分布。

scRNA-seq 是最常用的 SCS 技术之一，通过 scRNA-seq 测序技术，可以得到组织的细胞-基因表达矩阵 $\mathbf{X} = (x_{ij})_{m\times n} \in \mathbb{R}^{m\times n}$, m 表示细胞数, n 表示基因数。scRNA-seq 数据与传统的 bulk 测序数据的主要不同是 scRNA-seq 数据测量的是单个细胞分辨率的基因表达量，而 bulk 测序测量的是组织细胞的平均基因表达情况 (如图??所示)。

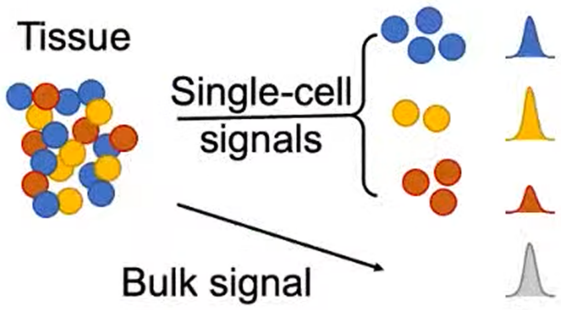

图 2: scRNA-seq 测序与 bulk 测序对比图。scRNA-seq 测序测量单个细胞的基因表达，而 bulk 测序测量平均基因表达。不同颜色表示不同的细胞类型。

基于 scRNA-seq 测序技术的特点，scRNA-seq 数据相比 bulk 数据具有高维度，高稀疏率以及强噪声的特点，这也为 scRNA-seq 数据分析任务带来了挑战。scRNA-seq 数据分析任务主要包括聚类任务，填补任务，整合任务及注释任务，主要内容包括:

- 聚类任务。通过对 scRNA-seq 数据的分析，对 scRNA-seq 数据中的细胞进行聚类分析，帮助划分具有相似基因表达的细胞群。

  分析难点：数据高纬度，高噪声。“维数灾难”影响分析结果。

- 填补任务。针对 scRNA-seq 数据高缺失率的特点，对 scRNA-seq 数据进行数据填补，根据模型恢复出缺失数据，提高 scRNA-seq 数据的质量。

  分析难点：区分 dropout 效应，即测序缺失的表达量和未表达数据的区分。

- 整合任务。对于同一组织的多个测序样本，可以通过整合任务进行统一分析，此时数据通常为 $\mathcal{X} = \{\mathbf{X}_i\}_{i=1}^{L}, \mathbf{X}_i \in \mathbb{R}^{m_i \times n}, m_i$ 表示第 i 个数据中的细胞数，通过对多个数据的联合分析，可以提高数据分析的效果。

  分析难点：克服批次效应，即不同测序数据之间由于不同样本，不同测序平台，不同测序时间等因素产生的系统误差。

- 注释任务。通过 marker 基因等信息，对不同的细胞进行细胞类型注释。

  分析难点：scRNA-seq 数据细胞数多，人工标注费时费力，难点在于如何设计准确的上游聚类算法降低注释人工成本。

通过对 scRNA-seq 数据的数据分析，可以发现一些新的生物学现象。如图3所示，通过 scRNA-seq 数据分析能够帮助发现新的罕见细胞类型，帮助构造细胞图谱以及帮助理解免疫机制。

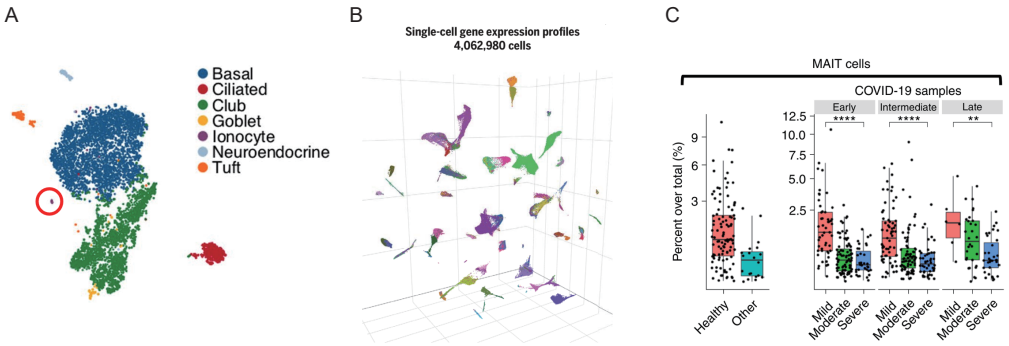

图 3: scRNA-seq 数据分析优势展示图。（A）scRNA-seq 数据分析能够发现罕见细胞类型。（B）scRNA-seq 数据分析能够帮助构建细胞图谱。（C）scRNA-seq 数据分析能够理解免疫机制。

