## 单细胞数据分析-数据质控

图表, 气泡图

AI 生成的内容可能不正确。

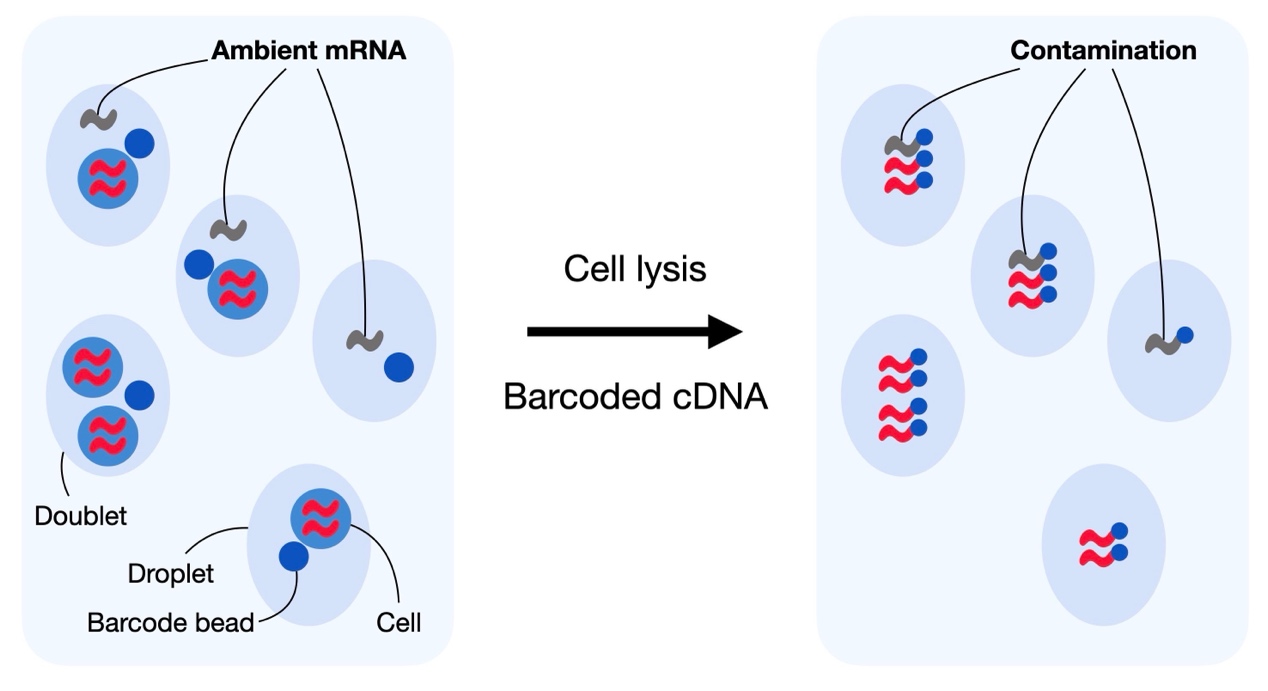
1. **Filtering low quality cells去除低质量细胞**

每个条形码的 UMI 计数深度 (count depth)

每个条形码检测到的基因数 (gene count per barcode)

每个条形码中线粒体基因 UMI 计数的比例 (fraction of mitochondrial counts per barcode)

