DOI:10.11721/cqnuj20210401

单细胞 RNA 测序数据分析方法研究进展*

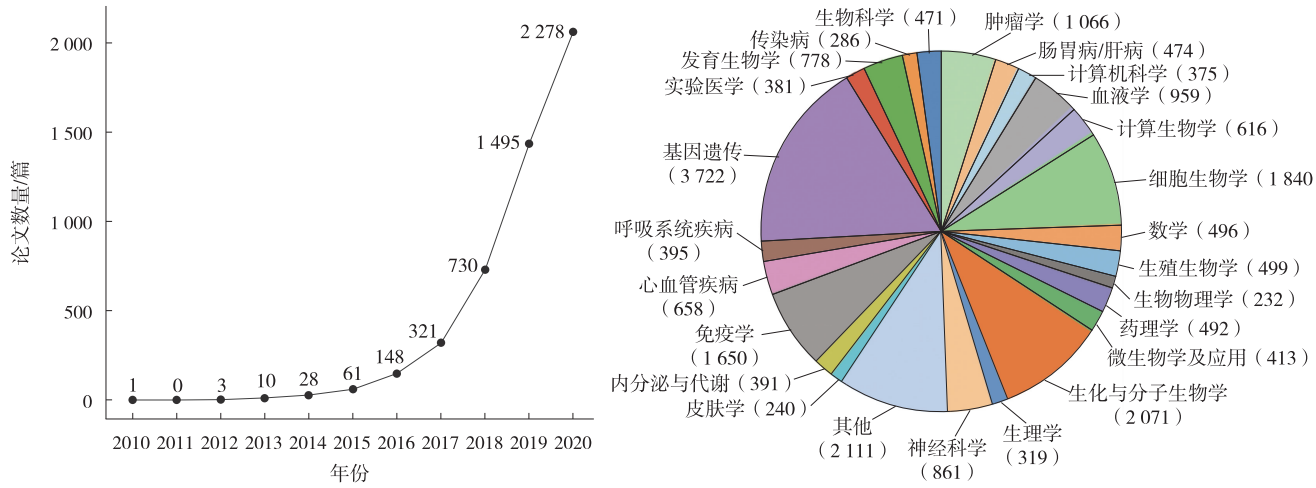
李　勃，朵泓睿
(重庆师范大学 生命科学学院，重庆 401331)

摘要:【目的】综述当前单细胞 RNA 测序数据分析的关键流程和环节，介绍完成不同分析任务所需的代表性方法及流行的工具。【方法】通过文献调研，总结了当前单细胞 RNA 测序数据分析的流程和代表性工具。【结果】许多针对单细胞 RNA 测序数据的分析流程和工具被陆续开发出来，用于从海量数据中发掘生物学知识，进而揭示复杂疾病或表型背后潜在的分子机制。【结论】单细胞 RNA 测序在生命科学研究中扮演了极为重要的角色，良好的数据分析策略是决定能否有效揭示单细胞表达谱数据背后所蕴含生物学信息的关键环节。目前单细胞 RNA 测序数据分析步骤和工具方法繁多，研究者应根据实际场景选择合适准确的分析方法与工具。

关键词:单细胞 RNA 测序；数据分析；降维；聚类；细胞轨迹；共表达网络
中图分类号:Q31　　**文献标志码**:A　　**文章编号**:1672-6693(2021)05-0129-07

单细胞 RNA 测序(Single-cell RNA sequencing，scRNA-seq)是一种在单细胞水平上利用 RNA 测序对特定细胞群体进行基因表达谱定量的高通量实验技术[1]。待测组织经过单细胞分离、RNA 提取、逆转录、文库构建和测序，便可利用数据分析获得多个细胞的基因表达谱。与传统的转录组学测序相比，scRNA-seq 技术可以描绘组织块(或细胞悬液)中单个细胞独特的基因表达模式，反映群体的细胞异质性。

scRNA-seq 技术由汤富酬等人[2]在 2009 年首次报道，该技术可以同时检测多个细胞中数千基因的转录水平。随后，Quartz-Seq、CEL-seq、Smart-Seq、Smart-Seq2、MARS-seq、Drop-seq 以及 inDROP 等不同平台的 scRNA-seq 技术陆续被开发，该领域迅速繁荣起来(图 1)。2013 年 scRNA-seq 被 *Nature methods* 杂志列为年度最主要的方法学进展，2019 年以 scRNA-seq 为核心代表的单细胞多组学方法再次被 *Nature methods* 评选为 2019 年度方法。显然，scRNA-seq 已经迅速成为当前生命科学领域最活跃和前沿的技术之一。



a　每年的发文数量　　b　单细胞 RNA 测序的应用领域与发表文章数量

图 1　当前单细胞 RNA 测序技术的发展状况调查

Fig. 1　A survey of current developments on single cell RNA-seq technology

* **收稿日期**:2020-12-23　**修回日期**:2021-01-20　**网络出版时间**:2021-06-30 09:22
资助项目:国家自然科学基金面上项目(No. 31871274)；重庆市自然科学基金面上项目(No. cstc2019jcyj-msxmX0527)
第一作者简介:李勃，男，副教授，博士，研究方向为生物信息学，E-mail：libcell@cqnu.edu.cn
网络出版地址:https://kns.cnki.net/kcms/detail/50.1165.N.20210629.1801.006.html

