·医学工程·

单细胞测序仪器平台与生物信息学分析体系的协同演进

王亚男[a,b,c,d] 晏菲[e] 印明柱[a,b,c,d] 李鑫[a,b,c,d*]

（重庆大学附属三峡医院 a.临床研究中心；b.医学病理中心；c.肿瘤早期诊治中心；d.转化医学研究中心；e.重庆市老年疾病临床医学研究中心，重庆 404000）

【摘要】 单细胞测序技术的发展使生命科学研究由群体平均测量逐步走向细胞分辨率，为理解组织内部的异质性及细胞状态变化提供了新的研究手段。该技术体系的持续进步不仅源于测序化学的改良，也与微流控液滴、分子条形码标记、长读长测序及空间分辨检测等仪器平台的不断优化密切相关。不同技术路线在测序通量、读长特征与噪声结构等方面存在差异，这些差异在很大程度上影响了后续数据处理方式，并推动生物信息学分析方法持续调整与发展。本文结合单细胞测序仪器平台的主要工程路径，对物理分选、微流控系统、长读长测序及空间技术等代表性方案进行梳理，同时讨论条形码与唯一分子标识（unique molecular identifier, UMI）所塑造的数据结构特征、高稀疏性带来的统计建模问题，以及大规模数据分析对计算框架提出的要求，并对单细胞测序与生物信息学平台在自动化、多模态整合及计算能力提升方面的发展趋势进行概述，以期为相关技术体系的优化提供参考。

【关键词】 单细胞测序；微流控；分子条形码；长读长测序；空间转录组；生物信息学

【中图分类号】 Q811.4 **【文献标识码】** A **【DOI】** 10.11876/mimt202603004

Co-evolution of single-cell sequencing platforms and bioinformatics analytical frameworks

Wang Yanan[a,b,c,d], Yan Fei[e], Yin Mingzhu[a,b,c,d], Li Xin[a,b,c,d]*

［*a. Clinical Research Center (CRC); b. Medical Pathology Center (MPC); c. Cancer Early Detection and Treatment Center (CEDTC); d. Translational Medicine Research Center (TMRC); e. Chongqing Municipality Clinical Research Center for Geriatric Diseases, Chongqing University Three Gorges Hospital, Chongqing 404000, China*］

【Abstract】 With the development of single-cell sequencing technology, life science research has gradually moved from population average measurement to cell resolution, which provides a new research method for understanding the heterogeneity and cell state changes within tissues. The continuous progress of this technology system is not only due to the improvement of sequencing chemistry, but also closely related to the continuous optimization of instrument platforms such as microfluidic droplets, molecular barcoding, long-read sequencing and spatially resolved detection. Different technical routes have differences in sequencing throughput, read length characteristics and noise structure, which largely affect the subsequent data processing methods and promote the continuous adjustment and development of bioinformatics analysis methods. Combined with the main engineering paths of the single-cell sequencing instrument platform, this paper combs the representative solutions such as physical sorting, microfluidic system, long-read sequencing and space technology. At the same time, we discuss the characteristics of data structure shaped by barcodes and unique molecular identifiers (UMI), the statistical modeling problem caused by high sparsity, and the

收稿日期：2026-02-27

基金项目：重庆市科卫联合医学科研项目青年（医师）科学家项目（2025QNKX004）；新重庆青年创新人才项目（CSTB2025YITP-QCRCX0005）

第一作者：王亚男，女，硕士研究生，助理研究员，研究方向：分子病理、药理学；晏菲，男，博士研究生，副研究员，研究方向：生物力学与康复工程

通信作者：李鑫，女，博士研究生，副研究员，研究方向：复杂疾病微环境调控机制研究

requirements of computational framework for large-scale data analysis. In addition, the development trend of single-cell sequencing and bioinformatics platforms in automation, multimodal integration and computing power improvement are summarized, in order to provide reference for the optimization of related technical systems.