1. **Correction of ambient RNA环境RNA**



2-1细胞破裂或凋亡：

在单细胞分离过程中，一些细胞可能受损、破裂或处于凋亡状态，导致其RNA内容物释放到周围的溶液中

溶液中的游离RNA：

这些RNA可能悬浮在细胞悬液中，并在液滴形成（如基于液滴的scRNA-seq方法，例如10x Genomics）时被意外封装到其他细胞的液滴中

2-2样本处理中的污染：

例如，取样或实验操作过程中引入的外源性RNA

2-3生信分析工具：

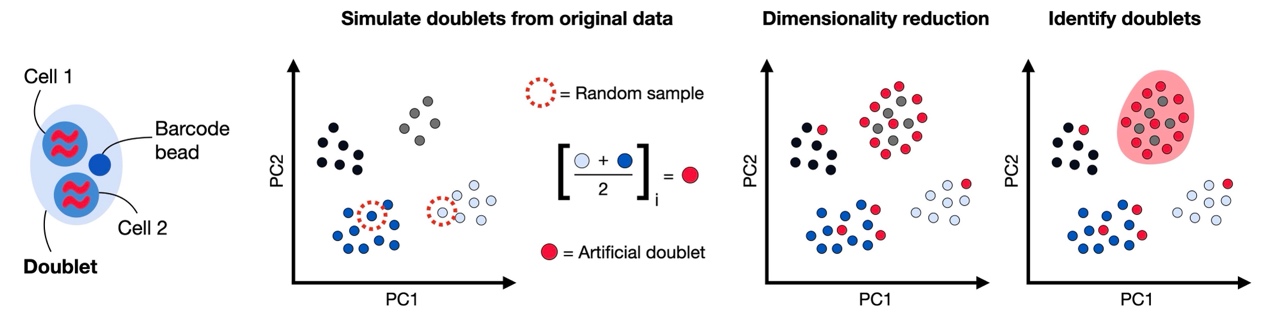
CellBender, DecontX, SoupX

CellBender provides the most precise estimates of background noise levels and also yields the highest improvement for marker gene detection.

Janssen P, Kliesmete Z, Vieth B, et al. The effect of background noise and its removal on the analysis of single-cell expression data[J]. Genome biology, 2023, 24(1): 140.

<https://kb.10xgenomics.com/hc/en-us/articles/360060415811-Can-I-remove-ambient-RNA-contamination-from-cells-in-my-gene-expression-data>

1. **Doublet Detection两个或多个细胞的 mRNA 被误认为是单个细胞的表达数据**



相同类型的细胞 (但来自不同个体) 组成，则称为同源双胞 (homotypic doublet)

由不同类型或不同状态的细胞组成，则称为异源双胞 (heterotypic doublets)

## 单细胞数据分析：从fastq到Count Matrix

‘

1. 关于参考基因组

* **two popular sources of assembly files:** UCSC (their assemblies are named hg19, hg38, mm10, etc), and GRC (GRCh37, GRCh38, GRCm38)这两个版本在染色体水平的序列区别不大，differ in additional contigs and so-called ALT loci,

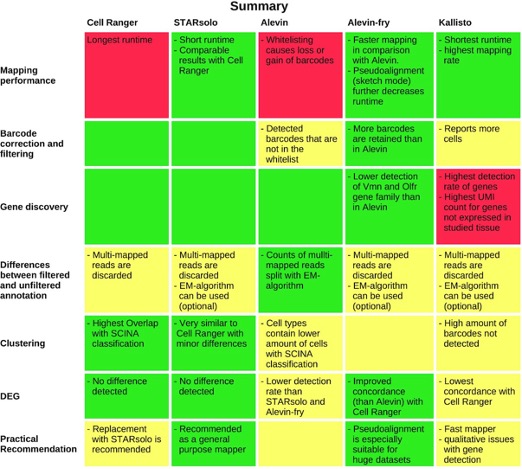
1. 关于注释文件

* human and mouse genome annotation are RefSeq, ENSEMBL, and GENCODE.
* Current ENSEMBL/GENCODE annotation of the human genome contains approximately 60k genes, 20k of which are protein coding, and 237k transcripts.
* 简单的来讲大多数基因可以分为：protein coding genes, long noncoding RNAs, short noncoding RNAs, and pseudogenes.
* 在cell ranger中保留了 protein coding, long noncoding RNA, antisense, and all biotypes belonging to BCR/TCR (i.e. V/D/J) genes (note that older Cell Ranger reference versions do not include the latter). All pseudogenes and small noncoding RNAs are removed.

