BIOF3 组学数据分析

07 多模态数据分析

导出日期：2026年5月12日

07 多模态数据分析

多模态单细胞技术在一个细胞里同时测两种信息：CITE-seq 同时测 RNA 和细胞表面蛋白，10x Multiome 同时测 RNA 和染色质可及性。相比纯 scRNA-seq，多模态数据的优势是可以相互验证和补充：RNA 层的聚类结果，往往能在蛋白层找到更清晰的 marker。

本节用 10x Genomics 的 5k PBMC CITE-seq 数据演示最常用的一条 CITE-seq 分析流程：分别标准化 RNA 和 ADT，再用 WNN（Weighted Nearest Neighbor）做联合聚类。

多模态技术对比

技术	测量内容	常见用途
CITE-seq	RNA + 表面蛋白（ADT）	免疫学、细胞分型
ASAP-seq	RNA + 染色质开放	调控机制
10x Multiome	RNA + ATAC（同细胞）	基因表达与调控联动
Perturb-seq	RNA + CRISPR 扰动	遗传学筛选

CITE-seq 门槛低，直接用标准 Seurat/Scanpy 就能做；Multiome 需要走 Cell Ranger ARC 流程，后续分析借助 Signac 或 MultiVI。

CITE-seq 的 WNN 分析

数据是 01 章里介绍过的 5k PBMC CITE-seq，下载后 `filtered_feature_bc_matrix/` 目录里同时包含 `Gene Expression` 和 `Antibody Capture` 两层矩阵。

先读入并分别对两层数据做标准化和降维：

```
library(Seurat)

# 读取 CITE-seq 数据（同时包含 RNA 和 ADT）
data_dir <- "~/biof3-data/pbmc5k-citeseq/filtered_feature_bc_matrix"
counts <- Read10X(data.dir = data_dir)

# 创建 Seurat 对象：RNA 作为主 assay，ADT 作为额外 assay
cbmc <- CreateSeuratObject(counts = counts$`Gene Expression`, project = "PBMC5K_CITE")
cbmc[["ADT"]] <- CreateAssayObject(counts = counts$`Antibody Capture`)

# RNA 标准化 + 高变基因 + PCA
DefaultAssay(cbmc) <- "RNA"
cbmc <- NormalizeData(cbmc)
cbmc <- FindVariableFeatures(cbmc)
cbmc <- ScaleData(cbmc)
cbmc <- RunPCA(cbmc)

# ADT 用 CLR 归一化（Centered Log Ratio），更适合抗体信号
DefaultAssay(cbmc) <- "ADT"
cbmc <- NormalizeData(cbmc, normalization.method = "CLR", margin = 2)
cbmc <- ScaleData(cbmc)
cbmc <- RunPCA(cbmc, reduction.name = "apca")
```

RNA 和 ADT 的归一化方式不同：RNA 常用 `LogNormalize`，ADT 用 CLR（`margin = 2` 意为按细胞做 CLR）。两层分别做完 PCA 后，用 WNN 把它们的邻居结构加权合并，再在合并图上聚类和降维：

```
# WNN：根据两个 reduction 同时找邻居
cbmc <- FindMultiModalNeighbors(
  cbmc,
  reduction.list = list("pca", "apca"),
  dims.list = list(1:30, 1:18)
)

# 在 WNN 图上做 UMAP 和聚类
cbmc <- RunUMAP(cbmc, nn.name = "weighted.nn", reduction.name = "wnn.umap")
cbmc <- FindClusters(cbmc, graph.name = "wsnn", resolution = 0.5)

# 可视化
DimPlot(cbmc, reduction = "wnn.umap", label = TRUE)
```

跑完得到的 `cbmc@meta.data` 里会新增一列 `seurat_clusters`，`Reductions(cbmc)` 里会多出 `wnn.umap`。比较三张 UMAP（只用 RNA、只用 ADT、WNN 联合）通常会看到：WNN 版本在分群边界上更干净，尤其是 T 细胞内部的 CD4/CD8 亚群。

真实示例：5k PBMC CITE-seq 走 WNN

配套脚本 `module08_cite_seq_sci.R` 把上面的流程完整跑了一遍，用的是 10x 官方 5k PBMC CITE-seq 数据（Gene Expression + 32 个抗体的 TotalSeq-B panel，~37 MB）：

```
Rscript scripts/single-cell/sc08_citeseq_sci.R
```

