姓 名	苏烨
学 号	201921080038

成 绩	
评阅人	

中南财经政法大学

本科课程论文

课程名称 ：	数据挖掘与商务智能
论文题目 ：	基于 3kPBMC 数据集的单细胞观测嵌入算法对比分析
学　　院 ：	统计与数学学院
专业班级 ：	大数据 1901
姓　　名 ：	苏烨
学　　号 ：	201921080038
序　　号 ：	02
完成时间 ：	2022 年 11 月 17 日

一、引言

（一）相关背景及研究目的

对于绝大多数生物而言，细胞核内（或者核区）的 DNA（Deoxyribonucleic acid，即脱氧核糖核酸）是其主要的遗传物质，引导着生物的生长发育与生命机能运作。但在细胞层面，细胞的生长、分裂、分化等生命活动则并非直接由 DNA 进行调控。根据中心法则，DNA 必须先通过转录产生编码 RNA（Ribonucleic acid，即核糖核酸），再由这些携带遗传讯息的 RNA 完成蛋白质的合成。此外，细胞中还有一系列没有编码蛋白质能力的非编码 RNA，它们通过催化生化反应，或调控、参与基因表达过程[①]而发挥相应的生理功能。所以，从微观角度来说，细胞的各种功能产物大都是由 RNA 序列直接作为前体合成的。通过研究在单个细胞，或特定类型细胞、组织、器官或发育阶段的细胞群内所产生的各类 RNA 分子的类型和数量，我们就能间接地研究这些细胞中各类基因的表达情况。

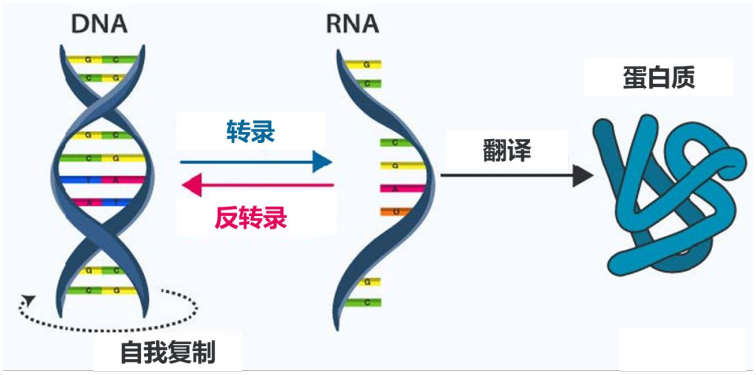


图 1　中心法则

以人体为例，人体完成一项或者多项生理活动需要各个系统的共同参与，而系统中的各类器官、组织的正常运转又离不开成千上万种细胞的通力协作。但这成千上万种人体细胞却共用同一套染色体。也就是说个体内各种细胞活动的差异，不是由细胞间染色体的差异引起的，而是由细胞内各类转录、调控活动的差异性造成的。这种转录、调控活动上的差异，即是基因差异表达的结果，它让各种细胞具有了不同的功能，并在人体复杂多样的生命活动中发挥着各自特殊的作用。再次回顾我们前面论述的中心法则，基因的差异表达，在微观层面上将体现为：在不同类型、组织、器官或细胞群中的细胞内，产生了不同类型、数量各异的

[①] 基因是指能指导产生具有某种生物功能产物（如 RNA，蛋白质分子）的一段特定的 DNA 序列（在某些生命体中，是一段特定的 RNA 序列）。基因表达即为细胞产生具有某种生物功能产物的生命过程。

RNA 分子。至此，一个自然的想法产生了，如果能测定细胞群内各个细胞中的 RNA 分子的类型和数量，我们将能揭示不同细胞在基因表达上的差异，进一步探究细胞异质性、多样性以及单细胞内分子间互作关系。

这种差异研究的意义是积极而深远的。如在癌症领域，在肿瘤组织中，肿块中心的细胞、肿块周围的细胞、淋巴转移灶的细胞，以及远端转移的细胞，其基因组和转录组[②]等遗传信息，是存在差异的[1]。通过单细胞 RNA 测序等技术手段揭示这种差异，在临床上，可以帮助我们判断针对该肿瘤的某种疗法是否有效；而在宏观上，这将直接推动精准医学的发展。

单细胞 RNA 测序（Single cell RNA sequencing，scRNA-seq）技术，是一种帮助我们理解单细胞水平上遗传信息表达异质性的技术，为我们弄清生命体遗传、发育、疾病机理打开了新的大门。单细胞测序技术自其问世以来[2]，已有十多年的发展历史。相关研究不断涌现，其重要性也不言而喻。根据 Luecken 等人的总结，测序后一般的分析流程如图 2 所示，包括：

