

PBMC 3k 单细胞 RNA 测序标准分析 流程技术报告

Standard scRNA-seq Analysis Pipeline for the 10x Genomics PBMC 3k Dataset

LitLit AI*

2026 年 8 月 3 日

摘要

本报告针对 10x Genomics 外周血单核细胞 (PBMC) 3k 数据集 (约 2,700 个细胞 $\times$ 约 32,738 个基因) 系统实施并完整记录了一套标准单细胞 RNA 测序 (scRNA-seq) 分析流程。流程涵盖质控 (QC)、对数归一化、高变基因 (HVG) 选择、主成分分析 (PCA)、 k 近邻图构建、Leiden 聚类与 UMAP 可视化, 并基于 Wilcoxon 秩和检验的标记基因与经典免疫标志物完成聚类注释。质控阶段按基因数 [200, 2500]、线粒体读占比 $< 5\%$ 的阈值剔除了 62 个低质量细胞, 保留 2,638 个细胞与 13,656 个表达基因; 归一化后选取 2,000 个高变基因并执行 50 维 PCA, 前 40 个主成分累计解释方差为 12.04%。在 $k=10$ 、 $n_{\text{pcs}}=40$ 的邻接图上以 Leiden (分辨率 0.5, 随机种子 42) 获得 6 个聚类, 对应六种免疫细胞类型: CD4⁺ T 细胞 (44.92%)、经典单核细胞 (18.12%)、CD8⁺ T 细胞 (16.22%)、B 细胞 (12.93%)、非经典单核细胞 (6.41%) 与树突状细胞 (1.40%)。分辨率敏感性分析 (0.3/0.5/0.7/1.0) 表明 0.5 在生物学合理性、最小簇规模与轮廓系数之间取得最佳折中, PCA 空间轮廓系数为 0.1970。结果验证了流程的可复现性, 并与人外周血免疫细胞图谱的既有认识一致。

关键词: 单细胞 RNA 测序; PBMC; Scanpy; Leiden 聚类; UMAP; 细胞类型注释; 标记基因

*联系邮箱: contact@litlit.ai

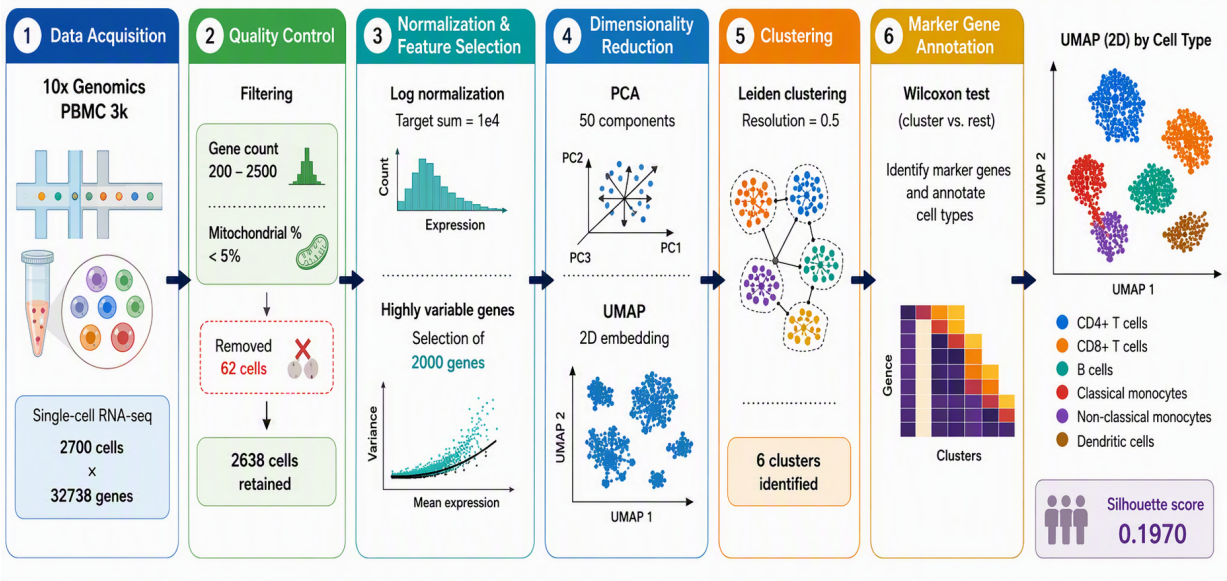


图 1 图形摘要。PBMC 3k scRNA-seq 标准分析流程：从数据获取、质控、归一化与高变基因选择、PCA 与 UMAP 降维、Leiden 聚类到基于标记基因的细胞类型注释，最终鉴定 6 种免疫细胞类型，PCA 空间轮廓系数为 0.1970。

目录

1	引言	3
2	数据与方法	3
2.1	数据来源	3
2.2	软件环境	4
2.3	分析流程	4
2.4	评估指标	5
3	结果	5
3.1	质控	5
3.2	归一化与高变基因选择	7
3.3	PCA 与 UMAP	8
3.4	Leiden 聚类与分辨率敏感性	9
3.5	细胞类型注释	11
3.5.1	标记基因表	14
3.6	标记基因表达在 UMAP 上的投影	16
3.7	细胞类型构成与综合评估	19
4	讨论	21
5	结论	22

1 引言

单细胞 RNA 测序 (single-cell RNA sequencing, scRNA-seq) 能够在单细胞分辨率下刻画基因表达谱, 从而揭示组织内细胞异质性、发现稀有细胞亚群并重建分化轨迹。自 Tang 等人首次实现单细胞全转录组测序以来 [1], 该领域在过去十余年经历了指数级增长 [2]。微流控与液滴条形码技术的成熟使通量从数百细胞跃升至数万乃至百万级细胞 [3, 4, 5], 其中 10x Genomics 的 Chromium 平台凭借稳定的高捕获率与商业化的可重复性, 已成为 droplet-based scRNA-seq 的事实标准之一 [5]。外周血单核细胞 (PBMC) 作为最易获取的人源免疫细胞样本, 是评估分析方法学、基准化工具与构建免疫细胞参考图谱的经典模型系统 [6, 7]。

然而, scRNA-seq 数据同时具有高维、稀疏、高噪声与强批次效应等特征, 其分析需经过一系列规范化的步骤才能得到可解释的细胞分群 [8, 9]。典型流程包括质控以剔除低质量细胞与双细胞 [10, 11]、归一化以消除细胞测序深度的尺度差异、高变基因选择以聚焦生物学信号 [12, 13]、降维以缓解维度灾难 [14]、图聚类以识别细胞群 [15, 16] 以及基于差异表达基因 (标记基因) 的细胞类型注释 [17]。Scanpy 框架 [15] 基于 Python 与 AnnData 数据结构, 提供了从质控到注释的一体化、可扩展的实现, 是目前使用最广泛的开源单细胞分析平台之一, 与 R 生态中的 Seurat 套件 [18, 19, 20, 21] 形成互补。

聚类是 scRNA-seq 流程的核心环节。基于图的社区检测算法 (如 Louvain 及其改进版 Leiden) 在单细胞邻接图上表现优异, 能够在保证社区连通性的同时高效处理大规模稀疏图 [16, 15]。Leiden 算法通过对划分的精细局部移动与细分 (refinement), 避免了 Louvain 算法中常见的断开社区 (disconnected community) 问题 [16]。聚类数目的确定高度依赖分辨率参数, 因此通常需对分辨率进行敏感性扫描并结合生物学先验加以权衡 [22, 8]。在降维与可视化方面, UMAP [23, 24] 因其在保留全局结构与计算效率上的优势, 已在很大程度上取代 t-SNE [25] 成为单细胞数据二维投影的首选。

细胞类型注释可基于已知标记基因的人工判读、参考图谱映射或差异表达统计检验完成 [7, 9]。对于 PBMC, 经典标记如 CD3D/CD3E 标识 T/NK 谱系、CD79A/MS4A1 标识 B 细胞、LYZ/CD14/S100A9 标识经典单核细胞、FCGR3A/MS4A7 标识非经典单核细胞、NKG7/GNLY 标识 NK、FCER1A/CLEC10A 标识树突状细胞, 已被多项研究反复验证 [6, 7, 20]。Wilcoxon 秩和检验作为非参数方法, 对计数数据的非正态性稳健, 是 Scanpy 默认且被基准测试认可的标记基因检验方法 [17, 15]。

本报告以 10x Genomics PBMC 3k 数据集为对象, 完整实施上述标准流程并系统记录各阶段参数与产出, 旨在为同行提供一份可复现、可审计的端到端分析参考。报告其余部分组织如下: 第 2 节描述数据与方法; 第 3 节呈现结果; 第 4 节进行讨论; 第 5 节给出结论。

2 数据与方法

2.1 数据来源

分析对象为 10x Genomics 公开提供的 PBMC 3k 数据集 (v1 化学), 通过 `scanpy.datasets.pbmc3k()` 自动获取已过滤的 feature-barcode 矩阵。原始数据包含约 2,700 个细胞与 32,738 个基因。该数据集是 Scanpy [15] 与 Seurat [20] 官方教程中长期使用的基准数据集, 已被广泛用于方法学评估与教学演示, 其免疫细胞构成已被多项研究刻画 [6, 7]。

2.2 软件环境

全部计算在 Linux 平台 (Linux-5.4.145, x86_64, glibc 2.41) 完成, Python 3.13.5。核心依赖及版本如下 (来源: `results/environment_versions.txt`): Scanpy 1.12.3 [15]、AnnData 0.13.2、NumPy 2.4.6、pandas 3.0.5、scikit-learn 1.9.0、matplotlib 3.11.1、seaborn 0.13.2。Leiden 聚类经由 `leidenalg` (Scanpy 后端) 实现 [16], 轮廓系数与 PCA 经由 `scikit-learn` [14] 计算。所有涉及随机性的步骤均固定随机种子为 42 以保证可复现性。

2.3 分析流程

分析按六个顺序步骤实施, 每步产出经校验的中间文件与汇总 (见 `results/` 目录)。流程参数概览见表 1 (对应 `results/pipeline_summary.json`)。

表 1 PBMC 3k 标准分析流程关键参数汇总 (`pipeline_summary.json`)。

参数	取值
数据来源	<code>scanpy.datasets.pbmc3k()</code> (10x Genomics PBMC 3k, v1)
QC 阈值	基因数 $\in [200, 2500]$; 线粒体读占比 $< 5\%$
归一化	<code>normalize_total(target_sum=1e4) + log1p</code>
高变基因	<code>seurat</code> flavor, $n_{\text{top}} = 2000$
回归与缩放	协变量 <code>[total_counts, pct_counts_mt]</code> , <code>scale(max_value=10)</code>
PCA	$n_{\text{comps}} = 50$, <code>random_state=42</code>
邻接图	$k=10$, $n_{\text{pcs}} = 40$, <code>random_state=42</code>
Leiden 聚类	分辨率 0.5, <code>random_state=42</code>
UMAP	<code>random_state=42</code>
标记检验	Wilcoxon 秩和, <code>use_raw=True</code> , $n_{\text{genes}} = 100$
敏感性扫描	分辨率 $\{0.3, 0.5, 0.7, 1.0\}$
种子	42

