文章编号：2095-2163(2020)07-0104-05 中图分类号：TP391 文献标志码：A

基于单细胞 RNA 测序数据的细胞类型聚类算法

何 睿，余 娜，李 淼，张峻巍，王浩杰，赵玉茗
（东北林业大学 信息与计算机工程学院，哈尔滨 150000）

摘 要：随着单细胞测序技术的发展，许多基于单细胞 RNA 测序数据的聚类算法被提出，用于单细胞分类，并取得较好的应用效果。但是到目前为止，单细胞聚类算法研究领域缺乏关于聚类模型的综述，缺乏对不同聚类模型的性能评估。本文从聚类模型的角度将常见的 11 种单细胞聚类算法分成了 K 邻近聚类、层次聚类、基于图形分类、基于模型分类、基于密度分类的 5 种类型，对相关算法的特点和研究进展进行总结，并选择了 10 组 scRNA-seq 数据集对这些聚类算法进行性能评价。实验结果表明，现有聚类方法中 SC3、Seurat 和 SIMLR 的性能较好，在 5 类模型中，基于密度模型的算法具有最优性能，体现出较好的应用价值。

关键词：细胞分类；聚类算法；单细胞测序

Cell type clustering algorithm based on single-cell RNA sequencing data

HE Rui, YU Na, LI Miao, ZHANG Junwei, WANG Haojie, ZHAO Yuming
(Information and Computer Engineering College, Northeast Forestry University, Harbin 150000, China)

【Abstract】Withthe development of single-cell sequencing technology, many clustering algorithms based on single-cell RNA sequencing data have been proposed and applied to single cell classification and achieve good application results. But so far, the research field of single-cell clustering algorithms still lacks summary research on clustering models and performance evaluation studies of different clustering models. Therefore, from the perspective of the clustering model, we divide 11 common single-cell clustering algorithms into five types: K-means clustering, hierarchical clustering, graphic-based clustering, model-based clustering, and density-based clustering. The characteristics and research progress of related algorithms are summarized, and ten scRNA-seq datasets are selected for evaluation of 11 clustering algorithms. The experimental results show that the performance of SC3, Seurat, and SIMLR are better in the existing clustering methods. Among the five types of models, the algorithms based on the density model have the best performance and reflect good application value.

【Key words】Cell classification; Cluster algorithm; Single cell sequencing

0 引 言

近年来，随着测序技术的不断提高和细胞研究的逐渐深入，单细胞 RNA 测序技术（scRNA-seq）成为当下生物领域研究的热点[1]。单细胞 RNA 测序技术是能够在单个细胞的水平上，对基因组进行高通量测序分析的一项新技术。传统高通量 RNA 测序是基于组织整体进行测序，所得到基因组信息是整体平均数据；单细胞 RNA 测序得到了反映单个细胞遗传信息的数据，用于研究相同表型细胞间遗传异质性，发现其特定生物功能[2]。单细胞测序数据的一个重要作用就是用于细胞分类，即根据单细胞测序数据，建立聚类模型，用以将具有相似基因表达模式的细胞聚类成相同的细胞类型，进而推断细胞功能，并理解疾病与基因组特征之间的相关性[3]。倘若可以对细胞进行更精确和无偏倚的分类，将会在肿瘤学、遗传学、免疫学等研究领域产生巨大的影响[4]。

目前，大多数细胞分类方法是基于传统 RNA 测序数据提出的，虽然可以被用于 scRNA-seq 数据，但是 scRNA-seq 数据具有明显区别于传统高通量 RNA 数据的特点，如数据量大、维度高以及噪声太多等。在 scRNA-seq 数据上直接使用基于组织 RNA-seq 数据开发的聚类方法具有很大的局限性[5]，因此设计与构建适用于单细胞数据特点的分类与可视化工具就成为单细胞分析领域的热点问题。

基金项目：国家级大学生创新创业训练计划（201810225173）；国家自然科学基金（61971119）。

作者简介：何 睿（1999-），女，本科生，主要研究方向：数据挖掘；余 娜（1999-），女，本科生，主要研究方向：数据挖掘；李 淼（1996-），女，本科生，主要研究方向：数据挖掘；张峻巍（1998-），男，本科生，主要研究方向：软件工程；王浩杰（1998-），男，本科生，主要研究方向：机器学习、Android 开发；赵玉茗（1978-），女，博士，副教授，主要研究方向：生物信息学、机器学习、自然语言处理。