1. 典型的单细胞数据预处理分析流程包含4个步骤：

* Mapping the cDNA fragments to a reference.
* Assigning reads to genes.
* Assigning reads to cells (cell barcode demultiplexing)
* Counting the number of unique RNA molecules (UMI deduplication)

1. 几款单细胞数据分析比对软件的比较



* STARsolo and Alevin-full\_decoy offer great computational speed-up and correct processing of multimappers, which reduces the quantification bias while retaining very high compatibility with Cell Ranger.
* 如果没生信经验就用cell ranger，能用最新版就用最新版本，如下图

图形用户界面, 图表

AI 生成的内容可能不正确。

* Cell Ranger 使用转录本注释 GTF 文件将 reads 分类为外显子区（exonic）、内含子区（intronic）和基因间区（intergenic），如果一个 read 至少有 50% 与外显子区域重叠，则被归类为外显子区；如果它不属于外显子区但与内含子区域重叠，则被归类为内含子区；否则，被归类为基因间区。如果某个 read 既能比对到单个外显子位点，同时也能比对到一个或多个非外显子位点，则优先考虑外显子位点，并将该 read 视为高置信度地比对到该外显子位点，并赋予最高的比对质量评分。默认情况下，属于转录组（在上图中标记为蓝色）的 reads 会被保留用于 UMI 计数。只有能够唯一比对到单个基因的 reads 才会被保留用于 UMI 计数，而多重比对的 reads 会被 Cell Ranger 丢弃。图片包含 图形用户界面

  AI 生成的内容可能不正确。

当测序样本为细胞核（nuclei）时，大量 reads 来自未剪接的转录本，并比对到内含子区域。为了将这些内含子 reads 计入 UMI 计数，可以在运行 cell ranger count 和 cell ranger multi 管道时使用 --include-introns 选项。如果启用了该选项，则所有以正义链方向比对到单个基因的 reads（包括上图中标记为转录组的蓝色 reads、外显子区的浅蓝色 reads 和内含子区的红色 reads）都会被保留用于 UMI 计数。使用 --include-introns 选项后，无需额外创建自定义的 “pre-mRNA” 参考基因组（该参考基因组通常会将整个基因体定义为外显子）。

Brüning R S, Tombor L, Schulz M H, et al. Comparative analysis of common alignment tools for single-cell RNA sequencing[J]. Gigascience, 2022, 11: giac001.

非模式生物注意两点：

好的参考基因组以及注释好的线粒体序列，MITOS2 is a specialized server that can be used to automatically generate good quality mitochondrial annotations for metazoans.

de novo sequenced genomes generate gene models that do not include UTR sequences这样会影响比对效果

## 单细胞数据分析：标准化

单细胞数据的标准化通常分为以下几个层次：

1. **细胞水平标准化（Cell-level normalization）**：

* 目标：消除由于测序深度（每个细胞的总计数差异）带来的技术偏差
* 方法：将每个细胞的基因计数标准化为相对值（如 CPM 或 TPM）

1. **基因水平处理（Feature-level processing）**：

* 目标：校正基因间的表达分布（如 log 转化）或筛选高变基因（HVG）。
* 方法：log 转化、缩放（scaling）等。

1. **可选的批次校正**：

* 在多批次数据中，可能会在细胞或基因标准化后额外进行批次效应校正。

1. **细胞水平标准化（Cell-level normalization）**：

Ahlmann-Eltze 和 Huber 在 2023 年发表的一项最新基准研究（如下文）对 22 种不同的单细胞数据转换方法进行了比较。结果表明，一种相对简单的方法——**加伪计数（pseudo-count）的对数转换（log transformation）结合主成分分析（PCA）**，在性能上与更复杂的方法相当，甚至更优。在单细胞RNA测序数据分析中，矩阵标志化（normalization）后进行log转化时，通常是以自然对数（底数为e）或以10为底的对数（log10）为主，而以2为底的对数（log2）较少见。

Ahlmann-Eltze C, Huber W. Comparison of transformations for single-cell RNA-seq data[J]. Nature Methods, 2023, 20(5): 665-672.

