[文章编号] 1003－4684(2026)01-0123-09

# 图聚类算法在单细胞 RNA 测序数据的分类问题中的应用

王景炜，郭依蓓，祝兆媚，李翰芳

（湖北工业大学理学院，湖北 武汉 430068）

[摘 要] 综合考虑数据处理的常规方法以及细胞和基因数据的独特性，对原始数据做预处理后用图聚类算法对预处理后的数据进行分类，最终将所有的细胞分为了 7 类。最后，分别用 UMAP 和 t-SNE 非线性降维技术，将数据降至 2 维，再对分类结果进行可视化分析。基于基因在不同处理下的表达量相同得到的分类结果，采用 FC 法和统计检验法相结合的方法选定阈值：$P<0.05$，同时 $\log_2 FC \geqslant 2$，$FDR<0.01$，来寻找差异表达的细胞基因，而后分同一类别、三个类别、所有类别等三种情况，分析了每种情况中差异化表达的基因。

[关键词] 单细胞 RNA 测序；非线性降维；图聚类；FC 法；差异统计检验

[中图分类号] O212 [文献标识码] A

单细胞 RNA 测序技术被广泛应用于基因表达谱的定量研究，它的优点不仅有高通量和高深度，同时也能精确地测定每个细胞的状况。如今这项技术已经快速发展，并且在当前的生命科学领域中处于前沿地位。随着单细胞 RNA 测序技术的飞速进步，数据量和数据复杂度陡增，给数据分析方法带来了巨大的挑战，因此科学的数据分析方法成为了重中之重。单细胞 RNA 测序是在单细胞水平上用 RNA 测序技术对特定细胞群体做基因表达谱定量的一类高通量实验技术[1]，在经过单细胞分离、RNA 提取、逆转录、文库构建和测序等步骤后，分析待测组织可得到多个细胞的基因表达谱，龚乐君等(2020)[2]就单细胞测序数据中的细胞类型提出了基于随机森林的细胞类型自动识别方法。Tang 等(2009)[3]首次报道了使得多个细胞中数千基因的转录水平可以同时被检测出的 scRNA-seq 技术。scRNA-seq 技术相较于传统的转录组学测序，通过单个细胞独特的基因表达模式可以展现出群体的细胞异质性。关于处理 scRNA-seq 的技术也陆续被开发，使得单细胞 RNA 测序迅速发展起来，如今，scRNA-seq 技术已经在生命科学的许多领域中被科学家们广泛研究应用，Philip (2013)[4]开发了一种在逐个基因的基础上观察到的细胞间表达强度变异的进行量化的定量统计方法。Rhonda(2016)[5]提出了用于 scRNA- seq 实验设计和分析的计算方法，Heng 等(2018)[6]使用单细胞 RNA 测序技术对小鼠乳腺上皮细胞进行分类和分层分析，通过分析不同细胞群集的基因表达模式，能够识别出乳腺上皮细胞的不同亚群，从而了解细胞层次结构和发育过程中的细胞转变，杨佳凤(2021)[7]等研究了单细胞转录组测序技术在细胞分类中的应用。

单细胞聚类分析是根据表达谱中细胞与细胞间的某些相似性来将细胞逐个分配给不同的集群的过程，即根据细胞间的关系揭示细胞亚型和推断细胞谱系。Raphael(2020)[8]等综合用于聚类 scRNA-seq 转录组的机器学习和统计方法，介绍了如何修改或定制传统的聚类技术以解决 scRNA-seq 数据分析中的独特挑战。由于近几年来单细胞 RNA 测序的飞速发展，以现有的数据分析方法处理复杂的数据和庞大的数据量还是当前的一大难题。就单细胞 RNA 测序数据样本庞大的数据量这一问题，Wang 等(2018)[9]提出了基于多核学习，从单细胞 scRNA-seq 数据中学习相似性度的分析框架 SIMLR，该方法弥补了 K-means 算法的不足，可以实现降维、聚类以及异构细胞群的可视化。Andrew 等(2018)[10]提出了整合 scRNA-seq 数据集，实现了跨数据集共享群体的识别和下游比较分析的 Seurat 无监督聚类算法，该算法是将图形分割法结合降维

[收稿日期] 2023－11－04

[基金项目] 湖北省教育厅哲学社会科学研究项目(22Y059)；湖北工业大学博士启动基金(BSQD2020103)

[第一作者] 王景炜(1998－)，女，安徽安庆人，湖北工业大学硕士研究生，研究方向为数理统计、贝叶斯分析、经验似然。

[通信作者] 李翰芳(1979－)，女，湖北麻城人，理学博士，湖北工业大学讲师，研究方向为计量经济建模、贝叶斯分析、高斯混合模型。

算法来对单细胞数据进行聚类。何睿(2020)[11]讨论了现有的单细胞聚类模型,并对不同聚类模型进行性能评估。本文在综合考虑各种方法后主要使用基于图聚类算法的 Seurat 无监督聚类算法对单细胞数据进行分析研究。

# 1 原始数据及预处理流程

## 1.1 原始数据

本文的数据来源于第五届应用统计专业学位研究生案例大赛,原始数据一共三个数据表,依次列出了单细胞 RNA 基因的 barcode 信息、单细胞信息,以及相应的细胞信息,基因信息和基因数量,需要分析以及进行后续处理的数据正如表 1 所示。

**表 1 原始数据**

| 指标 | cell | barcodes | Counts 总数 |
|---|---|---|---|
| 数量 | 8528 | 800 | 653064 |