【Key words】 Single-cell sequencing; Microfluidics; Molecular barcoding; Long-read sequencing; Spatial transcriptomics; Bioinformatics

传统组学研究多基于群体样本开展，其测量结果通常反映的是大量细胞信号的平均水平。这种研究范式在揭示整体分子特征方面发挥了重要作用，但同时也掩盖了细胞间潜在的功能差异[1]。尤其在肿瘤、免疫系统及神经组织等高度异质的生物体系中，不同细胞亚群在基因表达、调控网络及功能状态方面可能存在显著差异，而群体测序往往难以准确解析这种复杂性。随着研究问题逐渐由整体特征描述转向细胞状态解析，能够在单细胞分辨率下获取分子信息的技术需求日益凸显。

单细胞转录组测序技术的建立被认为是生命科学研究方法学上的重要突破。这一技术突破不仅拓展了基因表达研究的空间维度，也为理解细胞命运决定及发育过程提供了新的研究路径[2]。然而，早期单细胞测序通量和效率有限，制约了其在复杂组织研究中的应用。推动单细胞测序进入规模化阶段的关键因素很大程度上源于仪器工程领域的持续创新，微流控技术的引入、液滴分离策略、条形码标记体系等的发展显著促进了测序通量的提升，并奠定了复杂数据解析的基础[3]。随着通量的持续提升，单细胞数据在规模与结构上呈现出高稀疏性、技术噪声及批次差异等特征，对分析方法提出了新的要求，推动了生物信息学算法的快速发展。针对单细胞数据构建的降维与聚类方法可有效识别细胞亚群，提升复杂组织解析能力[4]。随后发展的整合分析策略进一步支持跨样本与跨实验数据的联合分析，能够在更大尺度上刻画细胞图谱[5]。

在此过程中，仪器平台与分析方法之间逐渐形成协同演进关系，工程创新推动数据形态变革，而数据复杂性的提升反过来促进算法与计算框架的持续优化。因此，从仪器工程视角系统梳理单细胞测序平台的技术路径，并探讨其对生物信息学分析体系的影响，对于理解该领域的发展逻辑具有重要意义。本文围绕单细胞测序仪器平台与分析方法的互动关系，综述主要技术路线及其数据特征，分析仪器与算法协同优化的应用模式，并对未来发展方向进行展望，旨在为单细胞测序领域的技术平台选择、算法开发及协同创新提供参考。

1 单细胞测序技术发展历程

单细胞测序技术经历了从低通量到高通量、从单一组学到多维度检测的演进过程，见图1。早期主要依赖显微操作和流式细胞分选实现单细胞分离，结合逆转录与扩增策略进行转录组分析，通量低、操作复杂，难以在复杂组织中大规模应用。微流控芯片技术的引入实现了纳升级反应体系中的细胞捕获与分子标记，显著提升了实验通量与可重复性。随后，基于液滴的高通量分离策略如Drop-seq、10x Genomics Chromium平台进一步实现了对数万级细胞的并行处理，使单细胞研究迈入大规模数据获取阶段。与此同时，条形码标记体系的发展使来自不同细胞的转录本能够在同一测序反应中被准确区分，为后续高通量测序与数据解析奠定了工程基础。

值得关注的是，单细胞技术正在由单一转录组测量向多维度分子检测扩展。空间转录组学通过保留组织位置信息，使研究者能够在空间背景下理解基因表达模式，被认为是连接分子信息与组织结构的重要技术路径[6]。与此同时，多组学单细胞技术逐步实现对基因表达、染色质可及性及蛋白质水平的联合测量，为系统解析细胞状态提供了更加全面的数据基础。技术路径的多样化不仅丰富了研究手段，也进一步增强了对分析平台整合能力的需求。在这一发展过程中，生物信息学逐渐由数据处理工具转变为单细胞研究的重要支撑体系，仪器平台与分析方法之间开始形成紧密的互动关系。