脚本顺序跟上一节完全对上：`Read10X` 自动拆出两层矩阵 → 分别建 RNA 和 ADT assay → RNA 做 LogNormalize + PCA、ADT 做 CLR + PCA → `FindMultiModalNeighbors` 合并两个 reduction → 在 WNN 图上做 Leiden 聚类和 `wnn.umap`，同时单独留一份 `rna.umap` 和 `adt.umap` 用来对比。QC 之后大概留下 4000 多个细胞。

每张图看什么

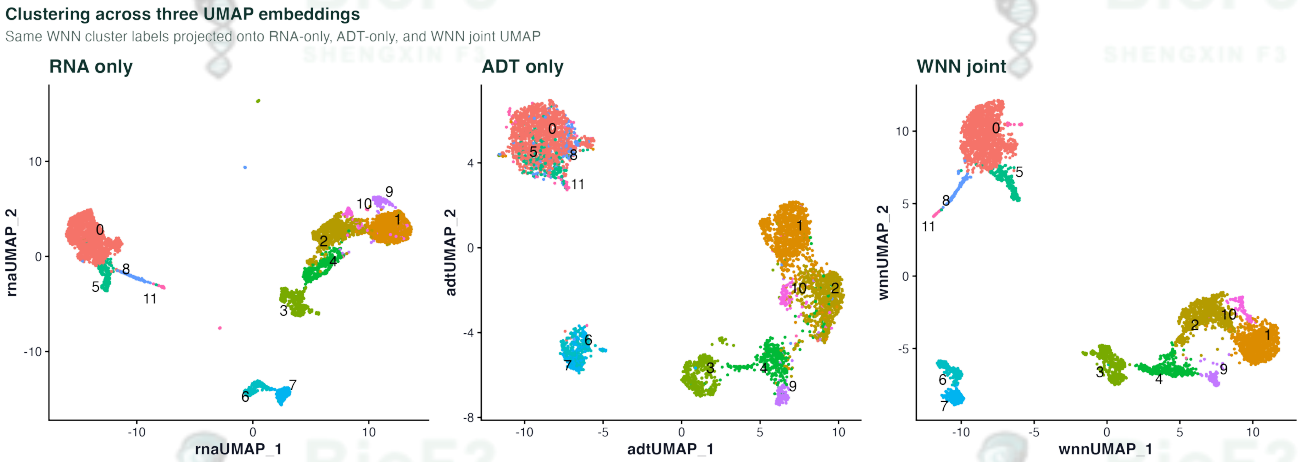

图 1：把同一份 WNN 聚类标签分别画到 RNA UMAP、ADT UMAP 和 WNN UMAP 上。RNA UMAP 在小群细胞（DC、浆细胞样）上分得更细，ADT UMAP 对 T/B/单核大类的边界更干净，WNN 结合两边的优势，最终的 cluster 边界最整齐。

