DOI:10.12113/202010004

单细胞测序方法研究进展

操利超[1],巴 颖[2],张核子[2*]
(1. 西北大学 生命科学学院,西安 710127;
2. 深圳市核子基因科技有限公司,广东 深圳 518071)

摘 要:近年来,高通量测序技术(Next-generation sequencing,NGS)快速发展,已广泛应用于生命科学各个领域,但传统的混合细胞测序(Bulk cell sequencing)检测的是细胞群体的总平均反应,无法反应每个细胞的真实情况,这会影响研究者对细胞功能认知的准确性。单细胞测序技术(Single cell sequencing,sc-Seq)的出现,从一定程度上解决了传统测序固有的缺陷。单细胞测序是针对单个细胞的RNA或DNA进行测序,能够准确测出单个细胞的基因结构和表达状态,从而分析相同表型细胞的异质性。本文首先介绍单细胞测序的原理、测序类型和测序平台,有助于理解单细胞测序和在进行科研项目时设计合适的项目方案。进一步介绍单细胞转录组测序的分析流程和各种常用的分析工具或软件,并重点阐述单细胞转录组测序分析中的细胞聚类和拟时序分析的原理和研究进展,为进行单细胞转录组测序数据分析提供参考。最后,本文简述了单细胞测序研究热度、单细胞测序的应用、挑战和展望等,有助于更全面地认识单细胞测序。

关键词:单细胞测序;生物信息分析;高通量测序
中图分类号:Q-3 **文献标志码**:A **文章编号**:1672-5565(2022)01-091-09

Recent progress in single cell sequencing

CAO Lichao[1], BA Ying[2], ZHANG Hezi[2*]
(1. *College of Life Sciences*, *Northwest University*, *Xi'an* 710127, *China*;
2. *Shenzhen Nuclear Gene Technology Co.*, *Ltd.*, *Shenzhen* 518071, *Guangdong*, *China*)

Abstract: In recent years, as the next-generation sequencing technology developed rapidly, it has been widely used in various fields of life science. However, The traditional bulk cell sequencing methods detect the total average reactions of cell groups and are unable to reflect the actual situation of each cell, which may affect the accuracy of understanding of the cell function. The emergence of single cell sequencing technology has solved the inherent defects of traditional sequencing to a certain extent. Single cell sequencing can accurately detect the gene structure and expression status of a single cell in RNA or DNA level, which is important for analyzing the heterogeneity of different cells with the same phenotype. In this study, the principle, sequencing method, and platform of single cell sequencing were briefly introduced, which is helpful to understand single cell sequencing and design the appropriate project scheme in scientific research projects. Then, the single cell transcriptome sequencing analysis pipeline and various commonly used analysis tools or software were summarized, and the principle and the research progress of cell clustering and trajectory analysis in single cell transcriptome sequencing were elaborated, which can provide reference for data analysis of single cell transcriptome sequencing. Finally, the research trends of single cell sequencing, as well as the application, challenge, and prospect of single cell sequencing were presented, which is helpful to understand single cell sequencing more comprehensively.

Keywords: Single cell sequencing; Bioinformatics analysis; High-throughput sequencing

收稿日期:2020-10-16; **修回日期**:2021-03-25.
基金项目:深圳市可持续发展专项(深科技创新〔2020〕180号,专2019N002)
作者简介:操利超,男,在读博士,研究方向:生物信息学. Email:13428941050@163.cm.
*通信作者:张核子,男,工程师,研究方向:肿瘤早筛技术和ctDNA精准医疗应用的研究. Email: zhz358@126.com.

1 单细胞测序原理

单细胞测序从单个细胞水平上对 DNA 或 RNA 进行扩增和测序,主要包括单细胞分选、核酸提取和文库构建、高通量测序和数据分析等。

单细胞分选是单细胞测序的第一步,如何低成本的获取大批量高质量完整的单细胞对于单细胞测序非常重要。Gross 等[1]详细介绍了 5 种单细胞分选方法,即有限稀释法(Limiting Dilution)、显微操作法(Micromanipulator)、荧光激活细胞分选(Fluorescence activated cell sorting, FACS)、激光显微切割(Laser capture microdissection, LCM)和微流控分选(Microfluidics),这 5 种单细胞分选方法各有利弊,其中,微流控分选方法由于通量高、成本低等原因而广泛应用于各种商业化单细胞测序平台,例如 10X Genomics 公司的 Chromium 系统就是利用微流控技术进行单细胞分选,通过控制流体流动来实现在微尺度上对目的细胞进行分离。

分离得到单细胞后经过细胞溶解获取 DNA 或 RNA。在传统的高通量测序流程中,提取的 DNA 或 RNA 需要进一步纯化后才能应用于扩增,而在单细胞测序中,为了避免 DNA 或 RNA 在纯化中丢失,目前大部分流程中去掉这一步骤。后述的文库构建前处理和文库构建流程在不同的测序平台和方法有所不同,但基本上都是基于单分子标签(Unique Molecular Identifier, UMI)和细胞标签(Cell label, CL)的技术,最终形成具有特异标签标记的单细胞文库,文库构建完成后即进行高通量测序和数据分析。在进行数据分析时,单分子标签为每个细胞甚至每个基因或转录本提供特异的识别码,使得精确分析每个细胞的功能和特性成为可能。

2 单细胞测序分类

单细胞测序主要包括单细胞基因组测序(Single cell DNA sequencing, scDNA-seq)、单细胞转录组测序(Single cell RNA sequencing, scRNA-seq)和单细胞表观组测序(Single cell epigenome sequencing),这三种测序类型可以从不同角度揭示细胞各个阶段的功能和特性。