1、数据预处理

①处理测序仪产生的输出，得到计数矩阵等原始的数据信息

②对计数矩阵等原始的数据信息进行质量控制、标准化、数据校正和特征选择（或者为基因选择）

③对留下的细胞观测在新的空间维度中进行可视化观察

2、下游分析

①对细胞观测进行聚类

②通过类簇中的高变基因等信息识别类簇中的标记基因（即 Marker 基因）

③对类簇进行注释

④单细胞轨迹推断

⑤单细胞组成分析

⑥单细胞差异表达分析

[②] 所有转录产物的集合，包括信使 RNA、核糖体 RNA、转运 RNA 及非编码 RNA。

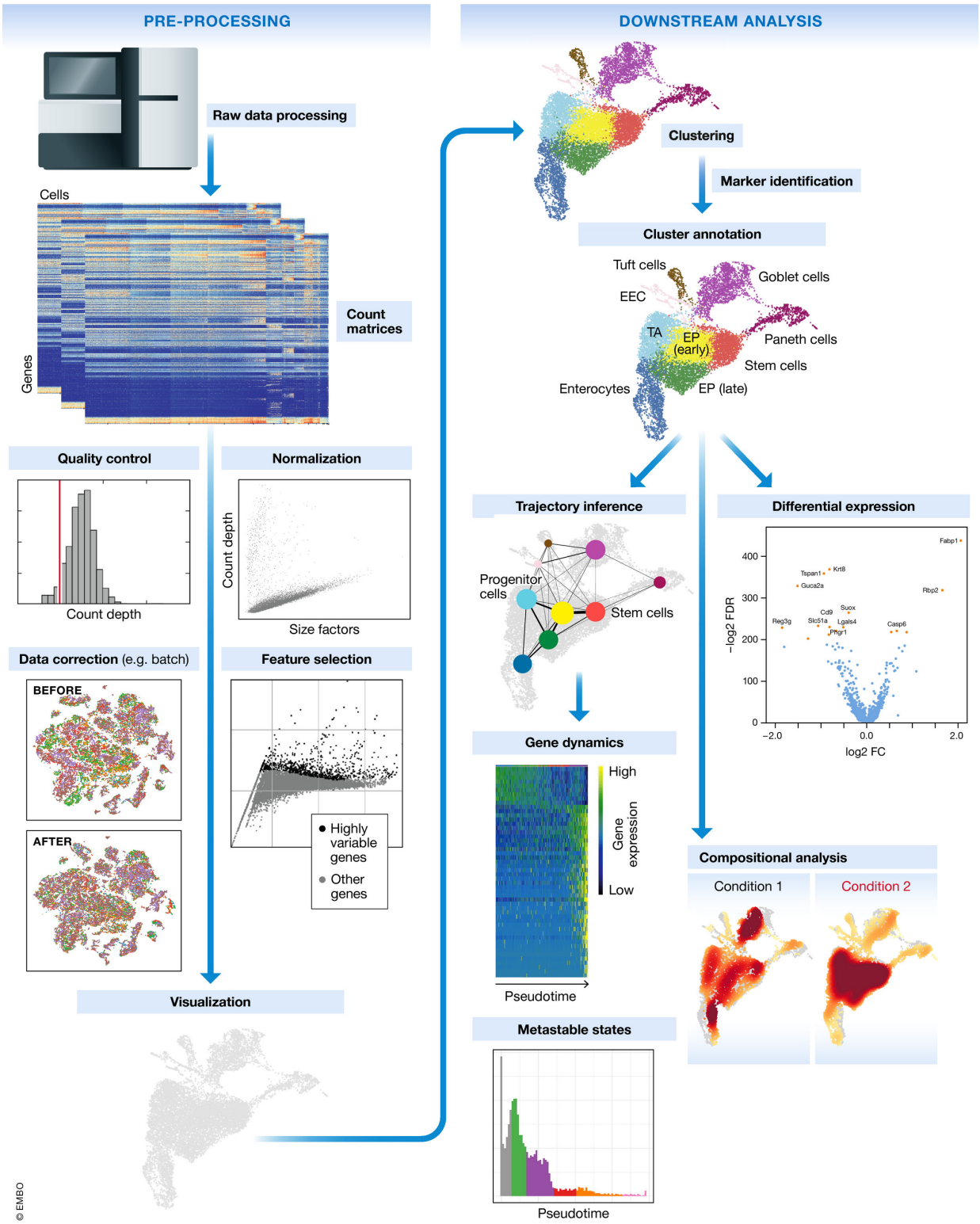


图 2　典型的 scRNA-seq 分析工作流程示意图

在数据预处理部分中，将细胞观测嵌入到新的空间维度中进行可视化观察，是一个非常重要的步骤。在绝大多数情况下，由测序仪输出整理得到的计数矩阵高达上万维，即使经过了初步的质量控制和特征选择，数据依然会是稀疏的高维数据（基因数高于观测数）。而为了可视化细胞的基因表达（也可以理解为分布），我们显然需要将细胞观测降维并嵌入到二维平面或者三维立体的新空间。并且在直观上，基因表达相近的细胞观测应当在可视化结果中聚在一起。

在当前 scRNA-seq 分析实践中，我们通常采用三种算法进行单细胞观测降维嵌入，即：PCA（Principal component analysis）[3]、t-SNE（t-distributed Stochastic Neighbor Embedding）[3,4,5]和 UMAP[6]（Uniform Manifold Approximation and Projection）。在本文中，我们希望通过 3kPBMC 数据集，对这三种单细胞观测嵌入算法进行仔细的对比分析，以确定 scRNA-seq 分析工作中可视化细胞基因表达的最佳实现方式。

（二）相关研究进展

降维算法在数据科学中扮演着重要的角色。在我们进行机器学习任务时，降维是对原始数据进行可视化及预处理的一类基础性工作。在 scRNA-seq 分析流程中，我们使用降维算法对单细胞观测进行重新嵌入的意义主要有二：原始数据的计数矩阵往往是高维的，即特征数（基因数）大于细胞观测数，在高维的原始空间内我们无法通过可视化的手段查看细胞基因表达的差异，也无法对不同类型细胞之间的差异性获得清晰直观的认识；原始数据的计数矩阵往往还是稀疏的，通过降维我们能找到更优的数据表示，重新定义细胞间的“距离”度量，而直接使用计数矩阵进行后续聚类分析时会带来非必要的计算开销。

1933 年 Hotelling 使用主成分分析法对多变量问题提取主成分进行统计分析[7]，这是一种将复杂的多变量观测投影到新的“主成分”坐标以实现方差最大的降维技术。它通过正交化线性变换，把数据观测变换到一个新的坐标系统中，使得数据的任何投影的第一大方差在第一个坐标方向上，第二大方差在第二个坐标方向上，以此类推。因为该方法是简单的线性变换，且从信息论的角度考虑它的降维是最优的（保留的信息最多），时至今日仍被广泛运用于各类需要降维的数据科学任务场景。

由于旋转、平移和翻转等变换不会改变数据本身的结构，对于高维空间也是如此，研究者考虑用点之间的距离代替点的坐标，转变降维问题的形式。1964 年 Kruskal 提出 MDS（Multidimensional scaling）方法[8]，1969 年 Sammon 进一步提出 Sammon Mapping 方法[9]。两种方法都将降维问题转化为优化问题的形式，通过优化“结构相似”函数，寻求数据在低维空间的最优表示。2000 年 Tenenbaum 等人[10]又对该优化问题进行改进，将其中的欧

几里得距离推广为“测地距离”（在离散空间中两点之间的最短路径长度），提出了 Isomap 方法。但是，MDS、Sammon Mapping 等方法依然不能很好地处理高维数据。

随着降维技术被用于越来越广泛的研究领域和各类规模不断扩大增长的数据集，研究者们希望寻找到一种既可扩展到大量数据，又能够处理数据内部多样性的算法。在此背景下最具代表性的两种算法当属 t-SNE[4]和 UMAP[6]方法。其中，t-SNE 将距离转换为概率分布，并使用 t 分布代替传统 SNE 中的标准正态分布。Hinton 等人对其降维后可视化的效果进行了对比实验[11]，结果表明，t-SNE 对 Sammon Mapping、Isomap、LLE 取得了压倒性的胜利。

2020 年 McInnes 等人基于黎曼几何和代数拓扑理论，提出一种新的降维算法 UMAP。作者在 COIL20, MNIST, Fashion MNIST 等标准数据集上进行了可视化效果的对比实验，并进行了嵌入稳定性的对比分析。结果表明，UMAP 在可视化质量上与 t-SNE 相当，并且在保留更多全局结构的同时还有着优越的运行性能。不仅如此，McInnes 及 Becht 等人将 UMAP 应用于单细胞数据[12]，基于三个特征良好的单细胞 RNA 测序数据集，将 UMAP 与其他五种降维可视化技术的性能进行比较，发现 UMAP 提供了最快的运行时间、最好的可复现性，呈现了最有意义的细胞簇结构。