步骤 1: 数据获取与质控。 以 MT- 前缀识别线粒体基因 (共检出 13 个), 计算每个细胞的 `n_genes_by_counts`、`total_counts` 与 `pct_counts_mt`。按基因数 $[200, 2500]$ 与线粒体读占比 $< 5\%$ 的双重阈值剔除低质量细胞, 随后剔除零细胞基因 [8, 9]。

步骤 2: 归一化与高变基因选择。 先以 `filter_genes(min_cells=3)` 剔除低覆盖基因, 再执行 `normalize_total(target_sum=1e4)` 与 `log1p` 对数归一化 [15]; 将归一化后的完整基因集保存至 `adata.raw` 以供后续标记检验。随后以 `seurat` flavor 选取前 2,000 个高变基因 [12], 对 `total_counts` 与 `pct_counts_mt` 进行协变量回归并 `scale(max_value=10)` 截断缩放 [26, 8]。

步骤 3: 降维、聚类与 UMAP。 在缩放后的高变基因矩阵上执行 50 维 PCA [14]; 基于前 40 个主成分构建 $k=10$ 的最近邻图; 在该图上运行 Leiden 聚类 (分辨率 0.5) [16, 15]; 以 UMAP [23, 24] 将细胞投影至二维。轮廓系数在 PCA 前 40 个主成分上、对 $n = 1,000$ 的子样本计算 [14]。

步骤 4：分辨率敏感性分析。 在复用的邻接图上以分辨率 {0.3, 0.5, 0.7, 1.0} 重新执行 Leiden，记录各分辨率下的簇数、最小/最大簇规模与轮廓系数，并计算成对 ARI 与 NMI [8]。综合准则为：簇数落在 PBMC 的生物学合理区间（5–9）、无微簇（最小簇 ≥ 30 细胞）、幸存者中轮廓系数最高 [8]。

步骤 5：标记基因分析与注释。 以 Wilcoxon 秩和检验 [17, 15] 在 `adata.raw`（完整对数归一化基因集）上逐簇进行差异检验，取每簇前 10 个标记基因。注释综合经典标记（CD3D/CD3E、CD79A、LYZ、NKG7、MS4A1、CD14、FCER1A 等）与扩展标记（CD4/CD8A 区分 T 亚群；FCGR3A/MS4A7 区分单核细胞亚群；GNLY/NCAM1 精修 NK） [6, 7]，并基于谱系签名分数定量佐证。

步骤 6：最终评估与汇总。 聚合全流程指标、重算并确认 PCA 空间轮廓系数、生成综合可视化与汇总工件。

2.4 评估指标

报告四类评估指标：(i) 簇数目；(ii) 每簇基因数中位数（衡量簇内细胞检测深度）；(iii) 标记基因表（Wilcoxon，每簇前 10）；(iv) PCA 空间轮廓系数 [14]，作为图聚类的补充诊断指标。

3 结果

3.1 质控

原始数据为 2,700 个细胞 $\times$ 32,738 个基因。剔除基因数落在 [200, 2500] 之外或线粒体读占比 $\geq 5\%$ 的细胞后，共移除 62 个细胞（其中基因数阈值剔除 5 个、线粒体阈值剔除 57 个），保留 **2,638** 个细胞；剔除零细胞基因后保留 **16,579** 个基因（表 2、图 2 与图 3）。质控后 `n_genes_by_counts` 中位数为 820（均值 850.09，范围 212–2455），`total_counts` 中位数为 2,214（均值 2,372.58），`pct_counts_mt` 中位数为 2.01%（均值 2.12%，最大 4.99%，严格落于阈值之内）。线粒体读占比的严格上限过滤有效移除了潜在死细胞或应激细胞，而基因数双向阈值排除了空液滴与高检测深度（疑似双细胞）的离群细胞 [8, 9]。

表 2 质控前后细胞与基因数及质控指标统计（`results/qc_statistics.csv`）。

指标	均值	中位数	最小	最大	标准差
<code>n_genes_by_counts</code>	850.09	820.00	212.00	2455.00	263.37
<code>total_counts</code>	2372.58	2214.00	561.00	8931.00	990.84
<code>pct_counts_mt</code>	2.12	2.01	0.00	4.99	0.84

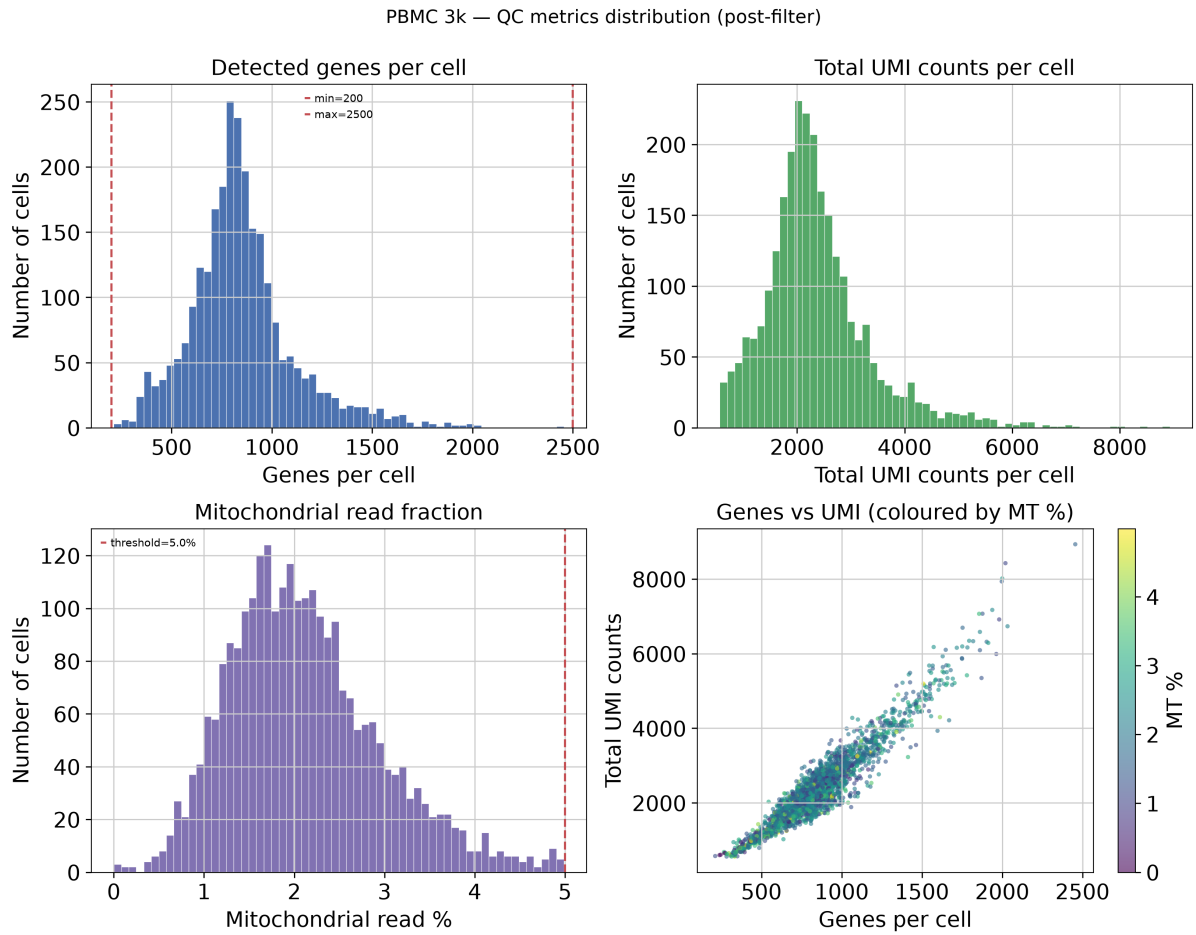


图 2 质控指标分布。左上：每细胞基因数；右上：每细胞总计数；下方：线粒体读占比。蓝色曲线为核密度估计，红色虚线标示质控阈值（基因数 200/2500、线粒体 5%）。质控后所有细胞均落于阈值内。

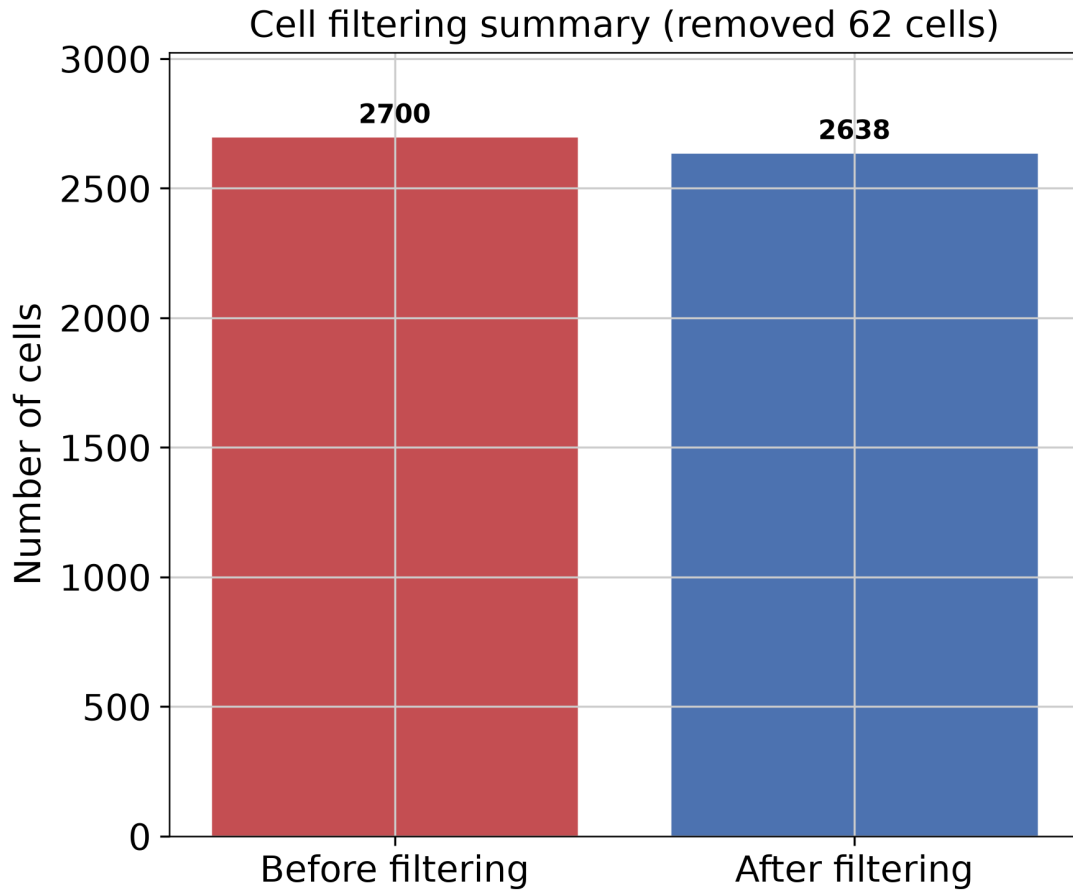


图 3 质控过滤汇总。展示原始/保留细胞数、各阈值剔除细胞数及其并集，以及质控前后关键指标的对比。

3.2 归一化与高变基因选择

基因级过滤 (`min_cells=3`) 将基因数由 16,579 降至 **13,656** (剔除 2,923 个低覆盖基因)。对数归一化后，以 `seurat` flavor 选取 **2,000** 个高变基因 (图 4)。协变量回归与缩放后的最终分析矩阵为 $2,638 \times 2,000$ ；`adata.raw` 保留 $2,638 \times 13,656$ 的完整对数归一化基因集，用于后续标记检验，避免缩放与回归对差异表达统计的扰动 [15, 17]。