从技术演进来看，单细胞测序已建立起从样本处理到数据产出的完整技术链条，仪器平台的创新

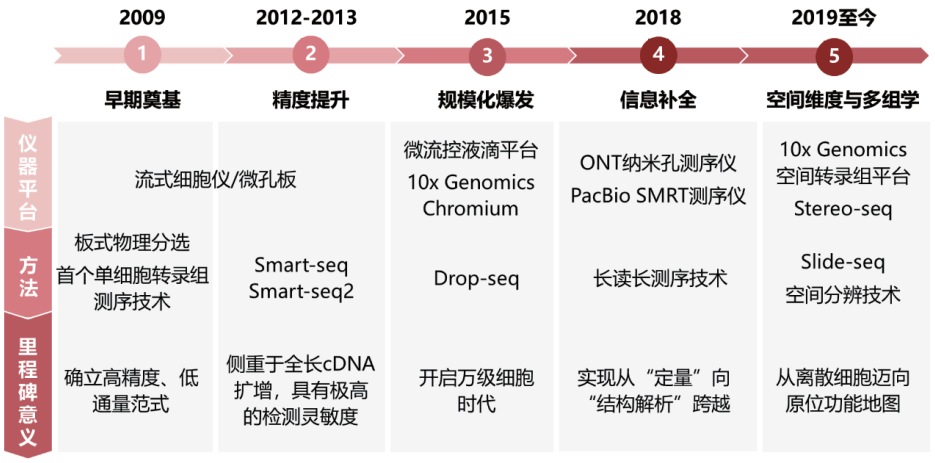


图 1 单细胞测序技术发展历程

与数据复杂性的提升相互促进。上述发展历程为后续深入探讨仪器平台的具体技术路径、单细胞数据特征及其对生物信息学分析的需求奠定了基础。

2 单细胞测序仪器平台的技术路径

2.1 基于物理分选的早期单细胞测序系统

单细胞测序技术的早期发展主要依赖于对单个细胞进行物理分离后再开展分子扩增与测序分析。这一阶段的技术路线通常结合显微操作或流式细胞分选等方式，将目标细胞逐一分配至独立反应体系中，从而避免来自群体样本的信号混合。尽管整体通量有限，但该类方法在单细胞研究起步阶段发挥了关键作用。代表性的技术路径包括基于微孔板的单细胞捕获策略。研究者通常借助流式细胞分选（fluorescence-activated cell sorting, FACS）将单个细胞准确分配至 96 孔板或 384 孔板中，随后完成反转录与全长 cDNA 扩增。早期建立的 Smart-seq 方法首次实现了对单细胞转录本的较高覆盖度测量，使研究者能够在单细胞尺度分析基因表达差异[7]。随后发展的 Smart-seq2 在扩增效率与检测灵敏度方面进一步提升，使低丰度转录本的识别能力得到改善[8]。

从仪器工程视角来看，此类技术的核心优势在于细胞分选精度较高，并能够获得接近全长的转录本信息，这为可变剪接分析及等位基因表达研究提供了重要基础。然而，由于每个细胞需在独立反应体系中完成建库，实验成本与操作时间随细胞数量增加而显著上升，从而限制了其在大规模细胞图谱研究中的应用。

值得注意的是，基于物理分选的技术路径在数据质量方面具有一定优势。相较于后续发展的高通量方法，该策略通常表现出更高的基因检测数量和更低的掉落率（dropout rate），在需要高灵敏度测量的研究场景中仍具有应用价值。研究表明，板式单细胞测序在转录本覆盖完整性方面优于部分液滴方法，使其在精细化转录调控研究中保持重要地位[9]。因此，当核心研究在于解析可变剪接、等位基因表达或对极其稀有的细胞样本进行深度挖掘时，基于微孔板的物理分选路径仍是首选，亦可作为高通量筛选后的精细验证工具，形成“广度由液滴覆盖，精度由板式定义”的互补格局。

总体而言，早期单细胞测序系统体现出以高精度、低通量为特征的仪器设计理念，尽管其规模化能力有限，但该阶段所建立的分子扩增策略与实验流程，为后续高通量单细胞平台的发展奠定了方法学基础。同时，这一路线也揭示了单细胞技术发展的关键工程矛盾，即如何在保持数据质量的同时提升实验通量。这一问题直接推动了微流控技术的引入，并最终促成高通量单细胞测序平台的出现。

2.2 微流控液滴驱动的高通量平台

随着单细胞研究需求的不断提升，早期依赖物理分选的低通量技术逐渐难以满足复杂组织解析与大规模细胞图谱构建的需要。微流控工程技术在样本处理环节的引入，成为推动单细胞测序进入规模化阶段的重要动力。该技术通过在微米尺度下精确控制液体流动，使单个细胞能够在封闭液滴中完成

捕获与分子标记，从而显著提升实验并行度。

基于液滴的单细胞分离策略通过将细胞悬液与带有条形码的微珠共同封装至纳升级反应体系，实现了单细胞与分子标签的高效配对。这种高度自动化的反应模式不仅降低了人为操作误差，也使单次实验能够处理数万乃至更大规模的细胞群体。从仪器架构角度来看，液滴平台的核心优势在于其突出的扩展能力[10]。通过连续流动反应与阵列化液滴生成机制，系统能够在保持较高捕获效率的同时实现稳定运行，这使单细胞研究由探索性实验逐步迈向标准化数据生产。此外，高度集成的设备设计将细胞分离、条形码标记及反应体系构建整合于同一流程中，在提升实验效率的同时增强了结果的一致性。