二、研究设计

（一）模型简述

1、主成分分析

假设我们的数据集为$D=\{x_1,x_2,\cdots,x_n\}$。每个样本x_i表示成d维向量，且每个维度均为连续型特征，则数据集D可以表示为一个$n\times d$的矩阵$\boldsymbol{X}$。为了简化描述，我们不妨认为每一维特征的均值为 0，降维过程即为使用线性方法将d维的数据降到l维（$l<d$），我们用$d\times l$的矩阵表示这个线性变化，则降维后的数据为：

$$\boldsymbol{Y}=\boldsymbol{XW}$$

整个方差为

$$Var(\boldsymbol{Y})=\frac{1}{n-1}\mathrm{tr}\left(\boldsymbol{Y}^{\mathrm{T}}\boldsymbol{Y}\right)=\frac{1}{n-1}tr(\boldsymbol{W}^{\boldsymbol{T}}\boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{XW})=\mathrm{tr}\left(\boldsymbol{W}^{\boldsymbol{T}}\frac{1}{n-1}\boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{XW}\right)$$

将原始数据集$\boldsymbol{X}$的协方差记为$\boldsymbol{\Sigma}=\frac{1}{n-1}\boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{X}$，则降维后的方差为$\mathrm{tr}(\boldsymbol{W}^{\boldsymbol{T}}\boldsymbol{\Sigma W})$。由于主成分分析得到目标是使得降维后的数据方差最大化，且降维后各特征不相关，则问题可表示为优化形式：

$$\max_{\boldsymbol{W}}\mathrm{tr}(\boldsymbol{W}^{\boldsymbol{T}}\boldsymbol{\Sigma W}),\qquad s.t.\ \boldsymbol{w}_{\boldsymbol{i}}^{\boldsymbol{T}}\boldsymbol{w}_{\boldsymbol{i}}=1,\qquad i\in\{1,2,\cdots,l\}$$

使用拉格朗日法求解：

$$L(\boldsymbol{W},\boldsymbol{\lambda})=\mathrm{tr}(\boldsymbol{W}^{\boldsymbol{T}}\boldsymbol{\Sigma W})-\sum_{i=1}^{l}\lambda_i\left(\boldsymbol{w}_{\boldsymbol{i}}^{\boldsymbol{T}}\boldsymbol{w}_{\boldsymbol{i}}-1\right),$$

其中λ_i为拉格朗日乘子。对参数求导并令导数等于 0，可得到：

$$\boldsymbol{\Sigma w}_{\boldsymbol{i}}=\lambda_i\boldsymbol{w}_{\boldsymbol{i}}$$

可见，我们要求的变换矩阵$\boldsymbol{W}$的每一个列向量$\boldsymbol{w}_{\boldsymbol{i}}$都是协方差矩阵$\boldsymbol{\Sigma}$的特征向量，而$\lambda_i$为对应的特征值。又$\boldsymbol{w}_{\boldsymbol{i}}^{\boldsymbol{T}}\boldsymbol{\Sigma w}_{\boldsymbol{i}}=\lambda_i\boldsymbol{w}_{\boldsymbol{i}}^{\boldsymbol{T}}\boldsymbol{w}_{\boldsymbol{i}}=\lambda_i$，进一步有

$$\mathrm{tr}(\boldsymbol{W}^{\boldsymbol{T}}\boldsymbol{\Sigma W})=\sum_{i=1}^{l}\boldsymbol{w}_{\boldsymbol{i}}^{\boldsymbol{T}}\boldsymbol{\Sigma w}_{\boldsymbol{i}}=\sum_{i=1}^{l}\lambda_i$$

则 PCA 最优化的数据方差等于原数据集$\boldsymbol{X}$的协方差矩阵的特征值之和。要使$Var(\boldsymbol{Y})$最大，我们只要先求得$\boldsymbol{\Sigma}$的特征值和特征向量，再取最大的l个特征值对应的特征向量即可。

2、t-SNE

对原始数据阵$X=\{x_1,x_2,\cdots,x_n\}$，t-SNE 基于梯度下降算法计算输出降维后的数据表示$Y^{(T)}=\{y_1,y_2,\cdots,y_n\}$。对于简化版的 t-SNE 算法[11]，我们需在迭代运算前，设定一些超参数。对于损失函数，我们需设定困惑度$Perp$；在优化方面，我们需设定迭代次数T，学习率η，下降动量$\alpha(t)$。

根据公式？？？，我们使用困惑度$Perp$计算数据点的成对相似度p_{ij}：

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

则成对相似度：

$$p_{ij}=\frac{p_{j|i}+p_{i|j}}{2n},$$

接着我们从多元正态分布$\boldsymbol{N}(0,10^{-4}\boldsymbol{I})$中生成初始解

$$Y^{(0)}=\{y_1,y_2,\dots,y_n\},$$

然后进行T次迭代，对于第t次迭代，我们需：

① 根据公式？？？计算低维相似度q_{ij}：

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

② 计算梯度$\frac{\delta C}{\delta Y}$：

$$\frac{\delta C}{\delta y_i}=4\sum_j\left(p_{ij}-q_{ij}\right)\left(y_i-y_j\right)\left(1+||y_i-y_j||^2\right)^{-1}$$

③ 更新解集：

$$Y^{(t)}=Y^{(t-1)}+\eta\frac{\delta C}{\delta Y}+\alpha(t)\left(Y^{(t-1)}-Y^{(t-2)}\right)$$

即可得到低维的数据表示。

3、UMAP

UMAP 是一种基于流形学习技术和拓扑数据分析思想的降维算法。

该算法基于对数据的三个假设：

①数据均匀分布在黎曼流形上；

②黎曼度量是局部常数（或可以近似）；

③我们关注的底层流形是局部连接的。

根据这些假设，可以用模糊拓扑结构对流形进行建模。通过搜索具有最接近的等效模糊拓扑结构的数据的低维投影来找到嵌入。具体的算法细节较为复杂，除了整体的 UMAP 算法以外，作者还提出了构造局部模糊单纯集算法、计算距离σ的归一化因子算法、初始化时图谱嵌入算法及嵌入优化计算算法。详细内容参见初始文献[6]。

（二）实验设计

在提出 UMAP 的原始文献[6]中，作者仅在一个生物数据集——**Mouse scRNA-seq** 数据集上进行了对比实验[13]。而其后他们又在另一篇文章使用 UMAP 测试了另外三个特征良好的单细胞数据集[14-16]，进行了单细胞观测嵌入的实验，并进行了诸如聚类分析的下游分析。在本文中，我们将使用 **3kPBMC** 数据集——一个健康供体的外周血单核细胞测序数据集，具体数据我们可以从 10X Genomics 官方网站下载 ③。

3kPBMC 数据集有 2700 个细胞观测，单个细胞基因数中值为 817，但最高单个细胞基因数达到了 32738，这使得原始数据呈现出高维、稀疏的特点。其基础信息如图 3 和表 1 所示。

表 1　3kPBMC 数据集计数信息

指标	值
细胞数	2,700
测序序列中细胞读段占比	96.8%
平均每个细胞的读段数	68,881
单个细胞基因数中值	817
平均每个细胞的 UMI 计数中值	2,197

③ 网址：https://support.10xgenomics.com/single-cell-gene-expression/datasets/1.1.0/pbmc3k