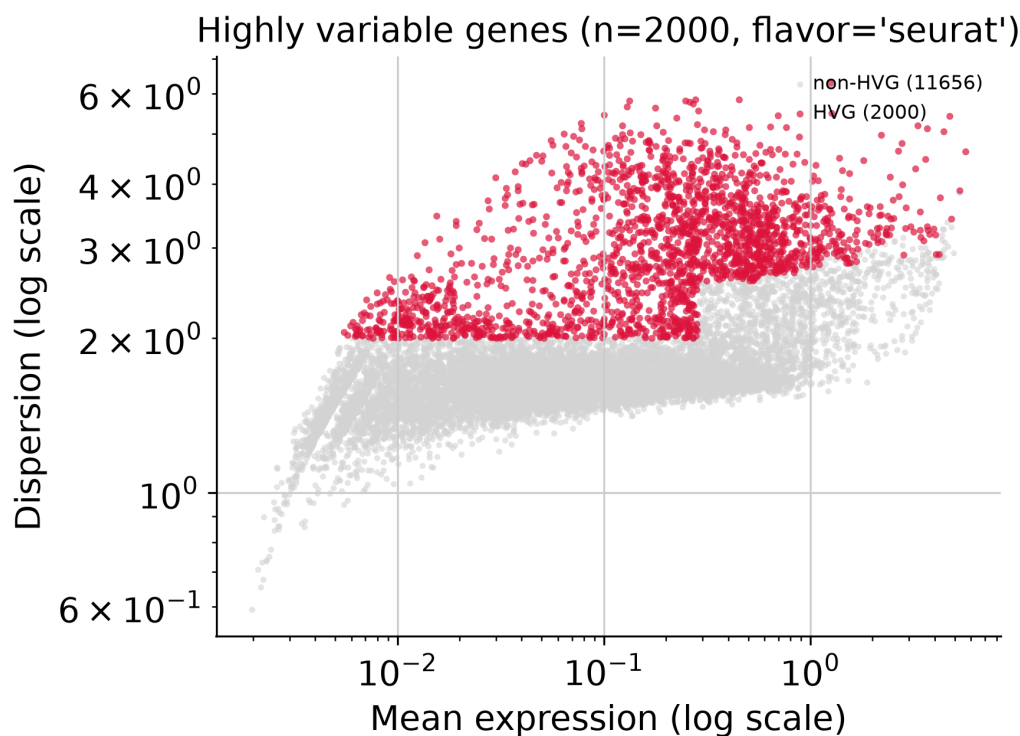


图 4 高变基因选择结果。红点为选定的 2,000 个高变基因，横轴为平均表达水平，纵轴为离散度。`seurat` flavor 在对数归一化数据上运行。

3.3 PCA 与 UMAP

50 维 PCA 的方差解释呈典型快速衰减：第 1 主成分解释 2.35% 方差，前 4 个主成分累计 5.49%，前 40 个主成分累计 **12.04%**，全部 50 个主成分累计 13.65%（图 5a、表对应 `results/pca_variance_ratio.csv`）。前 4 个主成分承载了主要的生物学信号，而后续主成分的边际增益趋缓，与单细胞数据方差分散、需以较多主成分重建结构的特征一致 [14]。PCA 散点图（图 5b）显示在 PC1/PC2 平面上可观察到初步的细胞分群结构。

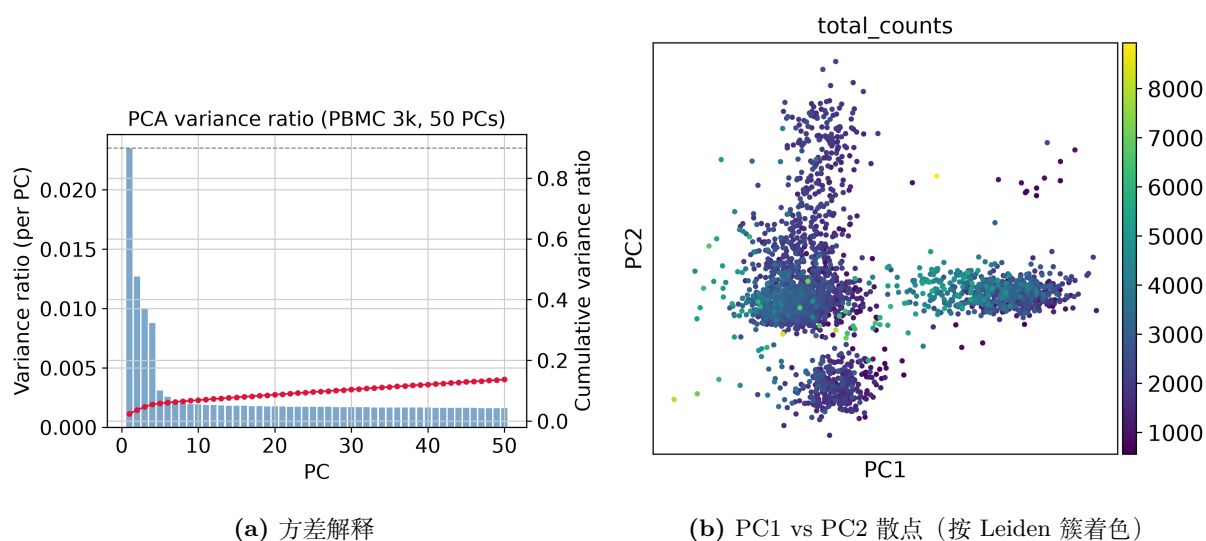


图 5 PCA 结果。左：各主成分方差解释比与累计方差；右：前两个主成分散点图，色标对应 Leiden 聚类（分辨率 0.5）。

UMAP 二维投影 (图 6a) 将六个 Leiden 簇清晰分离。以 `total_counts` 与 `pct_counts_mt` 着色的 UMAP (图 7) 表明这些技术协变量未在投影空间形成系统性偏倚, 提示协变量回归与质控有效抑制了技术噪声对分群的影响 [8, 9]。

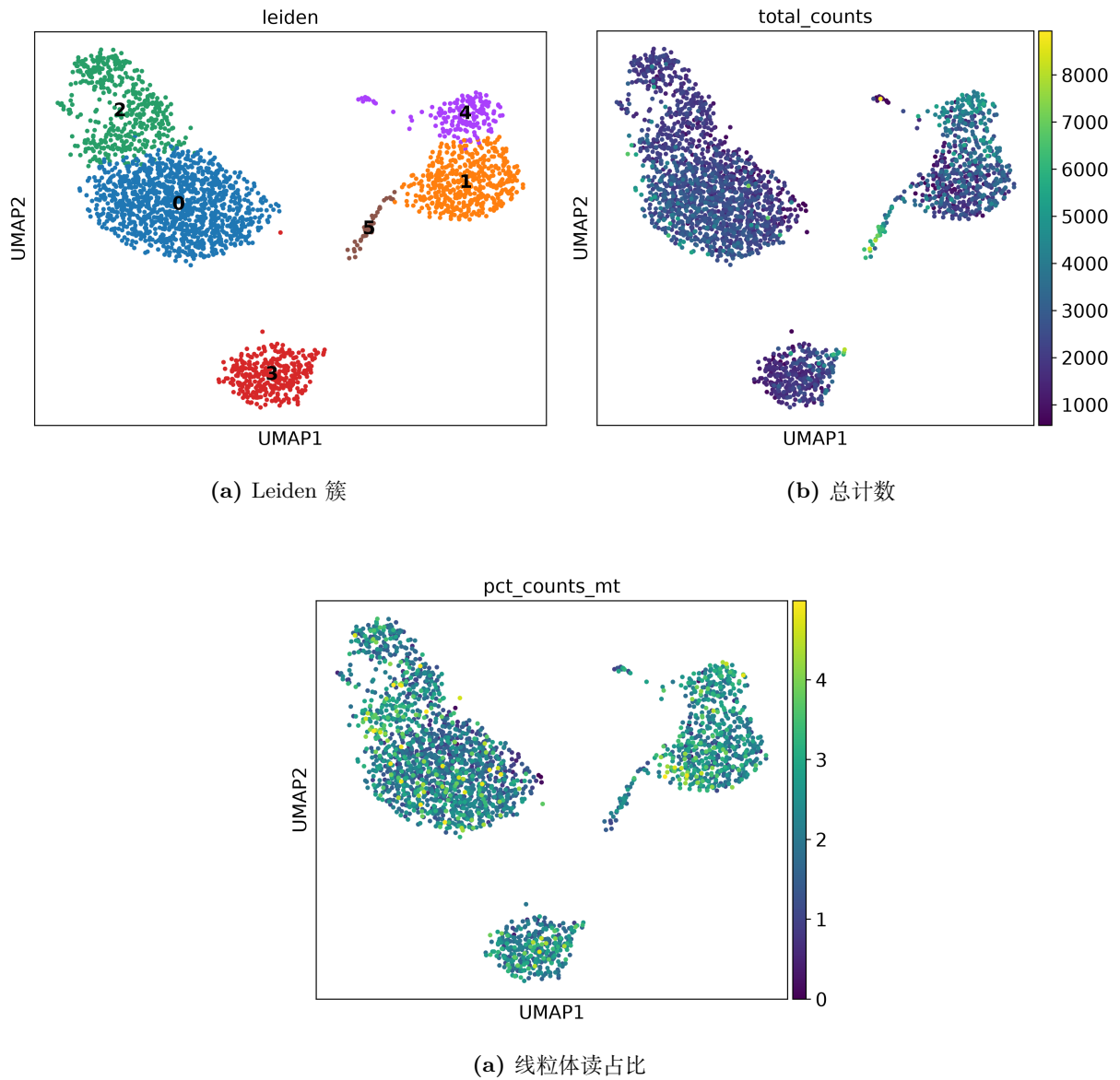


图 7 UMAP 投影结果。左上: 六个 Leiden 聚类; 右上: 以总计数着色; 左下: 以线粒体读占比着色。技术协变量在投影空间未见系统性聚集。

3.4 Leiden 聚类与分辨率敏感性

在分辨率 0.5 下获得 **6 个 Leiden 簇**, 规模自 37 至 1,185 个细胞不等 (表 3)。分辨率敏感性分析显示簇数随分辨率单调递增 (0.3→5 簇, 0.5→6 簇, 0.7→7 簇, 1.0→9 簇), 而轮廓系数总体下降 (0.2133、0.1970、0.1889、0.0843, 图 8a 与图 8b) [16, 14]。除 0.5 外各分辨率均产生 12–13 个细胞的微簇 (`min_cluster_size` 12–13), 被可靠性判据排除。成对 ARI/NMI 矩阵 (表 4, 图 9b) 显示相邻分辨率间高度一致 (0.5 与 0.7 的 ARI 达 0.981、NMI 达 0.970), 表明划分稳定; 而与 1.0 的差异较大 (ARI 0.66–0.73), 提示高分辨率下出现结构重组。综合生物学合理性、无微簇与最高轮廓系数三项准则, 0.5 **被推荐为最优分辨率**, 其同时复现了第 3 步的标准