然而，通量的大幅提升也在一定程度上改变了数据特征。与板式方法相比，液滴平台通常采用对转录本3’端进行测序的策略，这在支持大规模细胞分类的同时，也可能限制对全长转录本结构的解析能力。此外，由于单细胞RNA捕获效率存在差异，测序结果往往表现出更高的数据稀疏性。相关比较研究[9]表明，不同单细胞测序方法在基因检测数量与灵敏度方面存在差异，需要根据具体问题选择合适的技术路径。

微流控技术压倒性的通量优势使其成为构建单细胞图谱或挖掘罕见细胞亚群的重要手段，但鉴于其对转录本结构信息的覆盖不足，在涉及可变剪接、等位基因特异性表达等复杂基因调控机制的研究中，通常仍需结合板式方法或长读长测序技术，以弥补结构信息的缺失。

尽管如此，微流控液滴平台仍被广泛认为是推动单细胞测序进入高通量时代的关键工程突破，其所带来的数据规模跃迁不仅促进了细胞图谱计划的实施，也在客观上推动生物信息学方法向高维数据处理与并行计算方向发展。可以认为，该技术路线体现出以规模优先为特征的仪器设计理念，并在很大程度上重塑了单细胞研究的技术格局。

2.3 纳米孔与长读长单细胞测序技术的探索

微流控液滴平台虽然显著提升了单细胞测序的实验通量，但基于短读长测序的数据在解析转录本完整结构方面仍存在一定局限，例如，可变剪接事件、融合转录本及等位基因特异性表达等复杂分子特征，往往难以通过短片段拼接获得准确重建。因此，能够直接读取长片段序列的测序技术逐渐被引入单细胞研究，以期在保持细胞分辨率的同时提升转录本结构解析能力。长读长测序平台主要包括Pacific Biosciences公司开发的单分子实时测序（single-molecule real-time sequencing, SMRT）技术以及Oxford Nanopore Technologies公司推出的纳米孔测序体系。这类平台通常无需依赖聚合酶链式反应（polymerase chain reaction, PCR）扩增即可完成单分子读取，从而在一定程度上降低扩增偏倚，并支持对全长cDNA分子的直接测量。研究表明，基于SMRT技术的单细胞转录组测序能够有效识别新的剪接异构体，并改善转录本注释的完整性，为理解基因调控复杂性提供了更高分辨率的数据基础[11]。

从仪器工程角度来看，长读长测序改变的不仅是读取长度，更是数据生成方式。以纳米孔测序为例，该技术通过监测核酸分子穿过蛋白纳米孔时产生的电流变化实现碱基识别，使DNA或RNA序列能够被连续读取[12]。这种实时信号检测机制，使测序过程与数据生成更加紧密耦合，也为后续实时分析提供了潜在技术路径。然而，将长读长测序技术拓展至单细胞尺度仍面临多方面挑战：一方面，单细胞样本中可用核酸量有限，对文库构建效率提出了更高要求；另一方面，早期单分子测序较高的原始错误率也增加了数据分析复杂度。随着环形一致性测序（circular consensus sequencing, CCS）等策略的发展，高保真（high-fidelity, HiFi）读段在准确度方面已接近短读长测序水平，从而显著提升了长读长测序单细胞数据的可靠性[13]。

值得关注的是，长读长单细胞测序正在推动分析方法由表达定量向转录本结构解析拓展。相较于依赖推断的短读长测序策略，直接测序完整转录本有助于减少拼接误差，并提升对复杂转录调控事件的识别能力。因此，长读长测序技术在高分辨率结构解析、复杂基因组重构以及表观遗传研究中展现出独特优势。

需要明确的是，长读长测序并非旨在取代短读长测序，而是为了填补后者在异构体识别与融合基因鉴定上的天然缺陷。相关研究[14]指出，将长

读长测序与短读长测序数据进行联合分析，可在兼顾测序深度与结构分辨率的同时，提高转录组解析的整体准确性。可见，未来主流趋势将是“长短结合”的杂交测序模式，该模式在细胞分群的广度与转录组解析的深度之间达成了一种高效的工程学动态平衡。

总体来看，长读长单细胞测序代表着仪器平台在信息获取维度上的进一步延伸，该技术路线并非单纯追求读取长度，而是在高通量与结构完整性之间寻求新的工程平衡。随着测序准确度的持续提升及建库策略的不断优化，长读长平台有望在精细化转录调控研究与复杂基因组解析中发挥更加重要的作用。