高维数据不仅使得计算量陡然增加,加大了计算难度,而且较为抽象,不利于进行可视化分析。为了提高分析的准确性,需要对其进行合理的数据预处理。

## 1.2 数据预处理流程

除了一些常规方法如数据归一化法外,还考虑了 RNA 的独有的特征例如质控指标,高异质性基因检测方法等进行数据预处理。

**1.2.1 基于质控指标来筛选细胞** 筛选细胞考虑的质控指标主要有两个:线粒体的基因表达比例、nFeature_RNA。

1)要控制线粒体的基因含量不能过高,用线粒体 qc(quality control)指标衡量,当细胞开始凋亡时,线粒体基因会参与其中,在垂死细胞中线粒体基因的表达比例就会上升。因此,如果线粒体的基因含量很高,说明该细胞即将死亡,并不是我们要分析的正常细胞;在使用组织块制造单细胞悬液的过程中,难免会造成部分细胞的细胞膜破坏,而一旦细胞的细胞膜破坏,细胞质中的转录本就会大量流失,此时检测线粒体时就会表现出大量比例的线粒体基因表达。如果线粒体的基因含量很高,说明该细胞遭到了破坏,不能进行正常的基因表达,因此需要控制线粒体基因含量不能过高。

2)要控制 nFeature_RNA 不能过低也不能过高,其中 nFeature_RNA 代表每个细胞测到的基因数目。nFeature_RNA 过低,表示该细胞可能已死或是将死或是空液滴,因此只含有少量基因。过高则表明"细胞"实际上可以是两个或多个细胞,因此有异常的高基因数。但不管细胞是已死、将死、空液滴,或是两个或多个细胞都会对后续的分析造成影响,因此,需要控制 nFeature_RNA 保持在一个合理的范围内。

不管是基于线粒体的基因表达比例还是基于 nFeature_RNA,需要过滤掉的都是非正常的细胞。这些细胞不管是质量较低、濒临死亡、空液滴,还是两个或多个细胞,通常都表现出广泛的线粒体污染,所以可以结合线粒体基因的 count 数除去异常值。

利用 PercentageFeatureSet 函数计算出线粒体的 qc 指标。结合上述两个质控指标,将线粒体基因的表达比例设置为小于 5%,将 nFeature_RNA 设置为 200 到 2500 之间,从而将线粒体的基因表达超过 5%以及 nFeature_RNA 小于 200 和超过 2500 的异常细胞都过滤掉,得出初步筛选的结果见表 2。

**表 2 筛选后数据**

| 指标 | cell | barcodes |
|---|---|---|
| 数量(筛选前) | 8528 | 800 |
| 数量(筛选后) | 8315 | 796 |

过滤掉非正常细胞后,检测到 8315 个特征,和 796 个细胞,相对于原始数据的 8528 个特征和 800 个细胞,过滤掉了 213 个特征和 4 个细胞。

筛选之后的细胞可视化结果如图 1 所示。

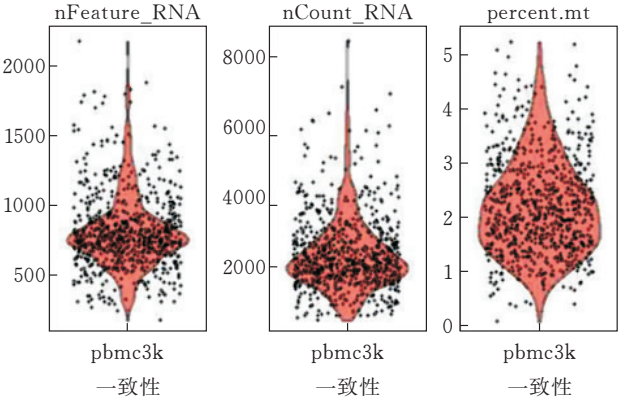


图 1 质控指标筛选细胞结果

图 1 中"nFeature_RNA"表示在每个细胞中检测到的基因个数,"nCount"表示在检测出的所有基因表达量的总和,"percent.mt"表示检测出的线粒体基因的比例。

通过可视化结果可以看到,超过 2500 或少于 200 个基因计数的细胞和线粒体比例超过 5%的细胞已经被过滤掉,只剩下符合要求的细胞。进一步查看特征与特征间的相关性可见,绝大多数细胞的基因表达量的总和与线粒体基因的比例在正常范围内;线粒体基因的比例异常高时,基因表达量的总和则异常低,反之亦然;基因表达量的总和与基因个数基本成正比。

基于筛选细胞的两个质控指标,可以通过 Seurat 对象里的单细胞数据集,对数据有进一步的

了解。

**1.2.2　数据归一化和高异质性基因检测**　从数据集中过滤掉可能会对结果造成干扰的细胞之后，还需要对高异质性基因进行检测和处理。高异质性基因是一类在某些细胞中具有高表达，但在另一些细胞中又会具有较低的表达的特殊基因。这些高异质性基因有助于在进行 scRNA-seq 的下游分析时找到单细胞数据集中的生物信号，因此需要计算出细胞中的高异质性基因子集。

由于高维数据具有多个特征，且特征的量纲和数值的量级不一样，无法进行统一处理，因为统一的处理会增大出错风险，且无法很好地展示不同特征的影响，因此需要对数据进行标准化处理，消除特征间的量纲关系，将数值的量级进行统一，从而进行统一处理，且具有可比性。

本文使用的是 log 函数标准化法进行全局缩放的归一化方法，该方法是通过总表达值对每一个细胞的基因表达值进行归一化，并将其乘以缩放因子（默认值 10000），最后再对结果进行对数变换。将数据进行标准化处理之后，就可以检测高异质性基因。在每个数据集中，识别 2000 个特征，并将前 10 个高异质性基因在图中标注出来。图 2 为检测出来的前 10 个高异质性基因。