目前scRNA-seq技术已广泛应用于生命科学的诸多领域,尤其在细胞种类鉴定、肿瘤异质性、细胞免疫微环境的识别、细胞谱系演化和组织空间重构等方面发挥着极为重要的作用[3]。以癌症为例,利用该技术可绘制细胞亚群图谱,揭示不同细胞的生理状态,解析不同细胞亚类的功能及它们在癌症发生中的作用[4]。Wang等人[5]利用该技术在胃癌病人体内证实肿瘤内异质性,并鉴定到肿瘤形成过程中的部分关键基因,为胃癌治疗提供了新方向。在免疫学领域,该技术能够在高分辨率和复杂度下解密免疫过程、识别炎症反应微环境的免疫细胞和追踪组织细胞成分在疾病发展过程中的变化[6]。此外,该技术也被用于细胞发育谱系的鉴定,如Zhong等人[7]利用该技术识别出人(*Homo sapien*)海马体的47种细胞亚型和两种祖细胞的演化过程,并揭示了调控此过程的发育阶段特异性基因,为深入阐述海马体发育过程与诊断相关疾病提供先验知识。

单细胞RNA测序的飞速发展、数据量的增长和数据复杂度的增加对数据分析方法提出了巨大挑战[8]。为最大限度发挥scRNA-seq技术的优势,挖掘数据背后所蕴藏的生物学信息、发掘新的有用的生物学知识,已有多种数据分析软件和流程被开发。封三彩图2展示了典型的scRNA-seq数据分析流程,包括从序列比对到质控、修正批次效应、降维、细胞分群以及更深入的分析等步骤,每一步都包含着多种分析和处理的方法与工具。但在实际的数据分析中,由于研究的侧重点和待解决的问题不同,数据分析的步骤和环节会有所调整。此外,随着分析工具和方法的不断涌现,研究者难免会在选择相关工具方法上犹豫不决。因此,本文回顾单细胞RNA测序数据分析的一般流程、使用的方法和工具,希望能为相关领域的研究者提供重要参考和技术指导。

1 序列比对和表达水平量化

基因表达谱矩阵是scRNA-seq数据后续分析的出发点,它来源于对高通量RNA测序数据的初步处理。带有双标签(barcode和UMI)的读长(reads)与参考的全基因组进行比对,而后对每个特定基因的转录本进行定量,即可获得表达谱矩阵。该环节是数据分析前最基本的步骤,保证后续分析的准确性。

当前流行的scRNA-seq序列比对工具有STAR、HISAT2、Tophat2、Rsubread等。Rsubread可单独进行序列比对与表达值定量,而其他软件需要与表达量化工具结合使用[9]。值得一提的是,尽管序列比对和表达值量化通常是由部分软件单独完成的,但目前部分新开发的工具(如scPipe和scAlign等)可同时提供包括序列比对、表达值定量和下游后续分析等一系列配套的功能[10]。

2 预处理

scRNA-seq数据的分析从一个描述每个细胞中基因丰度的稀疏矩阵开始,该矩阵的每行和每列分别代表一个细胞和一个基因,而矩阵的单个元素则代表特定转录本读长的数量(Count)。通过比对形成的原始矩阵不能直接用于下游分析,需要经过质量控制、缺失值填充等处理和优化。

2.1 质量控制

由于技术和环境等原因,细胞标签(Barcodes)可能会标记那些不需要的转录产物,如doublets、死细胞或外膜破损的细胞[11]。针对这些问题,目前有以下几种解决策略:1) 根据每个细胞中转录本的总量或库的大小进行合格细胞的筛选。2) 依据线粒体基因读长(Reads)所占的百分比来筛选合格细胞。3) 采用Spike-in占基因总表达量的比例来判断细胞是否符合标准。4) 根据每个基因在所有细胞中表达量的总和来筛选基因。在实际操作中,考虑到严格筛选可能删除具有实际意义的细胞,研究者通常会采用较为松弛的标准进行初筛,在后续分析的过程中再进行二次筛选或人工识别。

2.2 缺失值填充

由于技术限制在单细胞RNA测序过程中无法捕捉到数量较少的mRNA分子,导致了单细胞RNA测序数据高度稀疏且有很多低丰度的非零值。目前常采用填充缺失值的策略来解决单细胞RNA测序数据稀疏性问题。此外,近些年来一些专门用于scRNA-seq数据的缺失值填充方法和工具陆续被开发,如CIDR、MAGIC、ScImpute、SAVER、DrImpute、DCA等[12-13]。