通讯作者：赵玉茗 Email：zym@ nefu.edu.cn

收稿日期：2020-02-23

单细胞无监督聚类方法提供了一种通过相似性来聚类细胞的机制。虽然无监督的方法有优势，但是样本数目少，缺乏有关分组真实性的实验验证方法，缺少关于分组数目或类型的先验信息等，会为聚类带来问题。与此同时，单细胞数据的特征，如数据缺失、高维度和噪音，也增加了精确识别细胞集群的难度。尽管存在这些问题，研究者们仍然开发出了一些用于 scRNA-seq 数据的聚类方法，见表 1。

表 1　常见单细胞 RNA 测序聚类方法

Tab. 1　Common single-cell RNA sequencing clustering methods

方法	聚类模型
SIMLR	K-means
SC3	K-means
pcaReduce	K-means
CIDR	Hierarchical
SINCEAR	Hierarchical
SNN-Cliq	Spectral
DLA	Spectral
CountClust	Model
BISCUIT	Model
Seurat	Density
GiniClust	Density

本文以现有用于 scRNA-seq 数据的聚类方法为研究对象，参考经典聚类模型分类标准，对 11 种常用算法进行分类研究，选择了 10 组 scRNA-seq 数据集从聚类准确率、时间复杂度等方面对算法进行综合性能评估研究。

1　单细胞聚类方法分类

目前，基于单细胞测序数据的聚类算法有很多，每种算法均有其各自的特点。因此，在相同的数据下，各自的分类效果也有所不同。通过参考经典聚类模型分类理论[6]，将 11 种常用单细胞测序聚类算法分为基于 K-means 算法、基于层次聚类算法、基于图的聚类算法等 5 种聚类模型，下面主要对应用较多的 5 类聚类方法的算法原理、特点和优势进行概述。

1.1　基于 K-means 均值聚类的算法

K-means 均值聚类算法(K-means clustering algorithm)是一种经典的无偏聚类方法，已经被应用到许多方面。其核心思想是逐步对聚类结果进行优化、不断将目标数据集向各个聚类中心进行重新分配，以获得最优解。其中，判断是否是最优解的目标函数通常通过平方误差计算法得到[7]，其将 n 个样本分成 k 个簇，找出 k 个聚类中心 $c_1, c_2, \cdots, c_K$，使得每一个数据点 x_i 和与其最近的聚类中心 cv 的方差最小化。K-means 算法的主要优点是算法具有易用性、灵活性和高效性。由于其算法过于灵活，因此不同 k 值会导致不同的聚类结果，即不能保证一定可以获取全局最优值[8]。

针对单细胞 RNA 测序数据样本量较多的特点，Bo 等人提出了一种基于多核学习的分析框架 SIMLR，它可以从单细胞 RNA-seq 数据中学习相似性度量，用来中和 K-means 算法的缺点，实现降维、聚类和异构细胞群的可视化[9]。SIMLR 构造了具有多种超参数的多高斯核函数，再基于距离和相似性之间的反比关系，使用优化框架来计算细胞之间的相似性。其主要步骤是：假设聚类数为 C，理想情况下，相似度矩阵的秩与 C 相同；将相似性矩阵输入到 t-SNE 进行降维，使用 k 均值对细胞进行聚类。Vladimir 等人提出了一种无监督共识聚类方法 SC3，它通过重复使用 K-means 算法，使用不同的上游处理或者初始条件，求得一致性来克服贪婪的特性[10]。SC3 结合了 K-means 和层次聚类，由基因滤波、距离计算、变换结合、聚类和共识步骤 5 个步骤组成。其中，过滤具有低表达水平的基因，距离计算时除去 2 个细胞中没有表达的基因的两个步骤，都没有明显改善细胞聚类的效果。SC3 方法包含多个参数，因此其可以结合具体的数据集来灵活地调整参数范围。Justina 等人提出了一种基于 K-means 的迭代聚类方法 pcaReduce[11]。算法将原始基因表达矩阵 $Xn \times d$ 投影到顶部 $k-1$ 个主方向，使用 $k-$means 聚类投影数据以获得 K 个聚类。在聚类过程中，初始簇的数量 K 被设置为较大的值，例如 30，以确保捕获较多的数据类型。对于合并聚类，则使用多元高斯函数来计算每对聚类的合并概率。合并有两种方式：一种是选择合并概率最大的两个聚类；另一种则是两个聚类进行抽样。按照标准化的合并概率合并它们的一部分，并从现有的聚类中心和协方差矩阵中除去这个维度，将数据矩阵主要方向的数量减少到 $k-2$。重复以上步骤，直至只剩下一个集群。pcaReduce 算法融合了 K-means 和层次聚类的核心思想，其主要优势是可以自行设置详细的集群数量。

