单细胞转录组数据分析标准流程报告

pbmc3k 数据集 Seurat v5 教程与复现

2022 届河北医科大学 (由 Gemini 整理与增强)

June 16, 2025

Contents

List of Figures

List of Tables

Part I

认知准备：从宏观理解单细胞分析

在敲下任何代码之前，让我们先建立对单细胞分析的宏观认知。

1 为什么需要单细胞测序?

【核心理念说明】

一个生动的类比：果汁 vs. 水果拼盘
传统的 RNA 测序（Bulk RNA-seq）就像是把一篮子不同种类的水果（比如苹果、香蕉、橙子）全部丢进榨汁机。最后你得到了一杯混合果汁，你能尝出里面大概有哪几种水果的味道，但你无法知道这篮水果里到底有多少个苹果、多少根香蕉。

而**单细胞 RNA 测序（scRNA-seq）**则像是把这篮水果一个一个拿出来，仔细地识别、分类和计数。最终，你得到的不是一杯模糊的果汁，而是一份清晰的清单：“苹果 5 个，香蕉 3 根，橙子 4 个……”。

在生物学研究中，这“水果”就是我们体内的细胞。组织和器官正是由成千上万种功能各异的细胞组成的。单细胞测序使我们能够以前所未有的分辨率，看清单个细胞的基因表达特征，从而揭示细胞的异质性、发现新的细胞亚群、描绘细胞的发育轨迹。

2 本次分析的“作战地图”

我们的整个分析流程就像一场精心策划的战役，可以分为四个主要阶段：

1. **数据净化**：拿到原始数据后，先进行严格的质量控制，把“老弱病残”的细胞数据剔除掉，并进行标准化，让所有数据站在同一起跑线上。
2. **模式发现**：在干净的数据中，通过降维（PCA）抓住主要矛盾，再通过聚类（Clustering）将相似的细胞“抱团”，形成不同的细胞群落。
3. **生物学解读**：为每个细胞群落寻找独特的“身份标识”（Marker 基因），并通过这些标识来判断它们究竟是哪种类型的细胞（T 细胞、B 细胞、还是单核细胞?）。
4. **可视化呈现**：将最终的细胞身份注释结果绘制成一张精美的“细胞星图”（UMAP 图），直观展示样本中所有细胞的种类和关系。

3 分析环境与项目准备

此为基础准备工作，确保后续所有操作顺利进行。

- **安装 R 与 RStudio**: R 是分析引擎，RStudio 是操作界面。请确保两者均已正确安装。
- **获取项目框架与数据**: 从课程的 GitHub 仓库下载项目框架，并根据指导在 RStudio 中设置好工作目录，运行数据下载脚本。

Part II

数据净化：从原始数据到高质量矩阵

本部分我们将完成从加载数据到获得可用于分析的、高质量表达矩阵的全过程。

4 数据加载与 Seurat 对象创建

首先，我们将 10x Genomics 平台产出的三个核心文件（barcodes, features, matrix）读入 R，并创建一个 Seurat 对象。

```
library(Seurat)
data_directory <- "03_data/filtered_gene_bc_matrices/hg19/"
pbmc.data <- Read10X(data.dir = data_directory)
pbmc <- CreateSeuratObject(counts = pbmc.data, project = "pbmc3k",
                           min.cells = 3, min.features = 200)
```

【教学重点】

参数解读: `min.cells = 3` 表示一个基因至少要在 3 个细胞中被检测到才会被保留；`min.features = 200` 表示一个细胞至少要检测到 200 个基因才被视为有效细胞。这两个参数提供了第一道粗略的质控过滤。

5 质量控制 (QC)：细胞的“海关安检”

这是至关重要的一步。我们通过三个核心指标来识别并剔除低质量的细胞数据。

5.1 指标一：细胞内基因数量 (nFeature_RNA)