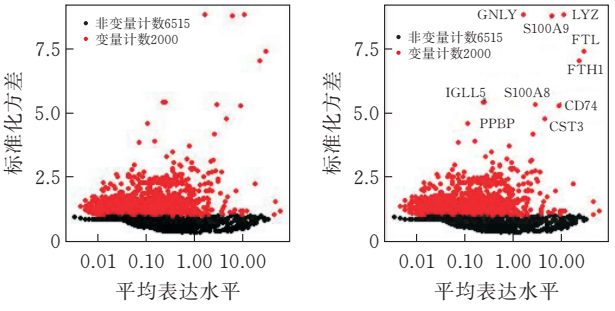


图 2　前 10 个高异质性基因

**1.2.3　缩放数据与降维**　归一化后的数据仍属于高维数据，不仅数据量很大而且难以解释。将数据减少到更少的维度通常会使分析算法更高效，结果更准确，也更加便于可视化。数据降维的主要方法有线性映射和非线性映射两大类。这里采用线性映射中的主成分分析方法，默认降维至 50 维。

而主成分分析的目的是试图得到方差最大的特征，对于高幅值的特征，方差较大，这使得主成分分析倾向于高幅值特征，低幅值的特征无疑就会被缩小，这将会导致结果有偏差，因此，在用主成分分析方法进行数据降维之前要先对数据进行缩放。结合本文细胞和基因数据的独特性，对数据进行线性变换（scaling）操作，具体的过程是：第一，需要改变所有基因的表达，使得所有细胞的平均表达水平保持在 0；再对所有的细胞表达进行缩放，使得所有细胞之间的差异为 1，对缩放好的数据，用主成分分析方法进行降维处理。

利用主成分分析方法将数据降维至 50 维的结果，其中每一个 PC 中各含有 25 个 positive 特征和 25 个 negative 特征，一共 50 个维度。

每个 PC 的 positive 和 negative 中的前 5 个维度见表 3。

细胞 PCA 的可视化分布情况见图 3。

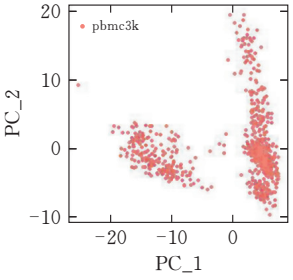


图 3　PCA 分布

**表 3　每个 PC 的前 5 个维度**

| PC | 特征 | 维度 | | | | |
|---|---|---|---|---|---|---|
| PC_1 | Positive | MALAT1 | RPS27A | LTB | CD3D | IL32 |
| | Negative | CST3 | TYROBP | LST1 | AIF1 | FCN1 |
| PC_2 | Positive | NKG7 | PRF1 | CST7 | GZMB | FGFBP2 |
| | Negative | CD79A | MS4A1 | LINC00926 | LTB | TCL1A |
| PC_3 | Positive | PF4 | ACRBP | PPBP | PTCRA | GP9 |
| | Negative | S100A8 | LGALS2 | RPS9 | VIM | CEBPD |
| PC_4 | Positive | CD3D | LDHB | CD3E | IL32 | IL7R |
| | Negative | CD79A | CD79B | MS4A1 | TCL1A | CD74 |
| PC_5 | Positive | S100A8 | LGALS2 | S100A9 | CD14 | MS4A6A |
| | Negative | MS4A4A | MS4A7 | LILRB1 | CXCL16 | CKB |