2.1 单细胞基因组测序

单细胞基因组测序可用来分析单细胞水平的点突变和拷贝数变异,用于揭示细胞群体差异、细胞进化关系等,可最真实的获得单克隆癌细胞的具体突变来源及精准的突变频率,以及区分癌症发生、发展、演化过程中的主动与被动突变等。其主要难点是获得高覆盖度高保真性的全基因组扩增产物,因此,在单细胞全基因组技术发展过程中,全基因组扩增技术(Whole-genome amplification, WGA)也经历了几次重大的变革。WGA 主要有三种方式,包括简并寡核苷酸引物 PCR(Degenerate Oligonucleotide - Primed Polymerase Chain Reaction, DOP-PCR)、多位点置换扩增(Multiple Displacement Amplification, MDA)、MDA 与 PCR 相结合的方法(Multiple Annealing and Looping-Based Amplification Cycles, MALBAC)等三种。DOP-PCR 方法的原理是在引物的 3'端含有 6bp 的随机序列,可以与基因组 DNA 随机结合,从而实现对全基因组高质量的扩增[2];MDA 方法引入了 phi29 DNA 聚合酶,使其与随机的六聚体发生反应,phi29 DNA 聚合酶具有很强的链置换特性,在等温条件下,能够扩增出的 50-100 kb 的 DNA 片段[3]。相对于 DOP-PCR,MDA 的覆盖度和均匀性有了明显提升,但是这种方法并不是无偏倚性、无误差的。鉴于 MDA 方法扩增得到的基因组覆盖度不均匀,Zong 等[4]开发了一种新的基因组 DNA 扩增方法,即 MALBAC。该方法将 MDA 与常规的 PCR 结合起来,利用部分碱基简并化的杂交引物与模板退火结合,在链置换酶的作用下进行扩增,扩增中间产物的 3'端带有特异性引物标记,经过一轮扩增后,互补的标记位于 5'端,两端的标记互补成环,成环后的扩增子通过常规 PCR 进行扩增。

2.2 单细胞转录组测序

单细胞转录组测序可对单细胞中 mRNA 进行基因表达定量、功能富集、代谢通路等分析,可以解决传统转录组测序技术在早期胚胎发育、干细胞、癌症、免疫等研究领域中存在的样品量极低或细胞异质性的问题,是在单细胞水平研究基因表达强有力的工具。单细胞转录组测序需要对获取的 RNA 进行逆转录,收集 cRNA 产物并扩增测序。各种成熟的单细胞转录组测序方法的标准操作规程(Standard Operating Procedure, SOP)在 2013 年前后均有文章发表(见表 1)。单细胞转录组测序方法很多,在实际的项目中,需要根据实际情况综合评估,选择合适的测序方法。Christoph 等[5]利用 6 种不同的单细胞转录组测序方法对 583 个小鼠胚胎干细胞进行分析。相对于 Smart-seq 和 Smart-seq2,其他检测方法 CEL-seq2、Drop-seq、MARS-seq 和 SCRB-seq 采用了单分子标签技术,这可以提高测序的准确性,而 Smart-seq 和 Smart-seq2 在建库时能获取全长转录本,这有利于检测到更多的基因,经过标准化的分析流程处理后,文章对这 6 种测序方法进行了

全面的对比,包括测序质量、测序深度、检测到的基因数、有效的细胞数、准确性和灵敏度等。通过综合考虑分析,给出的结论是当对大量细胞进行转录组定量分析时,采用 Drop-seq 方法性价比最高,而对于少量细胞,MARS-seq 和 SCRB-seq 和 Smart-seq2 方法更有效。

表 1　不同单细胞测序方法学统计表

Table 1　Different single cell sequencing methodologies

测序方法	年份	期刊	文献
Smart-seq	2013	Nature Methods	[6]
Smart-seq2	2014	Nature Protocols	[7]
CEL-seq2	2016	Genome Biology	[8]
Drop-seq	2015	Cell	[9]
MARS-seq	2014	Science	[10]
MARS-seq2.0	2019	Nature Protocols	[11]
mcSCRB-seq	2018	Nature Communication	[12]

2.3　单细胞表观组测序

单细胞表观组测序可从单细胞水平获得全基因组范围内的甲基化水平数据,对于表观遗传学的时空特异性研究具有重要意义。单细胞表观基因组测序主要是结合单细胞基因组测序和传统的表观组高通量测序方法(如 BS-seq 和 ChIP-seq 等)。如 Guo 等[13]利用单细胞全基因组简化胞甲基化测序(Single cell reduced representation bisulfite sequencing, scRRBS)的方法对小鼠胚胎干细胞进行测序,可在单个细胞中检测到高达 150 万个 CpG 位点的甲基化状态,同时,该技术可以检测单倍体精子细胞中单个 CpG 位点的甲基化状态,可探索单个细胞 DNA 甲基化的动态变化。RRBS 方法只能检测到全基因组上 5%左右的 CpG 位点的甲基化信息,且这些位点主要集中在 CpG 相对密集的区域,比如 CpG 岛、启动子等,但是在胚胎早期发育过程中,全基因组范围内的甲基化状态是变化的,为了检测到更全面的甲基化信息,Zhu 等[14]利用重亚硫酸盐处理后接头标记技术(Post-bisulfite adaptor tagging, PBAT)对人植入前胚胎发育的各个阶段进行了深度测序,平均每个细胞能捕获全基因组上的 20% 的 CpG 位点,分析发现数以万计的基因组位点表现出从头开始的 DNA 甲基化(De novo DNA methylation)。这一发现表明,全基因 DNA 甲基化重编程过程在植入前胚胎发育过程中甲基化和去甲基化是处于动态平衡的。同时,通过 DNA 甲基化分析,可以追溯早期胚细胞的遗传谱系,为破译早期人类胚胎中 DNA 甲基化重新编程的秘密铺平了道路。基于染色质免疫共沉淀技术(Chromatin Immunoprecipitation, ChIP)的测序方法,Rotem 等[15]结合微流控和 DNA 特征码技术进行了数千个单细胞测序,从单细胞水平收集染色质数据,进一步分析了表观遗传各方面的异质性,而这在转录水平是无法获取的。

