Advances in Applied Mathematics 应用数学进展, 2025, 14(4), 143-154
Published Online April 2025 in Hans. https://www.hanspub.org/journal/aam
https://doi.org/10.12677/aam.2025.144147

Hans汉斯

基于多种降维算法的单细胞差异化分析及可视化

杨泽明

青岛大学数学与统计学院，山东 青岛

收稿日期：2025年3月3日；录用日期：2025年3月26日；发布日期：2025年4月8日

摘　要

随着生物基因测序技术的发展，单细胞RNA测序(scRNA-seq)技术在细胞异质性研究中发挥了重要作用，单核细胞的转录组分析提供了丰富的基因表达数据。为了分析这些高维数据的潜在特征，本研究使用多种降维算法，包括主成分分析(PCA)、统一流形逼近与投影(UMAP)和t-分布随机邻域嵌入(t-SNE)，对单细胞基因表达的结构特征进行揭示。这些降维方法能够有效将高维数据映射至低维空间，从而帮助我们发现单细胞的亚群分布、动态变化和差异性表达基因的聚类特征。通过可视化展示，本文探索了基因表达的内在规律，进一步为单细胞在疾病诊断、治疗以及精准医疗中的应用提供了新的见解。本文强调了统计方法在生物数据分析中的核心作用，特别是与可视化技术的结合上，为后续研究提供了有力支持。

关键词

降维算法，单核细胞，可视化

Differential Analysis and Visualization of Single-Cell Based on Multiple Dimensionality Reduction Algorithms

Zeming Yang

School of Mathematics and Statistics, Qingdao University, Qingdao Shandong

Received: Mar. 3rd, 2025; accepted: Mar. 26th, 2025; published: Apr. 8th, 2025

Abstract

With the advancement of biological gene sequencing technologies, single-cell RNA sequencing (scRNA-seq) has played a significant role in studying cell heterogeneity. The transcriptomic analysis

文章引用: 杨泽明. 基于多种降维算法的单细胞差异化分析及可视化[J]. 应用数学进展, 2025, 14(4): 143-154.
DOI: 10.12677/aam.2025.144147

of mononuclear cells provides rich gene expression data. To analyze the underlying patterns of these high-dimensional data, this study applies multiple dimensionality reduction algorithms, including Principal Component Analysis (PCA), Uniform Manifold Approximation and Projection (UMAP), and t-Distributed Stochastic Neighbor Embedding (t-SNE), to reveal the structural features of single-cell gene expression. These dimensionality reduction methods map high-dimensional data to lower-dimensional spaces, helping to uncover the distribution of subpopulations, dynamic changes, and clustering features of differentially expressed genes. Through visualization, this study explores the inherent patterns of gene expression and provides new insights into the applications of single-cell data in disease diagnosis, treatment, and precision medicine. This paper emphasizes the central role of statistical methods in biological data analysis, particularly in the integration of dimensionality reduction and visualization techniques, offering strong support for future research.

Keywords

Dimensionality Reduction Algorithms, Mononuclear Cells, Visualization

1. 引言

近些年来，随着生物基因测序技术的快速发展，基因组学和转录组学研究进入了一个全新的时代，尤其是单细胞 RNA 测序(scRNA-seq)技术的应用，使得我们能够在单细胞维度下深入探索细胞异质性。基因测序技术[1]一共经历了三个主要阶段：第一代测序技术、第二代高通量测序技术(NGS)和第三代单分子测序技术，每一代技术的突破都极大推动了基因组学和生物信息学的发展。尽管现有的技术在解析基因组结构和功能方面取得了显著成就，但在数据处理和分析上仍面临诸多挑战，尤其是如何从海量的高维基因数据中提取有意义的生物学信息。

为克服现有这些挑战，统计学方法尤其是降维算法已成为现代生物数据分析的核心工具。降维技术通过将高维数据映射到低维空间，帮助研究人员发现潜在的模式和结构特征，从而揭示数据的内在规律。在单细胞转录组数据分析中，常用的降维算法包括主成分分析(PCA) [2]、统一流形逼近与投影(UMAP) [3]和 t-分布随机邻域嵌入(t-SNE) [4]等。这些方法不仅需要线性代数、概率论和信息论等数学理论，还通过优化算法来实现数据的有效降维和可视化，从而帮助揭示细胞群体的异质性和差异性基因表达。

随着单细胞 RNA 测序技术的应用，基因表达数据为疾病机制研究和精准医疗提供了丰富的信息，然而这些数据通常具有高维度和复杂的结构，还需要通过统计方法对数据进行有效的降维和可视化，以便提取出有价值的生物学结论。本文基于 PCA、UMAP 和 t-SNE 等降维算法，对单细胞基因表达数据进行系统分析，重点探索这些统计方法在揭示单细胞数据中的亚群结构[5]、动态变化和差异性基因表达方面的应用。通过这些降维结果的可视化，本文进一步探讨了基因表达的内在规律，并为疾病诊断和精准医疗中的潜在应用提供了新的数学统计视角。

2. 预备知识和数据来源

2.1. 主成分分析

主成分分析(Principal Component Analysis, PCA)是一种用于降维的无监督算法，因该算法不需要设置参数，所以便于理解和实现，广泛用于高维数据的分析和可视化。PCA 的核心思想主要是通过线性变换，