前 500 个细胞在前 6 个 PCA 中的热图可视化结果如图 4。

每一个 PC 其本质上是代表了一个“元特征”。它能够将相关特征的集中信息都组合在一起，越是

在顶部的主成分就越是有可能代表了数据集。采用 JackStraw 函数来选择尽可能包含原数据的 PC 的数目，结果如图 5 所示。

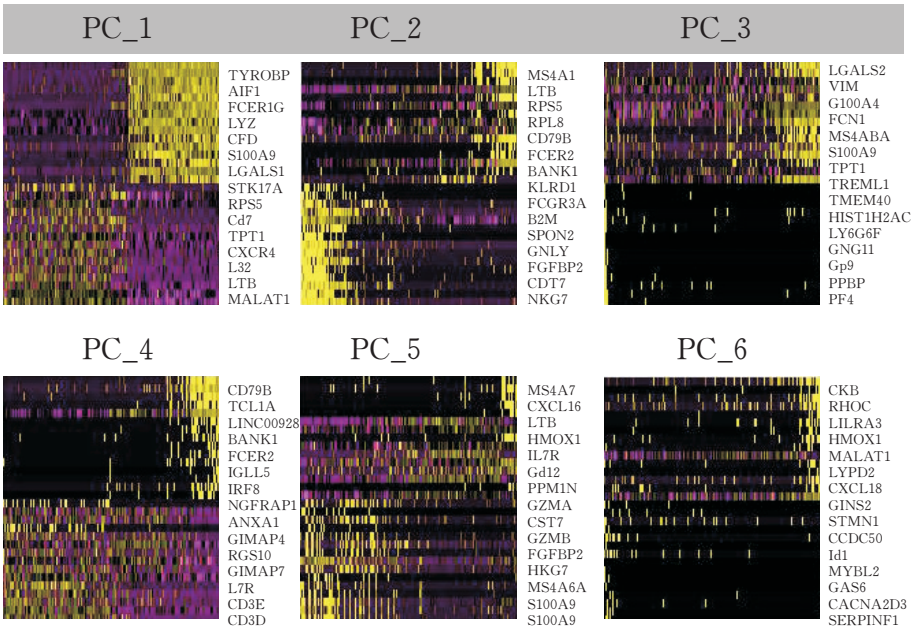


图 4 前 500 个细胞的 PCA 分布热图

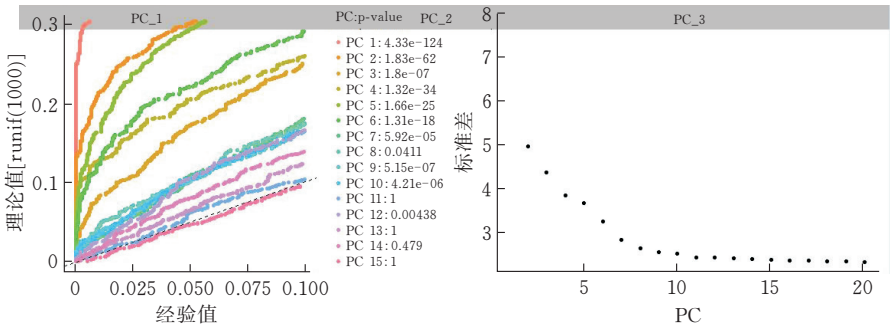


图 5 前 15 个 PC 均匀分布

如图 5 所示，左边的图是前 15 个 PC 的 $P$ 值和均匀分布，理论上，越是显著的 PC，其 $P$ 值越小，结合图形和 $P$ 值可知：在前 10－12 的 PC 之后，分布发生了剧烈的下降。而右边是一个肘型图，现 PC 在 9－10 附近出现一个“肘部”，也就是说大部分真实信息是在前 10 个 PC 中获取的，因此选择 10 个 PC 来进行后续的分析。

## 2 细胞分类

### 2.1 单细胞 RNA 聚类方法

**2.1.1 K-means 聚类** K-means 是聚类算法中最常用的一种，目标是要将数据点划分为 $k$ 个簇，找到其中每个簇的中心，并且最小化函数，要求每个数据点要与它们所属簇的中心尽量接近。由于数据量过大，首先采用随机抽选 2000，4000，8000 个样本，导入数据后，使用 wssplot 和 Nbclust 来确定当前聚类的个数，得到如图 6 所示。从图中可以看出 wssplot 建议聚类的个数为 $k=4$，划分为 4 个簇，而后找到每个簇的中心，并且最小化函数，旨在以整个样本中特征相似的基因或者以基因表达相似的样本对基因表达谱进行聚类而后将结果可视化，得到图 7。

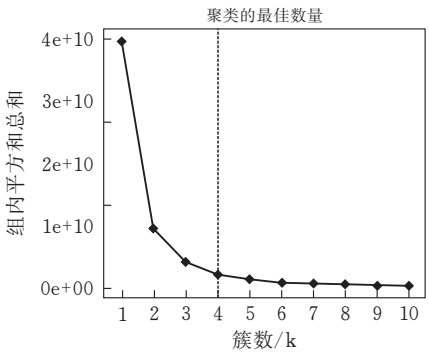


图 6 K-means 聚类最佳簇数

由上图可知，$K$ 的最优选择为 4，即可将细胞分为 4 类。

**2.1.2 层次聚类** 层次聚类法与 K-均值聚类具有本质的区别。层次聚类算法即将数据一层一层进行聚类，既可将大的簇从上而下进行分割，也可把小的

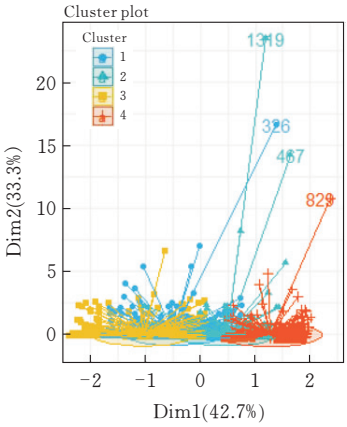


图 7　K-means 聚类集群图

簇从下而上合并聚集。这里采用 AL 层次聚类法，随机选取其中 500，1000，5000 来运行程序。首先让每个基因表达自成一类，再根据系统树图表示所有基因细胞个体间的相关性，同时利用 average-linkage，挑选出组间距离等于两组对象之间的平均距离的若干个基因表达样本，将这些样本合为一大类。再加以重复，自底而上通过合并不同类的基因表达以减少总类别数。在此方法中，确定应该取多少个 cluster 是聚类里面的一个非常重要的问题。用 NbClust 进行投票，根据结果可以选择最合适的聚类的个数，得到图 8 的投票结果。

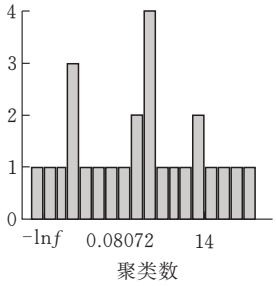


图 8　聚类簇数投票结果

由图 8 可知，在数据量有限的范围内，投票出来最合适的聚类个数是 4，得到的分类结果如图 9、10 所示。

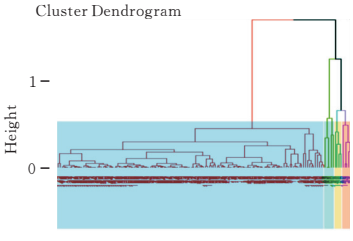


图 9　聚类集群系统树图

**2.1.3　图聚类**　K-means 和层次聚类都可以达到分类的效果，但能处理的数据量有限，只能对部分数据进行聚类，这使得这两种方法得出的聚类结果可能存在偏差，不能够保证分类结果的正确性。为了保证分类结果的有效性和正确性，需要对所有的数据进行分析聚类，从而得到正确且可靠的结果。