2.4　单细胞多组学测序

此外,单细胞多组学研究(Single cell multi-omics sequencing)也逐步在科学研究中得到应用。如 Angermueller 等[16]对 61 个小鼠胚胎干细胞同时进行单细胞转录组和表观组测序,分析了 DNA 甲基化异质性和转录水平异质性的关联。Macaulay 等[17]利用基因组和转录组平行测序(Genome and Transcriptome Sequencing, G&T-seq)方法对来源于小鼠和人类的 220 多个单细胞进行测序,通过关联分析发现了以前无法单独从 DNA 或 RNA 测序推断出的细胞特性。

3　单细胞测序技术平台

3.1　单细胞分离和标记平台

单细胞测序技术一般是结合单细胞分离技术和特异性分子标签技术将单个细胞标记上特定的标签,然后进行高通量测序的技术。目前,国内外大规模单细胞技术使用的单细胞分离和标记平台主要有基于 10X Next GEM 技术的 Chromium™ 系统、BD Rhapsody™单细胞分析系统、Illumina © Bio-Rad © 单细胞测序解决方案、ICELL8 单细胞系统、C1™单细胞全自动制备系统等。各个单细胞分离和标记平台的原理、特点及应用(见表 2)。在实际的项目中,需要结合项目的特点,考虑多种因素,选择一种最能满足实验且性价比高的平台。

3.2　单细胞高通量测序平台

单细胞测序技术使用的高通量测序平台有很多,如 illumina 系列、BGISEQ 系列、Roche 454、ABI solid、Ion Proton 等。目前,应用于单细胞测序的高通量测序平台主要是 illumina 系列,但其他测序平台也逐渐被证明可应用单细胞测序。Chen 等[18]在 illumina Hiseq2000 和 Ion Proton 平台上分别利用低覆盖度的单细胞全基因组测序的方法对相同的样本进行测序,并从可重复性、测序错误率、一致性、灵敏度和特异性等方面进行比较分析,结果表明,两种平台各有优缺点。

最近,有研究表明,BGISEQ 测序平台也可应用于单细胞测序。Natarajan 等[19]选取了 468 个单细胞对应1,297个 cDNA 样本,分别采用 SMARTer 和 Smart-seq2 的方法在 BGISEQ-500 和 Illumina HiSeq

平台上进行单细胞测序,本文首次从灵敏度和准确性上比较在两种平台在单细胞转录组测序上的应用,结果表明,BGISEQ-500 可完成高质量低成本的单细胞转录组测序。

表 2 不同单细胞分离和标记平台的原理、特点及应用

Table 2 Principle, characteristics, and application of different single cell separation and labeling platforms

单细胞分离和标记平台	原理	特点	应用
Chromium™ 系统	构建一个独特的试剂输送系统,将细胞或细胞核进行分离;并行地制备测序文库,使得每一个液滴产生的所有片段都标记上一个共同的分子标签。	可以在几分钟内生成大于 10 万个包含分子标签和样本的液滴。	用于基因组的组装,获得大片段的遗传信息;可利用油包水的微反应体系,通过分子标签区分群体中的不同细胞,实现数千甚至数万个单细胞群体分析。
BD Rhapsody™ 单细胞分析系统	基于微流控芯片技术,在磁性的分子条形码标记的微球上实现单细胞捕获,并为单细胞每个转录本标记上特异的分子标签,实现单细胞水平基因表达谱的绝对定量。	可实现单次实验制备 100-10,000 个单细胞文库,并可根据需求定制引物,将检测范围集中在目标基因,大幅降低后续测序成本,提高细胞捕获效率。	应用于细胞及细胞亚群表达特征聚类、标志物筛选等方面的分析。
Illumina © Bio-Rad © 单细胞测序解决方案	采用 Bio-Rad 的液滴分离技术,对单细胞进行隔离和标记分子标签,然后在 illumina 系列测序仪上进行测序,并可进行一站式数据分析服务。	该方案一次性可处理 8 个样本,每个样本可得到 500-10,000 个细胞。	应用于研究组织功能、病情进展和治疗反应等方面单个细胞的协同作用。
ICELL8 单细胞系统	有精密的 8 通道喷嘴,可以从细胞悬液中捕获待测细胞,并以纳升级别喷注到多孔纳升级芯片,使各种大小的细胞都待在芯片上的单个孔内。	芯片含有高达 5184 个纳米孔,15 min 即可抓到多至 1800 个单细胞。	适用于单细胞全长转录组研究。
C1™ 单细胞全自动制备系统	采用微流体技术,在同一张芯片上完成细胞捕获、裂解、逆转录和预扩增全过程,并自动回收单细胞基因产物用于单细胞表达谱和单细胞高通量测序实验。	可同时捕获 96 个单细胞,几个小时即可完成数以百计的单细胞中数百个基因的表达信息。	可获得全转录组信息,应用于生殖发育、干细胞分化、验证生物标志物和 RNA 干扰沉默基因表达等领域。

4 单细胞测序数据分析

单细胞测序分析中,对于不同的测序类型和研究目的,会采用不同的分析流程。对于单细胞基因组测序和单细胞表观组测序而言,数据分析流程与传统的高通量测序数据分析方法类似[13, 20]。在单细胞转录组测序数据分析中,常见的分析内容包括基因表达(gene expression)、可变剪切(alternative splicing)、T 细胞受体谱(T cell receptor profiling)或 B 细胞受体谱(B cell receptor profiling)、细胞聚类(Cells clustering)、拟时序分析(Trajectory analysis)等,常见的单细胞转录组测序数据分析流程(见图 1),一些主流的分析软件总结(见表 3)。