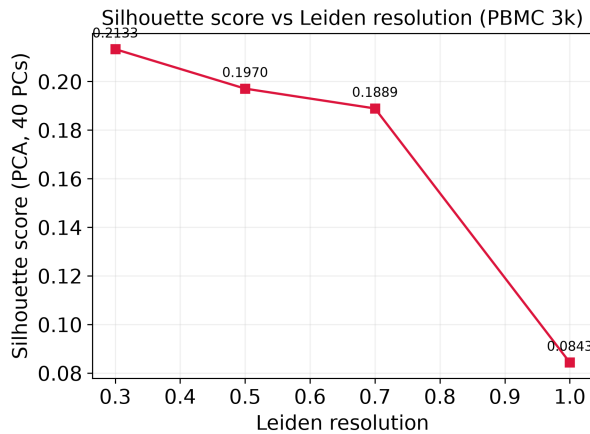
PBMC 3k 划分 [8]。

表 3 分辨率 0.5 下各 Leiden 簇的细胞数与占比 (results/cluster_sizes.csv)。

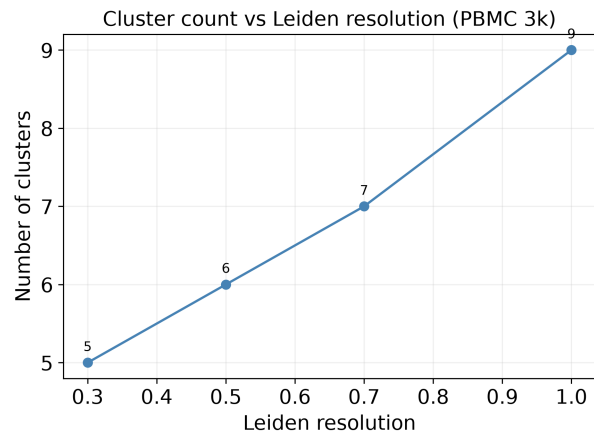
簇	细胞数	占比 (%)
0	1185	44.92
1	478	18.12
2	428	16.22
3	341	12.93
4	169	6.41
5	37	1.40

表 4 成对 ARI (下三角) 与 NMI (上三角) 矩阵 (results/ari_matrix.csv、nmi_matrix.csv)。

	0.3	0.5	0.7	1.0
0.3	1.000 / 1.000	0.903	0.897	0.774
0.5	0.914	1.000	0.970	0.836
0.7	0.902	0.981	1.000	0.842
1.0	0.659	0.728	0.737	1.000



(a) 轮廓系数 vs 分辨率



(b) 簇数 vs 分辨率

图 8 分辨率敏感性。左：轮廓系数随分辨率下降；右：簇数随分辨率上升。分辨率 0.5 在无微簇的候选中轮廓系数最高。

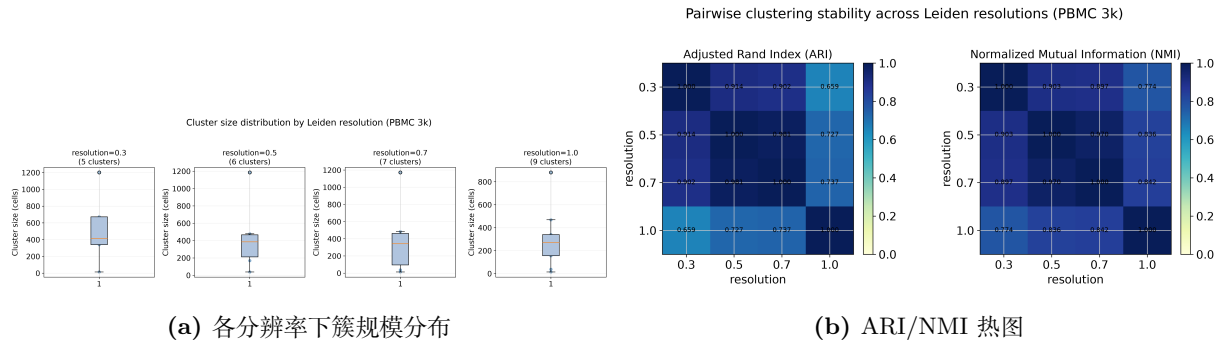


图 9 分辨率敏感性 (续)。左: 各分辨率下的簇规模箱线分布, 红色虚线标示 30 细胞可靠性阈值; 右: 成对 ARI 与 NMI 热图。

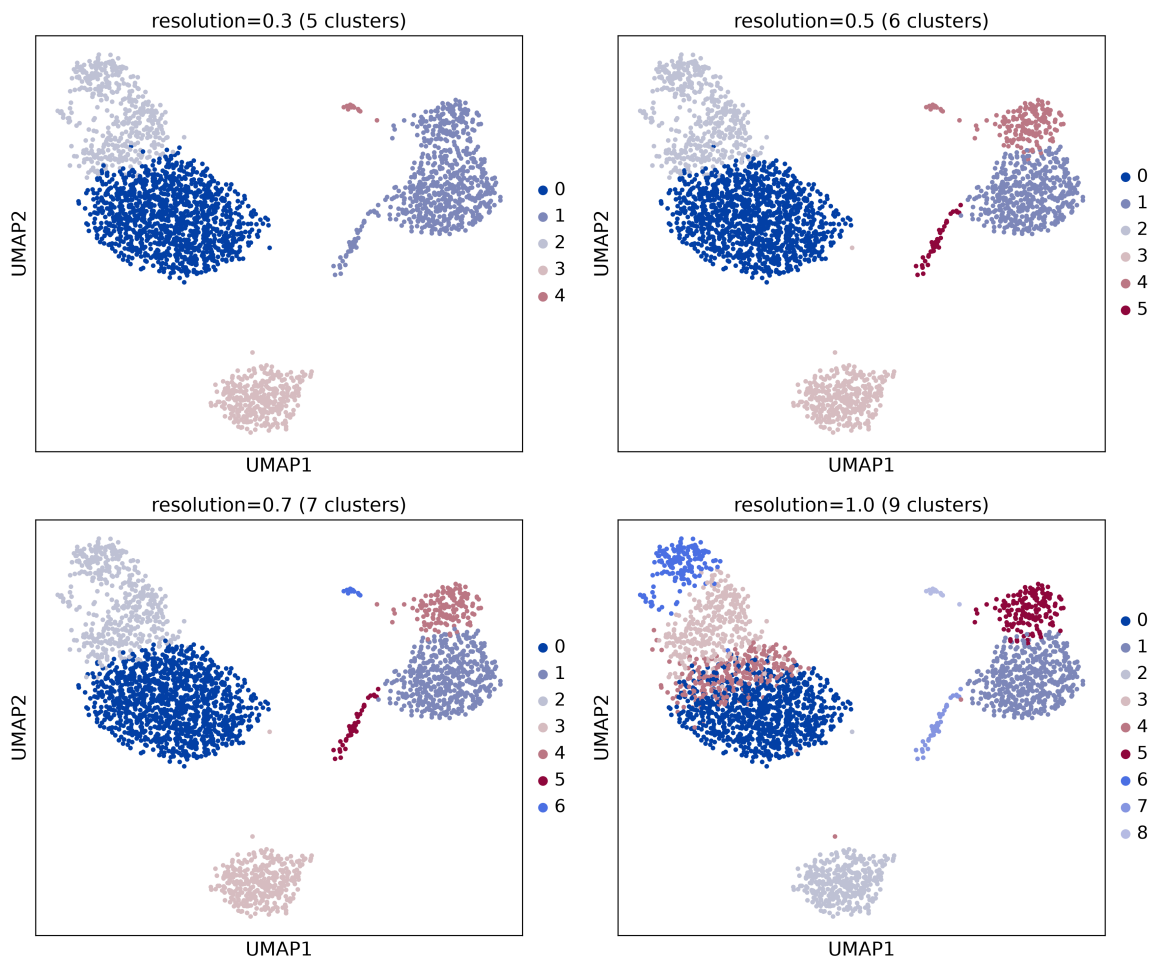


图 10 多分辨率 UMAP 对比。分辨率 0.3/0.5/0.7/1.0 下的 UMAP 投影并排展示, 可见随分辨率升高小簇不断从大簇中分裂出来。

3.5 细胞类型注释

基于 Wilcoxon 秩和检验的标记基因 (表 6、图 11) 与经典/扩展标记的点图 (图 12a、图 12b、图 13a), 六个簇被注释为六种免疫细胞类型 (表 5、图 13b)。注释依据经典标志物的簇间差异表达与谱系签名分数定量佐证 [6, 7]。

表 5 细胞类型注释汇总 (results/cell_type_summary_table.csv)。

簇	细胞类型	前 5 标记	细胞数	基因中位数	置信度
0	CD4 ⁺ T 细胞	LDHB, RPS12, RPS25, RPS27, RPS3	1185	812.0	高
1	经典单核细胞	LYZ, S100A9, S100A8, TYROBP, FTL	478	875.0	高
2	CD8 ⁺ T 细胞	NKG7, CST7, GZMA, B2M, CTSW	428	834.5	高
3	B 细胞	CD74, CD79A, HLA-DRA, CD79B, HLA-DPB1	341	676.0	高
4	非经典单核细胞	COTL1, LST1, SAT1, FTH1, FCER1G	169	1238.0	高
5	树突状细胞	HLA-DPB1, HLA-DPA1, HLA-DRA, HLA-DRB1, CD74	37	1567.0	中

簇 0: CD4⁺ T 细胞 (高置信度)。 标记基因富集核糖体蛋白 (RPS 家族) 与 LDHB, 并伴 CD3D (logFC +3.29)、CD3E (T 谱系), CD4/IL7R/CCR7/LEF1 签名均值 (0.81) 高于 CD8 效应签名 (0.22), 与辅助性 CD4⁺ T 细胞一致 [6, 7]。

簇 1: 经典单核细胞 (高置信度)。 LYZ (logFC +6.22)、S100A9 (+7.35)、S100A8 (+7.57)、TYROBP、FCN1、CST3 高表达, CD14/LYZ/S100A9 均值 3.70, 是典型的 CD14⁺ 经典单核细胞特征 [6, 7]。

簇 2: CD8⁺ T 细胞 (高置信度)。 NKG7 (+7.14)、CST7、GZMA、CTSW、CCL5、PRF1、GZMM 等细胞毒效应分子高表达, CD8A/B/GZMK/GZMA/PRF1 签名 (1.32) 高于 CD4/IL7R/CCR7 (0.37), CD3D/E 同时为阳性, 标识为效应/记忆型 CD8⁺ T 细胞 [6]。

簇 3: B 细胞 (高置信度)。 CD74、CD79A (+7.76)、CD79B、MS4A1 (+6.42)、HLA-DRA/DPB1/DQA1/DQB1/DRB1 等 MHC-II 分子高表达, CD3D 低表达 (0.15), CD79A/MS4A1/CD79B 均值 2.34, 确凿为 B 细胞 [6, 7]。

簇 4: 非经典单核细胞 (高置信度)。 FCGR3A (+5.16)、COTL1、LST1、SAT1、FCER1G、AIF1、IFITM3 高表达, FCGR3A/MS4A7 均值 2.11 高于 CD14/LYZ (1.78), 符合 CD16⁺ 非经典 (中间型) 单核细胞 [6]。