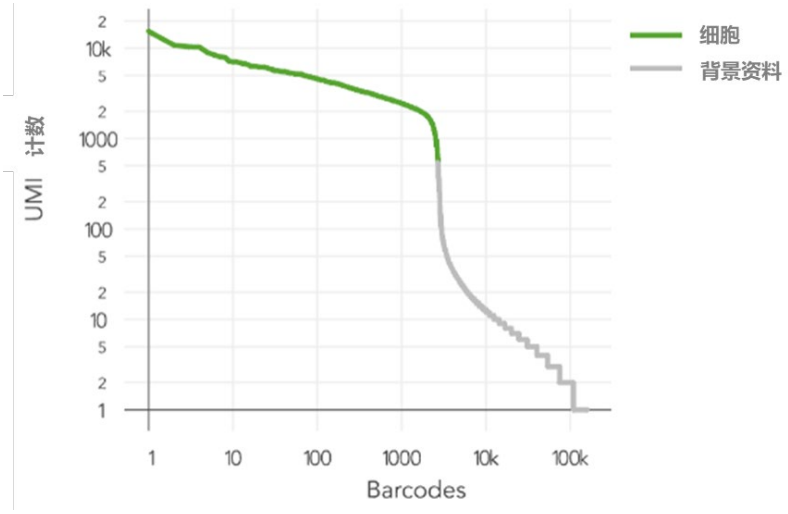


图 3　3kPBMC 数据集计数信息

我们将在 3kPBMC 数据集上从两个角度开展我们的单细胞观测嵌入算法对比分析。一方面，我们将通过分析探究 3kPBMC 测序数据中的高变基因，然后取 PCA、t-SNE、UMAP 降维后前两个维度的数据表示进行可视化，查看具有相同高变基因的细胞是否在二维平面上具有相近的距离；另一方面，我们将使用 Leiden 距离[17]对测序结果中的这近 3000 个细胞进行聚类，然后依然取 PCA、t-SNE、UMAP 降维后的前两个维度的数据表示进行可视化，查看在同一细胞簇中的细胞是否在二维平面上具有相近的距离。我们将通过这两方面的对比分析，来说明 scRNA-seq 分析工作中可视化细胞基因表达的最佳实现方式。

三、实证分析

（一）实验过程

1、寻找高变基因

高变基因（Highly Variable Genes，HVGs），是指细胞与细胞间进行比较后，基因表达量差别最大的那些基因。因为这些基因在不同的细胞之间的表达量差异很大，所以标记基因[④]（Marker Gene）通常是高变基因中很小的子集。但在寻找高变基因之前，我们需要先对测序数据进行一些预处理和探索性分析。

如图 4 所示，我们使用箱线图呈现了一些 3kPBMC 数据中的高表达基因，它们能反映每个基因在所有细胞中表达量的分布。我们进行了一些基本的质量控制，对表达基因数少于 200 的细胞和细胞观测数少于 3 的基因进行了剔除。

考虑到线粒体基因相对高表达的细胞可能发生了细胞穿孔，它们细胞质中的 RNA 发生了丢失，我们须将它们剔除。对此，我们需要计算每个细胞中有表达的基因数、基因总计数（总表达量）和线粒体基因表达量占比[⑤]，确定剔除细胞的阈值。如图 5 所示，我们注意到大部分细胞的线粒体基因表达量占比都在 5%以下，且每个细胞中有表达的基因数几乎都在 2500 个以下，所以我们对阈值以外的细胞进行剔除，使用剩余的 2638 个细胞观测进行后续的分析。

根据 Satija 等人的方法[18]，我们寻找 3kPBMC 数据集中的高变基因。图 6 显示了所有基因的平均表达和离散程度，我们用这些值计算得到高变基因，如 CST3、NKG7、PPBP。

