单细胞 RNA 测序数据分析方法研究进展

张森 孙祥瑞 徐春明

(北京工商大学轻工科学技术学院, 北京 100048)

摘要: 单细胞 RNA 测序 (Single cell RNA sequencing, scRNA-Seq) 已经广泛应用于细胞分化、肿瘤微环境及多种疾病病因学研究。目前, 由于 scRNA-Seq 具有低捕获率、高噪声、高变异性等特点, 通过优化数据分析方法提高测序结果准确性已经成为测序领域的研究热点。对近年来数据分析过程中利用的数学方法进行了总结, 讨论了数据分析的优势及存在的问题, 以期为新算法的开发和应用提供参考, 逐步提高测序结果的可靠性。

关键词: 单细胞测序; 数据分析; 质量控制; 差异表达; 算法

DOI: 10.13560/j.cnki.biotech.bull.1985.2020-0469

Research Progress of Approaches in Single Cell RNA Sequencing Data Analysis

ZHANG Miao SUN Xiang-rui XU Chun-ming

(School of Light Industry Science and Technology, Beijing Technology and Business University, Beijing 100048)

Abstract: Single cell RNA sequencing (scRNA-Seq) has been widely used in cell differentiation, tumor microenvironment and etiology of various diseases. Currently improving the accuracy of sequencing results by optimizing data analysis methods has become a research hotspot in the field of sequencing due to the characteristics of low capture rate, high noise, and high variability of scRNA-Seq. Here an overview of the mathematical methods used in the process of data analysis in recent years is summarized, and the advantages and issues in data analysis are discussed, aiming to provide a reference for the development and application of new algorithms, and to gradually improve the reliability of sequencing results.

Key words: single cell sequencing; data analysis methods; quality control; differential expression; algorithm

随着分子生物学技术的不断发展, 高通量测序已广泛应用于临床医学、基础医学、生物医学等众多领域, 因其具有通量高、速度快、灵敏度高等优势, 可在短时间内检测大量样本的基因变异及转录水平, 具有十分广阔的应用前景。以二代测序 (Next generation sequencing, NGS) 为典型代表的高通量测序已经广泛应用于多种疾病的诊断、治疗及预后评估^[1]。

NGS 可分为 DNA 测序和 RNA 测序, DNA 测序

以检测基因变异为主^[2], 如碱基替换、小片段插入及缺失等, RNA 测序主要以检测基因的 mRNA 丰度为主^[3]。根据所测细胞的群体区分, RNA 测序可分为全转录组测序 (Bulk RNA sequencing, bulk RNA-Seq) 和单细胞转录组测序 (Single cell RNA sequencing, scRNA-Seq)。Bulk RNA-Seq 是目前最常用的 RNA 测序方法, 但由于测序样本普遍具有异质性, bulk RNA-Seq 结果仅能代表大细胞群体中每一个基因的平均表达水平, 对比较转录组学研究有帮

收稿日期: 2020-04-21

基金项目: 国家科技重点研发计划项目 (2017ZX07301004)

作者简介: 张森, 女, 硕士研究生, 研究方向: 轻工技术与工程; E-mail: 792151970@qq.com

通讯作者: 徐春明, 男, 博士, 副教授, 研究方向: 轻工技术与工程; E-mail: xucm@th.btbu.edu.cn

助,但不利于异质性研究。2009年 scRNA-Seq 技术^[4]被首次引用,测序样本异质性问题得到一定程度解决,并逐渐成为近年来的研究热点。scRNA-Seq 原理是分离单个细胞并提取 RNA,经 PCR 扩增后进行高通量测序,主要工作流程包括细胞解离、单细胞分离、文库构建、上机测序及数据分析^[3],其中数据分析是整个 scRNA-Seq 过程中最重要的一个环节。

由于 scRNA-Seq 所使用的数据分析方法有别于 bulk RNA-Seq,越来越多的针对 scRNA-Seq 技术的数据分析方法不断涌现,但每种分析方法都有各自的优势及局限性。因此,本文对比了 scRNA-Seq 与 bulk RNA-Seq 技术在数据分析上的差异,对 scRNA-Seq 数据分析研究进展进行总结,探讨每种方法的优势与局限性,以期能够对 scRNA-Seq 数据分析方法进行系统了解。