这个指标反映了细胞的转录复杂性。

- **过低**: 可能意味着这个“细胞”其实是个空泡，或者细胞已经死亡，RNA 降解严重。
- **过高**: 可能意味着这是一个“双包体”（Doublet），即两个细胞被错误地包裹进同一个反应微滴中，导致基因数量异常翻倍。

```
> pbmc.data <- Read10X(data.dir = data_directory)
> pbmc.data
32738 x 2700 sparse Matrix of class "dgCMatrix"
  [[ suppressing 49 column names 'AAACATACAACCAC-1', 'AAACATTGAGCTAC-1', 'AAACATTGATCAGC-1' ... ]]
  [[ suppressing 49 column names 'AAACATACAACCAC-1', 'AAACATTGAGCTAC-1', 'AAACATTGATCAGC-1' ... ]]

MIR1302-10    . . . . . .
. . . .....
FAM138A       . . . . . .
. . . .....
OR4F5         . . . . . .
. . . .....
RP11-34P13.7  . . . . . .
. . . .....
RP11-34P13.8  . . . . . .
. . . .....
AL627309.1    . . . . . .
. . . .....
RP11-34P13.14 . . . . . .
. . . .....
RP11-34P13.9  . . . . . .
. . . .....
AP006222.2    . . . . . .
. . . .....
RP4-669L17.10 . . . . . .
. . . .....

 ..............................
 ........suppressing 2651 columns and 32718 rows in show(); maybe adjust options(max.print=, width=)
 ..............................
  [[ suppressing 49 column names 'AAACATACAACCAC-1', 'AAACATTGAGCTAC-1', 'AAACATTGATCAGC-1' ... ]]

KIR3DL2.1  . . . . . .
. . .......
AL590523.1 . . . . . .
. . .......
CT476828.1 . . . . . .
. . .......
```

Figure 1: Seurat 对象初始化后的摘要信息。

5.2 指标二：细胞内总分子数 (nCount_RNA)

这个指标主要与测序深度相关，但也反映了细胞的 RNA 总量。通常与 nFeature_RNA 呈正相关。

5.3 指标三：线粒体基因表达比例 (percent.mt)

这是衡量细胞健康状况的“金标准”。

【核心理念说明】

线粒体是细胞的“能量工厂”。当细胞处于应激或濒死状态时，细胞膜的通透性会增加，导致细胞质中的 mRNA 大量流失，而线粒体相对较大，其内部的 mRNA 会被保留下来。因此，线粒体基因转录本所占的比例异常升高，往往是细胞状态不佳的强烈信号。

```
% 计算线粒体基因比例
pbmc[["percent.mt"]] <- PercentageFeatureSet(pbmc, pattern = "^MT-")
% 可视化QC指标，为设定阈值提供依据
VlnPlot(pbmc, features = c("nFeature_RNA", "nCount_RNA", "percent.mt"), ncol =
    3)
% 根据观察执行过滤
pbmc <- subset(pbmc, subset = nFeature_RNA > 200 & nFeature_RNA < 2500 & percent
    .mt < 5)
```