簇 5: 树突状细胞 (中置信度)。 HLA-DPB1/DPA1/DRA/DRB1、CD74、CST3、FCER1A (+8.81) 高表达, FCER1A/CST3/CLEC10A 均值 2.58, 标识为树突状细胞/经典树突状细胞 (cDC); 因簇规模较小 (37 细胞) 置信度评为中等 [6, 7]。

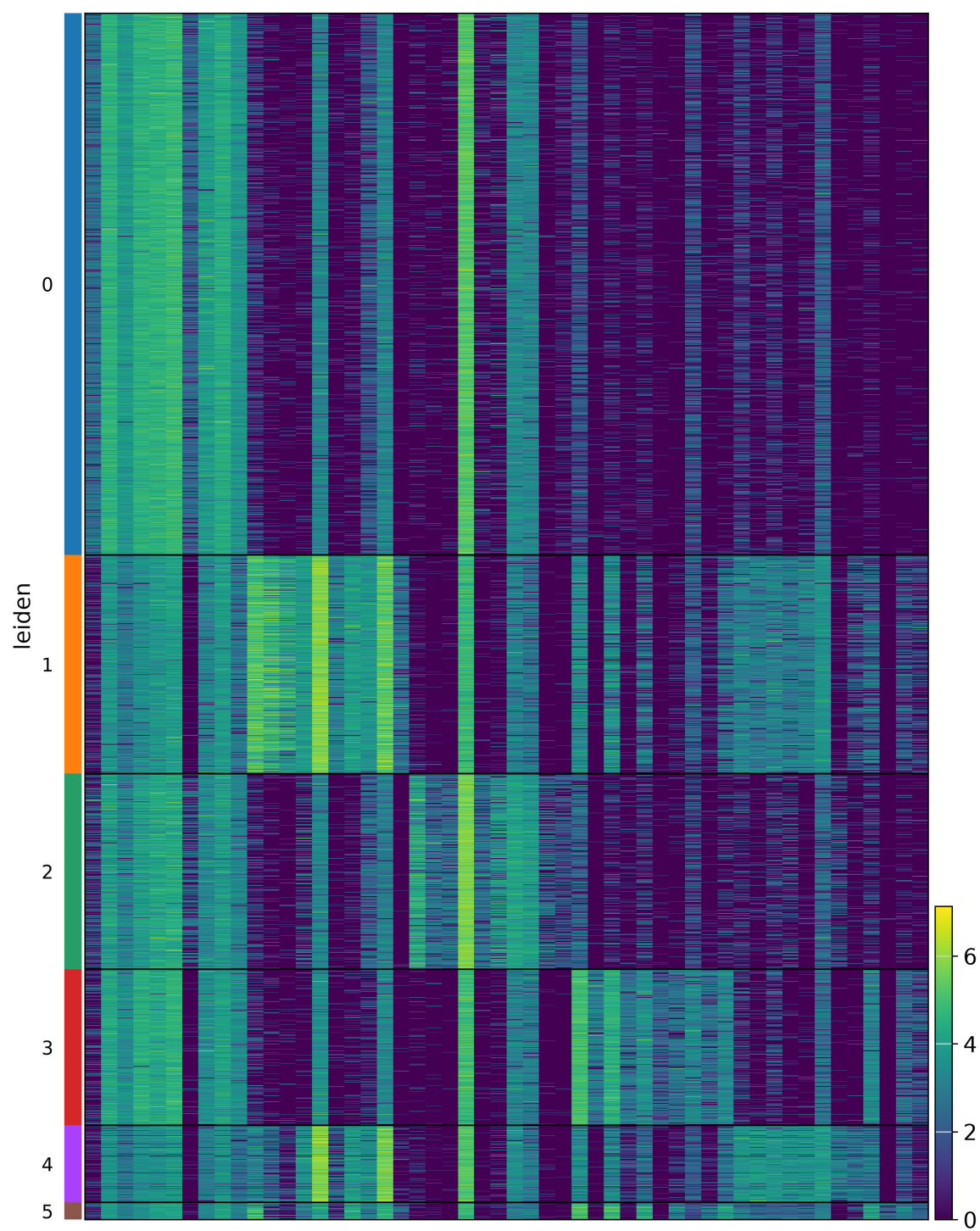


图 11 各簇前 10 标记基因热图（按簇分组，行中心化缩放）。可见各簇标记基因呈现明显的簇特异性高表达模式。

Classic marker genes by Leiden cluster

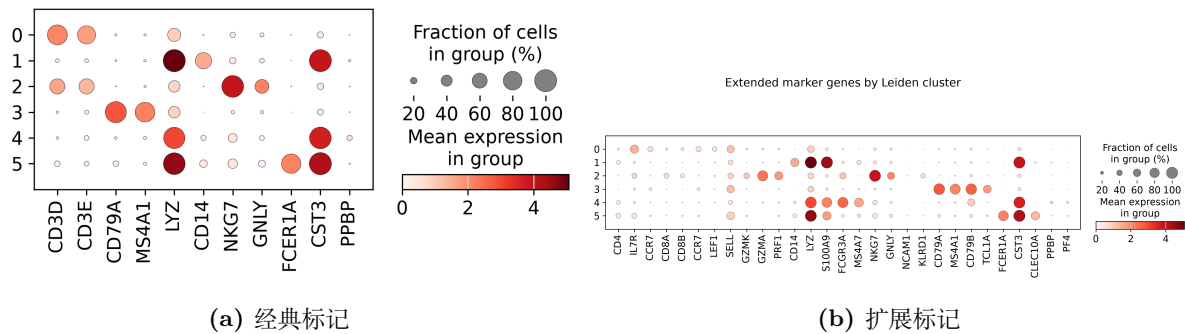


图 12 标记基因点图。点大小表示表达该基因的细胞比例，颜色表示平均表达水平。左：经典标记基因集；右：扩展标记基因集（含 CD4/CD8A、FCGR3A/MS4A7、GNLY/NCAM1 等）。

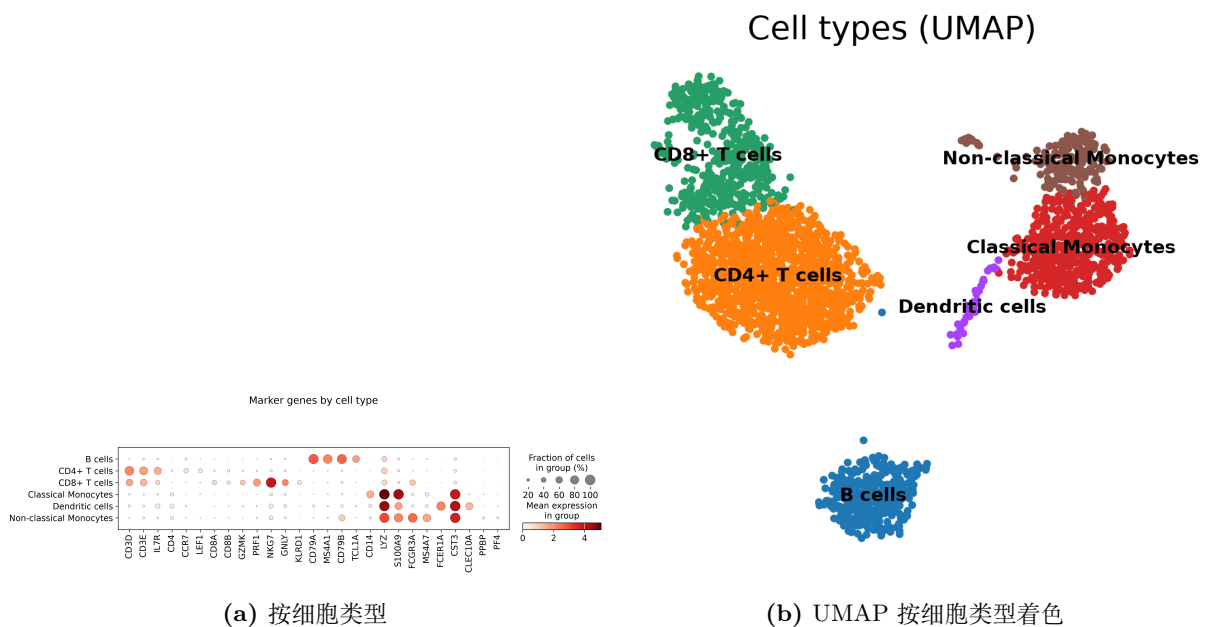


图 13 注释结果可视化。左：按注释细胞类型分组的标记点图；右：UMAP 按细胞类型着色，六种细胞类型在投影空间上清晰分离。

3.5.1 标记基因表

表 6 给出每簇前 10 标记基因的 Wilcoxon 检验统计量(来源:results/marker_genes_top10.csv)。所有标记均经 Benjamini-Hochberg 校正后 $p_{adj} \ll 10^{-12}$ ，统计学显著 [17]。

表 6 各簇前 10 标记基因 (Wilcoxon 秩和检验, use_raw=True)。

簇	排名	基因	得分	logFC	$-\log_{10}(p_{adj})$
0	1	LDHB	32.68	2.77	229.34
0	2	RPS12	32.35	1.06	225.95
0	3	RPS25	30.37	1.17	198.70

续下页

簇	排名	基因	得分	logFC	$-\log_{10}(p_{\text{adj}})$
0	4	RPS27	29.62	1.03	188.80
0	5	RPS3	29.46	0.89	186.77
0	6	RPS6	29.22	0.87	183.87
0	7	CD3D	28.47	3.29	174.43
0	8	TPT1	28.34	0.96	172.85
0	9	RPS14	28.09	0.78	170.07
0	10	RPL31	27.74	1.13	165.74
1	1	LYZ	33.67	6.22	243.68
1	2	S100A9	33.50	7.35	241.56
1	3	S100A8	32.89	7.57	233.12
1	4	TYROBP	31.73	5.49	216.74
1	5	FTL	31.25	3.49	210.24
1	6	FCN1	31.16	5.68	209.05
1	7	CST3	30.97	5.70	206.63
1	8	S100A6	30.34	2.99	198.30
1	9	FTH1	30.28	2.86	197.47
1	10	LGALS2	30.23	6.59	196.98
2	1	NKG7	31.50	7.14	212.67
2	2	CST7	28.02	6.40	168.24
2	3	GZMA	27.33	6.12	160.10
2	4	B2M	26.82	0.98	154.19
2	5	CTSW	26.70	4.63	153.04
2	6	CCL5	26.13	5.41	146.39
2	7	HLA-C	25.76	1.34	142.29
2	8	HLA-A	22.78	1.26	110.91
2	9	PRF1	20.74	5.60	91.34
2	10	GZMM	19.13	3.46	77.71
3	1	CD74	28.89	4.08	178.70
3	2	CD79A	27.84	7.76	166.03
3	3	HLA-DRA	27.54	4.89	162.61
3	4	CD79B	26.47	5.52	150.12
3	5	HLA-DPB1	25.84	4.05	143.08
3	6	MS4A1	25.18	6.42	135.92
3	7	HLA-DQA1	25.18	5.35	135.92
3	8	CD37	24.37	2.50	127.10
3	9	HLA-DQB1	24.31	4.98	126.58