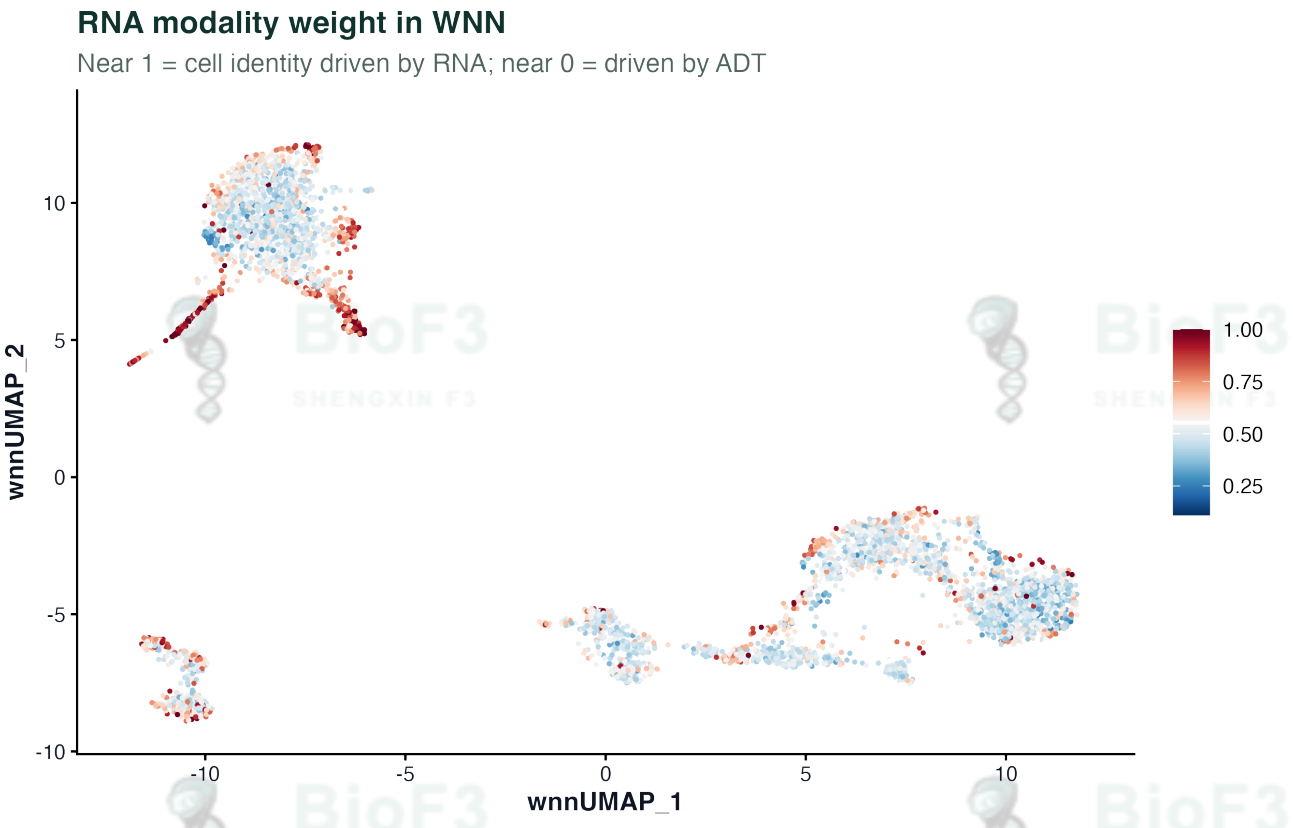

图 2：每个细胞在 WNN 邻居图里 RNA 的权重。接近 1 的区域主要靠 RNA 区分（像稀有亚群、发育状态相关的细胞），接近 0 的区域主要靠 ADT 区分（像成熟 T 细胞、单核细胞）。这张图是"WNN 不是简单平均"的最直观证据。