1.2　基于层次聚类的算法

层次聚类(Hierarchical clustering)又称为树聚类算法，是一种被广泛用于 scRNA-seq 的通用聚类算法。通过划分不同级别的数据，形成树状聚类结构，再通过计算不同类别的数据点间的相似性，构造分层嵌套聚类树。聚类树的构建基于层次分解的方向，共分为凝聚(agglome-rative)和分裂(division)两

大类。前者是将初始数据看成不同的簇,重复合并相似性最高的一对簇,直到所有的数据归为一簇为止。后者则将初始数据看成一个簇,重复细分,直到满足终止条件[12]。与 K-means 聚类算法相比,层次聚类算法的优点主要在于聚类结果可解释和簇的数量无需事先指定。

Peijie 等人提出了一种快速且准确的层次聚类算法 CIDR,其在对 scRNA-seq 进行分层聚类时,开发了一种新的类似主成分分析的降维算法[13]。这种算法通过在距离计算中加入隐式的归零,使得对低深度样本中细胞距离的估计更加稳定。根据单个细胞数据集中对数转换表达值的分布,CIDR 找到每个细胞的丢失候选阈值,并计算出丢失概率和表达值之间的关系。CIDR 通过将具有预期表达值的丢失候选者的表达值进行插入来建立插补过程。在插补过程后,计算这些细胞对之间的距离,使用 PCoA 来减小 CIDR 距离矩阵的维度,使用分层聚类来聚类单个单元。Minzhe 等人提出了一种用于单细胞 RNA-Seq 分析的计算模型 SINCEAR[14]。该模型可以识别主要细胞类型,鉴定细胞类型特异性基因,并对聚类结果进行分析解释。在预处理时,过滤细胞群中的低表达基因和非选择性表达基因,在基因和细胞两个层面分别运用 z 分数和修剪平均值对数据进行归一化。SINCEAR 的优势是可以使用间隙统计信息识别聚类编号。而其缺点是直接运用层次聚类方法,在没有降维的情况下,由于高维和噪声的影响,无法保证聚类精度。

1.3 基于图聚类的算法

一般来说,无监督聚类可以不需要任何先验信息,便可以将数据集划分为两个或更多的类。但是若采用这种划分方法,难以获得数据的真实全局相似性。对于此类问题,基于图形的聚类算法便可以进行解决[15]。该算法可以基于局部结构相似性构建图形,这对呈现两个对象之间的连接非常合适。

Chen 等人开发了基于图的单细胞聚类方法 SNN-Cliq[16]。这种算法利用了共享最近邻(SNN)来识别聚类。该方法首先根据距离测度确定每个单元的 k 近邻,用于计算每对单元之间共享的 SNN 的数量。如果两个单元至少有一个 SNN,则通过在两个单元之间放置一条边来构建一个图。使用“团”方法将集群定义为具有许多边的单元组。SNN-Cliq 需要手动定义几个参数。Vasilis 等人对基因表达数据进行二次挖掘,提出基于单细胞 RNA-seq 数据的转录物相容性计数(Transcript Compatibility read Counts, TCC)[17]。基因表达水平是指基因内的读数,然而,TCC 是指不同转录组内的读数。该方法基于 TCC 矩阵执行单细胞聚类,其列代表细胞,行代表 TCC。通过映射读数的总数归一化,获得 TCC 概率分布,计算每对细胞的 TCC 分布之间的 Jensen-Shannon 散度的平方根,以得到细胞的成对距离。如果已知聚类数,则使用谱聚类方法在成对距离矩阵上聚类细胞;如果未知聚类数,则使用亲和力传播聚类模型得到细胞分类。

1.4 基于模型聚类的算法

基于模型的聚类也是一种较为常见的聚类方法,其主要思想是假设数据由模型生成,并尝试从数据中恢复原始模型[18]。在基于模型的聚类方法中,数据被视为来自概率分布的混合,每个概率分布代表不同的聚类。即在该算法中,假设数据是由概率分布的混合生成的,其中每个分量代表不同的聚类。因此,当数据符合模型时,可以预期特定的聚类方法可以很好地工作。