许多流行的单细胞分析工具都实现了该方法，例如：

**Seurat标准化步骤NormalizeData()：**

Seurat 提供了一种默认的标准化方法 LogNormalize，计算方式为：

将每个细胞的 counts 除以该细胞的总 counts（即测序深度），得到一个归一化的值。乘以一个**缩放因子**（默认是 10,000，参数 scale.factor 可调）。对结果取自然对数使用 log1p，即 ln(1+x)

**Scanpy标准化步骤scanpy.pp.normalize\_total() 和 scanpy.pp.log1p()：**

normalize\_total()：将每个细胞的 counts 除以该细胞的总 counts（测序深度），并乘以一个目标和（默认是 1e4，类似 Seurat 的 10,000），生成归一化后的矩阵。log1p()：对归一化后的数据取自然对数ln(1+x)

单细胞 RNA-seq 数据中，每个细胞的总 UMI 计数通常在几百到几万之间（例如 1,000 到 50,000）。如果直接用占比（counts / total counts），结果会非常小（例如 10⁻⁵ 到 10⁻³），不便于直观理解和后续计算。乘以 10,000 后，标准化数据的值通常落在 0 到几百的范围内（经过 log 转化后为 0 到 5 左右），这与基因表达的生物学动态范围较为吻合，也方便可视化（如热图、散点图）。为什么不用 1,000 或 100,000？如果缩放因子太小（如 1,000），标准化后的值范围会偏小，可能导致 log 转化后的数值过于压缩，丢失分辨率。如果缩放因子太大（如 100,000），数值范围会过大，log 转化后可能放大噪声，尤其是在低表达基因中。10,000 是一个折中的选择，既不过分压缩也不过分放大，同时保持数据的动态范围适合大多数下游分析（如聚类、差异表达分析）。

1. **基因水平处理（Feature-level processing）**

单细胞RNA-seq 数据具有以下特点：

* **高维性**：每个细胞可能测定数万个基因的表达量。
* **稀疏性**：由于技术限制，许多基因在大多数细胞中表达为零（dropout）。
* **噪声**：数据中存在技术噪声（如测序深度差异）和生物学无关的变异（如管家基因的均匀表达）。

如果直接使用所有基因进行分析：

* 计算成本高。
* 噪声可能会掩盖生物学信号，导致降维或聚类结果不准确。

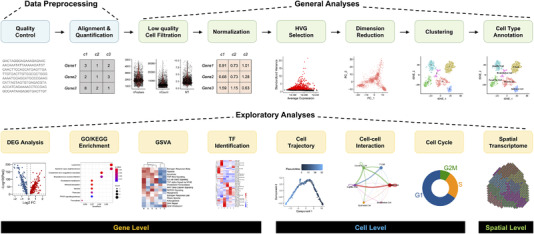
因此筛选Highly Variable Genes，HVG高变基因是单细胞数据中表达变异较大的基因，反映细胞间的生物学差异。

## 单细胞数据过滤-线粒体

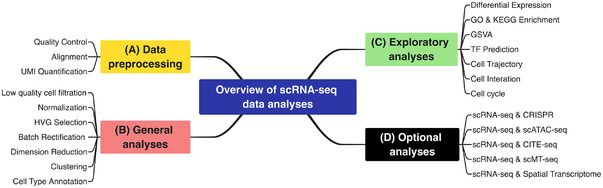
线粒体基因比例超过 5% 至 15% 的细胞被视为低质量细胞并予以剔除。然而，基于线粒体基因比例去除细胞的标准可能因物种、样本类型和实验条件等因素而有所不同。例如，与小鼠相比，人类样本通常表现出更高比例的线粒体基因，而新陈代谢高度活跃的组织（如肾脏）可能会显示出较强的线粒体基因表达。例如，在能量需求较低的组织中，30% 的线粒体 mRNA 可能表明细胞应激或凋亡，而对于能量需求较高的健康心肌细胞来说，这一比例是正常的。线粒体转录本并不会在细胞核中表达。然而，在 snRNA-seq 数据中，仍然检测到不同数量的线粒体转录本。

* Kim G D, Lim C, Park J. A practical handbook on single-cell RNA sequencing data quality control and downstream analysis[J]. Molecules and Cells, 2024, 47(9): 100103.

**Roadmap for typical single-cell RNA sequencing data analysis**



**Overview of the analysis modules for single-cell RNA sequencing data analysis.**

****

* Jovic D, Liang X, Zeng H, et al. Single‐cell RNA sequencing technologies and applications: A brief overview[J]. Clinical and translational medicine, 2022, 12(3): e694.

A recent systematic survey of scRNA-seq data suggested that a mitochondrial proportion threshold of 10% is appropriate to distinguish between healthy and low-quality cells in most human tissues, while in mouse tissues, the recommended threshold is 5%.