将高维数据投影到一个较低维度的子空间中，实现最大化数据方差并尽量减少信息损失。该算法首先计算数据每个特征的均值，使数据均值归一化，之后计算协方差矩阵并对协方差矩阵进行特征值分解，以得到特征值和特征向量，再之后进行降维，选择特征值较大的前几个主成分，将它们所对应的特征向量构成新特征空间的坐标轴，将原始数据投影到新特征空间中，从而实现降维。

假设首先有一个数据表达矩阵 $X \in R^{n\times d}$，其中每一行代表一个细胞样本，每一列代表一个基因特征，主成分分析算法主要分为以下步骤：

(1) 首先对数据进行中心化：$\tilde{X} = X - \bar{X}$

(2) 计算标准化数据的协方差矩阵：

$$C = \frac{1}{n-1}\tilde{X}^{\mathrm{T}}\tilde{X}$$

(3) 对协方差矩阵 C 进行特征值分解，得到特征值和特征向量，特征值 λ_i 和特征向量 v_i 满足方程：$Cv_i = \lambda_i v_i$。

(4) 选择最大的 k 个特征值对应的特征向量 $[v_1, v_2, \cdots, v_k]$，将所有特征向量标准化后组成特征向量矩阵 W。

(5) 将原始标准化数据投影到由前 k 个特征向量组成的矩阵 W 上得到降维后的数据矩阵 $X_k = \tilde{X}W$。

(6) 降维后的数据 X_k 包含了数据中方差最大的 k 个方向的信息，因此能够有效地表示数据的主要特征。

2.2. t-分布邻域嵌入

t-分布邻域嵌入(t-SNE, t-distributed Stochastic Neighbor Embedding)是一种非线性降维算法，特别适用于高维数据的可视化，这种方法主要是关注数据的局部结构，它由 Laurens van der Maaten 和 Geoffrey Hinton 在 2008 年提出。t-SNE 目标是将高维数据映射到低维空间中，同时尽可能保留相似数据点之间的局部结构关系，通俗来说，使高维空间中相似的数据点在低维空间中靠得更近，使高维空间中相对不相似的数据点在低维空间中分布得更远。t-SNE 需要将数据点之间的相似度转换为概率，原始空间中的相似度是由高斯联合概率表示，嵌入空间的相似度由“student-t 分布”[6]表示，t 分布(自由度为 1)比高斯分布有更长的尾部，因此使得低维空间中的远离点更容易分开。t-SNE 通过原始空间和嵌入空间的联合概率的 Kullback-Leibler (KL)散度[7]来评估可视化效果的好坏，也就是说用有关 KL 散度的函数作为损失函数，然后通过梯度下降最小化损失函数，最终获得收敛结果。需要注意的是该损失函数不是凸函数，即具有不同初始值的多次运行将收敛于 KL 散度函数的局部最小值中，以致获得不同的结果。

假设有一个数据表达矩阵 $X \in R^{n\times d}$，其中每一行代表一个细胞样本，每一列代表一个基因特征，t-SNE 算法主要分为以下步骤：

(1) 计算高维空间的相似度：对于每一对数据点 x_i 和 x_j，使用条件概率表示点在给定 x_i 时 x_j 出现的概率。定义为：

$$p_{ij} = \frac{\exp\left(-\frac{\|x_i - x_j\|^2}{2\sigma_i^2}\right)}{\sum_{k\neq i}\exp\left(-\frac{\|x_i - x_k\|^2}{2\sigma_i^2}\right)} \tag{1}$$

其中，σ_i 是 x_i 的高斯核宽度，可以动态调整以确保每个点的邻域包含一个固定数量的点(通过 perplexity 参数)，p_{ij} 越大表示 x_i 和 x_j 越相似。

(2) 对称化概率矩阵：

$$p_{ij} = \frac{p_{ij} + p_{ji}}{2n} \tag{2}$$

其中 n 是数据点总数。

(3) 计算低维空间的相似度：在低维空间中，t-SNE 通过自由度为 1 的学生 t 分布来计算数据点间的相似度，对于每一对低维空间的数据点 y_i 和 y_j，定义其相似度为：

$$q_{ij} = \frac{\left(1+\left\|y_i - y_j\right\|^2\right)^{-1}}{\sum_{k\neq l}\left(1+\left\|y_k - y_l\right\|^2\right)^{-1}} \tag{3}$$

其中，y_i 和 y_j 是低维空间中的数据点，q_{ij} 越大表示 y_i 和 y_j 越相似。

(4) 最小化分布差异：t-SNE 算法通过最小化高维空间分布 p_{ij} 和低维空间分布 q_{ij} 之间的差异来优化低维嵌入：

$$C = \sum_{i<j} p_{ij} \log \frac{p_{ij}}{q_{ij}} \tag{4}$$

这里的 C 是基于 Kullback-Leibler (KL)散度的目标函数，使用梯度下降优化 C，不断调整低维点的位置 y_j。

2.3. 均匀流形逼近与投影

均匀流形逼近与投影(UMAP, Uniform Manifold Approximation and Projection)是一种现代的非线性降维方法，广泛应用于高维数据的降维和可视化。UMAP 最早由 Leland McInnes 和 John Healy 在 2018 年提出，其基于流形学习(Manifold Learning)的理论，它继承了流形学习[8]的思想，且相比于传统的降维方法(如 t-SNE)，在处理大规模数据时具有更好的计算效率。UMAP 能够有效地保留数据的局部和全局结构，常用于数据可视化、聚类、分类等任务。