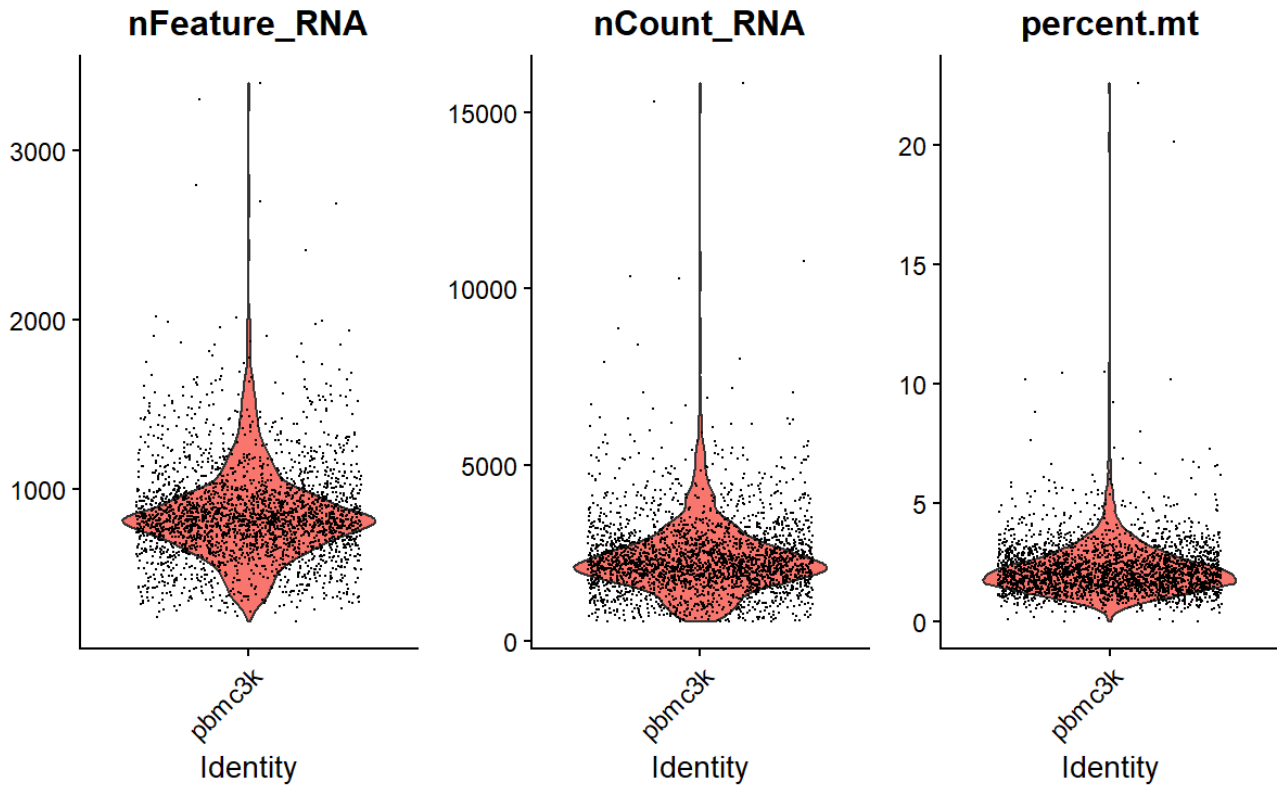


Figure 2: QC 指标小提琴图。我们根据这三个指标的分布，设定合理的阈值（图中虚线所示意的位置）来过滤掉两端的异常细胞。

6 数据标准化：消除技术偏差

完成质控后，我们需要对数据进行标准化，以消除细胞间因测序文库大小（即 `nCount_RNA`）不同而引起的技术误差，确保后续的比较是在公平的基础上进行的。

```
pbmc <- NormalizeData(pbmc)
```

Part III

模式发现：从高维数据中寻找结构

在净化后的数据中，我们将通过一系列降维和聚类算法，发现隐藏在其中的细胞群体结构。

7 特征选择与数据缩放

- **寻找高变基因 (HVGs)**: 从上万个基因中，找出在细胞间表达差异最大的 2000 个“明星基因”。这能帮助我们忽略背景噪音，聚焦于真正定义细胞身份的核心信号。

- **数据缩放 (Scaling)**: 对每个高变基因的表达值进行中心化和标准化处理，防止少数高表达基因在后续分析中“一家独大”。

```
pbmc <- FindVariableFeatures(pbmc, selection.method = "vst", nfeatures = 2000)
pbmc <- ScaleData(pbmc, features = VariableFeatures(object = pbmc))
```