2、高变基因表达可视化

我们对高变基因 CST3、NKG7、PPBP 的表达量分别在 PCA、t-SNE、UMAP 的 2 维嵌入中进行可视化，结果如图 7 所示。

3、细胞类簇可视化

使用 Leiden 距离对细胞进行聚类，并为各细胞簇打上标签，分别在 PCA、t-SNE、UMAP 的 2 维嵌入中进行可视化，结果如图 8 所示。

[④] 一种已知功能或已知序列的基因，能够起到特异性标记细胞类型（或者类簇）的作用。

[⑤] 这里的占比指每个细胞中，线粒体基因表达量占该细胞所有基因表达量的百分比。

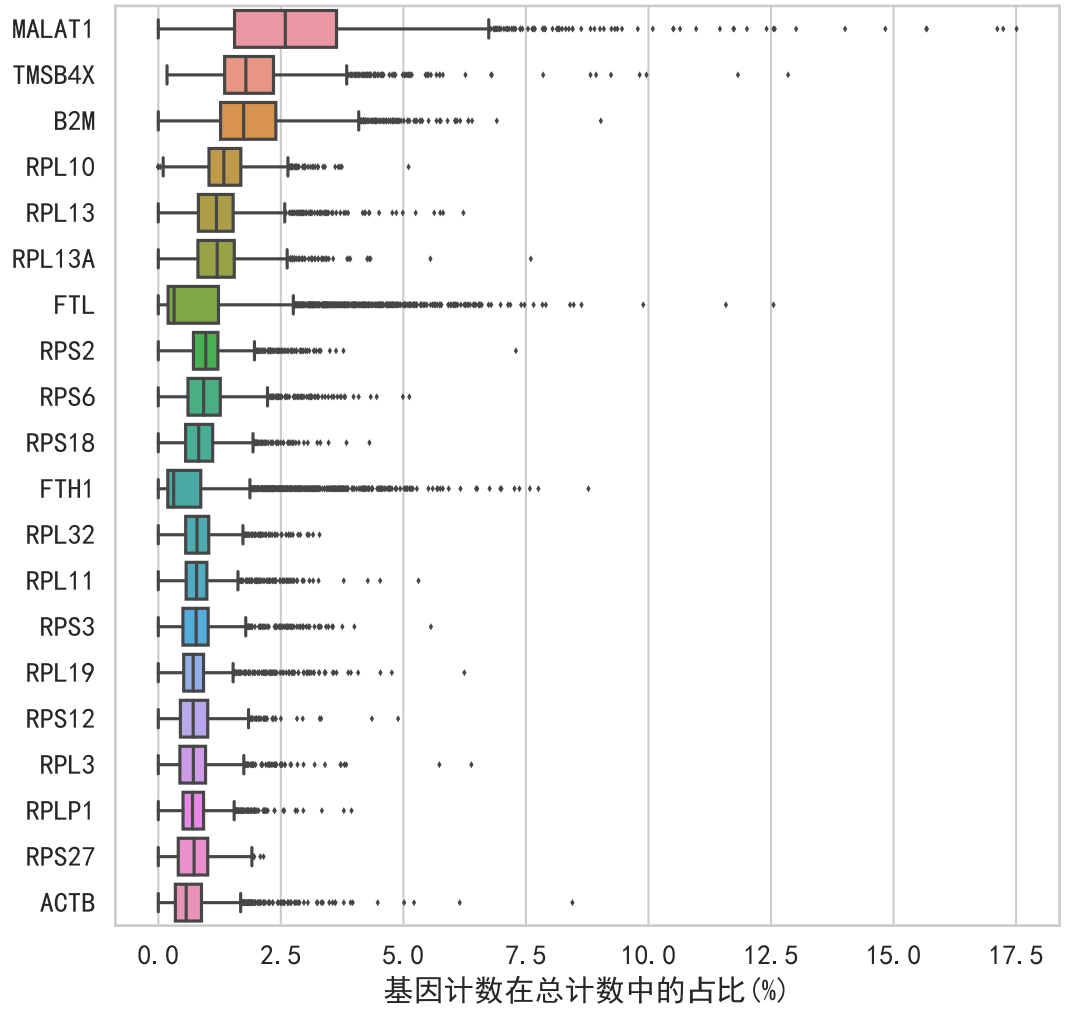


图 4　3kPBMC 中部分高表达基因的表达量分布

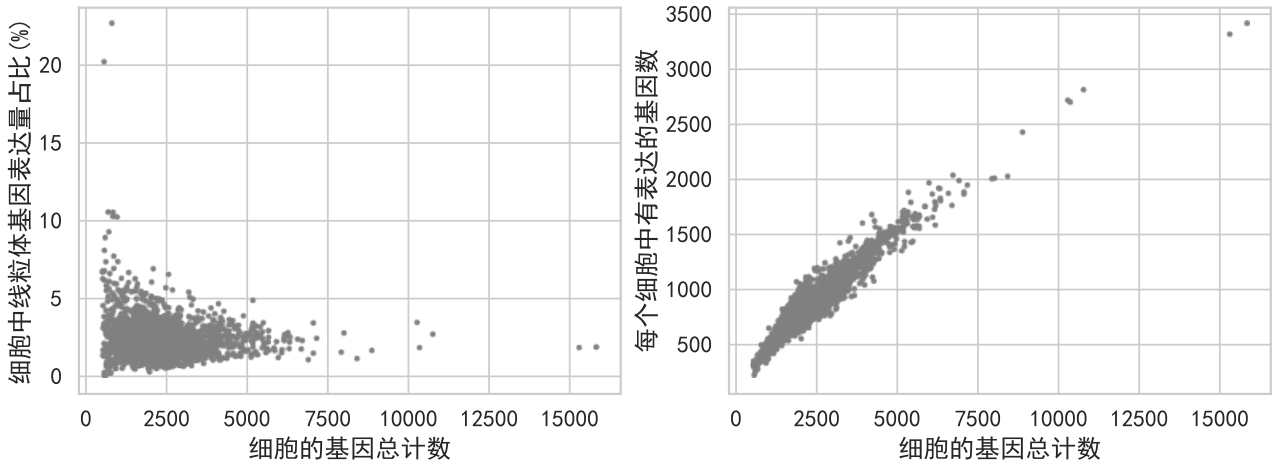


图 5　3kPBMC 中细胞基因表达的数值情况

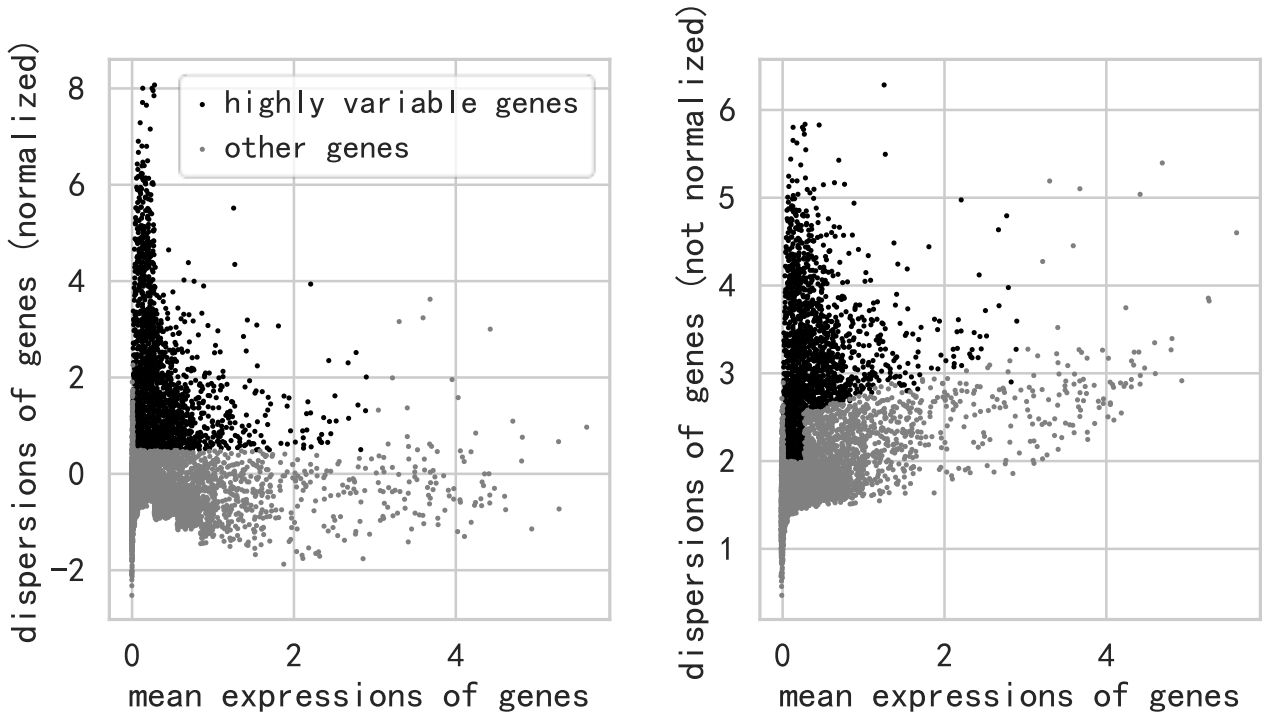


图 6　基因的平均表达和离散程度

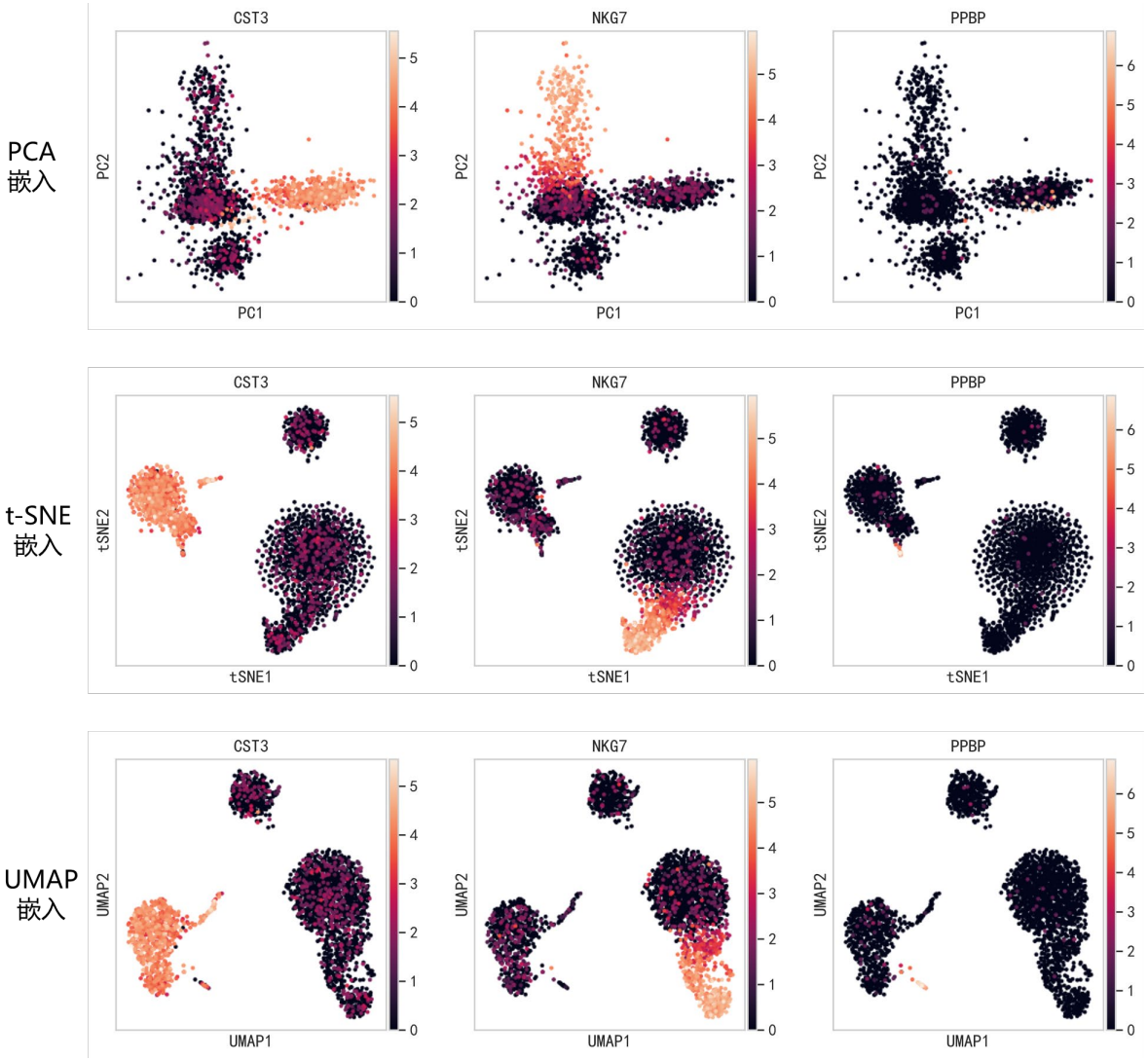


图 7　CST3、NKG7、PPBP 基因表达的可视化结果

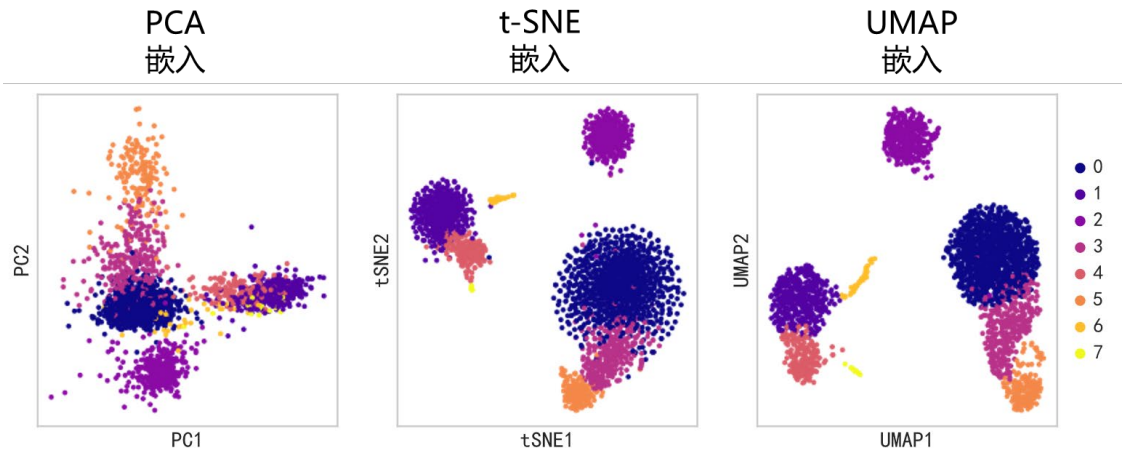


图 8　细胞聚类结果（基于 Leiden 距离）

（二）结果分析

1、定性分析

观察图 7 和图 8，我们可以明显发现，细胞观测在 PCA 下的嵌入是松散的，具有相同基因表达的细胞在平面距离上并不相近，且易与其他类型细胞混杂在一起；而 t-SNE 和 UMAP 几乎没有这样的问题——基因表达相同的细胞聚在一起，细胞类簇在平面中呈现清晰的边界和轮廓。同时，我们注意到，UMAP 和 t-SNE 具有相似的细胞簇分布结构，这一点在文献[6]中也有提到。

2、定量分析

我们更改部分随机数，在 3kPBMC 数据集上多次运行 PCA、t-SNE 和 UMAP，其降维运行性能如表 2 所示。我们发现，UMAP 相比于 t-SNE 有着优秀的运行性能。

表 2　算法的运行性能对比 ⑥