1 数据预处理

1.1 数据的比对

在 scRNA-Seq 过程中,测序数据下机后需要进行预处理,将基因转录序列转换为 fastq 格式并与参考序列比对,鉴定差异表达基因或寻找可变剪切位点。scRNA-Seq 数据整体质量评估的重要指标是比对率,即 reads 在参考基因组中所占的比例。比对率越高说明数据利用率越高^[5]。目前应用于数据比对的工具主要有 TopHat^[6], STAR^[7] 或 HISAT^[8]。TopHat 工具以 Bowtie 作为核心算法针对 75 bp 以上长度的 RNA 短序列与参考基因组进行比对,找到匹配的序列,对外显子进行选择性地拼接,具有内存小、准确性高、容错率低、可跨内含子比对等优势。Donbin 等^[7] 研究发现 STAR 工具的核心算法是 Maximal mappable prefix (MMP),其直接选用非连续序列进行比对,运行速度较 TopHat 工具快,但需要更大的内存。HISAT 工具基于 Burrows-Wheeler 变换 (BWT) 和 Ferragina-Manzini 索引 (FM) 结合的算法进行对齐。Kim 等^[8] 研究发现 HISAT 是第一个采用分层索引以及自适应策略进行对齐的工具,减少了内存需求,也是目前运行速度最快的工具,具有同其他比对工具相同甚至更高的精度。从总体来看,以上 3 种软件在运行速度和结果准确性方面均表现良好,但 Engström 等^[9] 进一步研究发现运行速度

较快的软件通常检测的准确性较低。

1.2 质量控制

数据识别和去除低质量细胞是 scRNA-Seq 质量控制 (Quality control, QC) 的关键步骤。首先,在细胞捕获过程中应避免参杂混合或死亡的细胞。其次,使用 FastQC^[10] 等工具检查测序数据的质量,根据 QC 值决定其在后续分析中是否被舍弃。异常的 cut-off 值可以人为定义,也可由程序自动定义,但需考虑被分析组织的多样性。此外,当细胞受损时,细胞质 RNA 会丢失,但线粒体 RNA 会保留在受损细胞中,线粒体 RNA 含量是 QC 的另一个指标^[11]。除上述方法外,Jiang 等^[12] 最新提出的一种针对单细胞 RNA 序列质量控制 (SinQC) 工具,通过整合基因表达模式和数据质量信息来检测并去除低质量细胞。

1.3 数据标准化

数据标准化是数据预处理的关键步骤。在 bulk RNA-Seq 中,DESeq2^[13] 和 TMM^[14] 是常用的标准化处理算法。然而,DESeq2 算法并不适用于 scRNA-Seq 数据分析,因为 DESeq2 算法假设所有样本 RNA 总量相等,reads 数仅与测序深度有关,根据不同样本的 reads 数来计算比例,但对于单个细胞可能会受到零值和高变异性的影响,结果不稳定。因此,Bacher 等^[15] 提出 SCnorm 工具,其使用分位数回归的方法对测序数据进行标准化,可以避免针对 bulk RNA-Seq 的传统标准化方法对 scRNA-Seq 数据进行标准化时所引入的错误,改善主成分分析和差异表达基因的识别,可以用于 scRNA-Seq 数据标准化。最常用的数据归一化方法为 Count depth scaling,又称为 Counts per million (CPM),它会根据每个细胞的总表达量计算一个 size factor,然后对其中各个基因表达量进行标准化。除 CPM 外,非线性标准化方法使用测序深度拟合的负二项模型,可解释更复杂的异质性。因此,数据标准化方法的使用需要根据不同细胞特性进行选择,单一的标准化方法不能适用于所有类型的 scRNA-seq 数据。