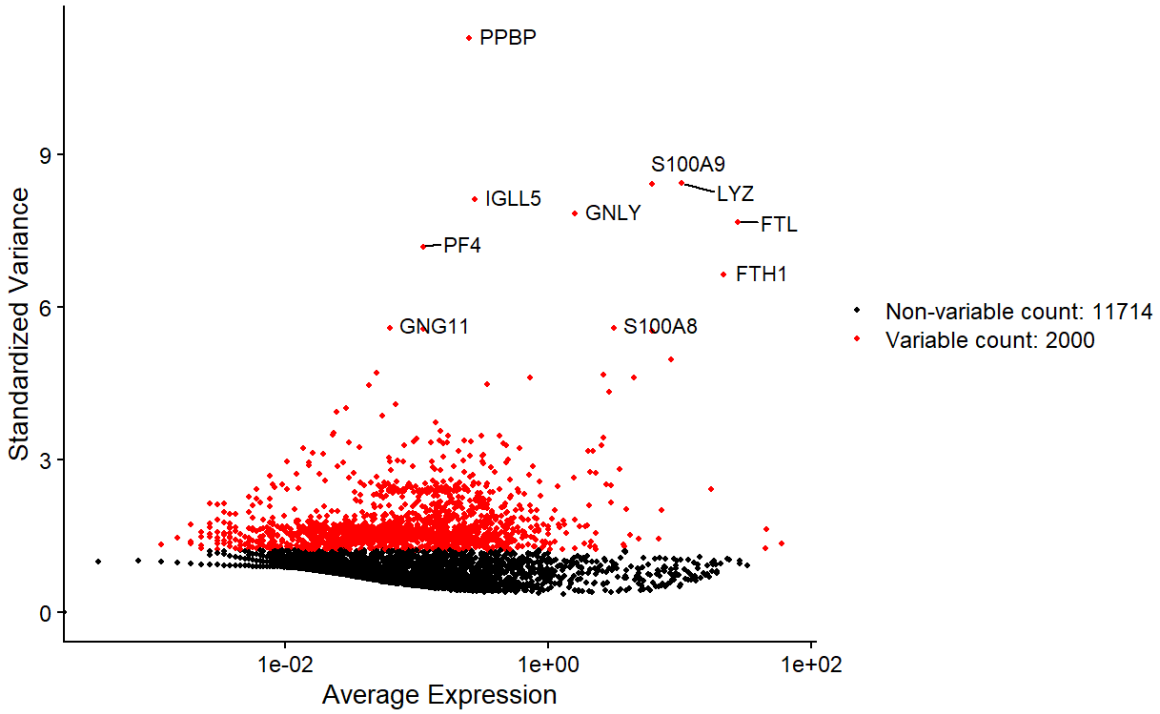


Figure 3: 高变异基因 (HVGs) 筛选图。

8 线性降维 (PCA)

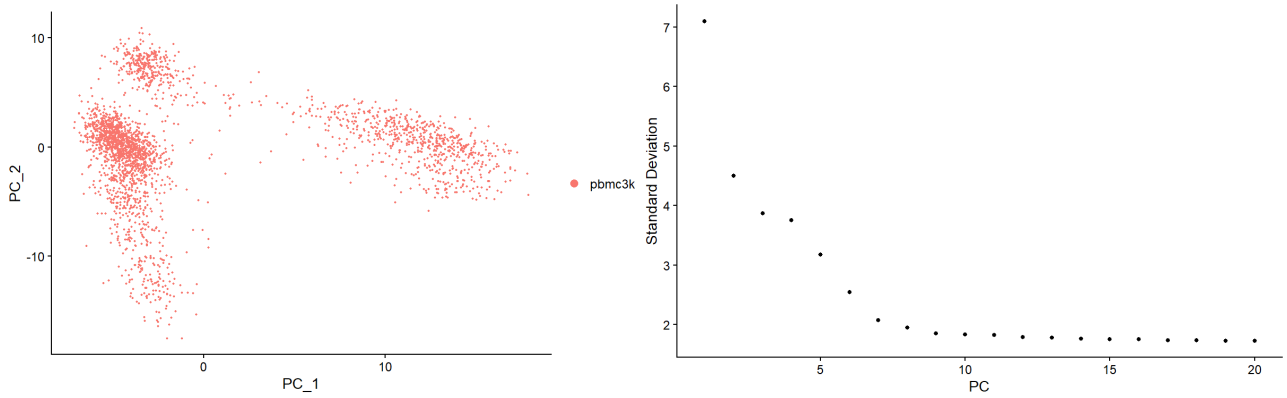
我们将 2000 个高变基因的复杂信息，浓缩到少数几个“主成分”(PCs) 中。

【核心理念说明】

想象一下，要描述一群人的特征，你不需要测量他们的身高、体重、臂展、腿长等所有 2000 个指标。你可能会发现，用“体型”（由身高和体重共同决定）和“身材比例”（由臂展和腿长等决定）这两个“主成分”就能很好地区分他们。PCA 做的就是类似的事情：它将高度相关的基因信息组合成少数几个 PCs，以抓住数据的主要变异方向。

```
pbmc <- RunPCA(pbmc, features = VariableFeatures(object = pbmc))
```

随后，我们使用“肘部图”(Elbow Plot) 来决定保留多少个 PC 用于下游分析，这是一个平衡“有效信号”与“随机噪音”的决策过程。根据图 4b，我们选择保留前 15 个 PC。



(a) PCA 降维散点图　　**(b)** PCA “肘部图”

Figure 4: PCA 结果可视化。左图显示细胞在 PC1 和 PC2 空间中已初步呈现分离趋势。右图帮助我们确定数据的“内在维度”，曲线的拐点（“肘部”）是理想选择，我们选择 15 个 PC。

9 细胞聚类与非线性降维 (UMAP)