续下页

簇	排名	基因	得分	logFC	$-\log_{10}(p_{\text{adj}})$
3	10	HLA-DRB1	24.27	3.89	126.25
4	1	COTL1	18.94	3.46	74.90
4	2	LST1	18.83	4.79	74.52
4	3	SAT1	18.72	3.62	74.17
4	4	FTH1	18.64	2.85	73.71
4	5	FCER1G	18.62	4.47	73.73
4	6	AIF1	18.20	4.47	69.93
4	7	OAZ1	17.93	2.12	68.16
4	8	FCGR3A	17.54	5.16	64.91
4	9	FTL	17.44	2.99	64.72
4	10	IFITM3	17.29	4.44	63.17
5	1	HLA-DPB1	9.44	4.41	15.31
5	2	HLA-DPA1	9.33	4.19	15.13
5	3	HLA-DRA	9.10	4.57	14.37
5	4	HLA-DRB1	9.00	4.18	14.15
5	5	CD74	8.99	3.62	14.15
5	6	HLA-DQA1	8.96	4.78	14.14
5	7	CST3	8.92	4.89	14.02
5	8	FCER1A	8.72	8.81	13.31
5	9	HLA-DRB5	8.54	3.87	12.69
5	10	HLA-DMA	8.03	3.49	11.11

3.6 标记基因表达在 UMAP 上的投影

图 14-18 展示了关键标记基因在 UMAP 空间的表达分布，直观印证了上述注释：T 谱系标记 CD3D/CD3E/NKG7 富集于簇 0 与簇 2，B 细胞标记 CD79A/MS4A1 严格限于簇 3，单核/髓系标记 LYZ/S100A9/FCGR3A/CST3 富集于簇 1 与簇 4，树突状细胞标记 FCER1A/HLA-DRA 在簇 5 表达 [6, 7]。