2 插补

在测序过程中,很多低表达或中度表达的基因无法有效检测到,导致表达值为零或减少,影响下

游后续分析, 增加细胞间变异, 甚至不能获得完整的单个细胞转录组信息, 这种情况称之为 Dropout。在实际操作过程中 Dropouts^[16]对数据分析影响较大, 合适的插补方法可以弥补 Dropout 产生的影响。目前针对插补开发的算法有 MAGIC^[17]、ScImpute^[18]、SAVER^[19]、DrImpute^[20]和 AutoImpute^[21], 各种算法的计算原理不同。MAGIC 算法使用基于 Markov 亲和力的矩阵确定细胞间的相似性, 对高度相似细胞中基因表达进行聚集, 以估算基因表达量。ScImpute 算法通过拟合 Gamma-Normal 混合模型估算基因缺失概率, 根据相似细胞信息估算可能的 Dropout。SAVER 和 MAGIC 算法分析可能会造成未受 Dropout 影响的基因表达发生变化, 但 ScImpute 算法可以利用其他类似细胞中不太可能受 Dropout 影响的相同基因信息, 在不引入新偏差情况下计算缺失值。有研究表明, MAGIC 和 scImpute 算法都依赖于相似细胞基因数据, 这会消除细胞间的随机性, 而 Huang 等^[19]提出的 SAVER 算法来接收具有唯一分子索引的矩阵后, 假定每个基因都遵循 Poisson-Gamma 模型, 使用多元广义泊松回归模型的贝叶斯分析还原基因表达水平, 消除技术差异的同时还可保留不同细胞间的生物学差异。DrImpute 是一种集群分析算法, 通过使用 Spearman 和 Pearson 相关系数计算距离矩阵, 可将 Dropout 从真正的零值中有效地分离出来。Gong 等^[20]将 DrImpute 算法与多种插补算法进行性能比较发现, 与 MAGIC 和 scImpute 算法相比, DrImpute 可以恢复更多的缺失值, 提高后续细胞类型识别和拟时间推断的准确性。受到上述软件的启发, Talwar 等^[21]提出了 AutoImpute 自编码分析算法, 通过学习 scRNA-Seq 数据的固有分布和模式来寻找缺失值。通过与现有的 9 种独立数据集的插补算法进行比较, AutoImpute 被证实是唯一能对最大数据集进行插补而不会消耗内存的算法。

3 批次效应校正

完整的 RNA 测序流程包括细胞分离、RNA 提取、文库构建、上机测序及测序后数据分析等多个环节, 但不同实验室、不同时间以及不同人员操作会造成批次效应, 影响结果可靠性。批次效应也成为 scRNA-Seq 技术中常见的变异来源^[22]。由于 scRNA-

Seq 与 bulk RNA-Seq 在数据特征上具有差异, 常规用于 bulk RNA-Seq 的批次校正算法, 如 RUVseq^[23]和 svaseq^[24]等算法可能并不适用于 scRNA-Seq。但是, 在 scRNA-seq 研究中, 批次之间的种群组成通常并不相同, 即使假设每个批次中存在相同的细胞类型, 数据集中每种细胞类型的丰度也会根据细胞培养或组织提取、解离等过程中的细微差异而变化, 因此造成变异的因子并非仅考虑技术性因素。为了校正单细胞测序中的批次效应, 多种 scRNA-Seq 数据校正算法被开发, 包括 ComBat^[25]、相互最近邻 (Mutual nearest neighbours, MNN)^[26]和 Scanorama^[27]等。当批次信息可用时, ComBat 算法使用参数和非参数经验贝叶斯框架通过批处理效应变量的加法组合来描述基因表达。Haghverdi 等^[26]提出的 MNN 算法是计算成对细胞的余弦归一化表达谱之间的欧氏距离, 再根据每个批次中共享种群的偏差来调整批次效果。尽管 MNN 和 ComBat 是常用的分析算法, 但研究发现 MNN 算法性能要优于 ComBat。Hie 等^[27]最近提出的 Scanorama 是采用一种可识别并准确整合数据集的算法, 利用匹配的信息进行批次效应校正, 相比于 MNN 算法, 该技术不需要依赖于数据集的顺序, 将邻近搜索优化为低维嵌入的基因表达谱, 极大减少了搜索时间。

4 降维分析

高维性是 scRNA-Seq 数据的显著特点, 数据分析时常常要用到降维分析法。主成分分析 (Principal component analysis, PCA) 作为一种经典的无监督降维算法, 借助正交变换使线性维数减少, 产生一组不相关的分量, 通过最大化投影数据的方差, 将高维数据投影到低维线性空间上。其具有两大主要优势: 第一, PCA 通过正交线性投影可以消除基因间的冗余, 被用作多种降维方法的预处理步骤。第二, PCA 可将高维数据投影到低维线性空间上, 可以预测多维数据的相关性。研究表明^[28]通过分散表达水平来过滤基因, 然后选择数百个最具可变性的基因来捕获整个种群的重要特征。PCA 已成功应用于 scRNA-Seq 数据分析中^[29-31], 以捕获细胞异质性的整体结构, 其局限性在于无法可视化细胞聚类 and 细胞类型识别所必须的局部结构。