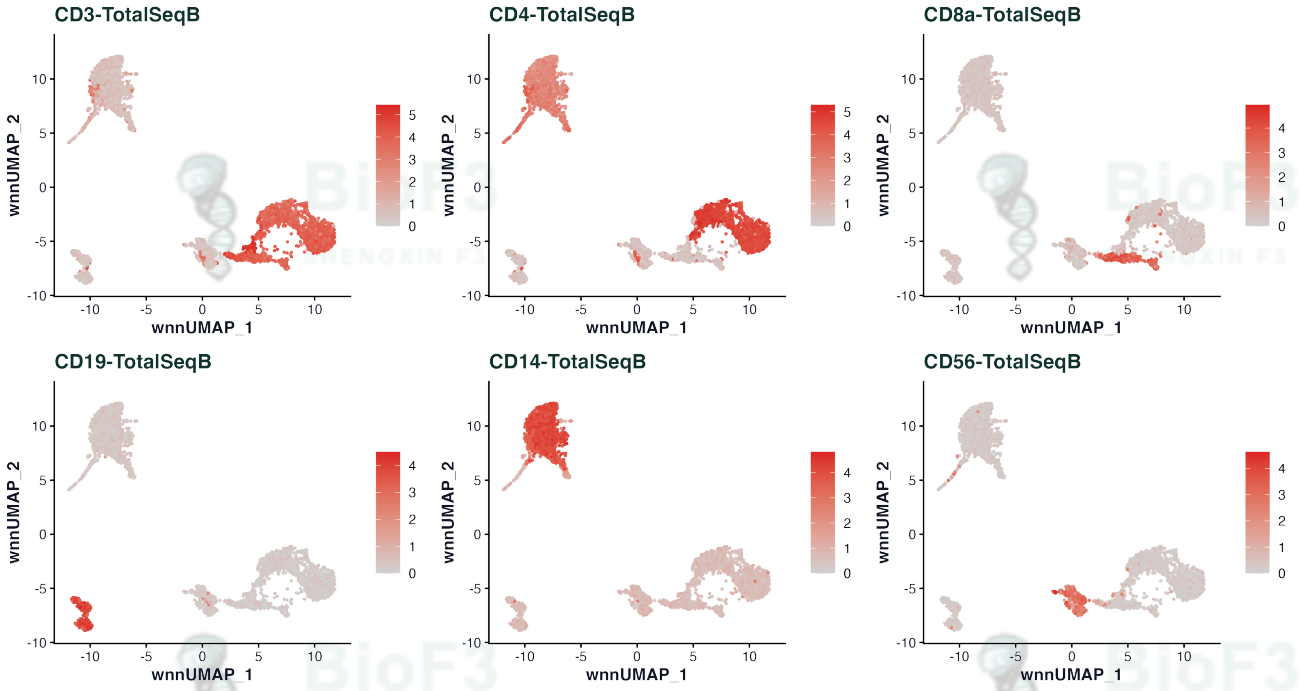

图 3：CD3、CD4、CD8、CD14、CD19、CD56 六个经典 marker 的 ADT 信号在 WNN UMAP 上的分布。$CD3^+$ 的大块就是 T 细胞，其内部 CD4 和 CD8 分得很干净；$CD19^+$ 对应 B 细胞；$CD14^+$ 对应单核；$CD56^+$ 对应 NK。不需要再额外做 marker 查询，这六张子图就能把 PBMC 的主要群体读出来。

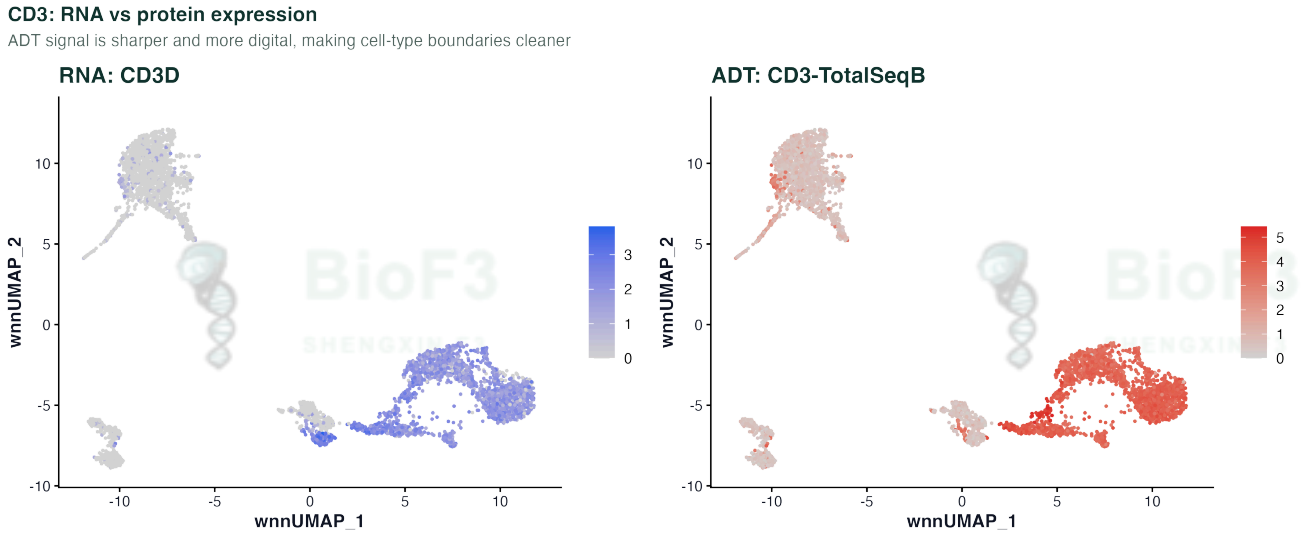

图 4：同一 WNN UMAP 上 `CD3D`（RNA 层）和 `CD3-TotalSeqB`（ADT 层）的表达对比。RNA 信号稀疏、有不少落在 T 细胞区的细胞读数为 0；ADT 信号连续、集中，边界非常锐利。这是 CITE-seq 最直接的价值：抗体信号稳定得多，做 T/非 T 的切分不会因为 dropout 误伤。