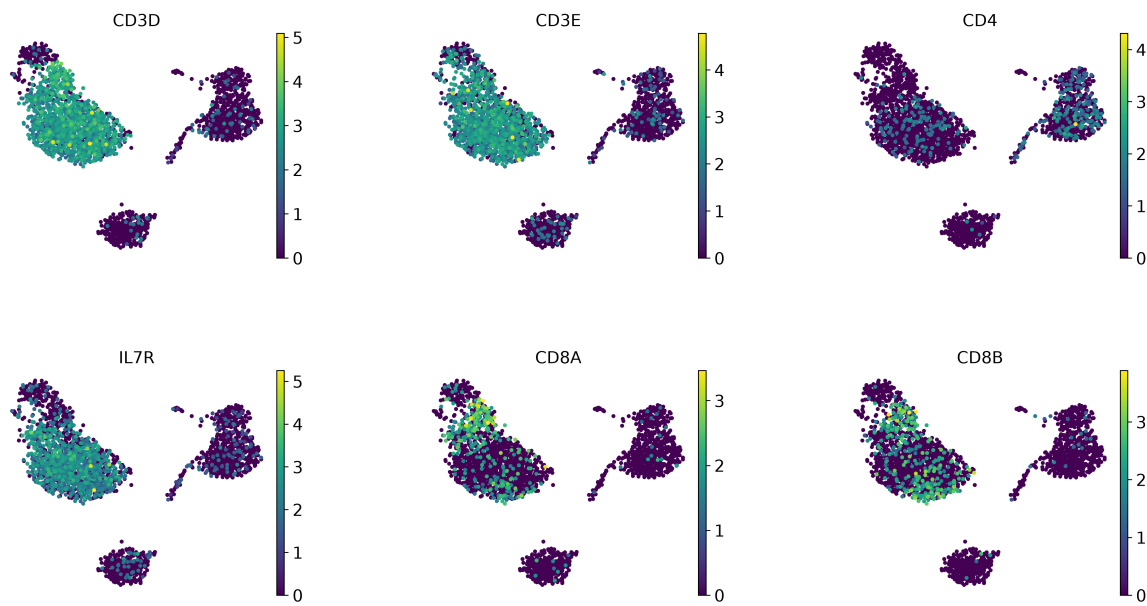


图 14 UMAP 标记基因表达投影（一）。

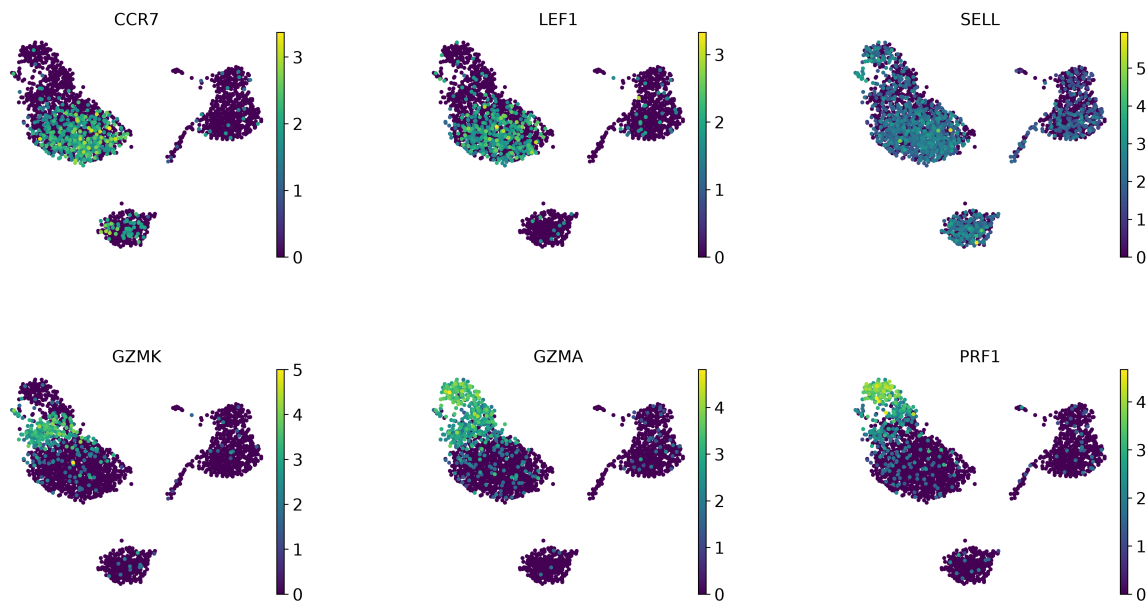


图 15 UMAP 标记基因表达投影（二）。

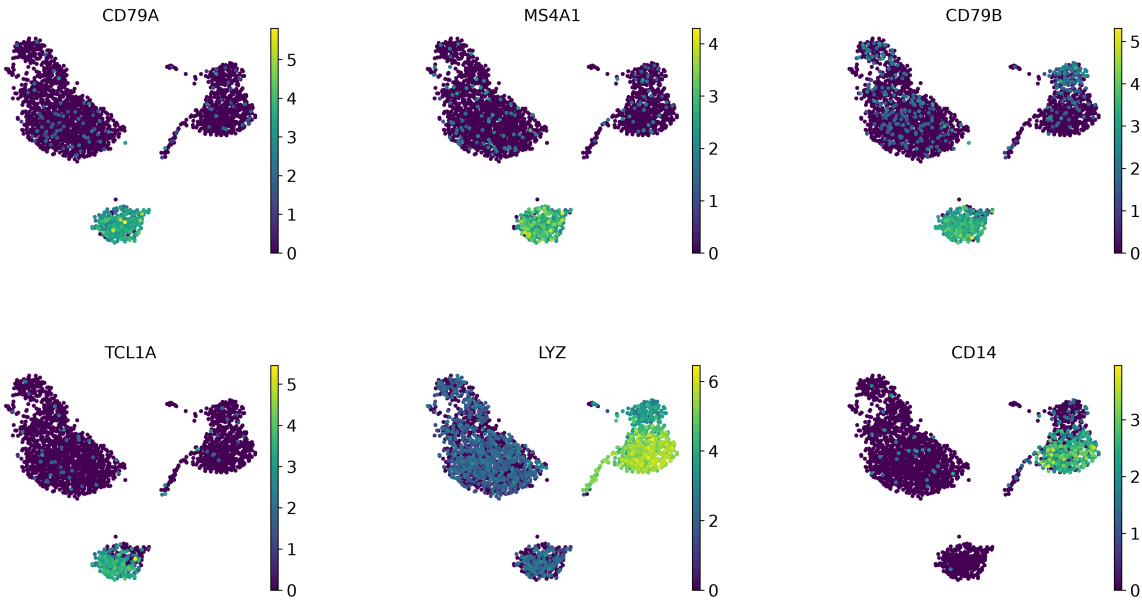


图 16 UMAP 标记基因表达投影（三）。

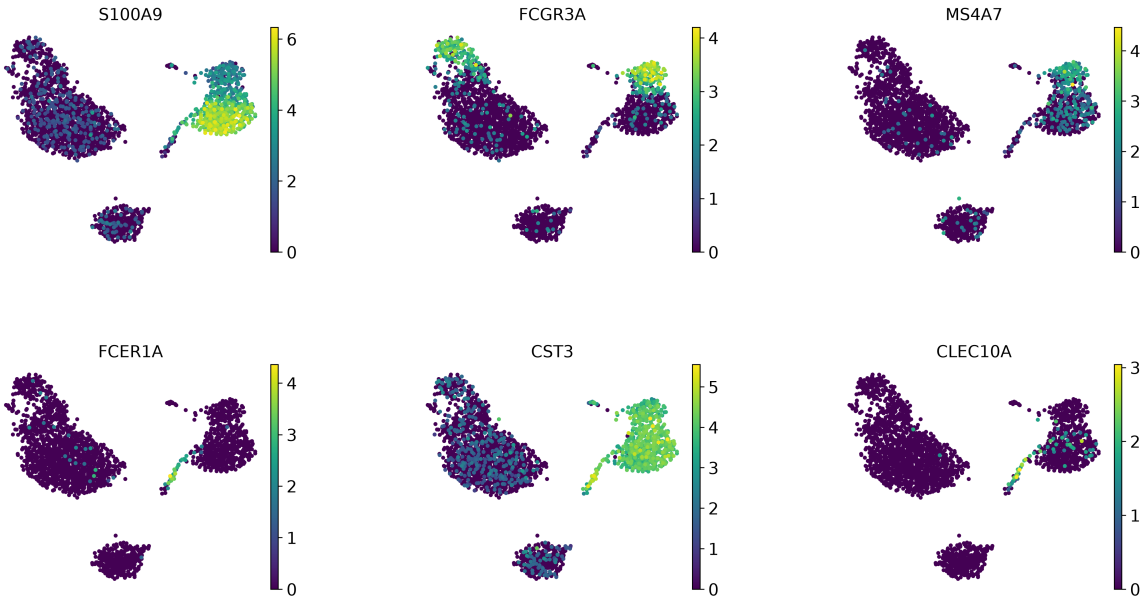


图 17 UMAP 标记基因表达投影（四）。

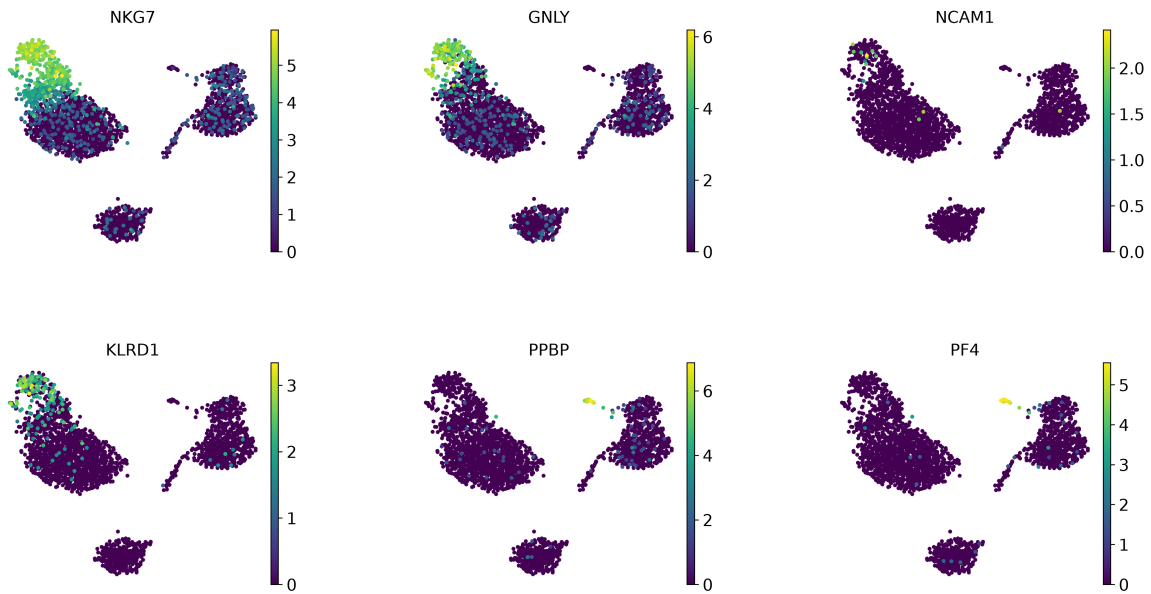


图 18 UMAP 标记基因表达投影（五）。

3.7 细胞类型构成与综合评估

各细胞类型占比与计数见图 19a: $CD4^+$ T 细胞占近半 (44.92%)，其次为经典单核细胞 (18.12%)、 $CD8^+$ T 细胞 (16.22%)、B 细胞 (12.93%)、非经典单核细胞 (6.41%) 与树突状细胞 (1.40%)，与人 PBMC 的预期免疫构成一致 [6, 7]。每簇基因数箱线图 (图 19b) 显示树突状细胞簇 (中位 1,567) 与非经典单核细胞簇 (中位 1,238) 的检测基因数显著高于其他簇，与其较高的转录复杂度相符；B 细胞簇基因数中位最低 (676)。综合 UMAP 面板 (图 20) 将 Leiden 簇、细胞类型、基因数与线粒体读占比四个维度并排呈现。最终评估指标见表 7。

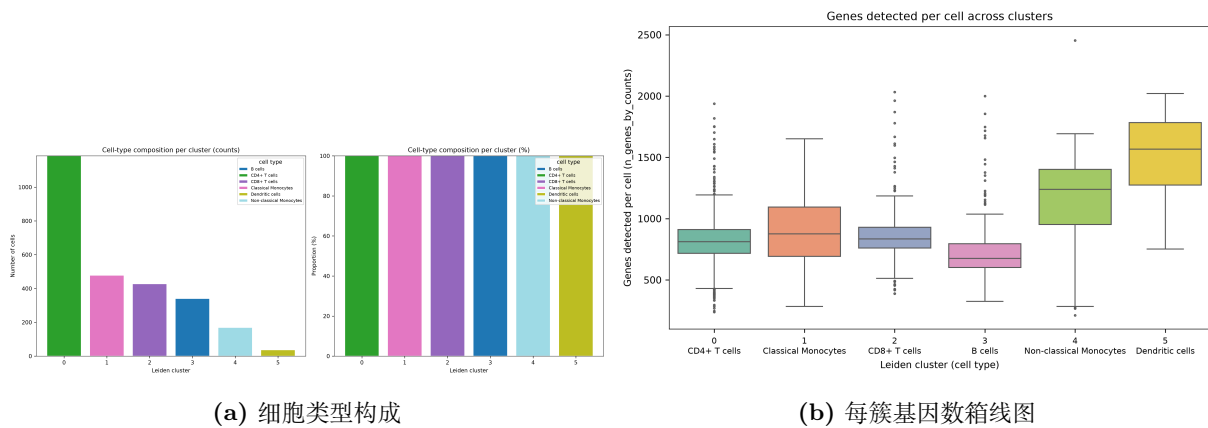


图 19 细胞类型构成与每簇基因数。左：各细胞类型的细胞数与百分比（堆叠柱状图）；右：各簇检测基因数分布。

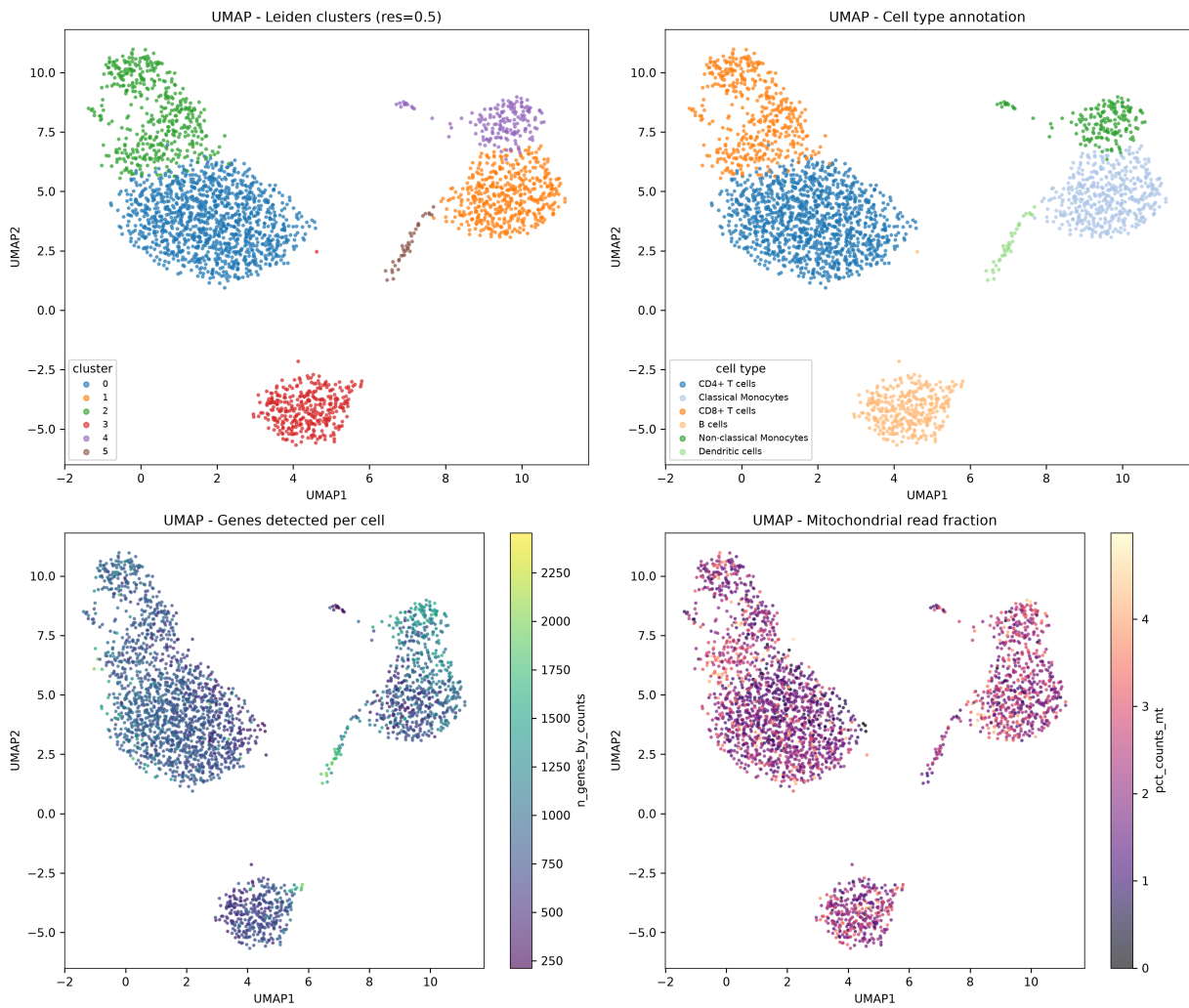


图 20 UMAP 综合面板。四面板分别展示 Leiden 簇、注释细胞类型、每细胞基因数与线粒体读占比，便于一站式审视聚类与质量指标。

表 7 最终评估指标汇总 (results/final_evaluation_metrics.csv)。

指标	取值	说明
n_cells_raw	2700	原始细胞数
n_cells_filtered	2638	质控后细胞数
n_genes_raw	32738	原始基因数
n_genes_after_cell_filter	16579	细胞级质控后基因数
n_genes_in_raw_layer	13656	adata.raw 基因数
n_hvg	2000	高变基因数
n_leiden_clusters	6	Leiden 簇数 (分辨率 0.5)
n_cell_types	6	注释细胞类型数
silhouette_score_pca40	0.197035	PCA 前 40 维轮廓系数 (n=1000)
pca_cum_var_40pc	0.120355	前 40 主成分累计方差
pca_cum_var_50pc	0.136514	前 50 主成分累计方差
median_n_genes_global	820.0	全局基因数中位数
recommended_resolution	0.5	推荐分辨率

4 讨论

本工作在 10x Genomics PBMC 3k 数据集上完整复现了标准 scRNA-seq 分析流程，并鉴定出六种主要免疫细胞类型，与人外周血免疫细胞图谱的既有认识高度一致 [6, 7, 20]。CD4⁺ 与 CD8⁺ T 细胞的拆分依赖于 NKG7/CST7/GZMA 等细胞毒效应分子与 CD4/IL7R/CCR7 签名的对比，而非单纯的 CD4/CD8 mRNA（二者在 scRNA-seq 中因低表达与 dropout 常不稳定），这一策略与 Scanpy/Seurat 官方教程一致 [15, 20]。单核细胞的经典/非经典二分以 CD14/LYZ/S100A9 与 FCGR3A/MS4A7 的此消彼长为据，与 Zilionis、Villani 等对单核细胞连续谱系的刻画相符 [6]。树突状细胞簇规模小（37 细胞）且置信度评为中等，提示其可能进一步细分（如 cDC1/cDC2/pDC），与 Villani 等报道的更精细的 DC 亚型构成对照 [6]。

分辨率敏感性分析揭示了一个值得强调的权衡。尽管分辨率 0.3 的轮廓系数（0.2133）略高于 0.5（0.1970），但 0.3 产生了仅 13 个细胞的微簇，其在统计上不可靠、易受双细胞污染 [10, 11]；而 0.5 的最小簇为 37 个细胞，所有簇均满足标记基因差异检验的统计功效要求，且复现了 Scanpy 官方对 PBMC 3k 的标准六簇划分 [15, 8]。这一结果说明，单一指标（尤其轮廓系数）不足以决定分辨率选择，必须结合簇规模可靠性判据与生物学先验 [8, 9]。需要指出，PCA 空间轮廓系数对基于图的聚类而言只是补充诊断——图聚类更直接的度量是模块度，而单细胞簇多为流形（manifold）而非凸分离，轮廓系数天然偏低 [16, 14]；故 0.197 这一数值应理解为”中等、可接受”而非”差”。