算法	运行时间（单位：秒）
PCA	2.6(±0.02)
t-SNE	7.6(±0.37)
UMAP	3.5(±0.13)

⑥ 硬件配置：处理器为 AMD Ryzen 7 4800H with Radeon Graphics 2.90 GHz；机载内存大小为 16GB。

四、结语

本文回顾了现有单细胞观测降维嵌入算法的相关原理，并在 3kPBMC 数据集上从两个角度开展了单细胞观测嵌入算法的对比分析。一方面，我们通过分析探究 3kPBMC 测序数据中的高变基因，然后取 PCA、t-SNE、UMAP 降维后前两个维度的数据表示进行可视化，查看具有相同高变基因的细胞是否在二维平面上具有相近的距离；另一方面，我们使用 Leiden 距离对测序结果中的这近 3000 个细胞进行聚类，然后取 PCA、t-SNE、UMAP 降维后的前两个维度表示进行可视化，查看在同一细胞簇中的细胞是否在二维平面上具有相近的距离，细胞簇是否有清晰的边界和轮廓。

我们发现，PCA 作为一种简单的线性降维技术，其在处理高维的单细胞测序数据时效果并不好。细胞观测在 PCA 下的嵌入是松散的，具有相同基因表达的细胞在平面距离上并不相近。当数据嵌入维数缩减到平面二维时，t-SNE 和 UMAP 方法展现出了更出色的可视化效果。而在运行性能上，PCA 是最快的，UMAP 只用了 t-SNE 一半的时间就完成了降维嵌入的过程，并和 t-SNE 保留了相似的细胞簇分布结构细胞簇分布结构。综合定性和定量的分析结果，我们认为 UMAP 是 scRNA-seq 分析工作中可视化细胞基因表达的最佳实现方式。

当然，我们仅在 3kPBMC 上进行了这样的对比分析，它仅仅是对 UMAP 可行性的一种补充。在附录中，我们给出了 UMAP 在标准数据集上的优秀表现作为补充的材料。在后续的研究中，我们还会通过一系列其他的下游分析进一步分析 UMAP 的效果。

参考文献

[1] Navin N E. Cancer genomics: one cell at a time[J]. Genome biology, 2014, 15(8): 1-13.