在单细胞转录组测序数据分析中,细胞聚类和拟时序分析是单细胞测序特有的分析内容,下面将重点阐述。细胞聚类是单细胞测序数据进一步分析的基础。细胞聚类的基本原理是根据细胞的特征(比如基因表达等)将大量的细胞,通过聚类算法将细胞分成不同的亚群的过程。Dai H 等[21]通过构建单细胞的 network,将 network 降维到二维的矩阵,并代替原有的不稳定的基因表达矩阵,用以细胞聚类和拟时序分析,同时,该种方法可以发现一些有意义的在传统差异基因表达分析中漏掉的认为不重要的基因。Elham 等[22]利用 Drop-seq 方法对45 000 个免疫细胞进行单细胞测序,通过自己开发的分析流程,通过贝叶斯聚类和标准化方法,提出差异表达分析不能完全体现亚群间差异,得到非常稳定的协方差模式(Covariance pattern)不同的亚群。Xie 等[23]用有监督的机器学习方法大大提高了细胞聚类的效率,在单核 2.3 GHz 的个人电脑上,训练 19 万个细胞的训练集只需要 5 分钟,而做 1 万个细胞的分类不到1 分钟,而常用的聚类软件 Seurat,在对 1W 个细胞进行聚类时,时长需要以小时来计算,内存用量约

15G-20 G,分析结果表明在准确性上还有待提升,这主要跟提供的数据训练集有关。同样,Feiyang Ma等[24]利用神经网络模型对小鼠细胞和人源免疫细胞数据集进行训练,利用得到的训练结果来预测小鼠白细胞、人源外周血单核细胞和人源 T 细胞亚型,分析结果快速而准确,表明该方法可以用来优化目前的单细胞测序分析流程。随着单细胞测序样本量的增大,传统的聚类分析方法需要耗费大量的资源,机器学习的引入会是一个很有潜力的解决方向,但需要进一步提高聚类的准确性和扩展应用场景。

拟时序分析是基于大量单细胞的基因组学数据,通过生物信息学算法来推断这些细胞的发育时序。拟时序分析对于研究某一特定细胞类型的转化,如 $CD8^+$ T 细胞的激活和耗竭、M1/M2型巨噬细胞极化等,往往具有一定的生物学意义。进行拟时序分析的软件有很多,Saelens 等[25]针对 110 个真实数据集和 229 个模拟数据集,利用 45 种拟时序分析工具进行分析和对比,分析结果表明工具和方法的选择,主要取决于数据集维度和拓扑结构。

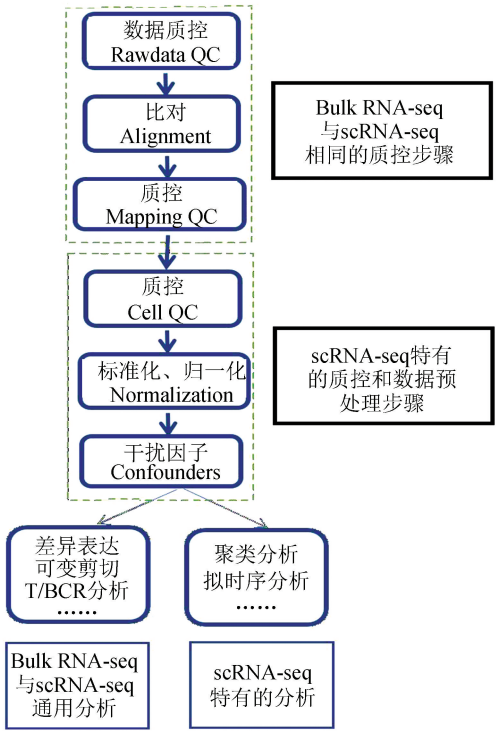

图 1　单细胞转录组测序分析流程

Fig.1　Analysis workflow of single cell transcriptome sequencing

表 3　单细胞转录组测序分析软件

Table 3　Single cell transcriptome sequencing analysis toolkits

软件名称	版本	语言	核心算法	特点	参考文献
scater	1.14.4	R package/ bioconductor	PCA t-SNE	专注于数据质量控制和数据可视化	[26]
monocle	2.14.0	R package/ bioconductor	密度聚类算法(densityPeak) Louvain 算法	无需对照实验即可准确完成基因或转录本的相对定量	[27]
Seurat	3.1.1	R package	PCA t-SNE	适用于多种测序方法,功能全面,版本更新较快。	[28] [29]
scran	1.14.5	R package/ bioconductor	PCA SCE 算法	适用于多种测序方法,功能全面	[30]
M3Drop	1.12.0	R package/ bioconductor	Michaelis-Menten 模型	数据兼容性好,适用于各种转录组测序数据	[31]
Scanpy	1.4.4	Python package	t-SNE UMAP PCA	对内存要求低,功能全面,支持 UMAP 降维	[32]
Wishbone	0.4.2	Python package	PCA t-SNE	使用分叉树(Bifurcating branches)来识别单细胞发育轨迹	[33]

5　单细胞测序研究热度分析

单细胞测序是为了解决传统高通量测序的局限性,比如无法获取单个细胞特有的遗传信息,可能会丢失低丰度的信息,无法检测难以培养富集的微生物等。因此,单细胞测序具有传统高通量测序无法比拟的优势。2011 年,《自然方法》杂志(Nature Methods)将单细胞测序列为年度值得期待的技术之一,2013 年,《科学》杂志(Science)将单细胞测序列为年度最值得关注的六大领域榜首,2018 年,《科学》杂志(Science)的年度十大科学突破之一就是单

细胞测序的重要一环“单个细胞分离并逐个测序(Development cell by cell)”的研究进展。由此可见,单细胞测序作为一种技术手段被广大科研工作者寄予厚望。