2.4 空间分辨单细胞测序技术的发展

尽管单细胞测序在解析细胞异质性方面取得了显著进展，但在组织解离过程中，细胞原有的空间位置信息往往随之丢失，从而限制了研究者对细胞间相互作用及组织结构特征的系统理解。为弥补这一不足，空间分辨转录组技术逐渐发展起来，通过在组织原位获取基因表达信息，使分子数据能够与组织形态结构建立直接关联。从仪器工程视角来看，空间转录组技术的核心突破在于将分子捕获体系与高精度空间定位策略相结合。早期建立的空间转录组方法通过在载玻片表面构建带有空间条形码的寡核苷酸阵列，实现对组织切片中转录本的原位捕获，随后结合高通量测序，得以在二维坐标框架下重建基因表达分布图谱[15]。这一设计标志着测序技术由细胞分离测量迈向原位分子检测，在数据维度上实现了重要拓展。

随着工程技术的持续进步，空间分辨率不断提升。例如，Slide-seq技术利用高密度条形码微珠阵列替代传统捕获表面，使空间解析度接近单细胞尺度，从而更精细地描绘组织内部分子结构[16]。进一步发展的高清空间转录组策略则通过优化阵列设计与信号检测体系，实现更高密度的分子捕获，为复杂组织微环境研究提供了更具分辨力的数据支持。除阵列捕获路线外，基于原位杂交与高分辨荧光成像的空间检测技术同样展现出重要价值。此类方法通常借助高灵敏光学系统对特定 RNA分子进行多轮成像，从而在亚细胞尺度实现基因表达定位。研究表明，该技术路径在分子检测精度方面具有明显优势，有助于揭示细胞内部的转录组织结构[17]。可以认为，空间技术的发展不仅依赖测序能力的提升，也深度融合了光学工程与成像系统的进步。

空间维度的引入显著改变了单细胞数据的结构特征。相较于传统表达矩阵，空间数据通常同时包含分子信息与坐标信息，使其在数据组织形式上更接近图结构。这一变化直接推动了空间聚类、邻域分析及空间轨迹推断等计算方法的发展。相关研究[18]指出，将空间信息纳入分析框架有助于提升细胞类型识别的准确性，并加深对组织微环境的理解。

空间分辨技术的引入不仅是增加了一个物理坐标维度，更是实现了单细胞研究由成分清单向功能地图的范式跃迁，其在揭示细胞微环境、肿瘤浸润机制及器官发育演变中的科研价值不可替代。尽管Slide-seq等技术已将解析精度推至单细胞尺度，但该领域目前仍面临视野大小与空间分辨率之间呈负相关的困境。随着光学成像与微阵列加工工艺技术的深度融合，空间分辨技术正加速突破现有局限，向着实现亚细胞分辨率和全基因组覆盖的终极目标迈进。

总体而言，空间分辨技术代表了单细胞测序仪器在信息维度上的进一步扩展，使研究由细胞状态解析迈向“细胞-空间”关系建模。这一工程路径不仅拓宽了单细胞研究的观察尺度，也对生物信息学平台提出了更高要求，即在处理高维表达数据的同时整合空间拓扑信息。随着空间分辨率与检测通量的持续提升，能够同时整合空间位置、转录信息及蛋白质组学数据的空间多组学有望成为推动系统生物学研究进展的重要技术方向。

综上所述，单细胞测序已形成多路径并行发展格局，通过对上述不同技术路径的梳理，可以发现单细胞测序仪器的发展本质上是在通量、深度与维度之间寻求动态平衡的过程，见表1。目前，没有任何单一技术路径能够同时满足高通量、全长覆盖与空间原位检测的所有需求，但各路径在通量、精度及信息获取维度上各具优势。这种多模态协同形成的互补格局，构成了当代单细胞研究的技术基石。