3 数据标准化

单细胞RNA测序中,由于细胞之间的异质性及技术因素,各单细胞文库大小和测序深度会有不同,需要通过统计学方法消除这种差异,即数据标准化(Normalization)。有代表性的方法和工具包括CPM、Seurat、Scater

和 Scran 等[14]。以 CPM 标准化为例，该方法主要基于如下假设：所有细胞中包含等量的 mRNA 分子，故所有的 Count 深度差异全部来自于抽样[11]。CPM 标准化后的 Count 矩阵需要进行对数转换，便于后续的差异表达分析。需要注意的是，对数转换过程中通常会添加一个极小值（如 1），以避免对数底数为 0。另一类常用的标准化方法主要基于概率模型，此类方法有 ZINB-WaVE、scVI、DCA 等[15]。

4 高变异基因的选择与降维分析

单细胞 RNA 测序数据通常是高维度的。很多基因不仅低表达，且在多个细胞间的表达值呈高度相关，因而导致大量信息冗余。特征选择（Feature selection）和降维（Dimension reduction）恰好为该问题提供了一个良好的解决方案。

4.1 高变异基因的选择

在 scRNA-seq 数据分析时，降维分析和细胞分群之前通常会先识别出高变异基因（Highly variable gene，HVG），即特征选择。当前应用于该领域的主要工具包括 Seurat 和 Scanpy，它们主要基于方差与平均值的比率来筛选 HVGs[11]。在 scRNA-seq 分析中保留的 HVGs 的数量主要取决于阈值的选择，通常为 500～5 000 个，对于复杂数据集可将 HVGs 的数量保持在 2 000～5 000[9]。此外需要注意的是，HVGs 的选取前需纠正技术误差，防止所选取的 HVGs 中部分变异是由批次效应产生的[11]。

4.2 线性降维方法

降维实际上是通过组合多个原有的特征到某个新的特征信息中，以得到一组压缩精炼的特征信息[16]。多数降维方法都基于线性分解模型，如 PCA、ICA、WNMF 等[9]。以 PCA 为例，一般要求尽量选取较少的主成分且又能获得较多的数据信息量。为确定主成分的数量，研究者通常会选取一个阈值，确保选取的主成分能够包含大部分有效的生物变异信息（比如可以捕捉到整个数据集 95％的方差）。简言之，使用线性分解的方法进行特征总结会得到更多有效的主成分结果，并能捕获到细胞之间更多的生物变异和减少离群噪音信号的影响[9]。

4.3 非线性降维方法与可视化

t-SNE 是 scRNA-seq 数据分析中应用最广泛的数据降维方法。但由于该方法主要是捕获细胞的局部相似性，因此可能会扩大细胞亚群之间的差别而缩小它们之间潜在的相似性。目前，有一些改良 t-SNE 算法将细胞的数量进行幂次缩放，以减少原始 t-SNE 算法的运行时间和提高运行的效率[17]。另一组非线性可视化方法是基于 k 最邻近算法（k-nearest neighbor algorithm），该算法将边缘的一个细胞与它最邻近的 k 个细胞相连接并根据力导向布局进行可视化[11]。此外，UMAP 是另外一种典型的非线性降维方法。与 t-SNE 相比，UMAP 不仅能够加快计算速度，而且在识别细微的细胞群体、保护全局结构和细胞子集连接性方面具有较大优势[18]。

5 聚类分析

聚类（Clustering）主要是依据细胞-细胞距离矩阵将细胞归属到数目不等的类群中，使高度相似的细胞最大限度地聚为一个类群。聚类的目标是探究或鉴定组织样本中细胞类型或亚型，揭示组织的复杂结构和潜在功能。研究者事先并不知原始的所有细胞应该归属于几类以及这些细胞是否具有聚类的意义，因此聚类前用户需要初步判断数据集的聚类趋势。霍普金斯统计量（Hopkins statistic）恰好是这样一种统计指标，它可用来测试数据在空间分布上的随机性，从而判断聚类趋势。若该统计量小于 0.5，则表明该数据集不太可能有聚类效应；相反，若数值接近于 1，则基本可以确定该数据集具有聚类效应。另外，对聚类的趋势判断也可通过可视化方法来实现。

目前在单细胞 RNA 测序数据分析中最流行的聚类策略有层次聚类（Hierarchical clustering）和分割聚类（Partitioning clustering）。对于层次聚类，代表性算法有 AGNES 和 DIANA 等。而分割聚类的典型代表则是 k-means 算法。此外，基于密度的聚类方法（如 DBSCAN）、基于神经网络的聚类方法（如 SOM）和基于模型的方法（如 EM 算法）等也已被应用于 scRNA-seq 数据分析领域[19]。

在聚类分析完成后，对聚类结果进行评估是一个不可或缺的环节，该过程对于筛选和确定针对特定数据集而言最佳的聚类方法和优化类别数目至关重要。目前，对聚类结果的评价主要基于“internal”、“stability”和“biological”准则[20]。最流行的聚类分析结果评估软件有 clValid 和 partition Comparison 包等。以 clValid 为例，“internal”准则主要使用数据的内在信息对聚类结果进行评估，通过计算细胞之间 Connective、Sihouette

width、Dunn index等参数,量化类群之间的距离和单个类群内部细胞的致密性,以评价聚类的程度。“stability”准则将一个原始矩阵的聚类结果与移除每一列后的矩阵得到的聚类结果进行比较,以评估聚类结果的一致性。而“biological”准则主要评估某种算法产生有生物学意义的聚类结果的能力。

