第三次作业

邹翔宇 | 2410833001

1. What’s the major difference between bulk and single cell sequencing data? What extra information can single cell data provide? (10’)

批量测序和单细胞测序数据在样本类型和提供的信息上有着根本的区别:

1. **批量测序**：
   - 在包含大量细胞（百万级别）的“批量”样本上进行。
   - 数据测量了许多细胞的平均信号（如基因表达、转录因子结合、甲基化等）。
   - 它忽略了细胞间的异质性，这意味着它无法捕捉不同细胞类型之间的差异或同一细胞类型内的变异。
2. **单细胞测序**：
   - 从多细胞生物中分离出细胞，并对每个细胞单独进行实验。
   - 可以在单细胞水平上进行不同类型的测序，包括 DNA-seq、ATAC-seq、ChIP-seq、BS-seq 和 RNA-seq。
   - 这种方法在研究中非常活跃，提供了对细胞异质性的详细见解。

总结来说，单细胞测序数据通过考虑细胞群体内的异质性，提供了更详细和细致的细胞过程视图，这是批量测序数据无法实现的。

2. For scRNA-seq technology, what’s the advantage of droplet-based method over plate-based method? (5’)

1. **高通量**： 基于微液滴的方法，能够同时处理成千上万个细胞。这种高通量能力对于需要大规模细胞分析的研究至关重要，如构建细胞图谱或分析细胞异质性。
2. **成本效益**： 由于可以同时处理大量细胞，基于微液滴的方法在成本上通常更具优势，这使得研究人员能够在有限的预算内获得更多的数据。
3. **自动化程度高**： 基于微液滴的方法通常与自动化工作流程兼容，这减少了人为操作的需要，提高了实验的可重复性和效率。
4. **数据稀疏性**： 虽然高通量带来了数据稀疏性问题，但这也意味着数据中包含了更多关于细胞亚群和罕见细胞类型信息的可能性。基于微液滴的方法通过技术进步，如使用独特的分子标识符（UMIs）来减少 PCR 扩增过程中的偏差，从而提高了数据的准确性和可靠性。
5. **技术灵活性**： 基于微液滴的方法允许研究人员根据实验设计灵活选择细胞数量和测序深度，从而可以针对不同的生物学问题进行优化。

3. Briefly describe the procedure for single cell RNA-seq (scRNA-seq) cell clustering. Explain the purpose of each step. (20’)

单细胞 RNA 测序（scRNA-seq）细胞聚类是分析过程中的一个关键步骤，其目的是将表达模式相似的细胞分组成簇。

1. **数据预处理**: 包括质量控制、标准化和数据校正。目的是去除技术噪声和批次效应，以便更准确地比较不同细胞的表达模式。
2. **特征选择**: 通常涉及筛选高变异基因（HVGs），这些基因在细胞间表达差异大，对后续分析贡献信息最多。这一步是为了减少数据维度并专注于最有信息量的特征。
3. **降维**: 使用如 PCA、t-SNE 或 UMAP 等算法，将高维数据降至二维或三维，以便可视化和进一步分析。降维有助于揭示数据中的潜在结构。
4. **细胞聚类**: 基于降维后的数据，通过计算细胞间的相似性，将细胞分组。常用的聚类方法包括 K-means 聚类、层次聚类和基于图的聚类算法，如 Louvain 方法。聚类的目的是将表达模式相似的细胞分到同一个簇中。
5. **聚类结果评估**: 评估聚类的质量，确定簇的数量或验证簇的生物学意义。这一步是确保聚类结果的可靠性和有效性。
6. **细胞标记基因分析**: 识别每个簇的标记基因，这些基因在特定簇中的表达水平显著高于其他簇，有助于进一步理解簇的生物学特性。
7. **可视化**: 使用 t-SNE、UMAP 或 PAGA 等方法，将聚类结果在低维空间中可视化，以直观展示不同细胞簇的分布和关系。

4. What’s the similarity and difference between cell clustering and pseudotime construction. (10’)

细胞聚类和伪时间分析都是单细胞 RNA 测序数据分析中用于理解细胞状态和动态的方法。

相似之处:

- 两者都用于分析单细胞数据，通过降维技术将细胞映射到低维空间，并基于表达模式的相似性进行分组。

不同之处:

- **细胞聚类**着重于将细胞分成离散的群体，主要用于识别不同的细胞类型或状态。
- **伪时间分析**则将细胞排列在一个连续的轨迹上, 以推断它们在某一生物学过程中的发展顺序, 适用于研究细胞分化或疾病进展等动态过程。

5. What’s the advantages and weaknesses of supervised cell type identification method? (10’)

优势:

1. **准确性和鲁棒性**：有监督的方法通常比无监督的方法表现更好，尤其是在数据包含不平衡的细胞类型比例时。
2. **不受样本大小影响**: 与无监督聚类方法相比, 有监督方法不依赖于目标数据的细胞数量, 因为它们是对每个细胞单独进行预测的。
3. **计算性能**：有监督方法在计算上通常更有效，尤其是当细胞数量较多时。
4. **易于处理技术差异**：有监督方法可以通过选择合适的特征和预测模型来减少不同数据集之间的技术差异对预测结果的影响。
5. **可扩展性**：有监督方法可以很好地扩展到大规模数据集，这对于大规模、群体水平的研究尤为重要。

局限性:

1. **依赖参考数据集**：有监督方法的性能在很大程度上依赖于参考数据集的选择
2. **对新细胞类型的识别能力有限**: 有监督方法主要设计用于识别已知的细胞类型, 对于目标数据集中可能出现的新细胞类型，这些方法可能无法有效识别。
3. **可能需要复杂的预处理**
4. **对参考数据集的质量和一致性要求高**

6. What’s the purpose of tSNE and UMAP? (5’)

1. **数据降维**：将高维的单细胞数据降至二维或三维，使其更易于可视化和分析。
2. **数据可视化**：通过在低维空间中展示数据点，帮助研究人员直观地观察和理解细胞间的相似性和差异性。
3. **揭示细胞异质性**：通过可视化揭示样本中细胞的异质性，识别不同的细胞群体或亚群。
4. **辅助聚类分析**：tSNE 和 UMAP 通常用于聚类分析之后，以图形的方式展示聚类结果，帮助验证聚类的准确性。
5. **探索性分析**：它们为研究人员提供了探索单细胞数据集结构和模式的工具，有助于发现新的生物学见解。

tSNE 注重保持数据点之间的相对距离，而 UMAP 则在保持局部结构的同时，还考虑了全局数据结构，两者都是单细胞数据分析中不可或缺的工具。

7. Based on the results obtained from the lab, write a short report to present the scRNA-seq analysis results. Explain the steps in data analysis and discuss the meanings of the figures. (40’)

Steps

1. **数据预处理和质量控制**

数据加载和初始化：将原始数据加载到 Seurat 对象中，这是 Seurat 用于 scRNA-seq 分析的主要数据结构。

质量控制指标: 我们计算了线粒体基因的百分比（percent.mt）作为细胞应激或损伤的度量。检查了前五个细胞的初始 QC 指标。

可视化 QC 指标：为 nFeature_RNA、nCount_RNA 和 percent.mt 生成 violin plot，以评估细胞间基因特征和计数的分布。