- **细胞聚类**: 基于选定的 15 个 PC 的“坐标”，我们首先构建一个细胞“社交网络”（‘FindNeighbors‘），然后利用图算法（如 Louvain 算法）在网络中寻找联系紧密的“小团体”（‘FindClusters‘），这些小团体就是我们的细胞群落。
- **UMAP 可视化**: UMAP 是一种强大的非线性降维算法，非常适合将高维的聚类结果可视化到二维平面上，同时能很好地保持细胞群落的局部和全局结构。

```
pbmc <- FindNeighbors(pbmc, dims = 1:15)
pbmc <- FindClusters(pbmc, resolution = 0.5)
pbmc <- RunUMAP(pbmc, dims = 1:15)
```

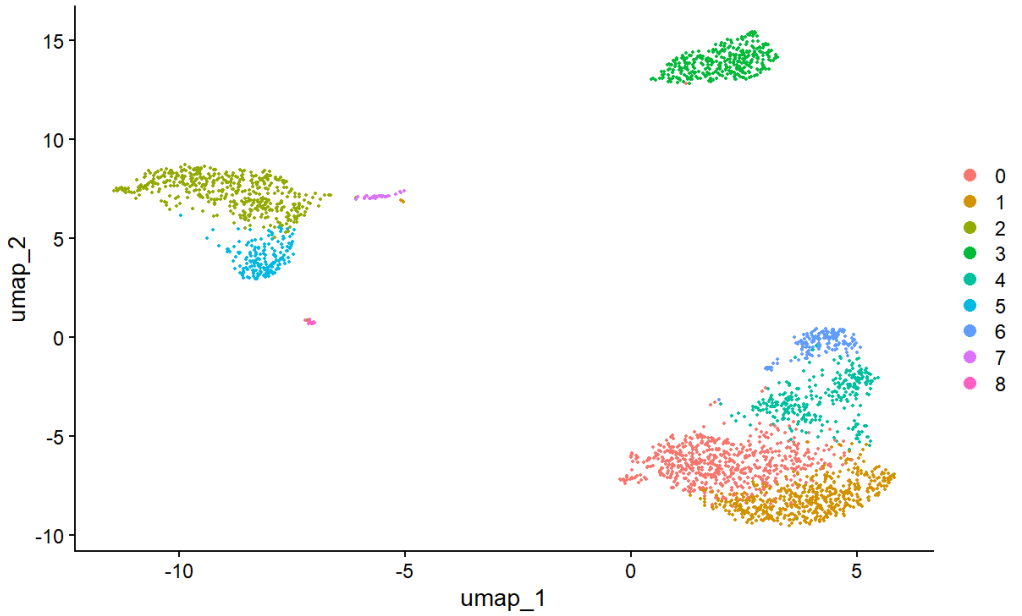


Figure 5: 细胞聚类与 UMAP 可视化初步结果。细胞被清晰地划分为 9 个不同的群落。

Part IV

生物学解读：从数字到细胞身份

这是整个分析流程中最激动人心的部分：我们将赋予这些数字化的细胞群落明确的生物学身份。

10 寻找各群落的 Marker 基因

Marker 基因，即差异表达基因，是指在某个特定细胞群落中表达水平显著高于其他所有群落的基因。它们是每个群落最独特的“身份名片”。

```
pbmc.markers <- FindAllMarkers(pbmc, only.pos = TRUE, min.pct = 0.25, logfc.
    threshold = 0.25)
```

【教学重点】

参数解读： `min.pct = 0.25` 要求一个基因至少在目标群落 25% 的细胞中被检测到；`logfc.threshold = 0.25` 要求该基因在目标群落中的平均表达量比在其他群落中高出 $\log_2(0.25)$ 倍。这两个参数帮助我们筛选出既普遍又特异的 Marker 基因。

11 注释细胞类型：严谨的“侦探工作”

这是一个结合生物学知识和数据证据的推理过程。

1. **导出 Marker 列表：** 我们通常会将每个群落中差异最显著的 Top 10 或 Top 15 的 Marker 基因列表导出为 CSV 文件，便于查阅。
2. **知识库比对：** 我们将列表中的关键基因（通常是 avg_log2FC 最大、p_val_adj 最小的那些）与已知的细胞类型 Marker 基因数据库（如 CellMarkerDB、PanglaoDB）或相关文献进行比对。
3. **逻辑推理与注释：** 例如，当我们发现 Cluster 3 的 Top Marker 是 `MS4A1`（编码 CD20 蛋白）和 `CD79A` 时，所有证据都强烈指向这个群落是 B 细胞。
4. **可视化验证（关键步骤）：** 注释完成后，我们可以通过在 UMAP 图上可视化某个关键 Marker 的表达，来直观地验证我们的注释是否正确。

```
% 示例：可视化B细胞的经典Marker：MS4A1 (CD20)
FeaturePlot(pbmc, features = "MS4A1")
VlnPlot(pbmc, features = "MS4A1")
```