6 细胞类型注释

在scRNA-seq数据分析中,通过对单个细胞类群的marker基因进行鉴定,可赋予每个类一个有生物学意义的标签,该过程即细胞类型注释。鉴定和注释细胞类群主要依赖于外部参考数据库,如Human Cell Atlas、CellMarker、CancerSEA、PanglaoDB等[21]。在无相关参考库的情况下,可通过现有细胞的标志基因(Marker gene)和文献报道的特定类型细胞的标志基因进行匹配来鉴定未知细胞的类型。目前有多种工具包能够进行自动化的细胞注释。以Garnett和scmap为例,它们可通过直接将注释过的参考细胞群的表达谱和未知种类细胞的基因表达谱进行比对,进而将注释信息对应到待测scRNA-seq数据集的细胞群上。此外,结合已发表的特定类型细胞的标志基因等信息,人工进行细胞类型的注释也是一种重要的解决方案[22]。相比而言,人工注释的方法在准确性上要优于自动注释,而自动注释在注释效率和灵敏度上要优于手动注释。对于较大的数据集来说,现阶段最好的办法是同时进行软件或数据库自动注释及人工注释;而对于细胞类型复杂度较低的数据集而言,人工注释更为经济有效。

7 差异表达基因的鉴定及富集分析

对已注释的细胞类群进行基因差异表达分析,可获得一组细胞类群间差异表达的基因(Differentially expressed genes,DEGs),以此为线索可以深入探讨细胞异质性,得出更加合理的实验结果。此外,在获得的不同细胞类群间DEGs的基础上进行基因集富集分析,则可进一步挖掘DEGs所反映的重要生物学意义,揭示数据背后所蕴含的可能的分子机制。

7.1 差异表达基因的筛选

基因差异表达分析是RNA测序后数据分析的一个重要环节,目标是鉴定出两组或者多组细胞类群间呈现稳定差异表达的基因集。筛选DEGs最便捷的方法是基于非参数检验思想,以wilcoxon秩和检验所得p值的校正错误发现率(False discovery rate)为准则,实现DEGs的筛选。除非参数方法外,还有多种参数方法和基于建模的方法,以消除技术因素和改善正确率。在scRNA-seq数据分析领域的ROC、MAST、DESeq2等,它们具有较好的灵活性和便捷性[23]。专于scRNA-seq数据的差异表达基因筛选方法与传统的方法相比并无绝对的优势,因此研究者在选择相应的差异表达基因筛选方法和工具时无需刻意回避传统的DEGs筛选方法和工具。

7.2 基因集富集分析

基因集富集分析(Gene set enrichment analysis)用来评估一个预先定义的基因列表特定基因集上的分布趋势,从而判断它对表型的贡献。在获得感兴趣的细胞群体间差异表达基因集后,进行基因集的富集分析将有利于生物学问题的解释和疾病(或表型)相关分子机制的阐明。代表性的基因集富集分析策略有GO富集分析、KEGG通路富集分析、转录因子富集分析和micro RNA调控模块富集分析等。基因集的富集通常采用过代表分析策略:首先筛选差异基因,然后通过预先构建的基因注释数据库对基因进行注释,最后通过统计检验算法(如超几何检验、Fisher精确检验等)得出基因富集的结果。此外,还有一类代表性的富集分析策略即GSEA,主要思想是使用预定义的基因集(基于注释信息),将基因在两类样本中的差异表达程度排序,然后检验预先设定的基因集合是否在这个排序表的顶端或者底端富集[24]。与超几何检验、Fisher精确检验等ORA策略相比,GSEA不需要制定明确的筛选阈值来筛选DEGs,不容易遗漏一些差异性不大但生物学意义极其重要的基因,因此产生的结果会更加准确可信。

8 进阶分析

8.1 细胞谱系追踪

多细胞生物个体都是由最初的单个细胞(受精卵)通过增殖和分化发育而来。由于基因的选择性表达,细胞分化的过程中会产生形态、功能各异的细胞。在发育生物学和遗传学中,细胞谱系分化推断(Cell lineage inference)是指推测由某一种细胞(例如祖细胞)经过逐级分化形成各类后代细胞的过程[25]。到目前为止,研究

者不仅能根据遗传学和发育生物学证据追踪细胞谱系分化的轨迹（封三彩图 3a），而且能从 scRNA-seq 数据出发实现细胞谱系的追踪。通常 scRNA-seq 技术只能描绘在某一时刻细胞中的基因转录表达状态，只能得到一张细胞的“快照”。近些年来人们提出了“伪时间（pseudotime）”的概念，它与真实时间最大的区别在于它可以近似作为衡量细胞分化发育的相对次序，而该次序是通过对细胞间表达谱的相似性计算和推测得到。随后，这些细胞将会按照这种相对次序被分配到一个一维空间中，代表着细胞发育分化进程中的一种独特状态。值得注意的是，伪时间的排序目前尚有一定的缺陷，分析结果不一定代表实际的细胞分化过程。