UMAP 的基本思想是：假设数据点所在的高维空间实际上是一个低维流形的高维嵌入，数据点的局部邻域关系反映了它们在低维空间中的关系，降维的目标是找到一个低维空间，使得数据在高维空间和低维空间中的局部结构尽可能一致。其基本流程包括：局部邻域建立、构建相似度图和优化嵌入空间三部分。首先是局部邻域建立，对每个数据点，UMAP 首先通过计算其与其他数据点的距离来构建局部邻域，通常使用 k 近邻(KNN) [9]来定义邻域。其次是构建相似度图[10]，通过欧氏距离、余弦相似度[11]等相似度度量构建一个加权图。权重越大，代表两个点在高维空间中的相似度越高。最后是优化嵌入空间，优化目标是使得低维嵌入空间的邻接关系尽可能保留原始数据的结构特征。

假设有一个数据表达矩阵 $X \in R^{n\times d}$，其中每一行代表一个细胞样本，每一列代表一个基因特征，UMAP 算法主要分为以下步骤：

(1) 构建高维空间中的相似度图：相似度度量通常是基于距离度量(如欧氏距离)或基于核密度估计的概率模型。UMAP 首先为每个数据点 x_i 定义一个局部邻域，通常使用 k-近邻(KNN)或者半径邻域方法。对于高维数据每一对数据点 x_i 和 x_j，定义其相似度概率 p_{ij}，可以通过以下公式来表示：

$$p_{ij} = \frac{\exp\left(-\frac{d\left(x_i, x_j\right)}{\sigma_i}\right)}{\sum_{k\neq i}\exp\left(-\frac{d\left(x_i, x_k\right)}{\sigma_i}\right)} \tag{5}$$

其中，σ_i 是高维空间中点的邻域大小，用来控制高维空间中邻居的密度，表示为 $\sigma_i = \text{mean}\left(\left\|x_i - x_j\right\|\right)$ 在其

邻域点 x_j 上的平均距离

(2) 构建低维空间中的相似度图：低维空间中的相似度是通过学生 t 分布来度量的。学生 t 分布具有较重的尾部，有助于处理高维数据中的噪声和离群点。低维空间中的相似度 q_{ij} 通过以下公式表示：

$$q_{ij}=\frac{\exp\left(-\frac{d\left(y_i,y_j\right)}{\sigma_i}\right)}{\sum_{k\neq i}\exp\left(-\frac{d\left(y_i,y_k\right)}{\sigma_i}\right)} \tag{6}$$

其中， y_i 和 y_j 是低维空间中的数据点。

(3) 优化损失函数：UMAP 通过最小化高维空间和低维空间中相似度分布之间的差异来优化低维嵌入。损失函数通过 Kullback-Leibler 散度(KL 散度)来度量两个概率分布之间的差异：

$$\text{Loss}=\sum_{i,j}\left(p_{ij}-q_{ij}\right)^2 \tag{7}$$

p_{ij} 和 q_{ij} 是我们上述分别求得的高维空间和低维空间的相似度。

(4) 优化算法：UMAP 就通过梯度下降算法[12]来优化损失函数。优化的目标是使得低维空间中的相似度分布 q_{ij} 尽可能接近高维空间中的相似度分布 p_{ij} 。

2.4. 数据来源

从 10X Genomics 单细胞测序技术平台(https://www.10xgenomics.com/)下载冷冻人外周血单个核细胞 Human Frozen PBMCs 的单细胞测序数据。

3. 模型建立与预测

3.1. 单细胞测序数据预处理

3.1.1. 基因指控

通过基因质控指标来筛选细胞，质控指标有以下三种：(1) 每个细胞中检测到的基因数。低质量的细胞和空油滴只有少量基因，两个及以上的细胞会有异常的高基因数，这两类细胞需要被筛选。(2) 每个细胞中的 UMI 总数。本文设定的标准为过滤 UMI 数大于 2500，小于 200 的细胞。(3) 线粒体基因组的 reads 比例。低质量或死细胞会有大百分比的线粒体基因组，使用 PercentageFeatureSet 函数来计数线粒体质控指标，本文设定的标准为过滤线粒体百分比大于 5%的细胞。