近年来,一些学者将基于模型的聚类方法应用于细胞亚群的识别。Dey Kushal 等人,开发了一种 CountClust 方法,利用隶属度(GoM)模型(集群模型的泛化)对单细胞 RNA-seq 数据进行聚类[19]。GoM 模型允许每个样本在每个簇中具有一定比例的隶属度,其通过映射到基因组中的每个基因的读数计数来总结 RNA-seq 样品,类似于文档聚类。通过与层次聚类方法进行比较,发现 GoM 模型更准确,也更能表示细胞在不同基因上可能聚类不同的情况;Elham 等人基于贝叶斯概率模型提出了一种数据驱动的聚类模型 BISCUIT[20]。以往将单细胞 RNA-seq 数据进行全局归一化的方法,不能解决数据丢失问题,并可能导致不准确的聚类和下游分析偏倚,而 BISCUIT 将归一化和聚类合并为一个模型,上述问题可以得到解决。该方法通过学习细胞特异性参数对细胞进行迭代归一化和聚类,并将其合并到一个分层的 Dirichlet 过程混合模型(HDPMM)中,该模型允许根据基因表达模式和共表达模式的相似性进行同步聚类,同时对技术伪制品进行校正。在综合和实验的 scRNA-seq 数据上,BISCUIT 均表现出色。

1.5 基于密度聚类的算法

由于大多数聚类算法都没有异常值的概念,因此这些算法均不会检测到异常值,这也会降低分类的准确性。为了解决这个问题,可以引入基于密度的聚类算法来进行解决[21]。该类算法通过识别“密集”点集

群，通过数据点的位置分布均匀度，来识别数据中的异常值，其中一个点的密度等于该点周围的特定半径内的点数。基于密度的聚类方法不预先指定聚类数，适用于没有明确聚类数的聚类问题[22]。

在大多数情况下，由于技术影响，scRNA-seq 数据具有很大的噪音，需要识别的细胞类型的数量不清楚。因此，目前已经开发了几种基于密度的聚类方法，用于对 scRNA-seq 数据分类。Seurat 是一种用于质量控制、分析和探索 scRNA-seq 数据的综合方法，旨在使用户能够识别和解释单细胞转录组测量的异质性来源，并整合不同的细胞类型[23]。对于单细胞的无监督聚类，Seurat 结合了线性和非线性的降维算法。与其他使用 scRNA-seq 数据中所有表达基因的聚类方法不同，Seurat 首先通过计算基因的 Fano 因子，即方差与均值之间的比率，来确定一组在单细胞数据集中变化最大的基因。为了确定有统计学意义的主成分，对可变基因基于主成分分析进行降维，在显著的 PC 分数上实施非线性降维（t-SNE）。最后，Seurat 应用密度聚类对 t-SNE 图上不同的细胞群进行分类。Jiang Lan 等人发现在稀有细胞类型中，基因的 Fano 因子值之间没有明显差异，这意味着 Fano 因子不适合选择针对稀有细胞类型的基因[24]。为了解决这个问题，他们提出了一种新方法 GiniClust，通过使用 Gini 指数来识别罕见的细胞类型特异性基因。结果表明，基因的 Gini 指数不受细胞类型比例的影响。选择具有较高 Gini 系数的基因作为细胞类型特异性基因。之后，利用基于数据密度的 DBSCAN 算法从 scRNA-seq 数据中识别细胞类型，尤其是稀有细胞类型。

2　仿真实验及结果分析

为了进一步比较与分析五类单细胞 RNA 测序数据聚类算法的性能，从相关论文中收集了 10 个 scRNA-seq 数据集，见表 2。使用改进的兰德指数（Adjusted rand index，简称 ARI）做性能指标，对算法进行了比较与分析。本文所有实验在 3.40GHz 主频，16.00GB 运行内存的 PC 机上进行。实验结果如图 1 所示。

从图 1 中可以看出，SC3、Seurat 和 SIMLR 在 2 个数据集上获得第一；GiniClust、BISCUIT、pcaReduce 和 DLA 分别在一个数据集上表现最佳。