以"single cell sequencing"[All Fields]为关键词去 NCBI-pubmed 数据库搜索,统计搜索到的文章数(见图 2),可以发现从 2011 年到 2020 年间,单细胞测序发表的文章数基本上呈指数增长(截止至 2020 年 10 月 16 日)。

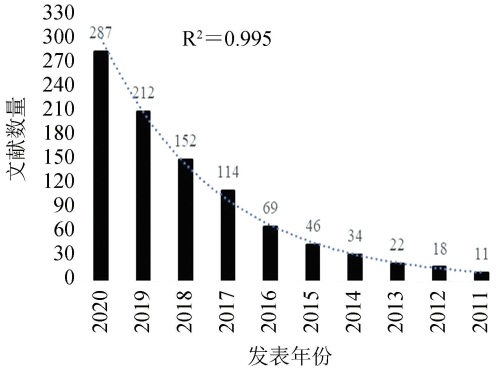

图 2 以“单细胞测序技术”为关键词在 NCBI-pubmed 库中的搜索情况(截止至 2020 年 10 月 16 日)

Fig.2 Statistics of published literature in NCBI-pubmed database using “single cell sequencing” as keyword (by October 16, 2020)

在科学网基金页面(http://fund.sciencenet.cn/),以“单细胞测序”为项目名称关键词搜索,查询结果有 1 218 项,累计金额为 75 772 万元,项目涉及学科分类广,主要集中在生命科学和医学科学领域,分布图(见图 3)(查询结果截止至 2020 年 10 月 16 日)。

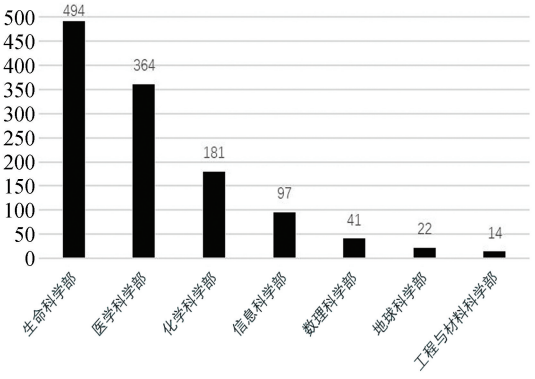

图 3 以“单细胞测序”为关键词在科学网基金页面查询得到的项目分布统计图(截止至 2020 年 10 月 16 日)

Fig.3 Statistics distribution of projects on the Science Net using “single cell sequencing” as keyword (by October 16, 2020)

在美国国立卫生研究院(NIH)官网上(https://projectreporter.nih.gov/),以“single cell sequencing”为关键词搜索,查询结果按照经费申请机构进行统计,我们通过经费申请机构可以简单了解其研究方向或研究领域,最终的统计结果(见图 4)(查询时间跨度为 2018 年至 2019 年)。从图中可以看出,在单细胞测序研究中,肿瘤和常见疾病相关研究机构申请到的项目或研究经费较多,由此可以粗略推断单细胞测序的热点研究领域。

由此可见,单细胞测序在近年来一直是科研界的研究热点,被广泛应用于各个领域。

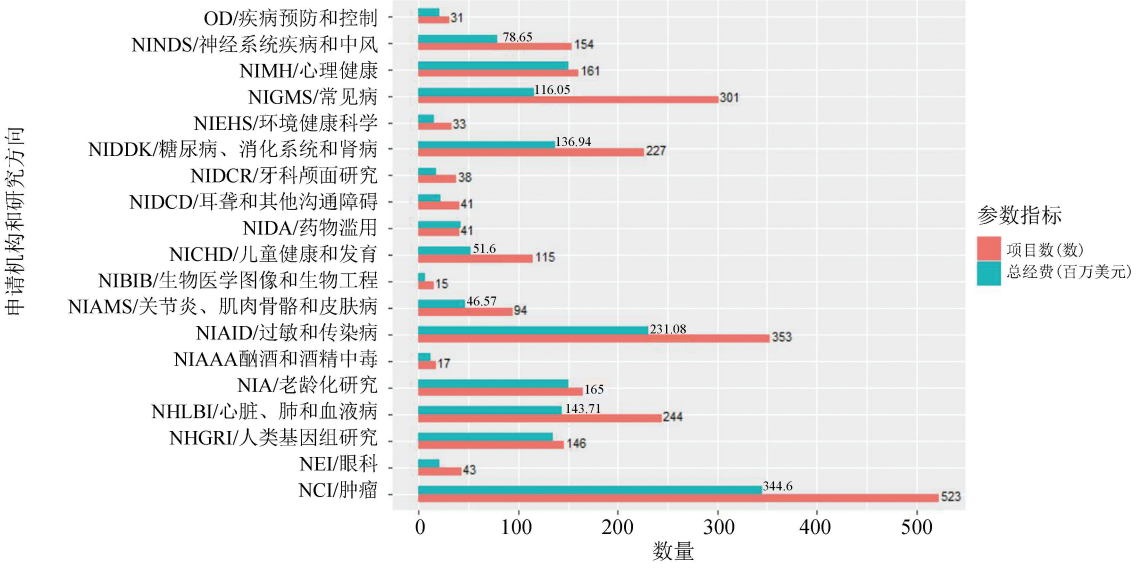

图 4 以“single cell sequencing ”为关键词在美国 NIH 页面查询得到的项目分布统计图(2018-2019 年)

Fig.4 Statistics distribution of projects on the NIH website using “single cell sequencing” as keyword (2018 - 2019)