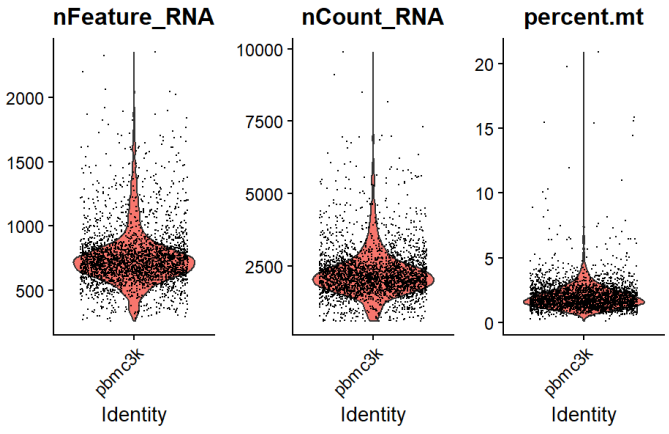

Figure 1. The percentage of mitochondrial read
图 1. 线粒体 read 的百分比

MT 是线粒体基因，通过图 1 我们查看线粒体 read 的百分比，一般情况下，我们认为细胞中线粒体含量较多，意味着细胞可能趋于死亡，所以大部分情况下线粒体的含量应该去除较大的，我们可以根据具体的情况进行分析，一般 10%以内可以接受。

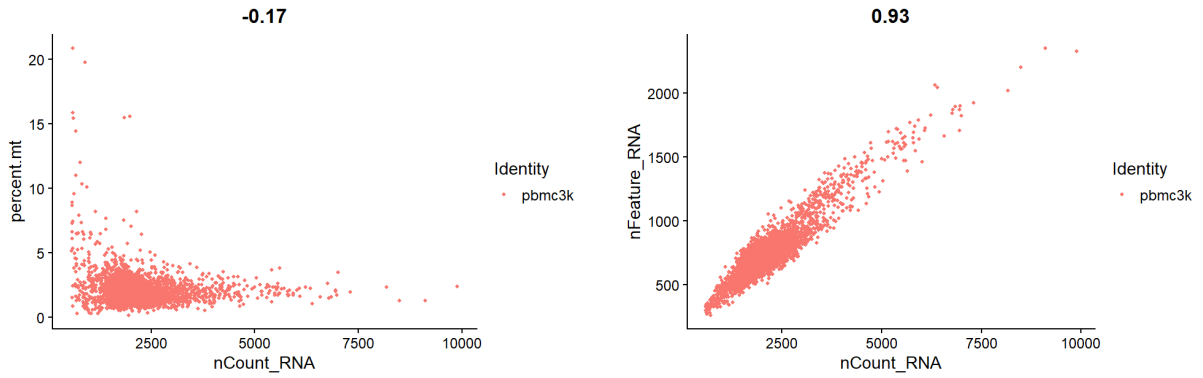

Figure 2. The correlation between features before filtering
图 2. 过滤之前的特征与特征间的相关性

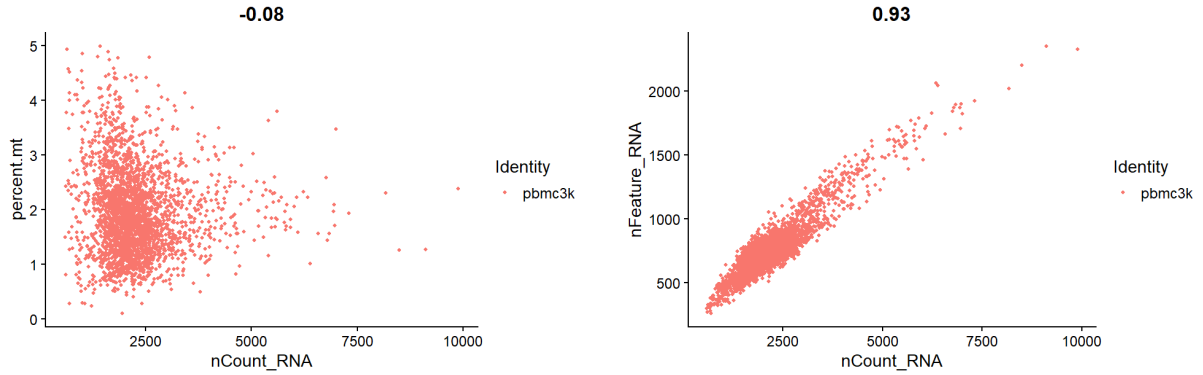

Figure 3. The correlation between features after filtering
图 3. 过滤之后的特征与特征间的相关性

我们过滤 UMI 数大于 2500，小于 200 的细胞和线粒体百分比大于 5%的细胞重新查看过滤之后的特征与特征间的相关性，再次用小提琴图展示(如图 2、图 3 所示)。

3.1.2. 归一化处理

我们每个细胞分别进行检测，所以要进行标准化，本文使用 NormalizeData 函数对上述处理的细胞数据进行归一化处理。

3.1.3. 识别高异质性特征

如果许多基因的表达在各个细胞之间表达是恒定的，那么要区分各个细胞就要使用变化差异大的基因，这就是高变基因，在细胞数据集中寻找高表达的基因特征，有助于找到单细胞数据集中的生物信号进行下游分析。使用 VariableFeatures 函数提取高变基因，之后对高变基因进行展示绘图。