SC3、Seurat 和 SIMLR 是单细胞分析领域最常用的 3 种方法，在实验中取得了最好的实验结果，再次证明了算法的先进性，但这 3 种算法也只是分别在 2 个数据集中取得了最佳表现。说明现有方法的性能和数据集相关，仍有很大的改进和完善空间，需要去探索新的方法与算法。现有聚类方法大都采用的是人工智能领域已有的聚类方法，如：k-means、层次聚类、DBSCAN、图划分和 NMF 以及统计混合模型等，提出新模型、新机制的很少。随着 scRNA-seq 技术的发展，scRNA-seq 实验和项目数量增加，scRNA-seq 数量呈指数增长。目前，一次实验测试上千万细胞已经成为现实，这对聚类算法的效率提出了很高要求，而现有方法只有 CIDR 考虑了聚类效率问题。

表 2　scRNA-seq 数据集

Tab. 2　scRNA-seq data set

数据集	细胞数	类别个数
GSE59892	49	3
GSE36552	90	7
GSE45719	268	10
SRP041736	301	11
GSE52583	80	5
GSE51372	187	7
GSE57872	430	5
GSE59739	622	11
GSE65525	2 717	4
GSE60361	3 005	9

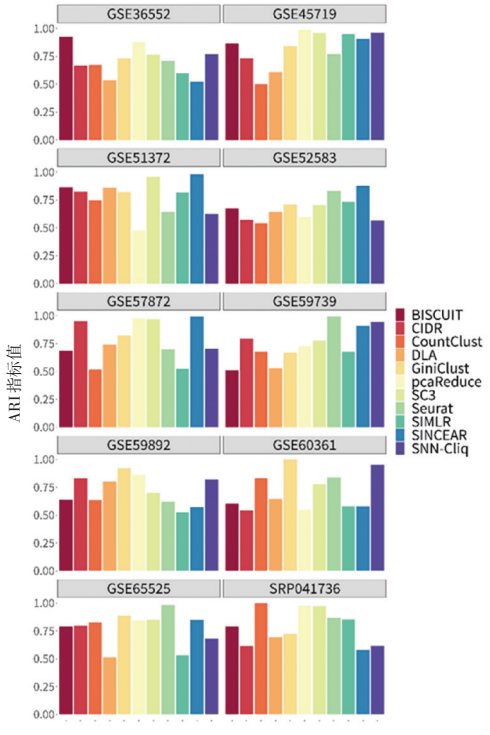

图 1　11 种常见 scRNA-seq 聚类算法 ARI 指标结果

Fig. 1　11 common scRNA-seq clustering algorithm ARI indicator results

从图 2 中可以看出，基于图的聚类算法整体表现较好，在 5 个测试函数上均取得最优性能；基于层次思想的聚类算法，在 3 个数据集上取得了最好的

精度;基于 K-means 的算法和基于模型的算法分别在 1 个数据集上取得了最优精度。虽然基于图的聚类算法取得了最优的性能,但是该类方法由于需要构建图模型,需要消耗大量计算资源与运行时间,随着单细胞测序深度的增加,样本个数的增长,单细胞测序数据会更加巨大。因此,如何降低算法时间与空间复杂度,成为提升聚类算法性能的关键。

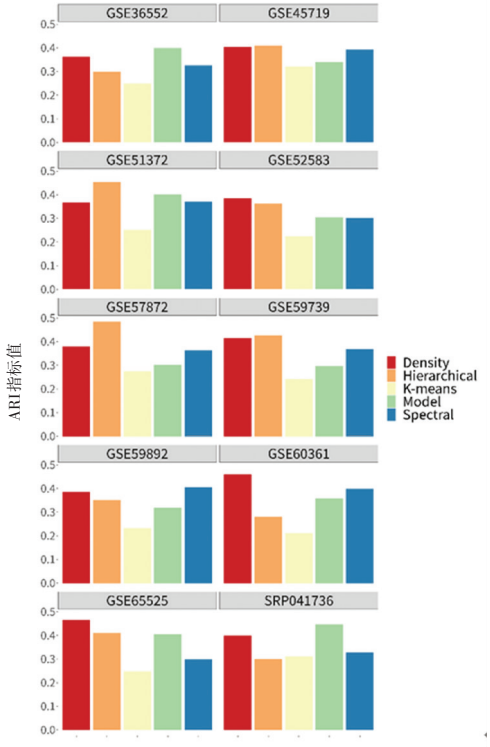

图 2 五类聚类算法 ARI 指标结果

Fig. 2 Results of the ARI indicator of the five-class clustering algorithm

3 结束语

为了进一步比较基于单细胞测序数据对细胞类型分类是一个较新的概念,目前所提出的五种聚类算法虽然各有千秋,但均已经较为成熟。不过,受限于算法本身,其应用到基因领域仍然需要进行适应性的改造。在实际操作过程中,由于技术原因,当前单细胞测序的结果准确度会直接影响到细胞分类结果的准确性。因此,对于细胞分类领域而言,提高单细胞测序的精度将会更好的促进聚类算法在基因领域的应用。同时,上述算法倘若合理加以运用,则会更加有效提升单细胞测序数据的实用价值,亦可在生物信息学领域为细胞研究提供理论依据。