2 单细胞多组学

单细胞单组学的分析揭示了新的生物学现象，但是针对单个组学数据的分析仍然一定的局限性，因此研究者考虑联合分析多个组学的数据，这也被称为单细胞多组学（多模态）分析。单细胞多组学分析可分为配对多组学和非配对多组学。配

对多组学即通过某些测序技术同时测得同一细胞的多个组学数据，例如 CITE-seq 能够同时测得 RNA 数据和表面蛋白数据，这两种数据天然的具有细胞维度的匹配性，但与单组学数据相比，配对多组学数据通常具有更低的测序通量和更高的稀疏率。非配对多组学即直接对现有的单组学数据进行联合分析，从而得到更深入的结果，但是不匹配的数据分布为数据分析带来了挑战。

单细胞多组学分析主要挑战就是不同模态数据间的整合与对齐问题，不同组学的整合问题具体如图4所示。

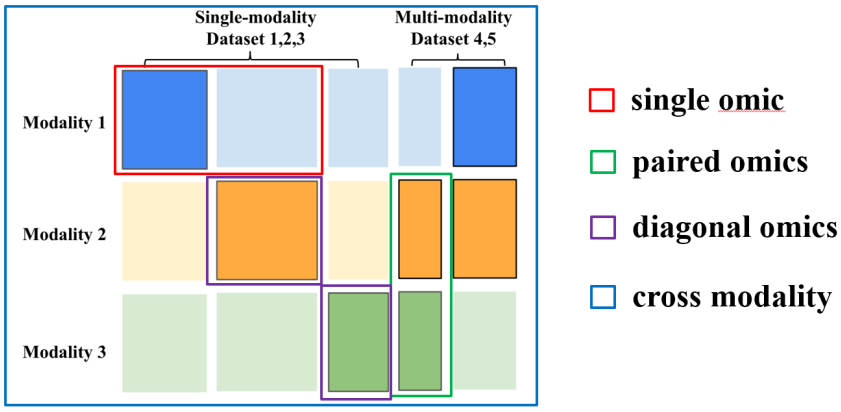

图 4: 不同模态数据整合分析任务概要图。

整合分析问题根据数据不同主要可分为单组学整合，配对多组学整合及非配对多组学整合三类，其中非配对多组学整合问题又可分为对角整合和交叉整合两种，具体问题如下：

- 单组学整合任务。同上节整合任务。
- 配对多组学整合。整合匹配多组学数据，此时数据细胞维度相同，但特征空间不同。

  分析难点：数据质量较低，且不同组学数据特征空间不同。
- 对角整合任务。整合两种（或多种）模态的单组学数据，此时数据细胞维度不匹配，同时特征空间也不匹配，

  分析难点：不同特征空间对齐，不同数据分布对齐。
- 交叉整合任务。整合单组学和配对多组学数据，此时数据细胞维度不匹配，同时特征空间也不匹配，

  分析难点：不同特征空间对齐，不同数据分布对齐，且数据分布更加混乱。

3 空间转录组学

在某些研究情况下，研究人员会对组织中某个区域的功能作用产生研究兴趣，这时需要利用空间转录组（Spatially Resolved Transcriptomes, SRT）数据来进行分析。SRT 数据相对于 scRNA-seq 数据，能够保留测序样本中细胞的空间位置，能够研究特定的功能区域的作用。通过 SRT 数据测序技术，通常可得到基因表达矩阵 $\mathbf{X} = (x_{ij})_{m\times n} \in \mathbb{R}^{m\times n}$ 和空间位置矩阵 $\mathbf{S} = (s_{ij})_{m\times 2} \in \mathbb{R}^{m\times 2}$，$m$ 表示细胞数，n 表示基因数。

现有的 SRT 数据测序技术主要分为基于荧光染色技术和基于测序技术两类。基于荧光染色的 SRT 技术，如 MERFISH, smFISH 等技术，通常能够获得单细胞分辨率的数据，但此类方法测得的细胞通量少，且测序深度低。基于测序的 SRT 技术通常具有高细胞通量和较深的测序深度，主要技术包括 10X Visium, Slide-seqV2, Stereo-seq 等。其中，10X Visium 数据分辨率较低，每个 spot 中通常会测到多个细胞。此外，尽管 Stereo-seq 等技术理论上能获得单细胞分辨率的表达量，但实际仍然无法保证每个 spot 测得单个细胞表达。