为了弥补 PCA 无法可视化的局限性, t 分布随机领域嵌入 (t-distributed stochastic neighbor embedding, t-SNE) 算法被引入单细胞测序分析。Alexander 等^[32]提出的 t-SNE 是一种用于高维数据可视化的非线性分析算法, 通过捕获局部结构, 将原始高维空间中不相似单元以大距离建模, 而相似单元则以小距离建模, 在不丢失数据点间相对距离的基础上, 将高维数据嵌入到二维或三维空间中进行可视化。通过降维与最近邻网络相结合来考虑数据点之间的局部距离, 目的是分离不同的群集。t-SNE 可以通过构造概率分布来描述数据集, 相似的单元格分配概率高, 相异的单元格分配概率低, 在高维空间中相似的单元将在低维空间中聚集在一起。t-SNE 在维持相似细胞群集能力方面优于 PCA。目前, t-SNE 还不能很好地捕获全局结构, 如群集之间的距离。尽管 t-SNE 在 scRNA-Seq 数据可视化方面取得了成功, 但仍存在两种算法的缺陷^[33]。首先, 由于 t-SNE 的随机性, 同一数据集在不同的运行中可能产生不同的可视化效果。为了获得对种群结构的认识, 可能需要对同一数据集进行多次 t-SNE 运行。其次, 虽然 t-SNE 将原始空间中相似单元格放置在低维空间中来维持簇, 但原始空间中不相似单元格不一定会在低维空间中按比例放置。最近, 一种基于黎曼几何和代数拓扑理论的 UMAP 工具^[34]被开发出来, 其性能和效率均优于 t-SNE。UMAP 工具能够沿着分化轨迹排列簇并保留瞬时细胞的分化连续体, 通过在二维或三维图上覆盖标记基因的表达或与生物过程有关的一组基因的活性, 捕获 scRNA-Seq 数据中局部和全局结构。

5 细胞亚型鉴定

在特定条件下, 对组织中的细胞亚群进行鉴定是 scRNA-Seq 数据分析的关键目标之一, 其结果可以揭示细胞异质性^[35]。结合降维分析方法, 通过聚类分析实现细胞亚群的鉴定。在无监督聚类分析中, 主要以分层聚类和 K-means 聚类为主。分层聚类无需预先定义聚类数量, 以聚集或分裂的方式进行连续合并或拆分, 目前常用的工具包括 SINCERA^[36]和 bigScale^[37]。其中, Iacono 等^[37]提出的 bigScale 工具框架构建了一个概率模型来定义所有可变性成

对细胞之间的表型距离。与在简单或混合概率模型中假设负二项式, 伽马或泊松分布的其他方法相比, bigScale 工具构建了一个高精度、全面的噪声数值模型, 通过将 P 值分配给每个基因来量化细胞间距离。而 K-means 聚类^[11]则是先确定簇中心, 再将细胞分配到最近的簇中心, 迭代优化质心位置, 将细胞分为 k 个簇, 根据质心聚类中细胞的平均值重新计算质心, 工作速度快于分层聚类。以上两种传统聚类方法都会受到数据规模和噪声的影响。为此, Lin 等^[38]在聚类前通过插补和降维进行聚类 (CIDR) 使用非线性最小二乘回归拟合数据, 并对零值进行插补来减弱 Dropout 影响。该算法可识别并评估 Dropout 与基因表达水平之间的关系, 计算基因表达谱之间的差异。通过实验证实 CIDR 运算速度远快于传统算法。近年来, 新的聚类算法不断被开发, 如 graph-based 聚类包括 SNN^[39]和 RaceID2^[40], 这些算法将单元格嵌入图形中, 每个边代表两个单元格之间的相似度, 将图形划分为高度互连的模块, 具有高效性和稳定性。Seurat 基于共享近邻 (Shared nearest neighbor, SNN) 聚类算法来识别细胞簇, 通过差异表达或方差分析来识别不同亚群标记物, 基于表达水平的相似度的不同构建共享近邻网络。为了证明 SNN 算法的有效性, Xu 等^[39]通过在不同结构的数据集上进行测试发现, 与原始数据的研究结论相同。为开发出一种可靠的方式推断分化轨迹, Grün 等^[40]提出 RaceID2 工具, 该软件适合测试微分动力学, 在增加集群数后可通过识别集群内的饱和点确定亚群数量, 使数据比 K-means 聚类更可靠, 且已在动物实验中证实。除上述方法外, 单细胞一致性聚类 (Single cell consensus clustering, SC3) 是特别为 scRNA-Seq 数据开发的聚类算法, 通过共识方法将多个聚类算法组合在一起, 具有高度的准确性和鲁棒性^[41], 相比于 K-means, SNN 和 SINCERA 算法分析, SC3 缺点是运行时间长, 但准确性最高^[42]。