6　单细胞测序的应用

单细胞测序在不同研究领域得到应用,比如干细胞和发育生物学、肿瘤、免疫学等领域。

发育生物学作为生命基本过程的基础学科,传统的高通量测序无法详细和系统的研究所有器官各种细胞的分子状态和分化途径。而单细胞测序可以分离单个细胞,对其中的遗传信息进行分析,构建高分辨率的不同组织发育阶段的基因表达谱。Takahashi K 等[34]研究发现,已分化的成纤维细胞可以在四种反转录因子 OCT3/4, SOX2, c-Myc 和 KLF4 的作用下重分化成多能干细胞,这一发现有望在多种疾病的机制研究和治疗中发挥重要的作用。Brunskill EW 等[35]利用单细胞测序研究了肾脏器官发育过程中的基因表达谱,发现在不同的发育阶段,结构相同的细胞群却有着不同的来源。

在肿瘤研究中,了解肿瘤细胞内异质性对肿瘤的发生发展、其耐药性的影响和重新定义细胞亚型等非常重要。传统的高通量测序方法只能针对大量细胞群体进行研究,只是细胞群体的总平均反应,无法反应每个细胞的真实情况。而单细胞测序能够揭示单个细胞的基因结构和基因表达状态,反应细胞间的异质性。Li H 等[36]利用单细胞测序技术对 11 个结直肠癌细胞和对应的正常粘膜细胞进行测序,利用参考成分分析(reference component analysis,RCA)算法进行聚类分析鉴别出两种不同的纤维母细胞瘤亚型,同时,在其中发现了一些与上皮间充质干细胞分化相关的基因表达上调,为肿瘤细胞异质性表征提供了一个很好的方法。Tirosh 等[37]从 19 个黑色素瘤患者中分离了 4 645 个细胞进行单细胞测序,研究发现,非恶性肿瘤细胞会根据细胞类型如 T 细胞、B 细胞、巨噬细胞、血管内皮细胞等聚类,而不同病人来源的恶性肿瘤细胞会被分成不同的聚类,表明肿瘤细胞异质性的存在。Baslan 等[38]综述了单细胞测序技术在研究肿瘤异质性和肿瘤细胞演化中的作用,并以肺癌为例说明了单细胞测序的应用场景。

在免疫学领域,由于传统的免疫学分析方法取样来自于大量细胞,导致分析结果低估了单个免疫细胞的多样性,单细胞测序可以更加精确地检测单个免疫细胞的遗传物质,从而理解机体复杂的免疫机制。Villani 等[39]通过对来源于健康献血者的 2 400 个细胞进行单细胞测序,发现了 6 个 DC 细胞和 4 个单核细胞亚型,该研究是对 pDCs 分类的重新鉴定和修订,这一研究成果将使得对疾病和健康的免疫检测更加精确。

7　挑战和展望

相对于传统的高通量测序,单细胞测序检测的技术难点不在于测序本身,而在于单个细胞中核酸物质太少,以至于难以直接检测。因此,单细胞测序的关键技术之一时需要对单细胞中极微量的 DNA 进行高质量、大幅度的扩增,目前已经有一些相对成熟的扩增方法,但是这些扩增技术都存在一些局限性,比如扩增区域不全而导致未扩增的区域无法被测序,而且扩增会存在偏向性(Bias),导致扩增不均一,这些问题都会给后续的生物信息分析带来很大的挑战。同时,FFPE 样本的实验处理和数据分析优化很重要[38],因为临床上有很多样本是 FFPE 样本,而这些样本很容易降解,从而使其进行单细胞测序存在一些挑战。除此之外,对于大量样本量的单细胞测序来讲,会产生海量的测序数据,这对于数据的存储、分析带来了挑战。

近年来,云平台在高通量测序领域得到广泛的应用,而数学、物理学等学科为生物信息学的发展提供了基础算法,这使得生命科学大数据的计算、存储和应用成本大大降低。同时,单细胞测序成本的降低,使其广泛应用于生命科学各个领域,这也积累了海量的单细胞测序数据,而缺少高效精准的单细胞测序数据分析方法或工具阻碍了单细胞测序的进一步发展。目前,结合人工智能和生物信息算法,已经在单细胞测序数据分析领域取得了突破。如 Xiong 等[40]利用人工智能深度学习算法,结合变分自编码器和高斯混合模型,提取单细胞 ATAC-seq(Assay for Transposase-Accessible Chromatin using sequencing)数据的隐层特征,将问题从复杂稀疏的高纬度的染色质开放图谱空间投射到了简单抽象的低纬度特征空间,这种处理不但可以发现与解析细胞特异性的染色质图谱模式,还通过相似细胞信息共享,填补了技术限制导致的缺失值,从而巧妙地解决了单细胞 ATAC-seq 数据中高维度、稀疏性、二值化等问题。而 Cheng 等[41]开发的基于机器学习的分析管道潜在细胞分析法(Latent Cellular Analysis,LCA),通过将隐元胞状态的余弦相似度度量与基于图的聚类算法相结合,为种群数量推断、降维、特征选择和技术变异控制提供启发式解决方案,且无需显式基因过滤。Xie 等[42]对正常肺中 6 种间充质类型细胞和纤维化肺中 7 种间充质类型细胞进行单细胞 RNA 测序,并通过机器学习的方法,确定了它们的分化轨迹,从而为了解成纤维细胞的结构和成纤维细胞在纤维化疾病中的作用提供了新的资源。Duan 等[43]

基于人工智能主题模型(Topic Modeling)的计算框架 MUSIC(Model-based Understanding of Single cell CRISPR screening),用以有效地对单细胞 CRISPR 筛选数据进行分析,用以揭示数据本身所体现的生物学意义。He 等[44]基于半监督学习的单细胞测序数据填补方法 DISC (Deep learning Imputation model with semi-supervised learning for Single Cell tranomes),该方法可以利用少量的表达出来的基因信息及数据庞大的缺失表达基因之间的表达结构,可以实现基因表达分布修复、差异基因预测、基因表达相关性预测和稀有细胞类型预测等,为单细胞测序数据分析提供了重要的技术支撑。