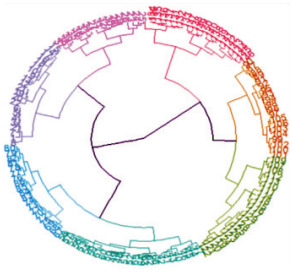

图 10　聚类集群系统网络图

Seurat 基于图聚类算法来对所有数据进行细胞的聚类分析。因此，采用图聚类算法，对所有数据进行分析，得出聚类结果。图聚类的主要思想是：PCA 空间中存在着各个细胞，首先依次计算出来，这个空间中距离每一个细胞最近的 $k$ 个细胞的欧氏距离，然后构造出来一个又一个的 KNN 图，接着，利用任意两个细胞在它们的局部邻域中的共享重叠来细化它们之间的边缘权重，并试着将该图划分为互联度较高的社区，最后运用模块化的优化技术（Louvain 算法）将细胞进行迭代，并分在一组中。

根据图聚类算法的结果，将细胞分为 7 个类别。由此也可以印证，前面通过 K-means 和层次聚类对部分数据进行聚类的结果确实是不准确的。

查看前 5 个细胞的聚类 id 如表 4 所示。

**表 4　前 5 个细胞的聚类 id**

| | |
|---|---|
| AAACCGTGCTTCCG-1 | 1 |
| AAACGCACTGGTAC-1 | 2 |
| AAACGCTGACCAGT-1 | 4 |
| AAACGCTGTAGCCA-1 | 5 |
| AAAGAGACGCGAGA-1 | 1 |

## 2.2　非线性降维

通过主成分分析法，将数据降维至 50 维，虽然可减少数据计算量，对提高计算效率起到了很大的作用，但还是不利于可视化分析，需要对数据进行进一步的降维处理，来达到可视化的目的。

使用 Seurat 中的 UMAP 和 t-SNE 来实现非线性降维技术，对前面选择的 PC 维度中的数据进行降维，将其降维至二维，然后进行可视化。

**2.2.1　UMAP 降维原理**　UMAP 一致流形估计与投影算法，是基于流形理论和拓扑数据分析理论提出的一种降维算法。可以用加权图的构造和操作来描述，第一步构造一个特定的加权 K 邻域图，接着计算该图的低维表示。输入 $X=\{x_1,x_2,\cdots,x_N\}$ 和度量矩阵 $d:X\times X\to\Re_{\geqslant 0}$。给定超参数 $k$，对于每一个 $x\in X$，计算其度量 $d$ 下的 $k$ 近邻集合 $KNN=\{x_1,x_2,\cdots,x_k\}$。最后用最近邻下降法进行计算：

$$\rho = \min\{d(x,y_j) \mid 1 \leqslant j \leqslant k, d(x,x_j) > 0\} \quad (1)$$

用平滑 KNN 距离设定 $\sigma$ 的值为：

$$\sum_{j=1}^{k} \exp\left(\frac{-\max(0, d(x,x_j)-\rho)}{\sigma}\right) = \log_2(k) \quad (2)$$

将局部模糊的单纯集合定义为 $fs_set_1 = \{[x,y],0 \mid y \in X\}$，其中每个 $y \in$ knn，有

$$d_{x,y} = \max\{0, \text{dist}(x,y)-\rho\}/\sigma \quad (3)$$

$$\text{fs_set} = \text{fs_set}_1 \cup ([x,y], \exp(-d_{x,y})) \quad (4)$$

合并 $X=\{x_1,x_2,\cdots,x_N\}$ 中每个 $x$ 对应的 fs_set，得到全局模糊拓扑表示 top_set，把 top_set 当成加权图，先用标准谱算法对标准化的 Laplace 算子进行谱嵌入，再根据最小化模糊集交叉熵让嵌入效果更好，最后完成降维。

**2.2.2 t-SNE 降维原理** t-SNE 是一种用于数据降维和可视化的机器学习算法。它将高维数据映射到一个低维空间，方便对数据进行分析和理解。计算流程如下：

首先定义高维空间中任意两点之间观测距离高维数据矩阵和低维数据矩阵的高斯概率：

$$p_{i \mid j} = \frac{\exp(-\| x_i - x_j \|^2/2\sigma_i^2)}{\sum_{k \neq i} \exp(-\| x_k - x_j \|^2/2\sigma_i^2)} \quad (5)$$

$$q_{i \mid j} = \frac{(1+\| y_i - y_j \|^2)^{-1}}{\sum_{k \neq i} (1+\| y_k - y_j \|^2)^{-1}} \quad (6)$$

定义 Kullback-Leibler 散度损失函数，

$$KL(Pi \| Qi) = \sum_i \sum_j p_{j \mid i} \log \frac{p_{j \mid i}}{p_{i \mid j}} \quad (7)$$

将高维概率投影到低维概率上，则梯度下降优化中使用的梯度表示为：

$$\frac{\partial C}{\partial y_i} = 4\sum_j (p_{ij} - q_{ij})(y_i - y_j)(1+\| y_i - y_j \|^2)^{-1} \quad (8)$$

分别用 UMAP 算法与 t-SNE 算法对聚类结果进行降维处理后，得到的可视化结果分别如图 11 和 12 所示：

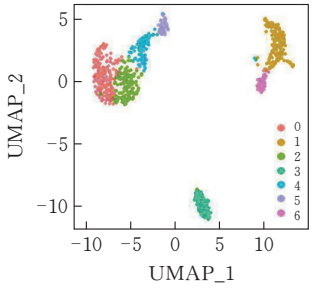


图 11 UMAP 降维可视化图

由两个可视化图像可知，两种算法都将 65 万个细胞分为了 7 类。

至此完成了对细胞的分类，根据提供的细胞和基因的信息，将细胞分为了 7 类。

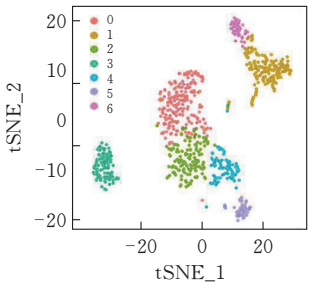


图 12 t-SNE 降维可视化图

## 3 差异化表达的基因

基因的表达差异(differential expression，DE)包括相同细胞内同一时间不同基因的表达差异；相同细胞内不同时间里相同基因的表达差异；不同细胞内相同基因的表达差异三种情况。对差异基因的筛选一般使用倍数变化法(fold change，FC)，它可以对预实验和实验进行初筛，对差异基因的参数分析包括 $t$ 检验、$F$ 检验以及 SAM 法等方法。