6 差异表达分析

差异表达基因分析可以检测不同细胞类型、不同细胞亚群间的 mRNA 丰度, 通过组间比对, 获得不同样本或不同处理方法对基因表达水平的影响, 或上调或下调^[43], 进而可对差异表达基因进行功能

分析, 如通过基因本体分析 (Gene ontology, GO), 确定基因所参与的生物学过程、分子功能及细胞组分, 通过 KEGG 分析差异基因参与的信号通路。尽管 scRNA-Seq 结果可以鉴定差异表达基因, 但其也存在一定的局限性。首先, 由于单细胞测序数据通常具有较高的背景噪音, 很多低表达或者中等表达水平的基因不能被有效检测到。所以, 针对 bulk RNA-Seq 数据开发的差异表达检测算法, 并不完全适用于 scRNA-Seq。针对 scRNA-Seq 的差异表达算法被陆续开发, 如 SCDE^[44], MAST^[45], Census^[46] 和 BCseq^[47] 等。其中, SCDE 是一种运用贝叶斯算法, 从单个测量中获得的不确定信息, 使用泊松过程来解释 Dropouts, 通过对比分析证明 SCDE 算法具有比传统方法更高的灵敏度^[44]。MAST 算法采用线性模型对转录阳性表达的平均值进行建模, 同时控制模型的离散性和技术因素。其采用广义相加模型 (Generalized additive models, GAMS) 与 Tobit 模型进行正态分布。Finak 等^[45] 通过比较发现 SCDE 算法检测的差异表达基因数量高于 MAST, 但 MAST 算法的特异性更高。Qiu 等^[46] 研究发现 Census 算法可将常规的相对表达量转换为相对转录本计数, 与标准的读取计数相比, 使用回归技术更容易建模, 而且显著提高准确度。BCseq 算法无需指定偏差的来源或格式即可校正表达量化中的偏差, 即以自适应的方式纠正固有偏差, 有效降低技术噪音。Chen 等^[47] 通过对比发现 BCseq 算法在细胞类型分类性能上优于 MAST 和 SCDE。尽管多种算法被用于差异表达基因分析, 但不同算法处理后的数据结果仍存在一定偏差。早期一项针对 36 种基因差异表达分析算法有效性的研究发现, 不同算法得到的差异表达基因在特征及数量上均存在显著差异^[48]。因此, 测序后数据分析算法的优化仍然是今后的一项重要工作。

7 拟时序分析

拟时序分析法是指根据单个细胞的基因表达模式推断出细胞发育或分化的动态路径^[49]。与 bulk RNA-Seq 不同, 单细胞测序可以沿着一个连续发展的过程对细胞进行排序, 在轨迹的开始、中间和结束状态对细胞类型进行识别, 进展越少越接近原