[2] Tang F, Barbacioru C, Wang Y, et al. mRNA-Seq whole-transcriptome analysis of a single cell[J]. Nature methods, 2009, 6(5): 377-382.

[3] Pedregosa F, Varoquaux G, Gramfort A, et al. Scikit-learn: Machine learning in Python[J]. the Journal of machine Learning research, 2011, 12: 2825-2830.

[4] Van der Maaten L, Hinton G. Visualizing data using t-SNE[J]. Journal of machine learning research, 2008, 9(11).

[5] Amir E D, Davis K L, Tadmor M D, et al. viSNE enables visualization of high dimensional single-cell data and reveals phenotypic heterogeneity of leukemia[J]. Nature biotechnology, 2013, 31(6): 545-552.

[6] McInnes L, Healy J, Melville J. Umap: Uniform manifold approximation and projection for dimension reduction[J]. arXiv preprint arXiv:1802.03426, 2018.

[7] Hotelling H. Analysis of a complex of statistical variables into principal components[J]. Journal of educational psychology, 1933, 24(6): 417.

[8] Kruskal J B. Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis[J]. Psychometrika, 1964, 29(1): 1-27.

[9] Sammon J W. A nonlinear mapping for data structure analysis[J]. IEEE Transactions on computers, 1969, 100(5): 401-409.

[10] Tenenbaum J B, Silva V, Langford J C. A global geometric framework for nonlinear dimensionality reduction[J]. science, 2000, 290(5500): 2319-2323.

[11] Van der Maaten L, Hinton G. Visualizing data using t-SNE[J]. Journal of machine learning research, 2008, 9(11).

[12] Becht E, McInnes L, Healy J, et al. Dimensionality reduction for visualizing single-cell data using UMAP[J]. Nature biotechnology, 2019, 37(1): 38-44.

[13] Campbell J N, Macosko E Z, Fenselau H, et al. A molecular census of arcuate hypothalamus and median eminence cell types[J]. Nature neuroscience, 2017, 20(3): 484-496.

[14] Han X, Wang R, Zhou Y, et al. Mapping the mouse cell atlas by microwell-seq[J]. Cell, 2018, 172(5): 1091-1107. e17.

[15] Samusik N, Good Z, Spitzer M H, et al. Automated mapping of phenotype space with single-cell data[J]. Nature methods, 2016, 13(6): 493-496.

[16] Wong M T, Ong D E H, Lim F S H, et al. A high-dimensional atlas of human T cell diversity reveals tissue-specific trafficking and cytokine signatures[J]. Immunity, 2016, 45(2): 442-456.

[17] Traag V A, Waltman L, Van Eck N J. From Louvain to Leiden: guaranteeing well-connected communities[J]. Scientific reports, 2019, 9(1): 1-12.

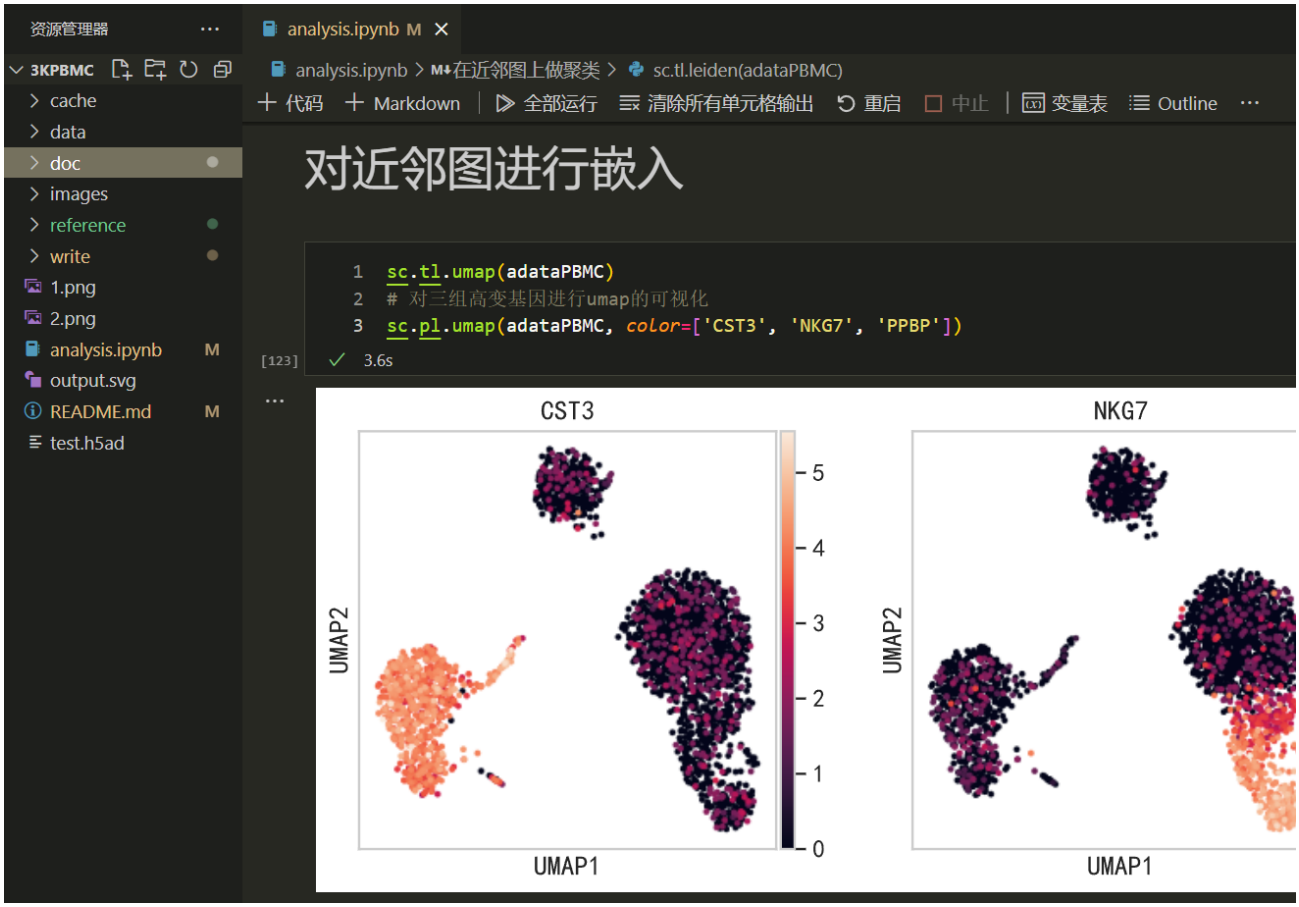
[18] Satija R, Farrell J A, Gennert D, et al. Spatial reconstruction of single-cell gene expression data[J]. Nature biotechnology, 2015, 33(5): 495-502.