表1 单细胞测序主要技术路径对比

技术路径	核心原理	代表性平台	技术优势	主要局限与挑战	适用场景
物理分选	单细胞分配至微孔板，在独立孔内完成全长cDNA扩增	①Smart-seq；②Smart-seq2	高灵敏度；全长转录本覆盖；掉落率低；稀疏性低	通量低；成本随细胞数线性上升；操作复杂	罕见细胞亚型、可变剪接、等位基因表达研究
微流控液滴	细胞与条形码微珠封装于纳升级液滴，在液滴内完成裂解与分子标记	①10x Genomics Chromium；②Drop-seq	通量极高；自动化程度高；实验一致性强	全长信息有限；数据具有高度稀疏性	大规模细胞图谱构建
长读长测序技术	单分子实时测序或纳米孔技术直接读取	①PacBio SMRT；②Oxford Nanopore	全长转录本直接测序；获取异构体、融合及等位变异信息；扩增偏倚低	文库构建效率要求高	异构体、融合转录本及等位基因等复杂分子特征研究
空间分辨技术	空间条形码阵列或原位杂交、高分辨荧光成像，保留分子空间坐标	①10x Visium；②Slide-seq；③MERFISH；④Stereo-seq	保留原位空间信息；解析细胞间互作	分辨率与视野大小呈负相关；计算算法复杂度高	细胞微环境、肿瘤浸润机制、器官发育演变等

3 单细胞数据特征与生物信息学分析需求

3.1 条形码体系与唯一分子标识驱动的数据结构变革

单细胞测序能够在高通量条件下同时解析大量细胞，其技术基础在很大程度上依赖于分子标识体系的建立。细胞条形码（cell barcode）的引入，使来自不同细胞的核酸分子能够在同一测序反应中被区分，从而突破了传统的“单样本建库”实验模式。与此同时，唯一分子标识（unique molecular identifier, UMI）的应用，使研究者能够在计算层面识别原始转录本并校正扩增偏倚，为分子定量提供了更加可靠的依据[19]。

从仪器设计角度来看，条形码体系本质上是一种面向高并行处理的工程解决方案。无论是板式建库还是液滴微流控平台，分子标识的核心作用均在于将大规模混合信号重新映射至单细胞来源。这一策略不仅显著提升了实验通量，也直接塑造了单细胞数据的组织形式，即以“基因 × 细胞”为基础的表达矩阵结构。与传统群体测序以样本为单位的数据格式相比，这种矩阵通常具有更高维度，并在统计特征上呈现出明显差异。UMI的引入进一步改变了转录组定量方式。早期RNA测序多依赖测序深度估计表达水平，而扩增过程中产生的重复读段可能导致表达量被高估。通过对每个原始分子赋予独立标签，UMI策略能够在数据处理中去除PCR重复，从而提升定量准确性[20]。

然而，分子标识体系在提升测量可靠性的同时，也带来了新的计算挑战。首先，条形码识别需要在允许一定测序错误的前提下完成精确匹配，因此相关算法通常需结合纠错策略以减少细胞误判。其次，大规模条形码集合使数据规模迅速扩展，对存储结构与计算效率提出更高要求。此外，UMI去重过程本质上涉及分子层面的聚类计算，其复杂度随测序深度增加而提升。值得关注的是，条形码驱动的数据结构在很大程度上促进了单细胞分析方法的体系化发展，例如，围绕表达矩阵构建的降维与聚类算法，使研究者能够在细胞层面识别潜在亚群，从而推动单细胞研究由基因表达比较转向细胞群体建模。可以认为，条形码与UMI不仅是实验设计中的关键组件，也在无形中确立了现代单细胞生物信息学的基本数据范式。总体来看，分子标识技术的出现标志着单细胞测序由低通量实验迈向大规模并行分析的重要阶段[19]。该工程策略在提升实验效率的同时，也深刻改变了转录组数据的统计属性，并在很大程度上推动了后续分析方法的创新。

3.2 高稀疏性与技术噪声带来的统计建模挑战

与传统群体转录组相比，单细胞测序数据最突出的特征之一在于其高度稀疏性。表达矩阵中大量零值的存在，并不完全代表基因未被表达，而往往源于分子捕获效率有限或反转录过程中发生的随机丢失。这种现象通常被称为掉落（dropout），是单细胞RNA测序数据的重要统计特征之一[21]。“掉