目前结合单细胞 RNA 测序数据进行细胞谱系分化推断的算法有很多，主要包括 3 类：1）基于降维的算法，如 Monocle、SLICE、TSCAN Waterfall、SCUBA 等。2）基于最近邻图的算法，如 Wanderlust、Wishbone 等。3）其他类细胞谱系追踪方法，如 RNA 速度算法等[26]。

8.2 生物网络构建

网络分析是单细胞 RNA 测序领域备受青睐的研究策略之一。由于细胞-细胞或基因-基因之间无时无刻不存在着互作和调节，影响着细胞的代谢、分化和其他生物学功能的执行，因此构建基因和细胞的相关互作网络对于理解复杂生物性状与微观水平的分子调控机制至关重要。

8.2.1 基因共表达网络　基因共表达网络（Gene co-expression network）是根据基因的表达值构建基因表达调控关系（封三彩图 3b），以探究特定的基因表达模式、基因共表达模块间的相互关系以及预测转录因子对基因表达的影响等，也可被用于潜在生物标志物的识别和药物作用靶点的鉴定。在共表达网络中，节点代表基因，基因与基因之间的连线代表两者存在互作或调控。节点大小表示与该基因相互调控的基因个数，节点越大，说明与该基因存在调控关系的基因越多。节点颜色则是按照聚类的结果所标记的，被划分到同类的基因有相同颜色的标记。每类基因中都有独特的表达模式，而通过对该类模块特征和核心基因的鉴定便可将每个模块与特定的生物学意义（甚至表型）相关联。加权基因共表达网络作为经典的网络分析方法，也被用于 scRNA-seq 研究领域[27]。

8.2.2 细胞-细胞互作网络　作为生命活动的基本单位，细胞与细胞之间能通过表面受体-配体蛋白识别结合，并以旁分泌和自分泌等方式传递信号，调控受体细胞的分化、凋亡以及有丝分裂等过程，因此细胞-细胞互作（Cell-cell interaction，CCI）网络在整个生命活动中发挥着重要作用。目前有很多软件可用来推测 CCI 网络，如 SingleCellSignalR、iTALK、NicheNet、celltalker、CellChat、CellPhoneDB 等。这些软件都遵循类似的思路：从单细胞基因表达矩阵出发，结合已有的配体-受体信息，然后量化配体-受体相互作用的强度，进而推测细胞间的互作关系。其中，CellChat 主要是以信号通路为单位来计算细胞间交流状态，并考虑了信号辅助因子在细胞通讯间的作用（封三彩图 3c）。而 CellPhoneDB 对配体、受体是否属于复合物、整合蛋白和分泌蛋白等注释信息都进行了精确的标注，并在构建细胞通讯网络时充分考虑了这些复杂的分子互作信息[28]。

8.3 空间转录组

基因表达具有空间特异性，常规单细胞 RNA 测序忽略了细胞在取样组织中的空间信息，而空间转录组（Spatial transcriptome）能够描绘基因表达的空间模式和细胞集群在组织空间的分布（封三彩图 3d），并在监测细胞癌变位点和肿瘤微环境等方面有重要意义[29]。基于这个背景，空间转录组在近些年来逐步成为研究者所偏好的研究方向。但由于空间转录组分辨率不高，因此空间信息还需要与单细胞 RNA 测序数据信息进行整合，才能够更全面地解释生物学问题和充分发挥空间转录组的优势[30]。

9 总结与展望

相较于传统的 RNA 测序，单细胞 RNA 测序的优势在于揭示组织块（或细胞悬液）中细胞间的异质性，解析单个细胞的功能与特性，因此在细胞生物学、免疫学与肿瘤学等方面应用极为广泛，成为现代生命科学发展最为迅速的领域之一。但也应看到，该技术发展的还面临着诸多挑战，尤其是在数据分析方法、流程和策略上。

随着单细胞 RNA 测序技术的飞速发展与繁荣，以单细胞为研究对象的其他组学（如单细胞基因组学、单细胞蛋白组学等）也在走向成熟，单细胞多组学整合或联合分析已成为当前单细胞组学领域一个重要发展方向。因此，未来的单细胞 RNA 测序数据分析方法和工具还应该考虑在可接受数据类型的多样化（如可同时接受 DNA 甲基化或蛋白质丰度等数据）、计算方法的高效性、分析流程的标准化、分析结果的可视化和可解释性等方面进行改进，以适应大数据时代多组学数据分析的需要。相信随着研究者对新技术、新方法、新环节和新思路的不断研究和深入，单细胞 RNA 测序和分析在个体发育、神经科学、临床医学、药物开发等方面发挥着越来越重要

的角色,以促进人类对生命现象和活动更深层次的理解。

参考文献:

[1] MEREU E,LAFZI A,MOUTINHO C,et al. Benchmarking single-cell RNA-sequencing protocols for cell atlas projects[J]. Nature Biotechnology,2020,38(6):747-755.