附录

（一）代码

项目代码托管于 GitHub：https://github.com/suye0620/3kPBMC

截图：



（二）附图和附表

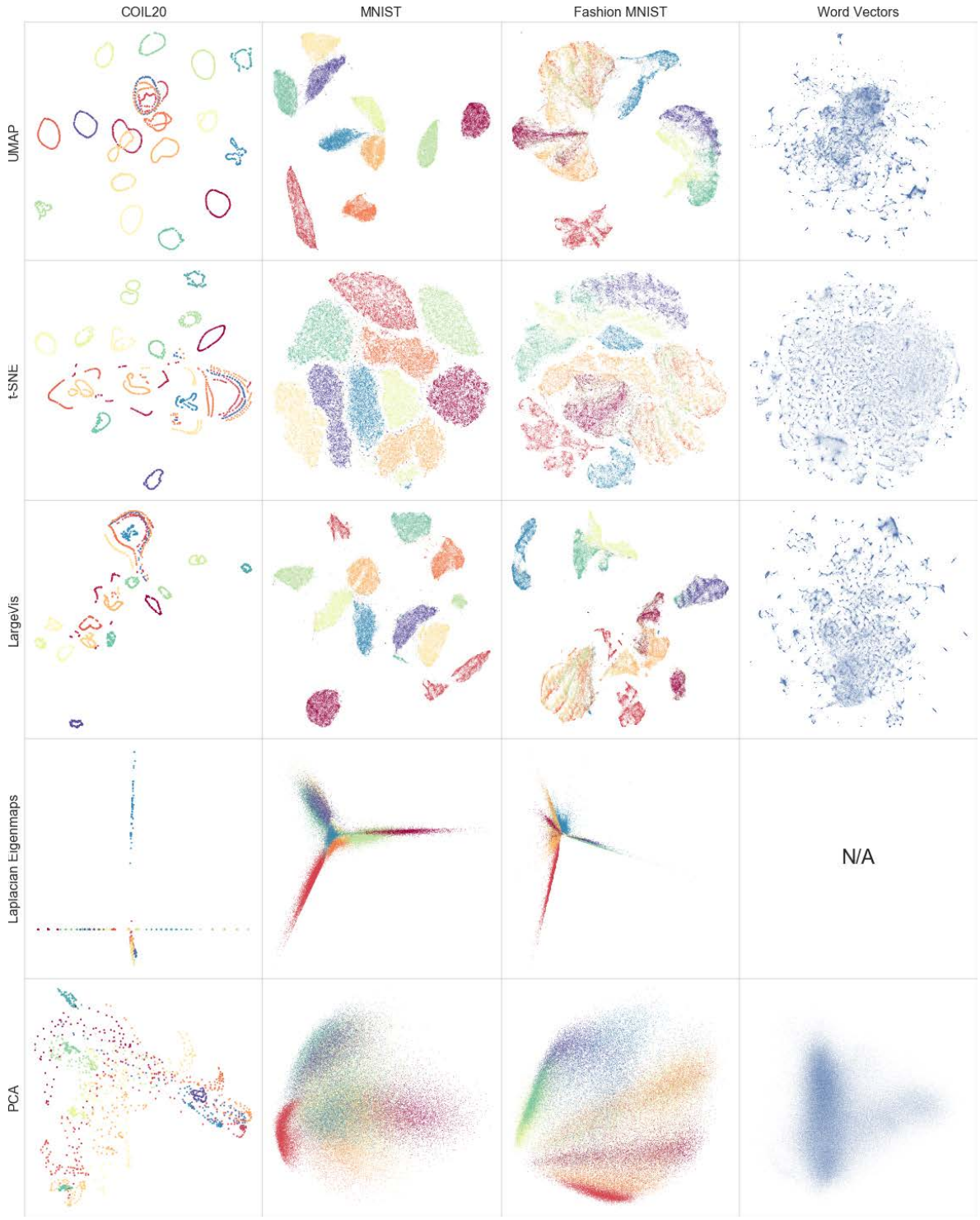


图 9　UMAP 在标准数据集上的降维效果

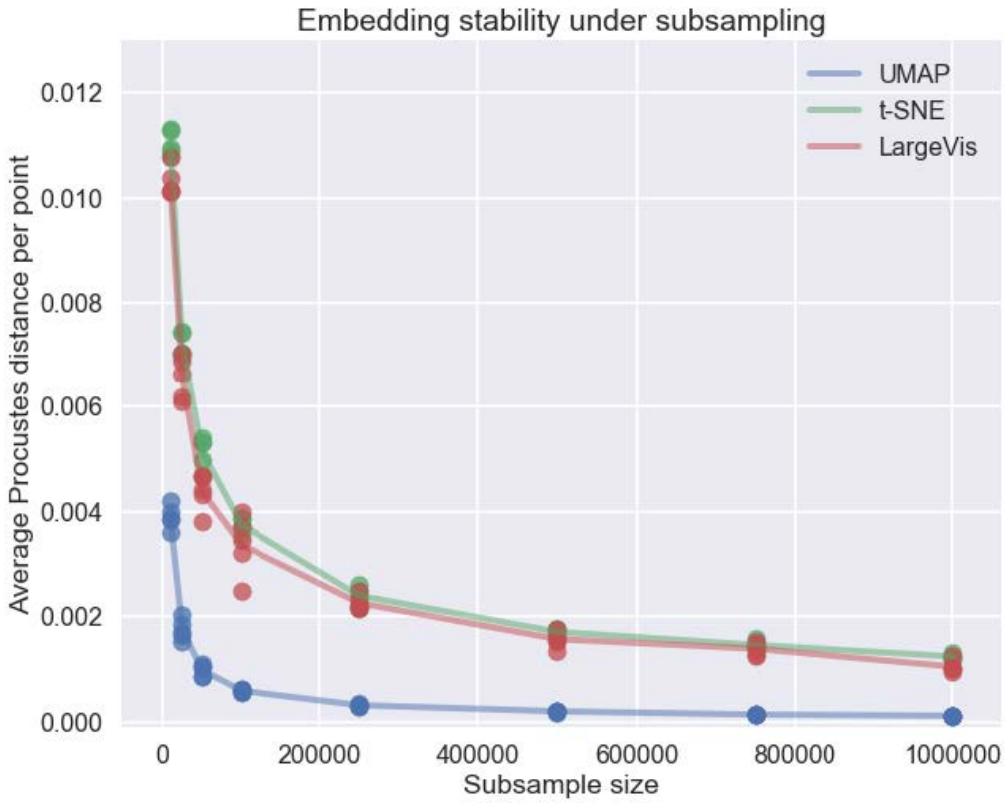


图 10　UMAP 降维算法的运行稳定性