通过图 4 结果展示我们可以看到 SA100A9、LYZ、IGLL5 等为高变基因。

3.1.4. 标准化处理

使用 ScaleData 函数准换每个基因的表达值，使每个细胞的平均表达值为 0，使细胞间方差为 1，使

每个基因具有相同的权重，有利于下一步分析。

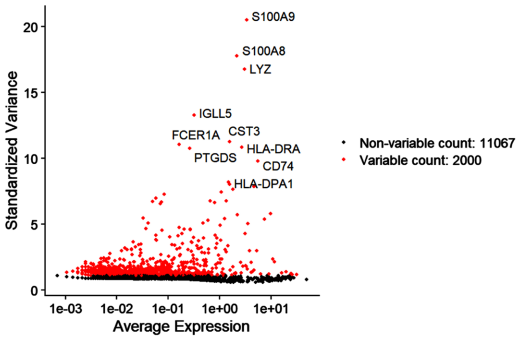

Figure 4. Visualizing genes with high variance
图 4. 展示高变基因

3.2. 使用 PCA 进行降维

可视化细胞与特征间的 PCA 有三种方式：(1) 使用 VizDimLoadings 函数，用于可视化与降维结果相关的基因，了解各个基因对主成分的贡献程度(如图 5 所示)。(2) 使用 DimPlot 函数，将降维技术的输出绘制在二维散点图上。每个点代表一个细胞，并且根据降维技术确定的单元嵌入进行定位(如图 6 所示)。(3) 使用 DimHeatmap 函数，绘制某个主成分的热图，这个热图会根据主成分得分对细胞和基因进行排序，从而允许用户直观地观察数据集中异质性的主要来源(如图 7 所示)。本文选取前五个主成分，打印每个主成分前 5 个贡献最大的基因，即最重要的特征基因。

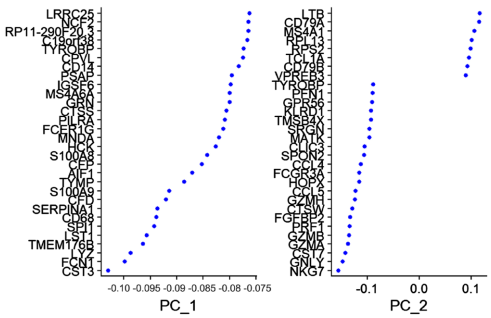

Figure 5. Contribution of each gene to the principal components
图 5. 各个基因对主成分的贡献程度

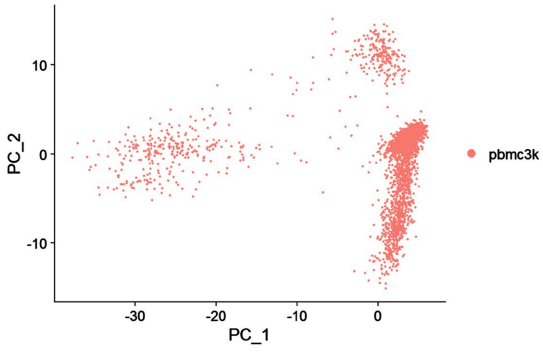

Figure 6. 2D embedding plot after PCA dimensionality reduction
图 6. PCA 降维后的二维嵌入图

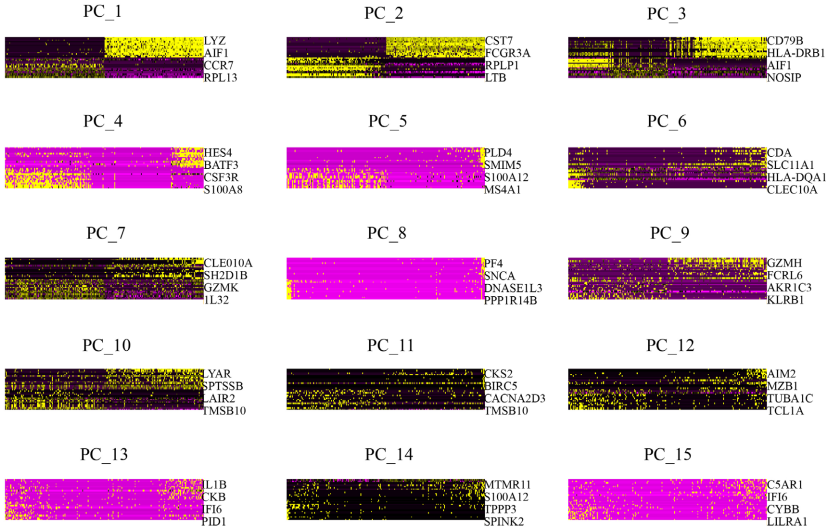

Figure 7. Heatmap after PCA dimensionality reduction
图 7. PCA 降维后的热图展示

3.3. 确定数据集的维度

为了克服在单细胞数据中在单个特征中的技术噪音，我们需要压缩数据集，并确定数据集的维数。JackStraw 分析[13]是一种用于评估单细胞测序数据中主成分分析中各个主成分显著性的方法。它是基于随机化方法，通过将原始数据中的样本标签进行随机重排，来估计每个主成分的 p 值，从而判断哪些主成分不是随机噪声而是真实的信号。JackStraw 分析首先随机置换数据集中的一部分(默认为 1%)样本标签，并重新进行 PCA 分析。对于这些“随机”基因，计算其 PCA 分数，并与观察到的 PCA 分数进行比较，以确定统计显著性，分析结果为每个基因在 PCA 分析中的 p 值。如果随后运行 ProjectPCA，那么这些 p 值代表了所有基因的显著性(图 8)。