8 结 论

相较传统的混合细胞测序,单细胞测序技术能够在复杂的群体组合(如肿瘤)中对不同细胞类型进行区分,为了解各种发育、生理和疾病系统背后的过程提供了一个强大的方法,这也使得单细胞测序成为科研界的一个研究热点。大量关注和资金的流入使得单细胞测序技术飞速发展,各种测序平台和数据分析算法层出不穷。然而由于单细胞测序数据量大的特点,使得数据分析的难度高,精度差成为阻碍该技术发展的主要因素。但随着人工智能的发展,越来越多的研究者将人工智能技术用于单细胞测序数据的分析,并取得了不错的成果。相信在未来,人工智能作为大数据分析的一个有效的解决方法而被广泛应用于单细胞测序数据分析领域。

参考文献(References)

[1] GROSS A, SCHOENDUBE J, ZIMMERMANN S, et al. Technologies for single-cell Isolation [J]. International Journal of Molecular Sciences, 2015, 16(8): 16897-16919. DOI: 10.3390/ijms160816897.

[2] TELENIUS H, CARTER N P, BEBB C E, et al. Degenerate oligonucleotide-primed PCR: General amplification of target DNA by a single degenerate primer [J]. Genomics, 1992, 133: 718 - 725. DOI: 10. 1016/0888 - 7543(92)90147-k.

[3] DEAN F B, NELSON J R, GIESLER T L, et al. Rapid amplification of plasmid and phage DNA using Phi 29 DNA polymerase and multiply-primed rolling circle amplification [J]. Genome Research, 2001, 11(6): 1095-1099. DOI: 10.1101/gr.180501.

[4] ZONG C, LU S, CHAPMAN A R, et al. Genome-wide detection of single-nucleotide and copy-number variations of a single human cell[J]. Science, 2012, 338(6114): 1622-1626. DOI:10.1126/science.1229164.

[5] ZIEGENHAIN C, VIETH B, PAREKH S, et al. Comparative analysis of single-cell RNA sequencing methods [J]. Molecular Cell, 2017, 65(4): 631-643 e4. DOI: 10.1016/j.molcel.2017.01.023.

[6] PICELLI S, BJORKLUND A K, FARIDANI O R, et al. Smart-seq2 for sensitive full-length transcriptome profiling in single cells [J]. Nature Methods, 2013, 10(11): 1096 - 1098. DOI: 10.1038/nmeth.2639.

[7] PICELLI S, FARIDANI O R, BJORKLUND A K, et al. Full-length RNA-seq from single cells using Smart-seq2[J]. Nature Protocols, 2014, 9(1): 171 - 181. DOI: 10.1038/nprot.2014.006.

[8] HASHIMSHONY T, SENDEROVICH N, AVITAL G, et al. CEL-Seq2: Sensitive highly-multiplexed single-cell RNA-Seq[J]. Genome Biology, 2016, 17: 77. DOI: 10. 1186/s13059-016-0938-8.

[9] MACOSKO E Z, BASU A, SATIJA R, et al. Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets [J]. Cell, 2015, 161(5): 1202 - 1214. DOI: 10.1016/j.cell.2015.05.002.

[10] JAITIN D. A, KENIGSBERG E, KEREN-SHAUL H, et al. Massively parallel single-cell RNA-seq for marker-free decomposition of tissues into cell types[J]. Science, 2014, 343(6172):776-769. DOI: 10.1126/science.1247651.

[11] KEREN-SHAUL H, KENIGSBERG E, JAITIN D A, et al. MARS-seq2.0: An experimental and analytical pipeline for indexed sorting combined with single-cell RNA sequencing [J]. Nature Protocols, 2019, 14(6): 1841 - 1862. DOI: 10.1038/s41596-019-0164-4.

[12] BAGNOLI J W, ZIEGENHAIN C, JANJIC A, et al. Sensitive and powerful single-cell RNA sequencing using mcSCRB-seq[J]. Nature Communications, 2018, 9(1):2937. DOI: 10.1038/s41467-018-05347-6.

[13] GUO H, ZHU P, WU X, et al. Single-cell methylome landscapes of mouse embryonic stem cells and early embryos analyzed using reduced representation bisulfite sequencing[J]. Genome Research, 2013, 23(12): 2126 - 2135. DOI: 10.1101/gr.161679.113.

[14] ZHU P, GUO H, REN Y, et al. Single-cell DNA methylome sequencing of human preimplantation embryos [J]. Nature Genetics, 2018, 50(1): 12 - 19. DOI: 10.1038/s41588-017-0007-6.

[15] ROTEM A, RAM O, SHORESH N, et al. Single-cell ChIP-seq reveals cell subpopulations defined by chromatin state[J]. Nature Biotechnology, 2015, 33(11): 1165 - 1172. DOI: 10.1038/nbt.3383.

[16] ANGERMUELLER C, CLARK S J, LEE H J, et al. Parallel single-cell sequencing links transcriptional and epigenetic heterogeneity [J]. Nature Methods, 2016, 13(3): 229-232. DOI: 10.1038/nmeth.3728.

[17] MACAULAY I C, HAERTY W, KUMAR P, et al. G&T-seq: parallel sequencing of single-cell genomes and transcriptomes[J]. Nature Methods, 2015, 12(6): 519-522. DOI: 10.1038/nmeth.3370.

[18] CHEN D, ZHEN H, QIU Y, et al. Comparison of single cell sequencing data between two whole genome amplification methods on two sequencing platforms [J]. Science Reports, 2018, 8(1): 4963. DOI: 10.1038/s41598-018-23325-2.

[19] NATARAJAN K N, MIAO Z, JIANG M, et al. Comparative analysis of sequencing technologies for single-cell transcriptomics[J]. Genome Biology, 2019, 20(1): 70. DOI: 10.1186/s13059-019-1676-5.