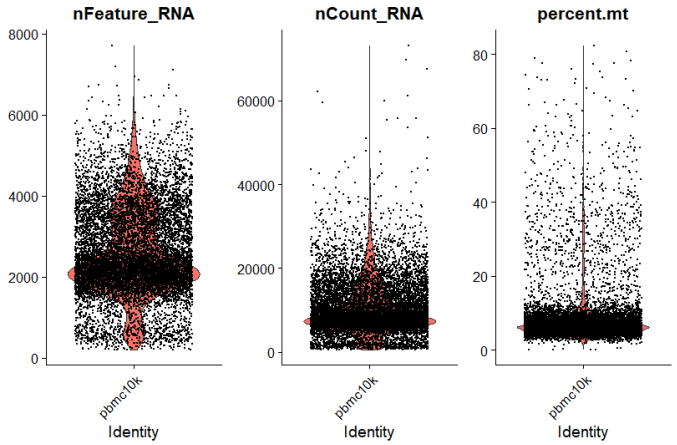

Figure 1: QC Metrics Violin Plot

过滤细胞：过滤掉具有少于 200 个或多于 5000 个独特基因特征的细胞，或线粒体基因含量高于 15% 的细胞，以确保用于进一步分析的高质量数据。

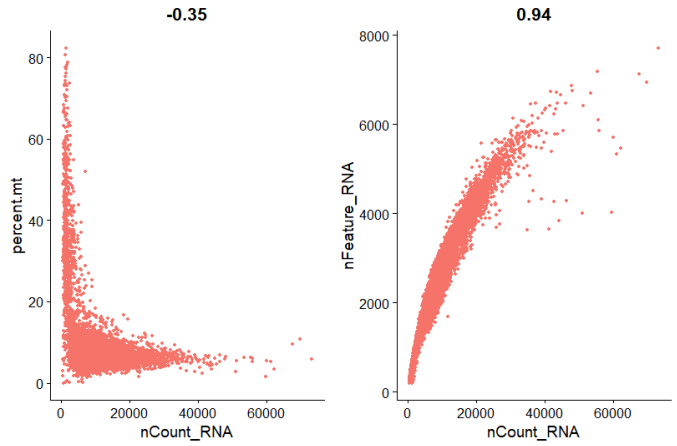

Figure 2: FeatureScatter Plots

第一个特征散点图（plot1）显示了每个细胞的基因计数与线粒体基因表达百分比之间的关系。第二个特征散点图（显示了每个细胞的基因计数基因数量之间的关系。这两个图用于初步评估数据的质量和细胞的异质性。

2. **标准化和特征选择**

标准化：使用 LogNormalize 方法对数据进行标准化，以减少高表达基因对下游分析的影响。
基因表达分布：绘制直方图以可视化标准化前后的基因表达分布。

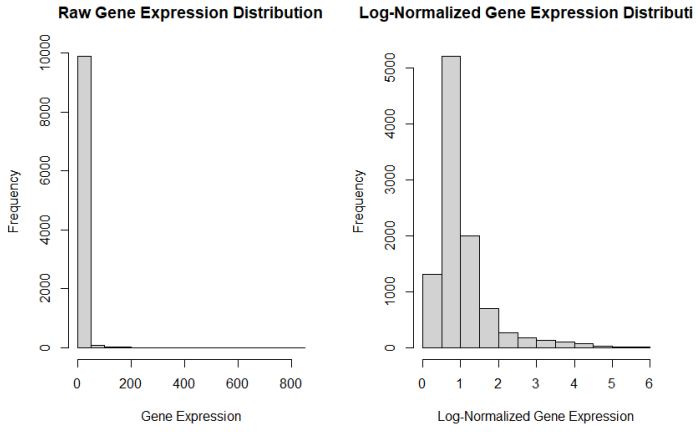

Figure 3: Raw and Log-Normalized Gene Expression Distribution Histograms

左边的直方图显示了原始数据中基因表达的分布情况。右边的直方图显示了对数归一化后基因表达的分布情况。

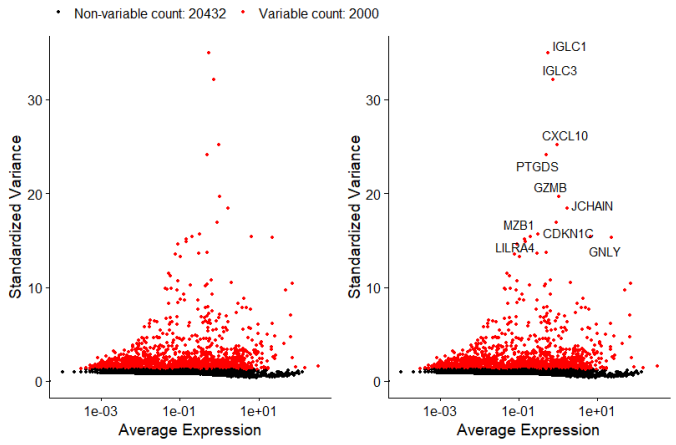

Figure 4: Variable Feature Plot

识别高度可变特征：使用 FindVariableFeatures 函数和 vst 方法识别高度可变基因。绘制了前 10 个最可变基因的图表，以突出显示变异性最大的基因。第一个图（plot1）显示了所有变异性高的特征。第二个图（plot2）在第一个图的基础上，标记了变异性最高的 10 个基因。