落”现象在很大程度上与检测灵敏度相关。单个细胞所含 mRNA 分子数量有限，而多步骤建库流程又进一步放大了分子损失的可能性。在单细胞 RNA 测序中，由于单个细胞中 RNA 含量极低，样本在测序前需要经过多轮扩增，这一过程会引入扩增偏倚，并导致部分实际表达的基因未被检测到，即所谓的“掉落”现象。由于这种“掉落”现象的存在，单细胞表达矩阵中往往出现大量零值，其比例往往高于常规计数模型所能解释的水平。为更好地描述这种数据特征，有研究[22]提出采用零膨胀负二项分布（zero–inflated negative binomial, ZINB）模型对表达数据进行建模。该模型在负二项分布的基础上引入额外的零值生成机制，从而能够同时刻画零膨胀（zero inflation）、过度离散（over–dispersion）以及计数数据的特性。随后发展的多种归一化方法，也尝试在测序深度差异与捕获效率波动之间建立平衡，从而提升跨细胞表达比较的可靠性。可以看到，这类模型的出现并非单纯源于算法创新，而是在很大程度上回应了单细胞实验体系所带来的数据不确定性。高通量平台在扩大细胞规模的同时，也加剧了批次效应对分析结果的潜在影响。不同实验时间、试剂批号或仪器运行状态均可能导致系统性偏移，使得技术差异在无形中被误认为生物差异。为缓解这一问题，整合算法逐渐成为单细胞分析流程的重要组成部分。通过在共享表达空间中对数据进行校正，这类方法能够在一定程度上恢复不同样本之间的可比性[5]。稀疏性所带来的另一个直接影响体现在细胞类型识别过程中。当大量基因呈现零表达时，传统基于距离度量的聚类策略往往难以准确反映细胞间真实关系。因此，能够在高噪声背景下提取稳定特征的降维方法逐渐成为分析核心。相关研究[4]表明，针对单细胞数据设计的降维框架有助于提升细胞亚群划分的稳健性，并增强对连续细胞状态变化的解析能力。

随着测序灵敏度与建库策略的持续优化，单细胞数据质量正在不断提升，但稀疏性仍将在相当长一段时间内成为该领域的基本属性。这一特征在某种程度上反映了单分子检测所面临的物理约束，也提示分析方法需要在统计推断与技术误差之间取得平衡。可以认为，高稀疏性既是单细胞数据的挑战，也是推动生物信息学方法持续演进的重要动力。

3.3 高维数据与计算框架的扩展需求

单细胞测序在持续提升实验通量的同时，也显著扩大了数据规模与维度。以常见的单细胞转录组实验为例，数万至数十万细胞的表达矩阵已成为常态，而每个细胞通常包含上万基因的表达信息。此类数据在规模与结构上均远超传统群体测序，计算负担随之迅速增加，尤其在细胞数量达到百万级别时，常规内存计算模式往往难以满足处理需求。高维表达矩阵带来的首要挑战体现在内存管理与计算效率上。矩阵规模随细胞数量线性增长，而降维、邻域构建与聚类等步骤通常涉及全局或局部距离计算，其时间复杂度在细胞规模扩展时显著提升。研究[23]表明，在大规模单细胞数据分析中，引入稀疏矩阵存储结构与并行计算策略，有助于显著降低内存占用并提升运行效率。这一趋势反映出单细胞分析正在由桌面级计算向高性能计算过渡。

与此同时，单细胞数据的高维特性也为模型设计带来新的要求。传统线性降维方法在处理复杂非线性结构时往往存在局限，因此基于图结构或概率模型的分析框架逐渐成为主流。图邻接关系的构建使细胞间局部相似性得以保留，为后续聚类与轨迹推断提供基础。然而，随着细胞规模增加，图构建与优化过程的计算复杂度迅速上升，对算法实现效率提出更高要求。近年来，基于深度学习的表示学习方法逐渐被引入单细胞分析。变分自编码器（variational autoencoder, VAE）等模型通过学习低维潜在空间结构，在保留表达特征的同时降低数据维度，从而提升大规模数据处理能力[24]。此类方法在批次校正与数据整合方面表现出较好性能，但其训练过程通常依赖图形处理器（graphics processing unit, GPU）等硬件加速资源。由此可见，算法形态演变与计算架构升级之间存在密切关联。

随着多模态单细胞技术的发展，数据结构进一步复杂化，除转录组矩阵外，还可能同时包含蛋白

表2 单细胞数据的技术成因与数据表现

数据特征	技术成因	数据表现
条形码体系	混合测序中为每个细胞分配唯一序列标签	从单样本变为“基因 × 细胞”的表达矩阵
UMI定量	赋予原始分子独立标签，区分原始分子与PCR冗余	消除扩增偏倚，表达量估计算法从Reads计数转为分子计数
高稀疏性	捕获效率限制或反转录过程中发生的随机丢失	产生典型的零膨胀分布
批次效应	不同实验批次或平台间的系统性差异	同类细胞在统计分布上出现系统性偏移
高维特性	条形码技术实现大规模并行分析（数十万级细胞）	维度极高，计算复杂度呈指数级增长
多模态整合	单次实验同时捕获转录组、蛋白质组及空间坐标等异质信息	形成多特征矩阵耦合