FC 法一般用于生物学不重复的样本，十分简单直接，首先计算得到基因在两个条件下表达水平的比值，根据所求出的比值来确定阈值，根据阈值把绝对值大于此阈值的基因定为差异基因。计算方法如下：

$$\text{FC} = \bar{M}_D - \bar{M}_B \quad (9)$$

式中 $\bar{M}_D$ 和 $\bar{M}_B$ 分别为不同分组下细胞基因表达水平的均值。通常对 FC 取对数同时选择两到三倍的基因作为结果，所以差异表达基因筛选标准一般为 $\log_2 \text{FC} \geqslant 2$。

寻找差异表达基因采用的是计算倍数变化(fold change)和统计检验相结合的方法。现有的 RNA-seq 是离散抽样，而其结果服从泊松分布，样本间的差异是服从 gamma-Poisson 分布或者负二项分布。因此可以用统计检验方法例如广义线性模型去分析 RNA-seq 前期。这样分析不需要参考样本，只需要处理随机取样，所有得到的离散型结果统计检验相对于 Fold Change 更具统计意义，但是该方法需要重复多次，这使得 $P$ 值会不准确，需要根据 FDR$<$0.01 对 $P$ 值进行矫正筛选。而对于 65 万个 counts 显然重复的次数过多，不是最佳方法。而如果只采用 FC 法将大于阈值的基因就判断为差异基因，得出的结论又不一定很准确。综合考虑以上情况，选择将两种方法相结合，以此尽可能地减少重复次数和提高准确率。

采用 $t$ 检验进行统计检验。$t$ 检验常用于两组独立样本数据的差异表达基因分析，第一步根据实验的可信度设定标准值，再分析对应的 $t$ 值与标准

差的关系判断基因差异，如果分析得到的相对应 $t$ 值超过标准差，那么该检验的基因存在差异表达。其中判断的关键指标就是 $P$ 值。用 $P$ 值作为衡量基因假阳性的概率，$P$ 值越低，说明基因出现假阳性的概率就越低，同时说明该基因是差异表达概率的可验证性就越高，通常 $P$ 值选择 0.05。

基因表达分析的零假设是：基因在不同处理下的表达量相同。由于不同的阈值或“合格”标准都会对最终得到的“显著基因”产生影响。结合上述两个筛选差异表达基因的方法，最终筛选阈值定为：$P<0.05$，$\log_2FC\geqslant 2$，FDR$<0.01$，即满足以上条件的就是差异表达的基因，以此来找到差异表达的基因。再使用 Seurat 通过分析表达差异就能够找得到每一个类别里面表达有差异的基因 markgene。

随后分成如下三种情况来做差异分析：某一个类别当中表达有差异的基因；某一个类别与另外两个类别之间表达有差异的基因；某一个类别相比于其余所有类别表达有差异的基因。

### 3.1　同一类别差异分析

通过前面的数据分析和处理，将所有的细胞分为了 7 类，在每一类细胞当中都会有一些表达有差异的基因，因此，可以分别考虑每一类当中表达有差异的基因。比如，表 5 所示为 cluster1 中表达有差异的基因。

**表 5　cluster1 中表达有差异的基因**

| 基因名 | $P$ 值 | avg_$\log_2$FC |
|---|---|---|
| LGALS2 | 6.92E-122 | 3.658661 |
| S100A8 | 1.40E-115 | 5.421745 |
| FCN1 | 4.96E-103 | 3.334768 |
| S100A9 | 1.22E-102 | 5.599597 |
| CST3 | 1.46E-94 | 3.400908 |

由上表的结果可知，$P$ 值均小于 0.05，同时 avg_$\log_2$FC 在 3.3～5.6 之间远大于 2，均可以被判定为差异表达基因。运用同样的方法，可以依次找到 cluster0 到 cluster6 每一类当中表达有差异的基因，结果如表 6 所示。

### 3.2　三个类别差异分析

实际情况中，某些细胞虽然属于不同的类别，但可能含有相同表达的基因，因此分析不同类别细胞之间的基因表达差异也具有很大的现实意义。这里，考虑某一个类别与另外两个不同类别表达差异的基因，如表 7 是找到能够将 cluster5 从 cluster0 和 cluster3 中区别出来的基因。

**表 6　每一类表达差异的基因**

| cluster 0 | cluster 1 | cluster 2 | cluster 3 | cluster 4 | cluster 5 | cluster 6 |
|---|---|---|---|---|---|---|
| RPS12 | LGALS2 | LTB | CD79A | CST7 | SP0N2 | MS4A4A |
| CYBA | S100A8 | IL32 | MS4A1 | CCL5 | XCL2 | LILRB1 |
| CD74 | FCN1 | LDHB | LINC00926 | NKG7 | FGFBP2 | MS4A7 |
| RPS25 | S100A9 | IL7R | CD79B | GZMK | PRF1 | CKB |
| RPL32 | CST3 | CD3D | TCL1A | GZMA | GZMB | FCGR3A |