[2] TANG F C,BARBACIORU C,WANG Y Z,et al. mRNA-seq whole-transcriptome analysis of a single cell[J]. Nature Methods,2009,6(5):377-382.

[3] KULKARNI A,ANDERSON A G,MERULLO D P,et al. Beyond bulk:a review of single cell transcriptomics methodologies and applications[J]. Current Opinion in Biotechnology,2019,58:129-136.

[4] POTTER S S. Single-cell RNA sequencing for the study of development,physiology and disease[J]. Nature Reviews Nephrology,2018,14(8):479-492.

[5] WANG B,ZHANG Y Y,QING T,et al. Comprehensive analysis of metastatic gastric cancer tumour cells using single-cell RNA-seq[J]. Scientific Reports,2021,11(1):1141.

[6] CONDE C D,TEICHMANN S A. Deciphering immunity at high plexity and resolution[J]. Nature Reviews Immunology,2020,20(2):77-78.

[7] ZHONG S J,DING W Y,SUN L,et al. Decoding the development of the human hippocampus[J]. Nature,2020,577(7791):531-536.

[8] LI W V,LI J Y J. An accurate and robust imputation method scImpute for single-cell RNA-seq data[J]. Nature Communications,2018,9(1):997.

[9] HIE B, PETERS J, NYQUIST S K,et al. Computational methods for single-cell RNA sequencing[J]. Annual Review of Biomedical Data Science,2020,3(1):339-364.

[10] AMEZQUITA R A,LUN A T,BECHT E,et al. Orchestrating single-cell analysis with bioconductor[J]. Nature Methods,2020,17(242):137-145.

[11] LUECKEN M D,THEIS F J. Current best practices in single-cell RNA-seq analysis:a tutorial[J]. Molecular Systems Biology,2019,15:e8746.

[12] HUANG M, WANG J S, TORRE E, et al. SAVER: gene expression recovery for single-cell RNA sequencing[J]. Nature Methods,2018,15(7):539-542.

[13] PATRUNO L,MASPERO D,CRAIGHERO F,et al. A review of computational strategies for denoising and imputation of single-cell transcriptomic data[J/OL]. (2020-10-01)[2020-12-20]. Briefings in Bioinformatics. https://doi.org/10.1093/bib/bbaa222.

[14] STUART T,BUTLER A,HOFFMAN P,et al. Comprehensive integration of single-cell data[J]. Cell,2019,177(7):1888-1902.

[15] WU Y,ZHANG K. Tools for the analysis of high-dimensional single-cell RNA sequencing data[J]. Nature Reviews Nephrology,2020,16(7):408-421.

[16] TOWNES F W,HICKS S C,ARYEE M J,et al. Feature selection and dimension reduction for single-cell RNA-seq based on a multinomial model[J]. Genome Biology,2019,20(1):295.

[17] LINDERMAN G C,RACHH M,HOSKINS J G,et al. Fast interpolation-based t-SNE for improved visualization of single-cell RNA-seq data[J]. Nature Methods,2019,16(3):243-245.

[18] BECHT E,Mc INNES L,HEALY J,et al. Dimensionality reduction for visualizing single-cell data using UMAP[J]. Nature Biotechnology,2019,37(1):38-44.

[19] 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.

[20] BROCK G N,PIHUR V,DATTA S,et al. clValid:an R package for cluster validation[J]. Journal of Statistical Software,2008,25(4):22.

[21] ZHAO T Y,LYU S X,LU G L,et al. SC2disease:a manually curated database of single-cell transcriptome for human diseases [J]. Nucleic Acids Research,2021,49(D1):1413-1419.

[22] SHAO X,LIAO J,LU X Y,et al. scCATCH:automatic annotation on cell types of clusters from single-cell RNA sequencing data[J]. iScience,2020,23(3):100882.

[23] HAFEMEISTER C,SATIJA R. Normalization and variance stabilization of single-cell RNA-seq data using regularized negative