* Osorio D, Cai J J. Systematic determination of the mitochondrial proportion in human and mice tissues for single-cell RNA-sequencing data quality control[J]. Bioinformatics, 2021, 37(7): 963-967.

There is no absolute standard for the setting of filter thresholds, which usually depends on the type of cell and tissue being analysed. Lambrechts et al. filtered out cells with ≤ 100 or ≥ 6000 expressed genes, ≤ 200 UMIs and ≥ 10% mitochondrial genes as described in their study. Fan et al. retained good quality cells using the following parameters: (1) 200 < total number of expressed genes per cell (nGenes) < 2500; (2) 300 < total number of UMIs per cell (nUMIs) < 15000; and (3) percentage of UMIs mapped to mitochondrial genes (MT%) < 10%.

* Lambrechts D, Wauters E, Boeckx B, et al. Phenotype molding of stromal cells in the lung tumor microenvironment[J]. Nature medicine, 2018, 24(8): 1277-1289.
* Fan X, Bialecka M, Moustakas I, et al. Single-cell reconstruction of follicular remodeling in the human adult ovary[J]. Nature communications, 2019, 10(1): 3164.

QC metrics vary by tissue. (*X*-axis) Fraction of mitochondrial reads (**A, B**), gene complexity (**C, D**), and percentage of ribosomal protein genes (**E, F**) per cell across human tissues (*Y*-axis) and technologies. Various human tissue scRNA-seq datasets generated by 10X droplet-based (**A, C, E**) and Microwell-seq (**B, D, F**) technologies. Each row in a panel is a density curve with the mean represented by a blue diamond. Red lines indicate conventional threshold values set at 10% for percentage of mitochondrial reads, and 200 for gene complexity.

图示, 示意图

AI 生成的内容可能不正确。

图示, 示意图

AI 生成的内容可能不正确。

* Subramanian A, Alperovich M, Yang Y, et al. Biology-inspired data-driven quality control for scientific discovery in single-cell transcriptomics[J]. Genome biology, 2022, 23(1): 267.

## 单细胞数据分析:PCA

在单细胞转录组分析（scRNA-seq）中，PCA（主成分分析）的输入基因通常是通过以下步骤筛选的，以确保降维结果能够有效地捕捉生物学信号，而不是技术噪音：

### **1. 过滤低表达基因**

许多基因在大多数细胞中几乎不表达，为了减少噪音，通常会去除低表达基因。例如：

* 过滤掉在少数细胞（如<1%或<3或5个细胞）中表达的基因
* **过滤低质量细胞**：同时也会移除那些总计数（UMI或reads）过低或线粒体基因比例过高的细胞，以确保数据质量。

### **2. 选择高变基因（Highly Variable Genes, HVGs）**

为了增强PCA对生物学差异的捕捉能力，常规方法会选择**表达变化较大的基因**（高变基因）：

* 选取高变基因（如前 2,000 个变异最大的基因）

HVG selection methods can be classified into four categories

文本

AI 生成的内容可能不正确。

### **3. 归一化和标准化**

在 PCA 之前，一般需要对数据进行标准化，以消除测序深度等技术因素的影响：

* **归一化**（Normalization）：如按总 UMI 计数归一化到每个细胞 10,000（CPM/TPM 标准化）
* **对数转换**（Log transformation）：如 log1p 变换（log(x + 1)）
* **Z-score 标准化**：将基因表达值减去均值后除以标准差

### Seurat中的ScaleData具体做什么？

在Seurat的工作流程中，ScaleData通常在NormalizeData（归一化）和FindVariableFeatures（筛选高变异基因）之后、运行RunPCA之前执行。它包括以下步骤：