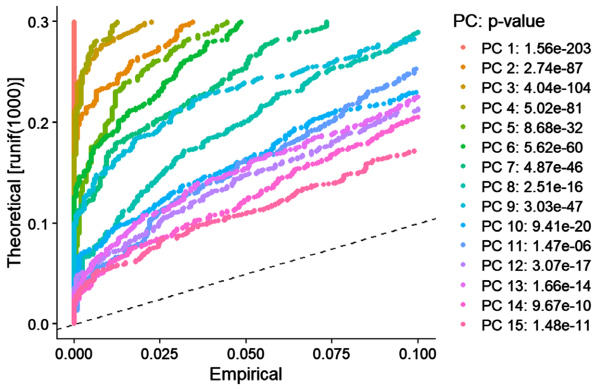

Figure 8. JackStraw-based null distribution visualization for PCA significance
图 8. 基于 JackStraw 方法的 PCA 特征显著性零分布图

3.4. 使用 UMAP 和 t-SNE 聚类细胞

使用 UMAP 和 t-SNE 算法对上述处理的细胞进行聚类。perplexity 参数的选择对于聚类效果至关重要，perplexity 参数影响局部和全局结构的平衡，较小的 perplexity，更关注局部结构，容易导致小群体紧

密聚集，较大的 perplexity 更关注全局结构，使不同类群更加分散。通常细胞数小于 1000 个 perplexity 参数选择 5 至 20，1000 个到 5000 个细胞 perplexity 参数选择 20 至 50，大于 5000 个细胞 perplexity 参数选择 30 至 100。由于本文数据超过了 5000 个，选择 perplexity 参数为 30 (见图 9，图 10)。

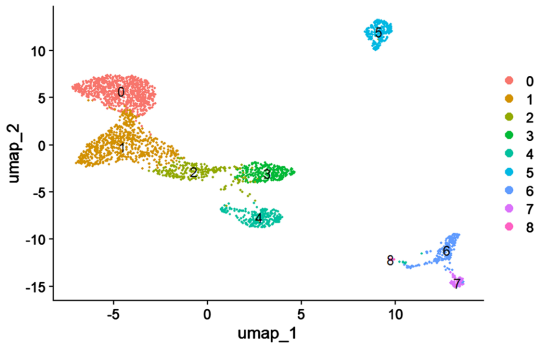

Figure 9. UMAP dimensionality reduction visualization
图 9. UMAP 降维可视化图

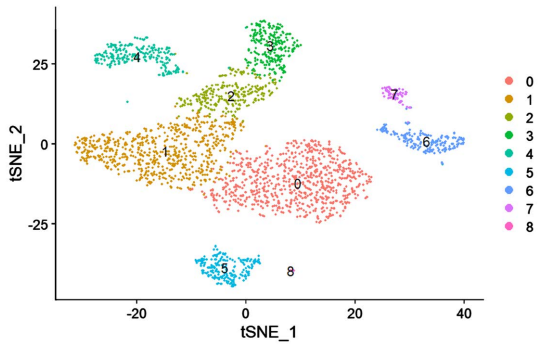

Figure 10. t-SNE dimensionality reduction visualization
图 10. t-SNE 降维可视化图

3.5. 发现差异表达特征，得到最终结果

在使用 UMAP 和 t-SNE 算法对细胞进行聚类后，可以对部分免疫相关基因进行分析，观察其在聚类中的表达分布，如图 11~14 表示。

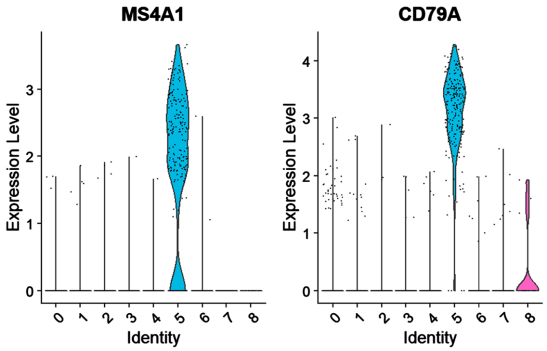

Figure 11. Expression distribution of MS4A1 and CD79A across different clusters
图 11. MS4A1 与 CD79A 在不同聚类中的表达分布

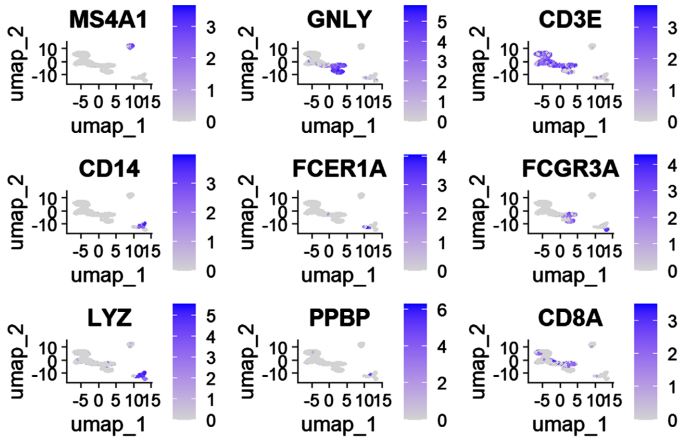

Figure 12. Expression distribution of immune-related genes across different cell populations
图 12. 免疫相关基因在不同细胞群中的表达分布图