始细胞状态, 进展越多越接近终点细胞状态。针对 scRNA-Seq 拟时序分析开发的算法有 Monocle^[50], Waterfall^[51], TSCAN^[52], Sincell^[53], SLICER^[54] 和 Wishbone^[55]。Monocle 算法将无监督的数据与反向图嵌入结合在一起, 通过分化进程对细胞进行排序, 揭示关键调控因子表达中的变化及细胞分化的新型调控因子。Perešini 等^[50] 将 Monocle 算法可应用于骨骼肌分化过程中, 明确了一系列形态学和转录组动力学。该算法将每个细胞的表达谱表示为高维空间的一个点, 高度相似的单元格之间添加连接边, 构建最小生成树图, 找到最长路径的对应转录序列, 即可找到分化过程中单个细胞的动态路径。当对体内连续生物学过程进行 scRNA-Seq 分析时, 由于缺少足够的信息不适合使用 Monocle 等传统方法, 因此 Jaehoon 等^[51] 开发了一种更为通用的算法 Waterfall, 它可以对连续生物学过程的多位单细胞数据集进行无偏差统计分析, 该算法使用 Hidden markov 模型以无偏差方式确定拟时序上每个基因的表达状态, 并量化为随时间变化的分子级联图, 最后将基因表达水平与拟时序相关联。尽管 Waterfall 算法中曾考虑过细胞聚类的影响, 但并未对细胞聚类对细胞有序性影响进行系统评估, 在此基础上, Ji 等^[52] 提出 TSCAN 工具, 基于构建最小生成树之前对细胞进行聚类可以降低复杂性, 解决 Monocle 由于高复杂性造成的稳定性差。Juliá 等^[53] 开发的 Sincell 实现了两种算法, 以区分稳定细胞和噪声影响的层次结构, 第一种依赖于基因重采样程序; 第二种系统随机生成的复制细胞, 这些复制细胞遵循原始细胞的随机模式。Sincell 可以提供细胞状态层次结构, 同时考虑 scRNA-Seq 序列中的随机因素。上述方法无法推断对于非线性基因表达的变化和与过程无关的基因的分析, Welch 等^[54] 开发 SLICER 使用局部线性嵌入来重建细胞轨迹, 该算法可以推断非线性的轨迹, 无需了解过程即可选择基因, 并自动确定分支的位置和数量, 通过对小鼠非细胞和神经干细胞的验证, 证实了此算法的有效性。如果要为多细胞构建分支轨迹, 上述方法在分辨率和准确性上较为不足, Manu 等^[55] 提出的 Wishbone 算法解决了此类问题, 关键技术在于通过重复采样边缘子集来确定轨迹, 并用动物实验证明了 Wishbone 的

准确性。因此，选择合适的方法应主要依赖于数据集特点。

8 结语

随着多种单细胞测序数据分析方法的开发与应用，在一定程度上促进了单细胞转录组学的发展，改善了高背景噪音和高变异性对数据产生的影响，为细胞异质性研究奠定了分子基础。但单细胞测序数据分析仍面临着新的问题和挑战。首先，随着单细胞测序数据集激增，如何提高软件运行速度和储存效率是目前需要解决的一项重要问题。其次，由于不同实验室在实验方案和数据处理流程方面存在差异，结果的室间比对较为困难。因此，我们仍然有必要对现有的分析方法进行优化，不断开发新的高效的数据分析方法，进一步提升单细胞测序结果的准确性和可靠性。

参考文献

- [1] Goodwin S, McPherson JD, McCombie WR. Coming of age : ten years of next-generation sequencing technologies [J] . Nature Reviews Genetics, 2016, 17 (6) : 333-351.
- [2] Shendure J, Balasubramanian S, Church GM, et al. DNA sequencing at 40 : past, present and future [J] . Nature, 2017, 550 (7676) : 345-353.
- [3] See P, Lum J, Chen J, et al. A Single-cell sequencing guide for immunologists [J] . Frontiers in Immunology, 2018, 9 : 2425.
- [4] 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.
- [5] Alessandra DM, Barbara DC . How to design a single-cell RNA-sequencing experiment : pitfalls, challenges and perspectives [J] . Briefings in Bioinformatics, 2018, 20 (4) : 1384-1394.
- [6] Trapnell C, Pachter L, Salzberg SL. TopHat : discovering splice junctions with RNA-Seq [J] . Bioinformatics, 2009, 25 (9) : 1105-1111.
- [7] Dobin A, Davis Carrie A, Schlesinger F, et al. STAR : ultrafast universal RNA-seq aligner [J] . Bioinformatics, 2012, 29 (1) : 15-21.
- [8] Kim D, Langmead B, Salzberg SL. HISAT : a fast spliced aligner with low memory requirements [J] . Nature Methods, 2015, 12 (4) : 357-360.
- [9] Engström PG, Steijger T, Sipos B, et al. Systematic evaluation of spliced alignment programs for RNA-seq data [J] . Nature Methods, 2013, 10 (12) : 1185-1191.
- [10] Crinier A, Milpied P, Escalière B, et al. High-dimensional single-cell analysis identifies organ-specific signatures and conserved NK cell subsets in humans and mice [J] . Immunity, 2018, 49 (5) : 971-986.
- [11] Darmanis S, Sloan SA, Croote D, et al. Single-cell RNA-seq analysis of infiltrating neoplastic cells at the migrating front of human glioblastoma [J] . Cell reports, 2017, 21 (5) : 1399-1410.
- [12] Jiang P, Thomson JA, Stewart R, et al. Quality control of single-cell RNA-seq by SinQC [J] . Bioinformatics, 2016, 32 (16) : 2514-2516.
- [13] Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2 [J] . Genome Biology, 2014, 15 (12) : 550.
- [14] Robinson MD, Oshlack A. A scaling normalization method for differential expression analysis of RNA-seq data [J] . Genome Biology, 2010, 11 (3) : R25.
- [15] Bacher R, Chu LF, Leng N, et al. SCnorm : robust normalization of single-cell RNA-seq data [J] . Nature Methods, 2017, 14 (6) : 584-586.
- [16] Hicks SC, Townes FW, Teng M, et al. Missing data and technical variability in single-cell RNA-sequencing experiments [J] . Biostatistics, 2018, 19 (4) : 562-578.
- [17] Van DD, Sharma R, Nainys J, et al. Recovering gene interactions from single-cell data using data diffusion [J] . Cell, 2018, 174 (3) : 716-729.
- [18] Li WV, Li JJ . An accurate and robust imputation method scImpute for single-cell RNA-seq data [J] . Nature Communications, 2018, 9 (1) : 997.
- [19] Huang M, Wang JS, Torre E, et al. SAVER : gene expression recovery for single-cell RNA sequencing [J] . Nature Methods, 2018, 15 (7) : 539-542.
- [20] Gong WM, Kwak HY, Pota P, et al. DrImpute : imputing dropout events in single cell RNA sequencing data [J] . BMC Bioinformatics, 2018, 19 (1) : 220.
- [21] Talwar D, Mongia A, Sengupta D, et al. AutoImpute : Autoencoder