binomial regression[J]. Genome Biology,2019,20(1):296.
[24] HOLLAND C H,TANEVSKI J,PERALES-PATÓN J,et al. Robustness and applicability of transcription factor and pathway analysis tools on single-cell RNA-seq data[J]. Genome Biology,2020,21(1):36.
[25] DING J,LIN C,BAR-JOSEPH Z. Cell lineage inference from SNP and scRNA-seq data[J]. Nucleic Acids Research,2019,47(10):e56.
[26] SAGA,GRÜN D. Deciphering cell fate decision by integrated single-cell sequencing analysis[J]. Annual Review of Biomedical Data Science,2020,3(1):1-22.
[27] XUE Z G, HUANG K, CAI C C,et al. Genetic programs in human and mouse early embryos revealed by single-cell RNA sequencing[J]. Nature,2013,500(7464):593-597.
[28] ARMINGOL E, OFFICER A, HARISMENDY O,et al. Deciphering cell-cell interactions and communication from gene expression[J]. Nature Reviews Genetics,2021,22(2):71-88.
[29] MONCADA R,BARKLEY D,WAGNER F,et al. Integrating microarray-based spatial transcriptomics and single-cell RNA-seq reveals tissue architecture in pancreatic ductal adenocarcinomas[J]. Nature Biotechnology,2020,38(3):333-342.
[30] ACHIM K,PETTIT J-B,SARAIVA L R,et al. High-throughput spatial mapping of single-cell RNA-seq data to tissue of origin [J]. Nature Biotechnology,2015,33(5):503-509.

Advances on Approaches Used in Single-Cell RNA Sequencing Data Analysis

LI Bo, DUO Hongrui

(College of Life Sciences, Chongqing Normal University, Chongqing 401331, China)

Abstract: [**Purposes**]To review the current key processes and aspects of single-cell RNA sequencing data analysis, and to introduce representative methods and popular tools required to accomplish different analysis tasks. [**Methods**] The current analysis processes and representative tools for single-cell RNA sequencing data are summarized through a survey of the literature and condensed summaries. [**Findings**] Many analytical procedures and tools for single-cell RNA sequencing data have been developed to uncover biological knowledge from the vast amount of data, in order to reveal complex diseases or phenotypes. [**Conclusions**] Single-cell RNA sequencing plays an extremely important role in life science field, and a good data analysis strategy is a key component to effectively reveal the biological information behind single-cell expression profiling data. There is a wide range of steps and tools available for single-cell RNA sequencing data analysis, and researchers should choose the appropriate and accurate analysis methods and tools, according to the practice scenarios.

Keywords: single cell RNA-seq; data analysis; dimension reduction; clustering; cellular trajectory; co-expression network

(责任编辑 黄 颖)