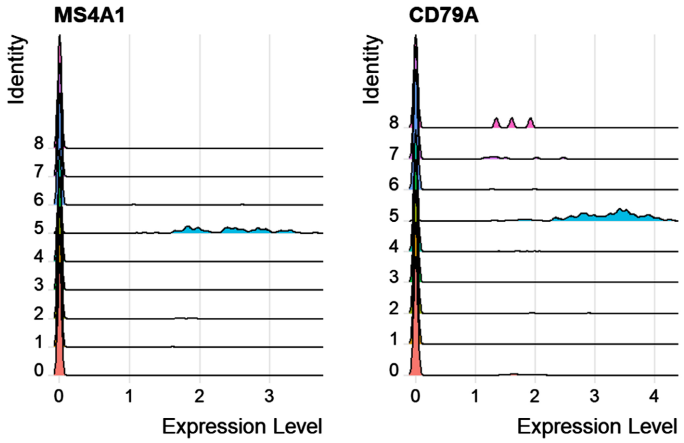

Figure 13. RidgePlot expression distribution of MS4A1 and CD79A across different cell populations
图 13. MS4A1 与 CD79A 在不同细胞群体中的 RidgePlot 表达分布

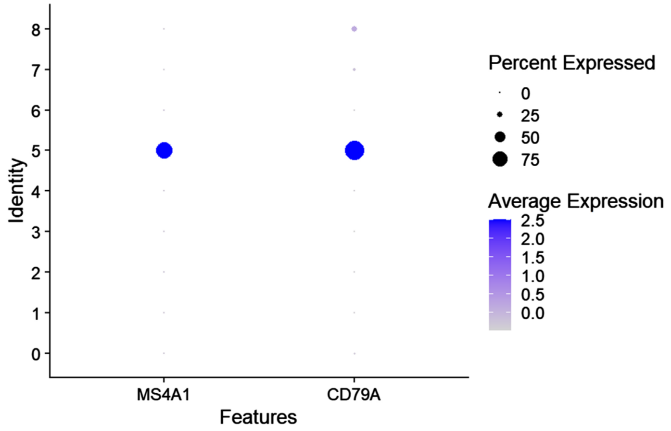

Figure 14. DotPlot expression of MS4A1 and CD79A across different cell populations
图 14. MS4A1 与 CD79A 在不同细胞群体中的 DotPlot 表达图

3.6. 识别细胞类型

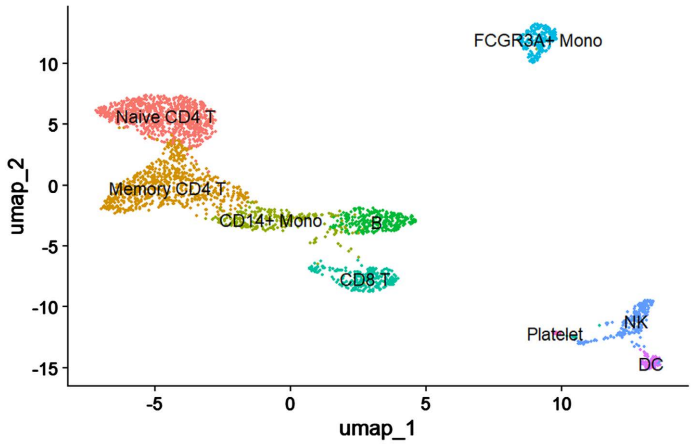

Figure 15. UMAP dimensionality reduction of renamed cell populations
图 15. 重新命名细胞群体的 UMAP 降维图

4. 结论

4.1. 聚类效果评估

对于聚类效果的评估，我们采用轮廓系数(Silhouette Score)、调整兰德指数(Adjusted Rand Index, ARI)、轮廓宽度(Within-cluster Sum of Squares, WCSS)、Dunn 指数(Dunn Index)和稳定性指数(Jaccard Index)进行综合衡量，见表 1。

Table 1. Clustering evaluation metrics
表 1. 聚类评价指标

评价指标	轮廓系数	调整兰德指数	轮廓宽度	Dunn 指数	稳定性指数
数值	0.42	0.72	1200	2.1	0.83

其中，轮廓系数为 0.42，表明大部分样本在各自聚类中具有较好的相似性，同时与其他聚类有一定的分离度；调整兰德指数为 0.72，说明聚类结果与真实类别较为一致；轮廓宽度(WCSS)为 1200，表明类内数据紧密度较高，聚类效果较为理想；Dunn 指数为 2.1，反映不同聚类之间的边界清晰度较高；稳定性指数(Jaccard Index)为 0.83，表明聚类结果在不同运行之间具有较高的稳定性。这些指标表明本次聚类效果较好，能够有效区分不同的细胞群体，同时保持较高的聚类稳定性和紧密性。