本流程采用 `seurat` flavor 的高变基因选择与对数归一化，这是 Scanpy 与 Seurat 长期默认的成熟方案 [15, 20]。但近年研究指出，对数归一化加方差选择在高变异度数据上可能引入均值-方差依赖偏差，基于广义线性模型（如 Pearson 残差、负二项回归）的归一化与特征选择在理论与部分基准上更优 [26, 13, 27]。此外，`seurat` 与 `cell_ranger` 等 flavor 在不同归一化路径下的高变基因集存在差异，使用时需注意其执行顺序 [14, 15]。本报告未实施双细胞显式检测（如 DoubletFinder、Scrublet） [10, 11]，仅以基因数上限间接过滤潜在双细胞，这是其局限之一。

局限性。其一，本数据集为单一供体、无批次，故未涉及批次校正（如 MNN、Harmony 等）

[28, 19]; 其二, 注释基于标记基因人工判读, 未使用自动注释或参考映射 (如 CellTypist、scANVI、Seurat 标签转移) [7, 20, 21], 对稀有亚群的细分能力有限; 其三, 未进行轨迹推断或下游通路分析 [29]。

下游方向。建议 (1) 纳入双细胞检测与质控的迭代精修 [10, 11]; (2) 以 Pearson 残差或 SCTransform 重做归一化以对比鲁棒性 [26, 13]; (3) 将注释扩展为自动/参考映射注释, 并细分 DC 与 T/NK 亚群 [6, 7, 21]; (4) 结合配对 V(D)J 或 CITE-seq 数据实现多模态整合 [30, 21]; (5) 在跨供体/批次数据上验证流程的迁移稳定性 [28, 9]。

5 结论

本报告在 10x Genomics PBMC 3k 数据集上完整实施并系统记录了标准 scRNA-seq 分析流程: 质控保留 2,638 个高质量细胞; 归一化与高变基因选择得到 $2,638 \times 2,000$ 分析矩阵; 50 维 PCA 前 40 维累计解释方差 12.04%; Leiden 聚类 (分辨率 0.5) 在 6 个无微簇候选中取得最佳折中, PCA 空间轮廓系数为 **0.1970**; 基于 Wilcoxon 标记基因与经典/扩展免疫标志物注释出六种细胞类型 (CD4⁺ T、经典单核、CD8⁺ T、B、非经典单核、树突状细胞), 构成与已知 PBMC 免疫图谱一致。分辨率敏感性分析表明单一轮廓系数不足以决定分辨率, 须结合簇规模可靠性与生物学先验。流程全部随机种子固定为 42、中间产物与汇总均落盘可审计, 具备良好的可复现性, 可作为 PBMC 类 scRNA-seq 端到端分析的参考实现。

数据与代码可用性

数据集经 `scanpy.datasets.pbmc3k()` 获取 (10x Genomics PBMC 3k, v1)。分析脚本 (步骤 1–6) 位于 `data/01-06_*.py`, 全部中间结果与汇总位于 `results/`, 图形位于 `figures/`, 元数据见 `results/environment_versions.txt` 与 `results/pipeline_summary.json`。

参考文献

- [1] Fuchou Tang, Catalin Barbacioru, Yu Wang, Ellen Nordman, Clarence Lee, Nan Xu, Xiaohui Wang, John Bodeau, Alan B. Todd, and M. Azim Surani. mRNA-Seq whole-transcriptome analysis of a single cell. *Nature Methods*, 6(5):377–382, 2009.
- [2] Valentine Svensson, Roser Vento-Tormo, and Sarah A. Teichmann. Exponential scaling of single-cell RNA-seq in the past decade. *Nature Protocols*, 13(4):599–604, 2018.
- [3] Evan Z. Macosko, Anindita Basu, Rahul Satija, James Nemesh, Karthik Shekhar, Melissa Goldman, Itay Tirosh, Allison R. Bialas, Nolan Kamitaki, Emily M. Martersteck, et al. Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets. *Cell*, 161(5):1202–1214, 2015.
- [4] Allon M. Klein, Linas Mazutis, Ilke Akartuna, Naren Tallapragada, Adrian Veres, Vlad Li, Leonid Peshkin, David A. Weitz, and Marc W. Kirschner. Droplet barcoding for single-cell transcriptomics applied to embryonic stem cells. *Cell*, 161(5):1187–1201, 2015.

- [5] Grace X. Y. Zheng, Jeffrey M. Terry, Phillip Belgrader, Paul Ryvkin, Zachary W. Bent, Ryan Wilson, Salomé S. Ziraldo, Tomas D. Wheeler, Geoff P. McDermott, Junjie Zhu, et al. Massively parallel digital transcriptional profiling of single cells. *Nature Communications*, 8(1):14049, 2017.
- [6] Alexandra-Chloé Villani, Rahul Satija, Gary Reynolds, Miranda Sarkizova, Karthik Shekhar, Irene Fletcher, Jörgen Griesbeck, Andrew Butler, Scott Weinstein, Andrew Jia, et al. Single-cell RNA-seq reveals new types of human blood dendritic cells, monocytes, and progenitors. *Science*, 356(6335):aah4573, 2017.
- [7] Gianni Monaco, Bennett Lee, Weili Xu, Seri Mustafah, You Yi Hwang, Christophe Carré, Nicolè Burdin, Luciane Möller, Andrew Pearson, Alan de Souza, et al. RNA-Seq signatures normalized by mRNA abundance allow absolute deconvolution of human immune cell types. *Cell Reports*, 26(6):1627–1640.e7, 2019.
- [8] Malte D. Luecken and Fabian J. Theis. Current best practices in single-cell RNA-seq analysis: a tutorial. *Molecular Systems Biology*, 15(6):e88746, 2019.
- [9] Lukas Heumos, Anna C. Schaar, Christopher Lance, Anastasia Litinetskaya, Fidel Drost, Luke Zappia, Malte D. Luecken, Daniel C. Strobl, Jesus Henao, Fabian Curion, et al. Best practices for single-cell analysis across modalities. *Nature Reviews Genetics*, 24(8):550–572, 2023.
- [10] Christopher S. McGinnis, Lyndsay M. Murrow, and Zev J. Gartner. DoubletFinder: Doublet detection in single-cell RNA sequencing data using artificial nearest neighbors. *Cell Systems*, 8(6):481–489.e5, 2019.
- [11] Samuel L. Wolock, Romain Lopez, and Allon M. Klein. Scrublet: Computational identification of cell doublets in single-cell transcriptomic data. *Cell Systems*, 8(3):281–291.e4, 2019.
- [12] Philip Brennecke, Simon Anders, Jong Kyoung Kim, Aleksandra A. Kołodziejczyk, Xingming Zhang, Valentina Proserpio, Bian Baying, Vladimir Benes, Sarah A. Teichmann, John C. Marioni, and Marcus G. Heisler. Accounting for technical noise in single-cell RNA-seq experiments. *Nature Methods*, 10(11):1093–1095, 2013.
- [13] F. William Townes, Stephanie C. Hicks, Martin J. Aryee, and Rafael A. Irizarry. Feature selection and dimension reduction for single-cell RNA-Seq based on a multinomial model. *Genome Biology*, 20(1):295, 2019.
- [14] Shiquan Sun, Jiejun Zhu, Ying Ma, and Xiang Zhou. Accuracy, robustness and scalability of dimensionality reduction methods for single-cell RNA-seq analysis. *Genome Biology*, 20(1):269, 2019.
- [15] F. Alexander Wolf, Philipp Angerer, and Fabian J. Theis. SCANPY: large-scale single-cell gene expression data analysis. *Genome Biology*, 19(1):15, 2018.

- [16] V. A. Traag, L. Waltman, and N. J. van Eck. From louvain to leiden: guaranteeing well-connected communities. *Scientific Reports*, 9(1):5233, 2019.
- [17] Charlotte Soneson and Mark D. Robinson. Bias, robustness and scalability in single-cell differential expression analysis. *Nature Methods*, 15(4):255–261, 2018.
- [18] Rahul Satija, Jeffrey A. Farrell, David Gennert, Alexander F. Schier, and Aviv Regev. Spatial reconstruction of single-cell gene expression data. *Nature Biotechnology*, 33(5):495–502, 2015.
- [19] Andrew Butler, Paul Hoffman, Peter Smibert, Efthymia Papalexi, and Rahul Satija. Integrating single-cell transcriptomic data across different conditions, technologies, and species. *Nature Biotechnology*, 36(5):411–420, 2018.
- [20] Tim Stuart, Andrew Butler, Paul Hoffman, Christoph Hafemeister, Efthymia Papalexi, William M. Mauck, Yuhao Hao, Marlon Stoeckius, Peter Smibert, and Rahul Satija. Comprehensive integration of single-cell data. *Cell*, 177(7):1888–1902.e21, 2019.
- [21] Yuhao Hao, Stephanie Hao, Erica Andersen-Nissen, William M. Mauck, Shiwei Zheng, Andrew Butler, Madison J. Lee, Andrew J. Wilk, Chiara Darby, Max Zager, et al. Integrated analysis of multimodal single-cell data. *Cell*, 184(11):3553–3569.e29, 2021.
- [22] Vladimir Yu Kiselev, Kristina Kirschner, Michael T. Schaub, Tallulah Andrews, Andrew Yiu, Tamir Chandra, Kedar Natarajan, Wolf Reik, Mauricio Barahona, Anthony R. Green, and Martin Hemberg. SC3: consensus clustering of single-cell RNA-seq data. *Nature Methods*, 14(5):483–486, 2017.
- [23] Leland McInnes, John Healy, Nathaniel Saul, and Lukas Großberger. UMAP: Uniform manifold approximation and projection. *Journal of Open Source Software*, 3(29):861, 2018.
- [24] Etienne Becht, Leland McInnes, John Healy, Charles-Alexandre Dutertre, Immanuel W. H. Kwok, Lai G. Ng, Florent Ginhoux, and Evan W. Newell. Dimensionality reduction for visualizing single-cell data using UMAP. *Nature Biotechnology*, 37(1):38–44, 2019.
- [25] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-SNE. *Journal of Machine Learning Research*, 9(86):2579–2605, 2008.
- [26] Christoph Hafemeister and Rahul Satija. Normalization and variance stabilization of single-cell RNA-seq data using regularized negative binomial regression. *Genome Biology*, 20(1):296, 2019.
- [27] Saket Choudhary and Rahul Satija. Comparison and evaluation of statistical error models for scRNA-seq. *Genome Biology*, 23(1):27, 2022.
- [28] Laleh Haghverdi, Aaron T. L. Lun, Martin D. Morgan, and John C. Marioni. Batch effects in single-cell RNA-sequencing data are corrected by matching mutual nearest neighbors. *Nature Biotechnology*, 36(5):421–427, 2018.

- [29] Wouter Saelens, Robrecht Cannoodt, Helena Todorov, and Yvan Saeys. A comparison of single-cell trajectory inference methods. *Nature Biotechnology*, 37(5):547–554, 2019.
- [30] Marlon Stoeckius, Christoph Hafemeister, William Stephenson, Brian Houck-Loomis, Priyanish Chattopadhyay, Harold Swerdlow, Rahul Satija, and Peter Smibert. Simultaneous epitope and transcriptome measurement in single cells. *Nature Methods*, 14(9):865–868, 2017.