（接正文130页）

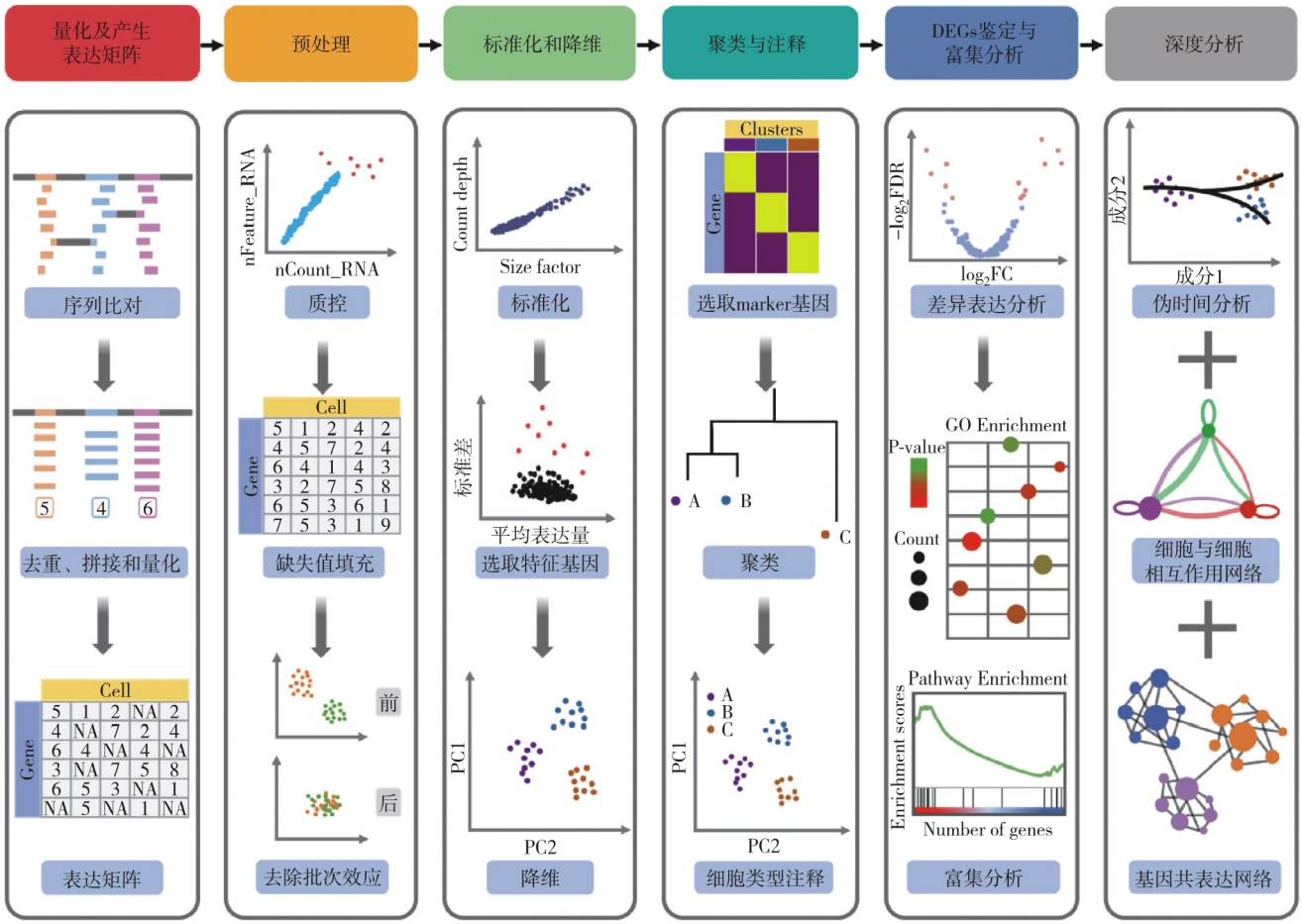
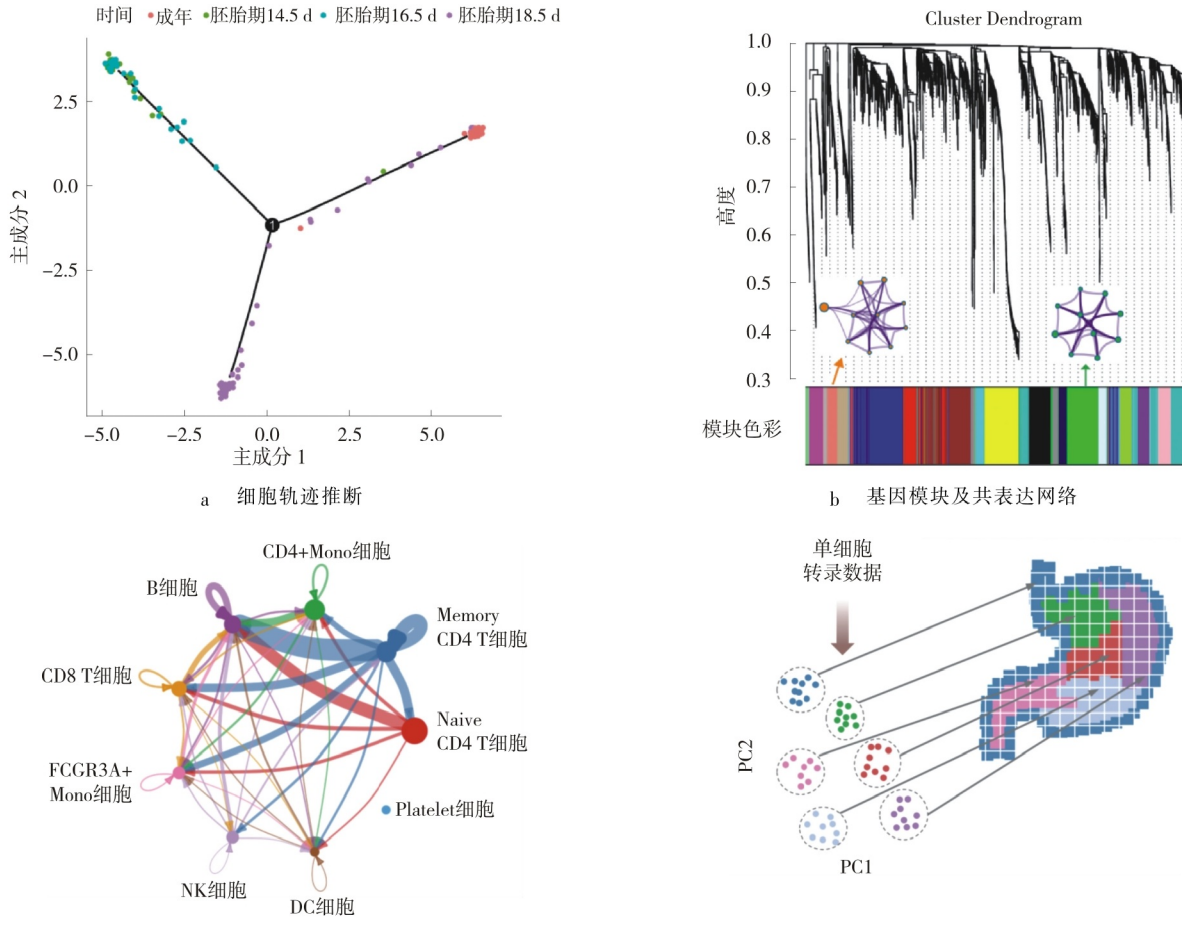


图2　单细胞RNA测序数据分析的流程和环节

Fig. 2　The pipeline and steps of single cell RNA-seq data analysis

（接正文133页）



a　细胞轨迹推断　　b　基因模块及共表达网络

c　细胞通讯网络　　d　空间转录组

图3　单细胞RNA测序数据高级分析示意图

Fig. 3　Schematic diagram of advanced analysis of single-cell RNA sequencing data