4.2. 生物学解释与意义

由图 15，我们可以看到 Naive CD4 T 细胞是初始型 $CD4^+$ T 细胞，尚未被抗原激活，具有分化潜能，可转化为 Th1、Th2、Th17、Treg 等亚型。Memory CD4 T 细胞是具有免疫记忆的细胞，能快速响应相同抗原的再次入侵。$CD14^+$ Mono 是经典单核细胞，主要通过吞噬作用清除病原体，并可分化为巨噬细胞或树突状细胞。B 细胞负责适应性免疫中的抗体生成，并能作为抗原呈递细胞。CD8 T 细胞是细胞毒性 T 细胞，能直接杀死受感染或癌变细胞。$FCGR3A^+$ Mono 是非经典单核细胞，主要参与炎症反应、趋化因子分泌、抗体依赖性细胞介导的细胞毒性。NK 细胞是自然杀伤细胞，不依赖抗原特异性，可直接杀死病毒感染细胞或肿瘤细胞。DC 细胞是树突状细胞主要作用是抗原呈递，能激活 T 细胞并诱导适应性免疫

反应。Platelet 细胞是血小板细胞，主要负责止血和血栓形成，同时参与炎症和免疫调节。

Table 2. Key biological pathways related to cells
表 2. 细胞相关的关键生物学通路

相关通路	相关细胞	重要通路	作用(或相关疾病)
适应性免疫相关通路	Naive CD4 T, Memory CD4 T, CD8 T	T 细胞受体信号通路	介导 T 细胞活化和分化，调节免疫应答
	CD8 T	细胞毒性 T 细胞介导的细胞毒性	通过 Fas-FasL 和 Perforin-Granzyme 途径杀死靶细胞
	B	B 细胞受体信号通路	介导 B 细胞活化，促进抗体生成
固有免疫相关通路	$CD14^+$ Mono, $FCGR3A^+$ Mono, NK, DC	TLR 信号通路	识别病原体 PAMPs，激活先天免疫反应
	NK	NK 细胞介导的细胞毒性	通过 MHC 识别异常细胞并杀伤
	DC	抗原处理和呈递	促进适应性免疫应答的启动
炎症与疾病相关通路	$CD14^+$ Mono, $FCGR3A^+$ Mono	NLRP3 炎性小体通路	炎症性疾病，如克罗恩病、类风湿性关节炎。
	Platelet	凝血系统	参与血栓形成，关联动脉粥样硬化和心血管疾病
	T, B	自身免疫通路	相关疾病包括系统性红斑狼疮(SLE)、多发性硬化(MS)。

我们使用基因集富集分析(GSEA)可以找到这些细胞相关的关键生物学通路，如表 2 所示。

参考文献

[1] 盛杰, 王玉欣, 齐江发, 等. 单分子测序技术在肿瘤诊断中的应用研究进展[J]. 生物工程学报, 2020, 36(2): 180-188.

[2] 王晶晶, 俞海国, 樊志丹. 单细胞核转录组测序揭示阿司匹林抑制川崎病小鼠模型心脏组织的血管新生[J]. 药学学报, 2024, 59(10): 2809-2819.

[3] 傅东升, 艾克热木江·木合热木. GEO 公共数据库中 4 例多指畸形患者细胞间质和上皮细胞的单细胞转录组分析[J]. 中国组织工程研究, 2025, 29(20): 4379-4388.

[4] 安宁, 张福燕, 秦波. 单细胞转录组测序在年龄相关性黄斑变性研究中的应用[J]. 国际眼科杂志, 2022, 22(6): 964-968.

[5] 许淑静, 陈国洪. 癫痫患者血清中 T 细胞亚群结构与患者病情及认知功能的关系[J]. 中国卫生检验杂志, 2019, 29(12): 1491-1493.

[6] 王宁, 郭梓昱, 田淑珂, 等. 基于融合特征 t-SNE 降维的控制图质量异常模式识别[J]. 系统工程理论与实践, 2024, 44(7): 2381-2393.

[7] 邓建国, 张素兰, 张继福, 等. 监督学习中的损失函数及应用研究[J]. 大数据, 2020, 6(1): 60-80.

[8] 温兆琦, 王晓哲, 侯艳芳, 等. 基于 sc RNA-seq 数据的单细胞非线性降维方法[J]. 数学建模及其应用, 2023, 12(3): 33-44.

[9] 杨品, 任振华, 袁增强. 多模态组学数据整合方法的性能评测[J]. 基因组学与应用生物学, 2024, 43(7): 1196-1213.

[10] 田英, 郝兆才. 基于增强加权共现图和图核相似性的文本分类方法[J]. 计算机工程与设计, 2023, 44(5): 1434-1440.

[11] 皇站飞, 赵桂华. 空间解析单细胞转录组的优化算法研究[J]. 上海理工大学学报, 2024, 46(6): 698-707.

[12] 郭雨茜, 李华玲. 基于标签噪声鲁棒学习的疾病风险预测[J]. 计算机系统应用, 2023, 32(10): 184-191.

[13] 陈仲扬, 马艳妮, 余佳. 基于单细胞转录组测序分析描绘人类早期胚胎红细胞发育图谱[J]. 基础医学与临床, 2022, 42(5): 776-781.