* **中心化（Centering）**：将每个基因的表达值减去其均值，使均值为0。
* **缩放（Scaling）**：将每个基因的表达值除以其标准差，使方差为1。
* **可选的回归（Regress out）**：可以选择移除不需要的变异来源，例如细胞周期效应、线粒体基因比例或批次效应（通过参数vars.to.regress指定）。
* # Seurat流程

seurat\_obj <- NormalizeData(seurat\_obj) # 归一化

seurat\_obj <- FindVariableFeatures(seurat\_obj) # 筛选高变异基因

seurat\_obj <- ScaleData(seurat\_obj) # 缩放数据

seurat\_obj <- RunPCA(seurat\_obj) # 运行PCA

### **4. 计算 PCA**

* 在上述筛选出的高变基因上进行 PCA
* 选择前 10-50 个主成分（PCs）用于后续聚类或 UMAP/tSNE 降维

## Dimensionality Reduction: PCA+Cluster the cells:UMAP+tSNE

在scRNA-seq数据分析中，我们通过寻找与已知细胞状态或细胞周期阶段相关的细胞身份来描述数据集中的细胞结构。这一过程通常被称为细胞身份注释。为此，我们将细胞组织成簇，以推断相似细胞的身份。聚类本身是一个常见的无监督机器学习问题。我们可以通过在降维后的表达空间中最小化簇内距离来得出簇。在这种情况下，表达空间决定了细胞在降维表示下的基因表达相似性。例如，这种低维表示可以通过主成分分析（PCA）确定.PCA：用于前期降维、去噪或输入下游分析（如聚类），不建议直接用于最终可视化。在我们运行PCA之后，需要决定将哪些主成分（PCs）纳入下游分析。我们希望纳入足够多的PCs以保留生物学信号，但又要尽量减少PCs的数量，以避免数据中的噪声干扰。根据每个主成分解释的方差百分比对主成分进行排名。In this example, we can observe an ‘elbow’ around PC 9-10, suggesting that the majority of true signal is captured in the first 10 PCs.

形状

AI 生成的内容可能不正确。

KNN图由反映数据集中细胞的节点组成。我们首先在PCA降维后的表达空间中为所有细胞计算欧几里得距离矩阵，然后将每个细胞与其K个最相似的细胞连接起来。通常，K的值根据数据集的大小设置为5到100之间。KNN图通过在图中将表达空间中的密集区域表示为密集连接区域，反映了表达数据的底层拓扑结构

卡通人物

AI 生成的内容可能不正确。

UMAP（均匀流形逼近和投影）和t-SNE（t-随机邻域嵌入）是单细胞数据集常用的降维和可视化技术。UMAP最近已成为这类分析的黄金标准，因为它具有更高的计算效率并且能更好地保持全局结构；尽管与t-SNE一样，它在局部距离上的准确性可能更高。Suggest a resolution of 0.4-1.2 for data sets of ~3,000 cells.

图表, 散点图

AI 生成的内容可能不正确。

tSNE is slow. tSNE doesn’t scale well to large numbers of cells (10k+)  
– UMAP is quite a bit quicker than tSNE  
– UMAP can preserve more global structure than tSNE\*  
– UMAP can run on raw data without PCA preprocessing\*  
– UMAP can allow new data to be added to an existing projection

## 单细胞学习笔记：cell annotation

大体上分为手动和自动注释。如果小型数据集且自己具有专业知识可以人工基于标记基因和知识推断基于常规的Seurat、Scanpy就可以实现。下面主要说说自动方法，总体上可以分为三类：

**marker genes for manual annotation**

* Smaller gene sets (e.g., size < 20) are more likely to yield cells with unstable scores

相关软件：**Garnett**（适合小规模数据≤50K 细胞，CPU 处理足够快，可自定义 marker）、CellAssign

**a larger set of genes**  
several thousands or more (e.g., size > 100)   
  
相关软件：

**CellTypist**：

CPU 运行：适用于 中等规模数据（10K~100K 细胞）：GPU 运行（使用 PyTorch 或 TensorFlow）；适用于 百万级别数据>1M 细胞内置大规模的细胞类型参考数据库：人类和小鼠，支持Scanpy和Seurat整合）

Clustifyr、**SingleR**（中~大型数据：≥10K 细胞）

**annotation by mapping to a reference**  
相关软件:**Azimuth**是 Seurat 开发团队提供的一种 基于参考数据库的自动化单细胞注释工具。它使用 Seurat label transfer（标签转移） 方法，将新的单细胞数据集投影到一个 预训练的参考数据库 上，以实现快速、自动的细胞类型注释，超大规模数据10K~百万细胞

参考学习链接：

https://www.sc-best-practices.org/