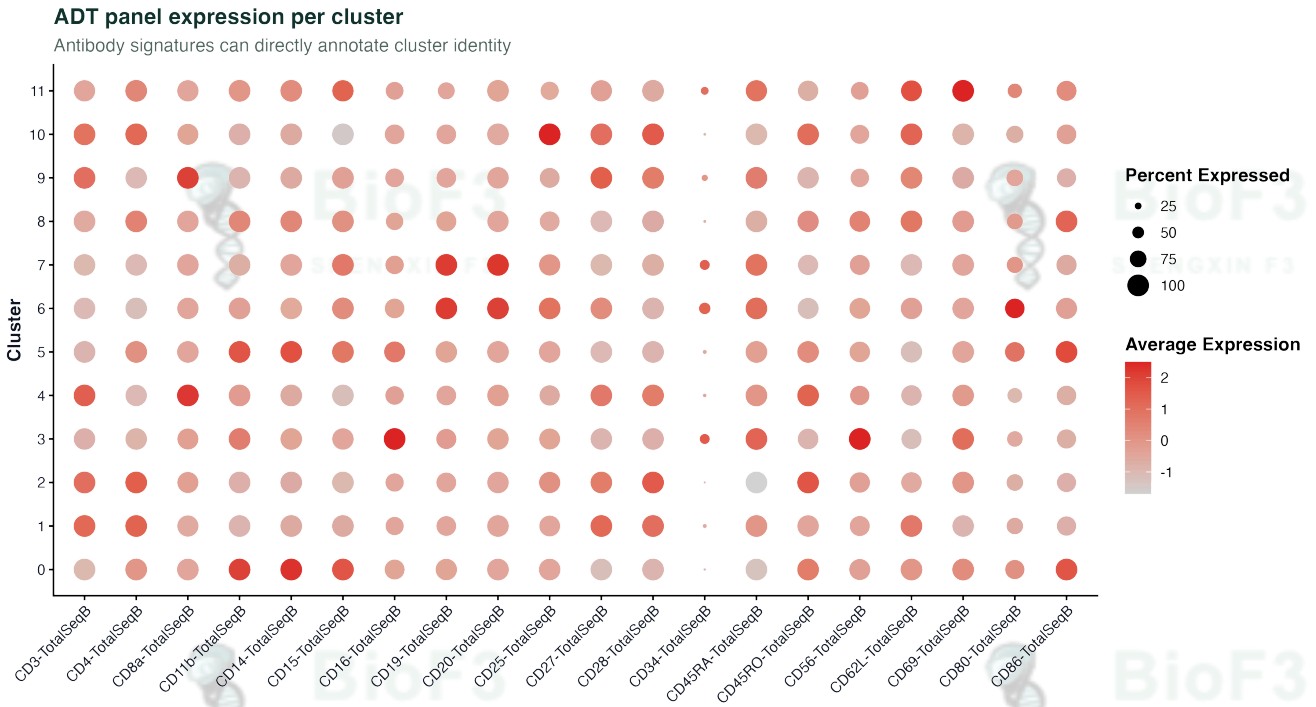

图 5：ADT panel 前 20 个抗体在每个 cluster 里的平均表达（颜色）和阳性细胞比例（点大小）。每个 cluster 对应的阳性抗体组合一看就懂：CD3/CD4/CD45RA 是 naive CD4 T，CD3/CD8/CD45RO 是 memory CD8 T，CD14/CD11b 是经典单核，CD19/CD20 是 B，依此类推。真实项目里直接看这张图就能把大部分 cluster 命名掉。