由表 7 的结果可知，$P$ 值均小于 0.05，同时 avg_$\log_2$FC 在 4.0～6.5 之间远大于 2，所以均可以被判定为差异表达基因。由于分类得到的类别比较多，不同类别的细胞组合相应地也会很多，就会产生巨大量级的数据，但事实上分析的原理是相同的，这里不再逐个进行分析。

**表 7　将 cluster5 从 cluster0 和 cluster3 中区别出来的基因**

| 基因名 | $P$ 值 | avg_$\log_2$FC |
|---|---|---|
| PRF1 | 4.08E-64 | 4.654064 |
| GZMB | 1.99E-61 | 5.415289 |
| CST7 | 2.17E-57 | 4.176487 |
| FGFBP2 | 2.42E-57 | 4.640882 |
| GNLY | 1.17E-54 | 6.231188 |

### 3.3　所有类别差异分析

前面两种情况，不管是单个类别，还是三个类别，都只是分析了部分类别的基因，没有将所有的基因考虑在内。因此，最后我们考虑整个分类空间，将所有的类别包含在内，从整个基因整体当中找到表达差异的基因。如图 13 所示为找到的与剩余所有细胞相比表达差异的基因。

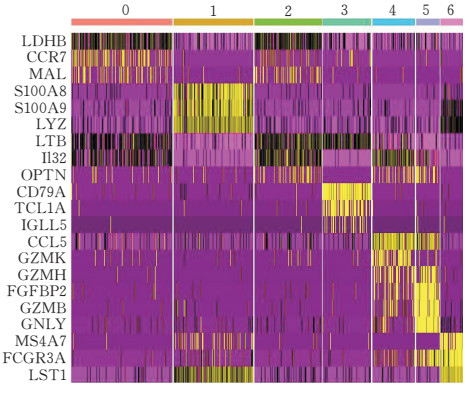


图 13　所有类别差异化表达的基因热图

差异表达基因以火山图、MA 图、韦恩图、聚类热图、蛋白互作图等形式呈现，通过聚类热图可以快速地查看基因在表达水平上的差异，以及差异的统

计学显著性。下面用可视化分析，将不同基因在不同分群的表达量差异情况在所有基因集后绘制热图，可以更直观地看出不同类别的细胞与基因的联系。如图 14 是举例说明差异化表达的基因所属的类别。

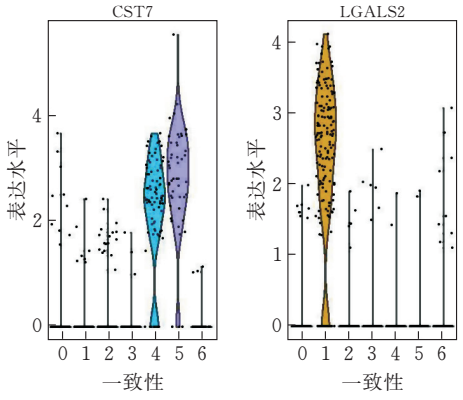


图 14 差异化表达基因所属类别

在图 14 中，以“CST7”“LGALS2”这两个基因为例查看其表达情况，可以看出来，“CST7”这个基因在 cluster4，cluster5 中具有较高表达，而“LGALS2”这个基因在 cluster1 中具有较高表达。

也可以将表达差异的基因映射到细胞当中去，使用 UMAP 聚类方式，有如图 15 所示结果。

而使用 t-SNE 聚类方式，则有如图 16 所示结果。

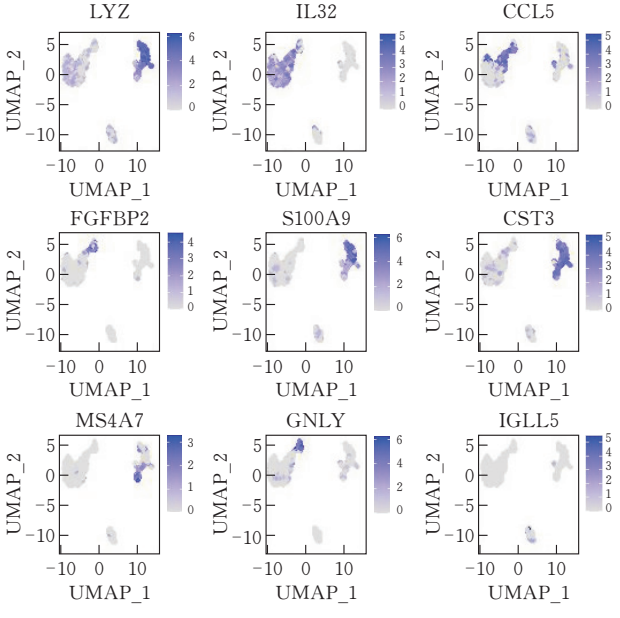


图 15 UMAP 聚类下差异化表达基因

## 4 结束语

本文针对细胞信息数据量庞大的问题，综合常规数据处理方法以及细胞和基因数据的独特性，提出了一套完整的单细胞测序数据预处理、分类和差异化基因表达分析流程。首先，通过基于线粒体的