- based imputation of single-cell RNA-seq data [J]. *Scientific Reports*, 2018, 8 (1): 16329.
- [22] Tran HTN, Ang KS, Chevrier M, et al. A benchmark of batch-effect correction methods for single-cell RNA sequencing data [J]. *Genome Biology*, 2020, 21 (1): 12.
- [23] Risso D, Ngai J, Speed TP, et al. Normalization of RNA-seq data using factor analysis of control genes or samples [J]. *Nature Biotechnology*, 2014, 32 (9): 896-902.
- [24] Leek JT. Svsseq: removing batch effects and other unwanted noise from sequencing data [J]. *Nucleic Acids Research*, 2014, 42 (21): 161.
- [25] Ritchie ME, Phipson B, Wu D, et al. limma powers differential expression analyses for RNA-sequencing and microarray studies [J]. *Nucleic Acids Research*, 2015, 43 (7): 47.
- [26] Haghverdi L, Lun ATL, Morgan MD, et al. Batch effects in single-cell RNA-sequencing data are corrected by matching mutual nearest neighbors [J]. *Nature Biotechnology*, 2018, 36 (5): 421-427.
- [27] Hie B, Bryson B, Berger B, et al. Efficient integration of heterogeneous single-cell transcriptomes using Scanorama [J]. *Nature Biotechnology*, 2019, 37 (6): 685-691.
- [28] Zhu S, Qing T, Zheng Y, et al. Advances in single-cell RNA sequencing and its applications in cancer research [J]. *Oncotarget*, 2017, 8 (32): 53763-53779.
- [29] Aljanahi AA, Mark D, Dunbar CE. An introduction to the analysis of single-cell RNA-sequencing data [J]. *Molecular Therapy - Methods & Clinical Development*, 2018, 10: 189-196.
- [30] Fletcher RB, Diya D, John N. Creating lineage trajectory maps via integration of single-cell RNA-sequencing and lineage tracing [J]. *Bioessays*, 2018, 40 (8): 56.
- [31] Byungjin H, Hyun LJ, Duhee B. Single-cell RNA sequencing technologies and bioinformatics pipelines [J]. *Experimental & Molecular Medicine*, 2018, 50 (8): 96.
- [32] Alexander P, Vladimir B. Visualization of SNPs with t-SNE [J]. *PLoS One*, 2013, 8 (2): e56883.
- [33] Gate D, Saligrama N, Leventhal O, et al. Clonally expanded CD8 T cells patrol the cerebrospinal fluid in Alzheimer's disease [J]. *Nature*, 2020, 577 (7790): 399-404.
- [34] Becht E, McInnes L, Healy J, et al. Dimensionality reduction for visualizing single-cell data using UMAP [J]. *Nat Biotechnol*, 2019, 37: 38-44.
- [35] Wei Z, Shu C, Zhang C, et al. A short review of variants calling for single-cell-sequencing data with applications [J]. *The International Journal of Biochemistry & Cell Biology*, 2017, 92: 218-226.
- [36] Minzhe G, Hui W, Steven PS, et al. SINCERA: A pipeline for single-cell rna-seq profiling analysis [J]. *PLoS Computational Biology*, 2015, 11 (11): e1004575.
- [37] Iacono G, Mereu E, Guillaumet-Adkins A, et al. bigScale: an analytical framework for big-scale single-cell data [J]. *Genome Research*, 2018, 28 (6): 878-890.
- [38] Lin PJ, Troup M, Ho JWK. CIDR: Ultrafast and accurate clustering through imputation for single-cell RNA-seq data [J]. *Genome Biology*, 2017, 18 (1): 59.
- [39] Xu C, Su ZC. Identification of cell types from single-cell transcriptomes using a novel clustering method [J]. *Bioinformatics*, 2015, 31 (12): 1974-1980.
- [40] Grün D, Muraro MJ, Boisset JC, et al. *De novo* prediction of stem cell identity using single-cell transcriptome data [J]. *Cell Stem Cell*, 2016, 19 (2): 266-277.
- [41] Levine JH, Simonds EF, Bendall SC, et al. Data-driven phenotypic dissection of AML reveals progenitor-like cells that correlate with prognosis [J]. *Cell*, 2015, 162 (1): 184-197.
- [42] Kiselev VY, Kirschner K, Schaub MT, et al. SC3: consensus clustering of single-cell RNA-seq data [J]. *Nature Methods*, 2017, 14 (5): 483-486.
- [43] Ankur S, Elaine YC, Vibhor KAS, et al. Longitudinal single-cell RNA sequencing of patient-derived primary cells reveals drug-induced infidelity in stem cell hierarchy [J]. *Nature Communications*, 2018, 9 (1): 4931.
- [44] Kharchenko PV, Silberstein L, Scadden DT. Bayesian approach to single-cell differential expression analysis [J]. *Nature Methods*, 2014, 11 (7): 740-742.
- [45] Finak G, McDavid A, Yajima M, et al. MAST: a flexible statistical framework for assessing transcriptional changes and characterizing heterogeneity in single-cell RNA sequencing data [J]. *Genome Biology*, 2015, 16: 278.
- [46] Qiu X, Hill A, Packer J, et al. Single-cell mRNA quantification and differential analysis with Census [J]. *Nature Methods*, 2017, 14 (3): 309-315.