运行以上代码后，您将在 RStudio 的绘图窗口看到两幅图。一幅是 FeaturePlot，它会在 UMAP 图上用颜色深浅展示 `MS4A1` 基因的表达高低；另一幅是 VlnPlot（小提琴图），它会展示 `MS4A1` 在每个 Cluster 中的表达分布。您应当会观察到，无论是 FeaturePlot 还是 VlnPlot，`MS4A1` 的

表达都被完美地限制在了我们注释为 B 细胞的 Cluster 3 中，这为我们的结论提供了强有力的视觉证据。

【动手练习与思考】

截图练习: 这一步是很好的练习机会，请您亲自运行代码，并将生成的 FeaturePlot / VlnPlot 截图保存下来，可以命名为 $\texttt{Validation}_M S4A1.png$

12 生成最终细胞图谱

最后，我们将注释好的细胞身份赋予 Seurat 对象，并绘制出最终的“细胞身份地图”。

Table 1: 各细胞群落生物学注释汇总

Cluster ID	注释细胞类型	关键 Marker 基因及理由
0	**记忆 CD4+ T 细胞**	IL7R, LTB。IL7R (CD127) 是记忆 T 细胞的关键受体。
1	**初始 CD4+ T 细胞**	CCR7, LEF1, TCF7。这三个是公认的初始 T 细胞“三驾马车”。
2	**CD14+ 单核细胞**	CD14, LYZ, S100A9。是经典单核细胞的强烈信号。
3	**B 细胞**	MS4A1 (CD20), CD79A。B 细胞最特异、最核心的 Marker。
4	**CD8+ T 细胞**	CD8A, CD8B。CD8+ T 细胞身份的“金标准”。
5	**FCGR3A+ 单核细胞**	FCGR3A (CD16), MS4A7。典型的非经典单核细胞特征。
6	**NK 细胞**	GNLY, GZMB, NKG7。高表达一系列强大的细胞毒性基因。
7	**树突状细胞 (DC)**	FCER1A, CLEC10A。经典的树突状细胞 Marker。
8	**血小板**	PPBP, PF4。血小板特异性趋化因子。

```
% 创建新旧ID的对应关系并重命名
new.cluster.ids <- c("Memory CD4 T", "Naive CD4 T", "CD14+ Monocyte", "B cell",
    "CD8 T cell",
                     "FCGR3A+ Monocyte", "NK cell", "Dendritic Cell", "Platelet")
names(new.cluster.ids) <- levels(pbmc)
pbmc <- RenameIdents(pbmc, new.cluster.ids)
% 绘制最终带标签的UMAP图
DimPlot(pbmc, reduction = "umap", label = TRUE, pt.size = 0.5) + NoLegend()
```