3. **缩放和降维**

缩放：使用线性变换对数据进行缩放，为 PCA 等降维技术做准备。

PCA：在缩放数据上执行主成分分析（PCA）以降低基因表达数据的维度。

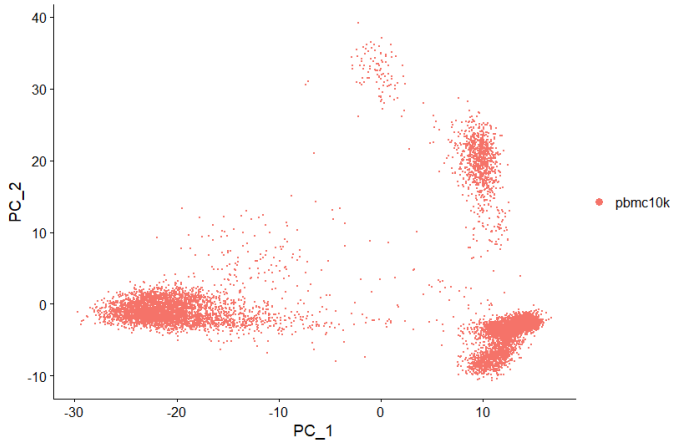

Figure 5: PCA Results

聚类：使用 PCA 结果，通过 FindNeighbors 和 FindClusters 函数对细胞进行聚类。

非线性降维 UMAP 和 tSNE：应用非线性降维技术，包括均匀流形近似和投影（UMAP）和 t 分布随机邻域嵌入（tSNE），以在二维中可视化数据。 DimPlot 展示了使用 UMAP 和 tSNE 方法进行非线性降维后的数据点在二维空间中的分布。这些图用于更直观地观察细胞之间的相似性和差异性。

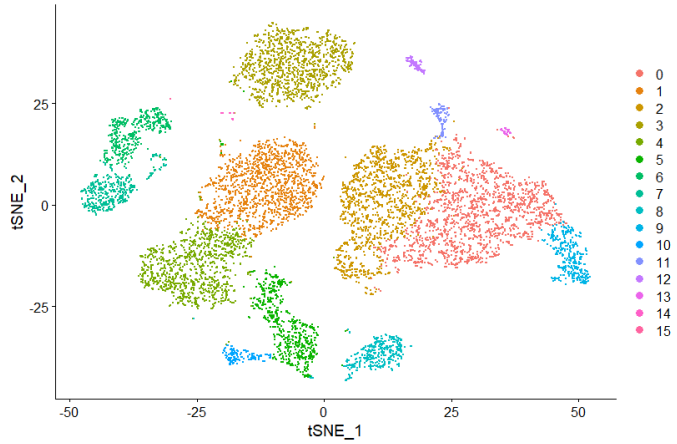

Figure 6: tSNE Results

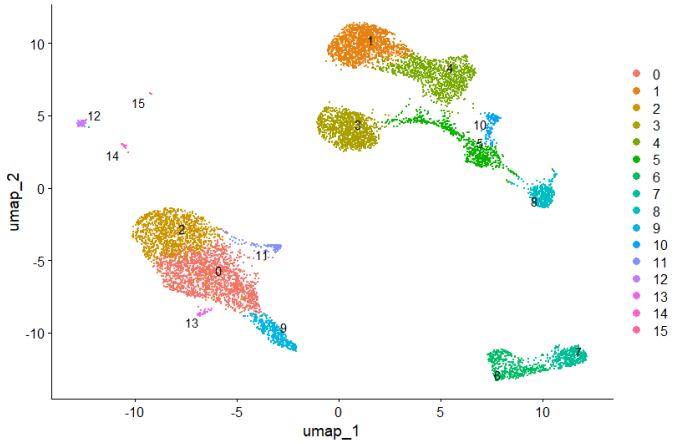

Figure 7: umap Results