针对 SRT 数据的特点，SRT 数据分析主要包括聚类分析，整合分析，反卷积分析和高变基因（Highly Variable Gene, HVG）检测四个分析任务，具体任务如下：

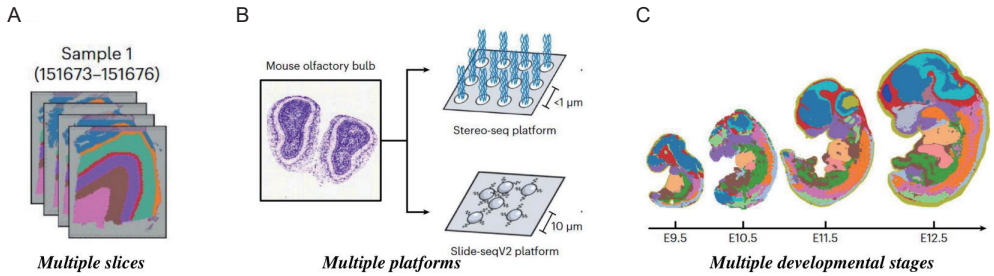

图 5: SRT 数据整合任务。（A）多切片整合。（B）跨平台整合。（C）跨发育阶段整合。

- 聚类任务。与 scRNA-seq 数据聚类分析类似，此时反应为空间区域识别。

  分析难点：如何结合表达谱信息和空间位置信息。

- 整合任务。与 scRNA-seq 数据整合分析类似。根据 SRT 测序数据的特点，整合任务包括多切片整合，跨平台数据整合，跨发育阶段整合等多种整合方式（如图5所示）。

  分析难点：如何结合表达谱信息和空间位置信息。

- 反卷积任务。对于低分辨率的 SRT 数据，因其单个 spot 中包含多个细胞，故需通过反卷积任务，根据单细胞 reference 数据集解析出每个 spot 中的细

胞类型比例。

分析难点：单细胞 reference 数据集与 SRT 数据的特征对齐问题。

- HVG 识别任务。识别特定区域内的高变异基因。

  分析难点：基因表达可能在空间区域层面和细胞类型层面体现出不同的表达差异。

4 空间多组学

与单细胞多组学分析类型，已有空间转录组测序技术能够同时测量同一细胞内的多个组学数据，如图6所示。空间多组学分析与单细胞多组学分析类似，其难点主要在于如何结合空间位置信息提高下游分析的效果。

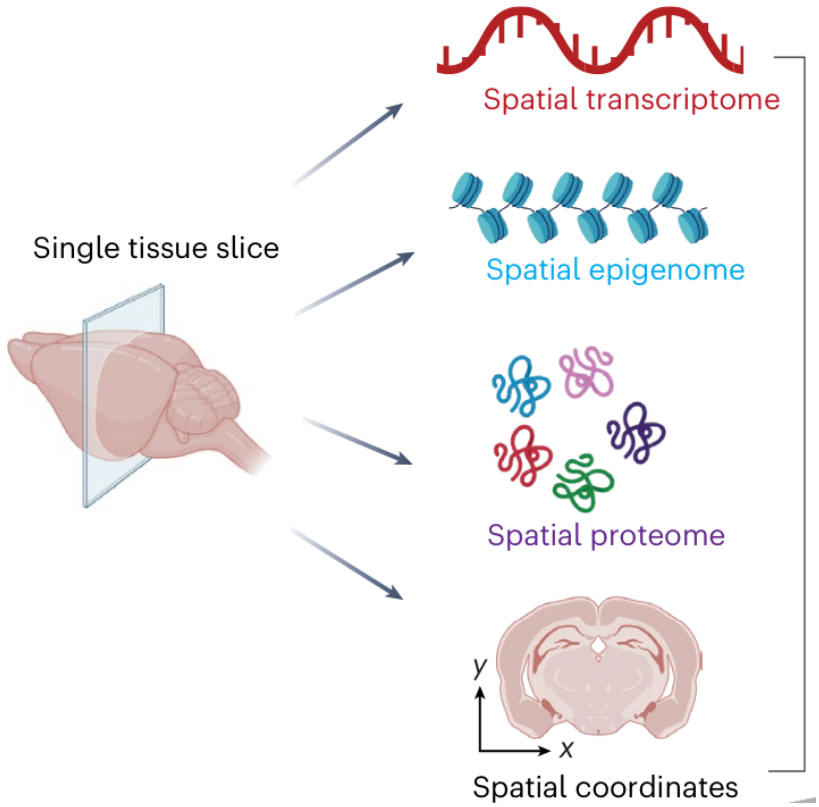

图 6: SRT 多组学数据示意图。