[20] CHEN C, XING D, TAN L, et al. Single-cell whole-genome analyses by Linear Amplification via Transposon Insertion (LIANTI) [J]. Science, 2017, 356(6334): 189-194. DOI: 10.1126/science.aak9787.

[21] DAI H, LI L, ZENG T, et al. Specific-cell network constructed by single-cell RNA sequencing data[J]. Nucleic Acids Research, 2019, 47(11): e62. DOI: 10.1093/nar/gkz172.

[22] 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. DOI: 10.1016/j.cell.2018.05.060.

[23] XIE P, GAO M, WANG C, et al. SuperCT: A supervised-learning framework for enhanced characterization of single-cell transcriptomic profiles[J]. Nucleic Acids Research, 2019, 47(8): e48. DOI: 10.1093/nar/gkz116.

[24] MA F, PELLEGRINI M. ACTINN: Automated identification of cell types in single cell RNA sequencing[J]. Bioinformatics, 2020, 36(2): 533-538. DOI: 10.1093/bioinformatics/btz592.

[25] SAELENS W, CANNOODT R, TODOROV H, et al. A comparison of single-cell trajectory inference methods[J]. Nature Biotechnology, 2019, 37(5): 547-554. DOI: 10.1038/s41587-019-0071-9.

[26] MCCARTHY D J, CAMPBELL K R, LUN A T, et al. Scater: Pre-processing, quality control, normalization and visualization of single-cell RNA-seq data in R[J]. Bioinformatics, 2017, 33(8): 1179-1186. DOI: 10.1093/bioinformatics/btw777.

[27] 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. DOI: 10.1038/nmeth.4150.

[28] 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. DOI: 10.1038/nbt.4096.

[29] STUART T, BUTLER A, HOFFMAN P, et al. Comprehensive integration of single-cell data [J]. Cell, 2019, 177(7): 1888-1902. DOI: 10.1016/j.cell.2019.05.031.

[30] LUN A T, MCCARTHY D J, MARIONI J C. A step-by-step workflow for low-level analysis of single-cell RNA-seq data with Bioconductor[J]. F1000Research, 2016, 5: 2122. DOI: 10.12688/f1000research.9501.2.

[31] ANDREWS T S, HEMBERG M. M3Drop: Dropout-based feature selection for scRNASeq[J]. Bioinformatics, 2019, 35(16): 2865-2867. DOI: 10.1093/bioinformatics/bty1044.

[32] WOLF F A, ANGERER P, THEIS F J. SCANPY: Large-scale single-cell gene expression data analysis[J]. Genome Biology, 2018, 19(1): 15.

[33] SETTY M, TADMOR M D, REICH-ZELIGER S, et al. Wishbone identifies bifurcating developmental trajectories from single-cell data [J]. Nature Biotechnology, 2016, 34(6): 637-645. DOI: 10.1038/nbt.3569.

[34] TAKAHASHI K, YAMANAKA S. Induction of pluripotent stem cells from mouse embryonic and adult fibroblast cultures by defined factors[J]. Cell, 2006, 126(4): 663-676. DOI: 10.1016/j.cell.2006.07.024.

[35] BRUNSKILL E W, PARK J S, CHUNG E, et al. Single cell dissection of early kidney development: multilineage priming[J]. Development, 2014, 141(15): 3093-3101. DOI: 10.1242/dev.110601.

[36] LI H, COURTOIS E T, SENGUPTA D, et al. Reference component analysis of single-cell transcriptomes elucidates cellular heterogeneity in human colorectal tumors [J]. Nature Genetics, 2017, 49(5): 708-718. DOI: 10.1038/ng.3818.

[37] TIROSH I, IZAR B, PRAKADAN S M, et al. Dissecting the multicellular ecosystem of metastatic melanoma by single-cell RNA-seq[J]. Science, 2016, 352(6282): 189-196. DOI: 10.1126/science.aad0501.

[38] BASLAN T, HICKS J. Unravelling biology and shifting paradigms in cancer with single-cell sequencing[J]. Nature Reviews Cancer, 2017, 17(9): 557-569. DOI: 10.1038/nrc.2017.58.

[39] VILLANI A C, SATIJA R, REYNOLDS G, et al. Single-cell RNA-seq reveals new types of human blood dendritic cells, monocytes, and progenitors[J]. Science, 2017, 356(6335): eaah4573. DOI: 10.1126/science.aah4573.

[40] XIONG L, XU K, TIAN K, et al. SCALE method for single-cell ATAC-seq analysis via latent feature extraction [J]. Nature Communications, 2019, 10(1): 4576. DOI: 10.1038/s41467-019-12630-7.

[41] CHENG C, EASTON J, ROSENCRANCE C, et al. Latent cellular analysis robustly reveals subtle diversity in large-scale single-cell RNA-seq data [J]. Nucleic Acids Research, 2019, 47(22): e143. DOI: 10.1093/nar/gkz826.

[42] XIE T, WANG Y, DENG N, et al. Single-cell deconvolution of fibroblast heterogeneity in mouse pulmonary fibrosis [J]. Cell Reports, 2018, 22(13): 3625-3640. DOI: 10.1016/j.celrep.2018.03.010.

[43] DUAN B, ZHOU C, ZHU C, et al. Model-based understanding of single-cell CRISPR screening[J]. Nature Communications, 2019, 10(1): 2233. DOI: 10.1038/s41467-019-10216-x.

[44] HE Y, YUAN H, WU C, et al. DISC: A highly scalable and accurate inference of gene expression and structure for single-cell transcriptomes using semi-supervised deep learning[J]. Genome Biology, 2020, 21(1): 170. DOI: 10.1186/s13059-020-02083-3.