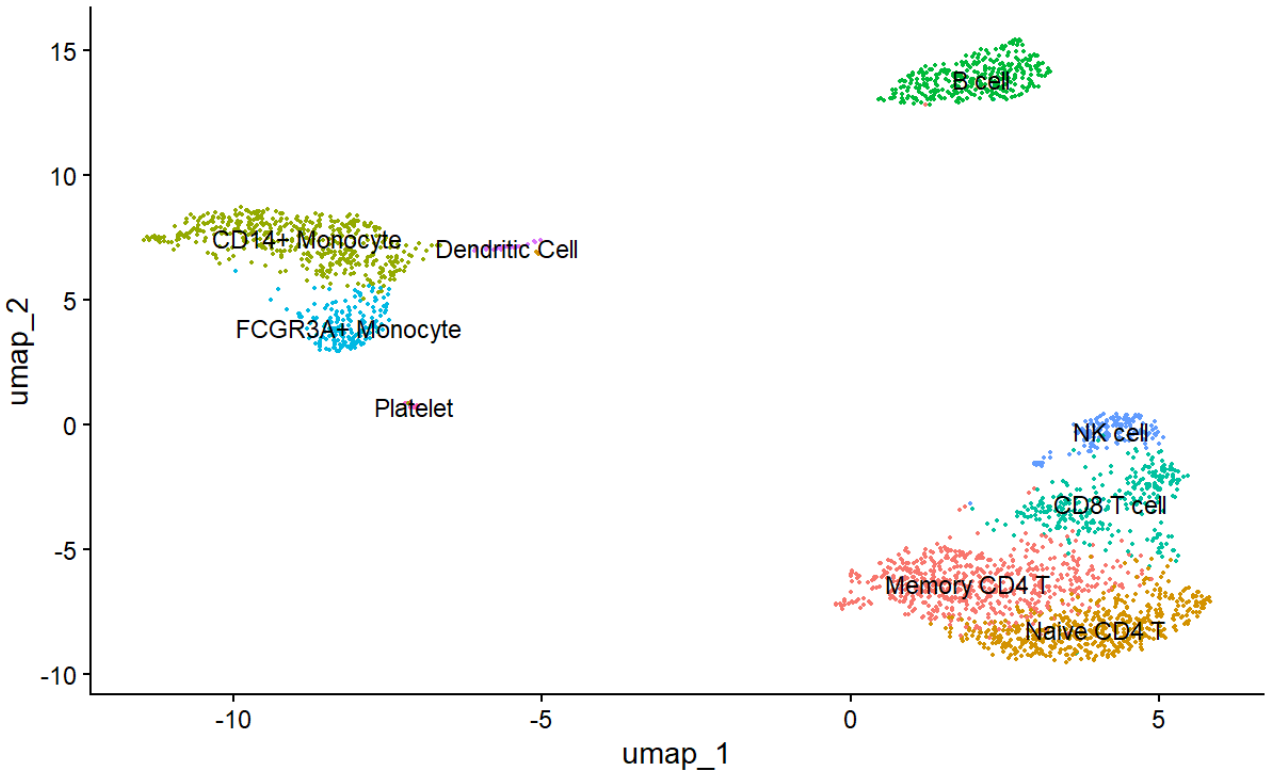


Figure 6: 最终细胞类型注释的 UMAP 图。该图是本次分析的核心成果，直观地展示了 pbmc3k 数据集中主要的免疫细胞亚群及其在转录组水平上的相对关系。

附录：完整分析 R 脚本

A scRNA-seq Seurat V5 标准流程脚本

```r
# 1. 环境设置与数据加载
library(Seurat)
library(dplyr)
data_directory <- "03_data/filtered_gene_bc_matrices/hg19/"
pbmc.data <- Read10X(data.dir = data_directory)
pbmc <- CreateSeuratObject(counts = pbmc.data, project = "pbmc3k", min.cells =
    3, min.features = 200)

# 2. 质量控制 (QC) 与标准化
pbmc[["percent.mt"]] <- PercentageFeatureSet(pbmc, pattern = "^MT-")
pbmc <- subset(pbmc, subset = nFeature_RNA > 200 & nFeature_RNA < 2500 & percent
    .mt < 5)
pbmc <- NormalizeData(pbmc)

# 3. 特征选择与数据缩放
pbmc <- FindVariableFeatures(pbmc, selection.method = "vst", nfeatures = 2000)
all.genes <- rownames(pbmc)
pbmc <- ScaleData(pbmc, features = all.genes)
```

```r

# 4. 线性降维 (PCA)
pbmc <- RunPCA(pbmc, features = VariableFeatures(object = pbmc))

# 5. 细胞聚类与非线性降维 (UMAP)
pbmc <- FindNeighbors(pbmc, dims = 1:15)
pbmc <- FindClusters(pbmc, resolution = 0.5)
pbmc <- RunUMAP(pbmc, dims = 1:15)

# 6. 寻找差异表达基因 (Marker Genes)
pbmc.markers <- FindAllMarkers(pbmc, only.pos = TRUE, min.pct = 0.25, logfc.
    threshold = 0.25)
top15.markers <- pbmc.markers %>% group_by(cluster) %>% slice_max(n = 15, order_
    by = avg_log2FC)
write.csv(top15.markers, file = "top15_markers_per_cluster.csv", row.names =
    FALSE)

# 7. 重命名细胞群落并最终可视化
new.cluster.ids <- c("Memory CD4 T", "Naive CD4 T", "CD14+ Monocyte", "B cell",
    "CD8 T cell",
                     "FCGR3A+ Monocyte", "NK cell", "Dendritic Cell", "Platelet")
names(new.cluster.ids) <- levels(pbmc)
pbmc <- RenameIdents(pbmc, new.cluster.ids)
# DimPlot(pbmc, reduction = "umap", label = TRUE, pt.size = 0.5) + NoLegend()
```