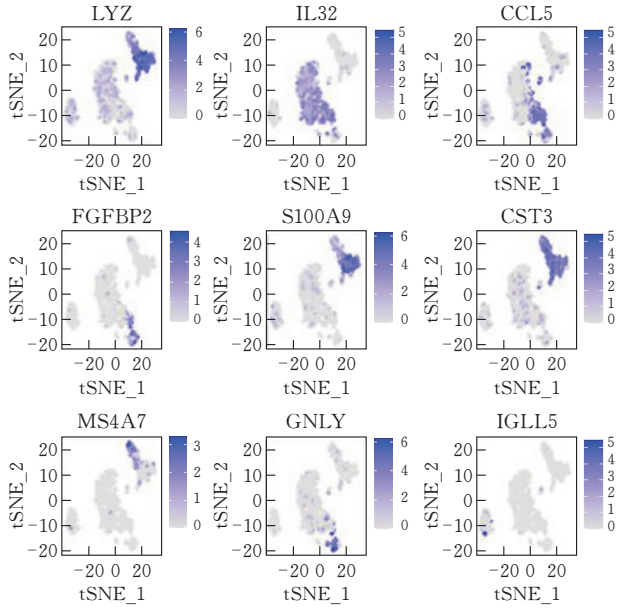


图 16 t-SNE 聚类下差异化表达基因

基因表达比例和 nFeature_RNA 质控指标，过滤掉非正常细胞，确保数据的质量。随后，对数据进行归一化处理，并检测出高异质性基因，通过线性变换对数据进行缩放，以便进一步分析。接着，采用主成分分析方法将数据降维至 50 维，以便在保留主要信息的同时降低数据复杂性。在数据分类方面，本文利用图聚类算法对预处理后的数据进行分类，成功将细胞分为 7 类。随后，通过 UMAP 和 t-SNE 非线性降维技术，将数据降至 2 维，实现了对分类结果的直观可视化分析。在差异化基因表达分析中，本文基于基因在不同处理下表达量相同的假设，结合 FC 法和统计检验法，寻找差异表达的细胞，并设定了相应的阈值。针对同一类别、三个类别以及所有类别三种情况，分别进行了差异化表达基因的分析，为后续的生物学研究提供了有力支持。

随着单细胞测序深度和样本数量的不断增加，数据规模愈发庞大，使得图聚类算法在构建图模型时面临巨大的计算资源与时间挑战。因此，降低算法的时间与空间复杂度成为了提升聚类算法性能的关键问题，也是未来研究的重要方向。

综上所述，本文在单细胞测序数据处理、分类和差异化基因表达分析方面取得了一定的成果，但仍需在算法优化方面继续探索，以适应日益庞大的单细胞测序数据规模。

[ 参 考 文 献 ]

[1] DEEPA N，PRAVEEN S. Benchmarking single image dehazing methods[J]. SN Computer Science，2021，3(01)：479-487.

[2] 龚乐君，周佘海，程逸飞，等.单细胞 RNA 序列数据的

PBMC 相关细胞的识别[J].中国科学技术大学学报,2020,50(07):1013-1018.
[3] TANG F, BARBACIORU C, WANG Y, et al. mRNA-Seq whole-transcriptome analysis of a single cell [J]. Nat Methods, 2009,6(05): 377-382.
[4] PHILIP B, SIMON A,KYOUNG J K, et al. Accounting for technical noise in single-cell RNA-seq experiments[J]. Nature methods,2013,10(11):1093-1095.
[5] RHONDA B,CHRISTINA K . Design and computational analysis of single-cell RNA-sequencing experiments.[J]. Genome biology,2016,17(01):63.
[6] HENG S, ZHENGQIANG M, XIN Z, et al. Single-cell RNA-Seq reveals cell heterogeneity and hierarchy within mouse mammary epithelia.[J]. The Journal of biological chemistry,2018,293(22):8315-8329.
[7] 杨佳凤,陈鹏璐,龚熹.单细胞转录组测序技术在细胞分类中的应用[J].中国细胞生物学学报,2021,43(02):476-483.
[8] RAPHAEL P, ZHULIU L, RUI K . Machine learning and statistical methods for clustering single-cell RNA-sequencing data.[J]. Briefings in bioinformatics,2020,21(04):1209-1223.
[9] WANG B, RAMAZZOTTI D, DE SANO L, et al. SIMLR: A tool for large - scale genomic analyses by multi - kernel learning[J]. Proteomics,2018,18(02):1700232.
[10] ANDREW B, PAUL H, PETER S, et al. Integrating single-cell transcriptomic data across different conditions, technologies, and species.[J]. Nature biotechnology,2018,36(05):411-420.
[11] 何睿,余娜,李淼,等.基于单细胞 RNA 测序数据的细胞类型聚类算法[J].智能计算机与应用,2020,10(07):104-108.

# Application of Graph Clustering Algorithm in the Classification of Single-cell RNA Sequencing Data

WANG Jingwei, GUO Yibei, ZHU Zhaomei, LI Hanfang
(*School of Science, Hubei Univ.of Tech., Wuhan* 430068, *China*)

**Abstract**: Single cell RNA sequencing technology is widely used in quantitative research of gene expression profiles. Its advantages include not only high throughput and high depth, but also precise determination of the condition of each cell. Nowadays, this technology has developed rapidly and is at the forefront of the current field of life sciences. With the rapid progress of single cell RNA sequencing technology, the amount and complexity of data have sharply increased, posing huge challenges to data analysis methods. Therefore, scientific data analysis methods have become a top priority. This article mainly solves the following problems: considering the conventional methods of data processing and the uniqueness of cell and gene data, the original data is preprocessed and then classified using graph clustering algorithm. Afterwards, all cells are divided into 7 categories. Finally, UMAP and t-SNE nonlinear dimensionality reduction techniques were used to reduce the data to 2D, and the classification results were visualized and analyzed. Based on the classification results obtained from the same expression level of genes under different treatments, a combination of FC method and statistical test method was used to select the threshold: $P$-value$<0.05$, while $\log_2 FC \geqslant 2$, FDR$<0.01$, to search for differentially expressed cell genes, and then analyze the differentially expressed genes in each of the same category, three categories, and all categories.

**Keywords**: single cell RNA sequencing; nonlinear dimensionality reduction; graph clustering; FC method; difference statistical test

[责任编校: 闫　品]