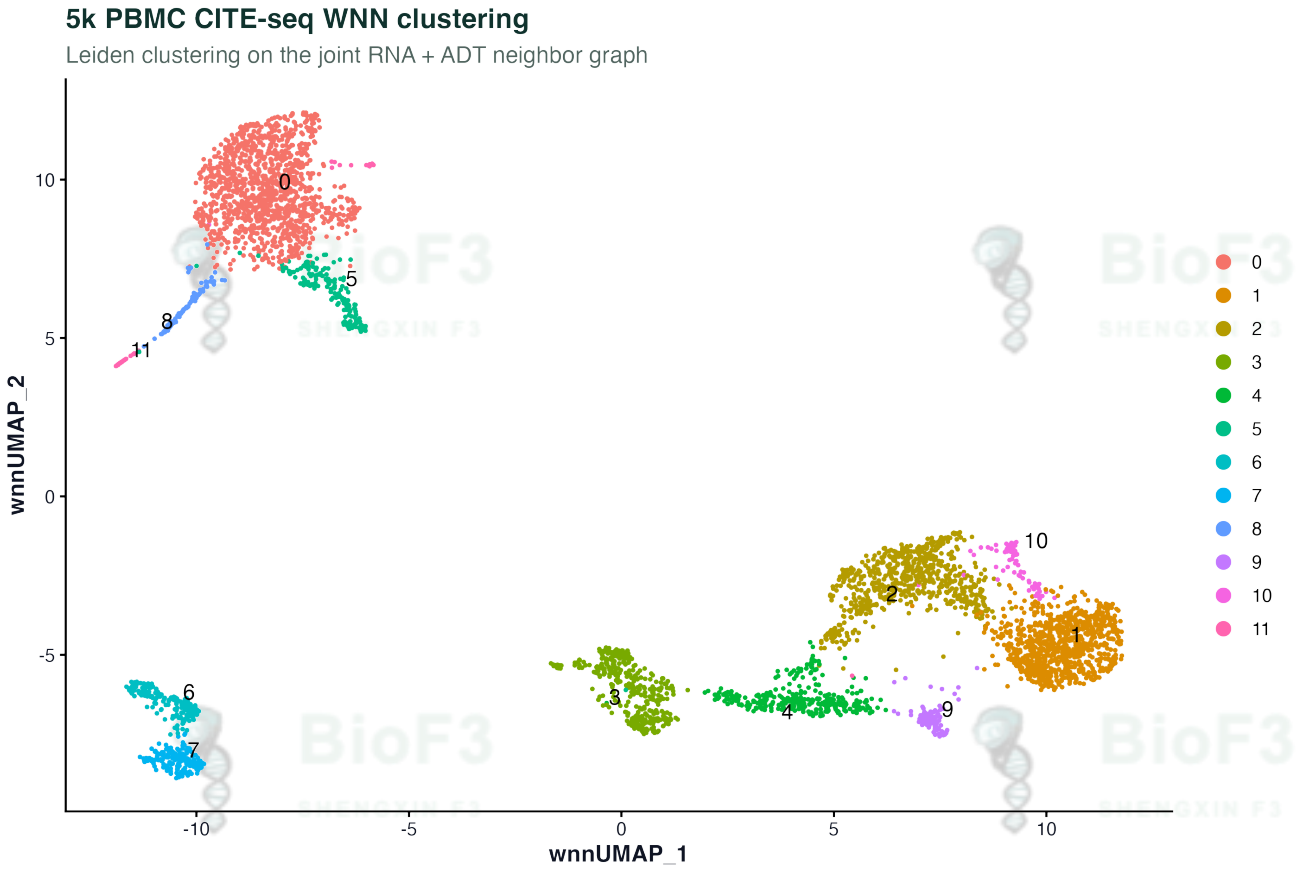

图 6：WNN UMAP 加聚类标签的完整视图。这张和单独做 RNA-only 的 UMAP 放在一起看差别最大：某些在 RNA 里贴得很近的群体（比如 CD4 memory 和 CD4 naive、CD8 effector 和 CD8 central memory）在 WNN 里通过 CD45RA/CD45RO 和 CD62L 的蛋白信号区分出来了。

套到自己数据上

脚本对 ADT 抗体名做了宽松匹配（`^CD3[-_]` 同时兼容 `CD3_TotalSeqB` 和 `CD3-TotalSeqB` 两种写法）。自己的 panel 只要还是 CD3/CD4/CD8/CD14/CD19/CD56 这类常规抗体名，直接跑通图 3 和图 4 的概率很高。panel 完全不同（比如肿瘤免疫的 checkpoint panel）的话，把 `adt_markers` 里的 CD 改成自己关心的抗体即可。`nfeatures` 、`dims.list` 、`resolution` 这三个参数按样本量调，其余不用动。

后续可做的分析

- **按蛋白 marker 重新注释聚类**：PBMC 的 ADT panel 一般包含 CD3/CD4/CD8/CD19/CD14 等经典标志，比起光靠 RNA 标记，注释会更可靠。
- **找差异蛋白**：`FindMarkers(cbmc, assay = "ADT", ...)` 可以直接在 ADT 层做差异分析。
- **与原章节对比**：把这份数据的 WNN 聚类结果和 04 章里只做 RNA 的整合结果画到一起，能看到蛋白信号如何修正 RNA-only 的分群。

10x Multiome（RNA + ATAC）分析思路类似，但 ATAC 层要用 Signac，归一化方法是 TF-IDF + SVD，后面在 10 scATAC-seq 分析 里会再讲。

下载资源

module08_cite_seq_sci.R
16 KB

下载 5k PBMC CITE-seq WNN 完整脚本 ↗

参考资源

- Seurat 多模态教程
- Seurat WNN 教程
- CITE-seq 官方网站
- 10x Genomics Multiome

扫码关注微信公众号【生信F3】
获取文章完整内容，分享生物信息学最新知识。