- [47] Chen L, Zheng S. BCseq : accurate single cell RNA-seq quantification with bias correction [J] . *Nucleic Acids Research*, 2018, 46 (14) : e82.
- [48] Sonesson C, Robinson M. Bias, robustness and scalability in single-cell differential expression analysis [J] . *Nature Methods*, 2018, 15 (4) : 255-261.
- [49] Halpern KB, Shenhav R, Matcovitch-Natan O, et al. Single-cell spatial reconstruction reveals global division of labour in the mammalian liver [J] . *Nature*, 2017, 542 (7641) : 352-356.
- [50] Perešini P, Kuzniar M, Kostic D. Dynamic, Fine-grained data plane monitoring with monocle [J] . *IEEE/ACM Transactions on Networking*, 2018, 26 (1) : 534-547.
- [51] Jaehoon S, Daniel AB, Yunhua ZS, et al. Single-cell RNA-Seq with waterfall reveals molecular cascades underlying adult neurogenesis [J] . *Cell Stem Cell*, 2015, 17 (3) : 360-372.
- [52] Ji ZC, Ji HK. TSCAN : Pseudo-time reconstruction and evaluation in single-cell RNA-seq analysis [J] . *Nucleic Acids Research*, 2016, 44 (13) : e117.
- [53] Juliá M, Amalio T, Antonio R. Sincell : an R/Bioconductor package for statistical assessment of cell-state hierarchies from single-cell RNA-seq [J] . *Bioinformatics*, 2015, 31 (20) : 3380-3382.
- [54] Welch JD, Hartemink AJ, Prins JF. SLICER : inferring branched, nonlinear cellular trajectories from single cell RNA-seq data [J] . *Genome Biology*, 2016, 17 (1) : 106.
- [55] Manu S, Michelle DT, Shlomit RZ, et al. Wishbone identifies bifurcating developmental trajectories from single-cell data [J] . *Nature Biotechnology*, 2016, 34 (6) : 637-645.

(责任编辑 狄艳红)