参考文献

[1] 晁珊珊,卜鹏程.单细胞转录组测序技术发展及应用[J].中国细胞生物学学报,2019,41(5):834-840.

[2] VIETH B, PAREKH S, ZIEGENHAIN C, et al. A systematic evaluation of single cell RNA-seq analysis pipelines[J]. Nature communications, 2019, 10(1): 1-11.

[3] KISELEV V Y, ANDREWS T S, HEMBERG M. Challenges in unsupervised clustering of single-cell RNA-seq data[J]. Nature Reviews Genetics, 2019, 20(5): 273-282.

[4] PAPALEXI E, SATIJA R. Single-cell RNA sequencing to explore immune cell heterogeneity[J]. Nature Reviews Immunology, 2018, 18(1): 35.

[5] 金开秀. 针对高维稀疏单细胞 RNA 测序数据的聚类研究[D]. 杭州:浙江大学,2018.

[6] 孙吉贵,刘杰,赵连宇.聚类算法研究[J]. 软件学报,2008(1):48-61.

[7] JAIN A K, MURTY M N, FLYNN P J. Data clustering: a review[J]. ACM computing surveys (CSUR), 1999, 31(3): 264-323.

[8] 王昊雷. K 均值聚类算法研究与应用[D]. 哈尔滨:哈尔滨工程大学,2015.

[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(2): 1700232.

[10] KISELEV V Y, KIRSCHNER K, SCHAUB M T, et al. SC3: consensus clustering of single-cell RNA-seq data[J]. Nature Methods, 2017, 14(5):483-486.

[11] YAU C. pcaReduce: hierarchical clustering of single cell transcriptional profiles[J]. BMC bioinformatics, 2016, 17(1): 140.

[12] 段明秀. 层次聚类算法的研究及应用[D]. 长沙:中南大学, 2009.

[13] LIN P, TROUP M, HO J W K. CIDR: Ultrafast and accurate clustering through imputation for single-cell RNA-seq data[J]. Genome biology, 2017, 18(1): 59.

[14] GUO M, WANG H, POTTER S S, et al. SINCERA: a pipeline for single-cell RNA-Seq profiling analysis[J]. PLoS computational biology, 2015, 11(11).

[15] 张涛. 基于图的聚类分析研究[D]. 昆明:云南师范学院, 2018.

[16] XU C, SU Z. Identification of cell types from single-cell transcriptomes using a novel clustering method[J]. Bioinformatics, 2015, 31(12): 1974-1980.

[17] NTRANOS V, YI L, MELSTED P, et al. A discriminative learning approach to differential expression analysis for single-cell RNA-seq[J]. Nature Methods, 2019, 16(2): 163-166.

[18] 宋浩远. 基于模型的聚类方法研究[J]. 重庆科技学院学报(自然科学版), 2008 (3):71-73.

[19] DEY K K, HSIAO C J, Stephens M. Visualizing the structure of RNA-seq expression data using grade of membership models[J]. PLoS genetics, 2017, 13(3): e1006599.

[20] AZIZI E, CARR A J, PLITAS G, et al. Single-cell map of diverse immune phenotypes in the breast tumor microenvironment[J]. Cell, 2018, 174(5): 1293-1308. e36.

[21] 刘青宝, 邓苏, 张维明. 基于相对密度的聚类算法[J]. 计算机科学, 2007, 34(2):192-195.

[22] 王莉. 数据挖掘中聚类方法的研究[D]. 天津大学, 2004.

[23] BUTLER A, HOFFMAN P, SMIBERT P, et al. Integrating single-cell transcriptomic data across different conditions, technologies, and species[J]. Nature biotechnology, 2018, 36(5): 411-420.

[24] JIANG L, CHEN H, PINELLO L, et al. GiniClust: detecting rare cell types from single-cell gene expression data with Gini index[J]. Genome biology, 2016, 17(1): 144.