注：UMI为唯一分子标识；PCR为聚合酶链式反应。

表3 单细胞数据分析挑战与方法学响应

数据特征	分析挑战	方法学响应
条形码体系	测序误差导致条形码无法识别或细胞误判	部署容错匹配与纠错算法
UMI定量	大规模数据下分子层面的复杂聚类与计算冗余	开发高效分子去重流程与归一化算法
高稀疏性	零膨胀分布掩盖低表达基因的生物学信号	引入零膨胀负二模型与稳健降维框架
批次效应	导致生物学亚群分类错误；无法进行跨样本对比	采用整合算法对数据进行整合与校正
高维特性	维度灾难导致的计算资源瓶颈与内存溢出	引入稀疏矩阵存储、图结构聚类及深度表示学习
多模态整合	异构数据量纲不一与统计分布失配	跨模态对齐算法、共享潜在空间建模与分布式计算框架

注：UMI为唯一分子标识。

表达、染色质可及性及空间坐标信息。多源数据的联合建模不仅增加了数据维度，也提高了计算复杂度。为适应这一变化，分布式计算框架与云端资源调度逐渐被引入单细胞分析流程，使研究者能够在更大规模数据集上开展整合研究[25]。可以看到，单细胞测序仪器能力的持续提升，使数据规模与信息维度不断扩展，而计算体系则在响应这一变化的过程中逐步演进。单细胞数据的技术成因与数据表现见表2，单细胞数据分析挑战与方法学响应见表3。由高维矩阵处理到图结构建模，再到深度表示学习与分布式计算，单细胞生物信息学正在形成与实验通量相匹配的计算基础设施。这种由仪器扩展驱动计算升级的趋势，构成了单细胞技术协同演进的重要组成部分。

4 总结与展望

单细胞测序技术的发展在很大程度上体现为仪器平台与分析方法之间的相互促进。从早期依赖物理分选的低通量体系，到微流控驱动的规模化细胞捕获，再到长读长测序与空间分辨技术对信息维度的拓展，仪器工程的持续进步不断改变数据生成方式，并进一步塑造生物信息学分析的核心问题。从条形码与UMI所确立的分子计数模式，到围绕高稀疏性数据建立的统计模型，再到面向大规模矩阵的计算框架扩展，分析体系的演进始终与实验技术能力保持紧密关联。

值得注意的是，单细胞测序在提升通量与分辨率的同时，也持续暴露出若干基础性挑战：首先，分子捕获效率的不稳定仍可能影响定量精度，不同平台之间的数据差异亦在一定程度上增加了跨研究整合的复杂性；其次，随着测序规模不断扩大，计算资源需求与数据管理压力同步上升，这使得单纯依赖算法优化已难以完全应对未来的数据增长。如何在仪器设计阶段即兼顾数据结构与分析需求，正逐渐成为单细胞技术发展的重要议题。从发展趋势来看，单细胞测序有望在多个工程方向上继续深化。一方面，自动化建库与一体化测序系统的推进，可能进一步降低实验操作复杂度，并提升数据获取的一致性；另一方面，长读长测序与空间多组学技术的结合，有望在保持细胞分辨率的同时提供更完整的分子结构信息。此外，面向超大规模数据集的计算基础设施正在逐步由本地环境向分布式与云端架构延伸，这一变化不仅关系到分析效率，也将在一定程度上影响未来研究组织方式。

总体而言，单细胞测序的发展已不再局限于测序能力本身，而是逐步走向实验系统与计算体系的整体优化。可以预见，随着仪器性能、数据结构与分析方法之间的耦合程度不断加深，单细胞研究将更加依赖跨领域的技术协同。在这一过程中，保证在提升检测能力的同时维持数据质量，并构建可持续扩展的分析框架，仍将是推动该领域持续前行的重要方向。

利益冲突：作者声明没有利益冲突。

[参 考 文 献]

[1] Lim J, Park C, Kim M, et al. Advances in single–cell omics and multiomics for high–resolution molecular profiling [J]. Exp Mol Med, 2024, 56(3): 515–526.

[2] A focus on single–cell omics [J]. Nature Reviews Genetics, 2023, 24: 485.

[3] Zheng G X Y, Terry J M, Belgrader P, et al. Massively parallel digital transcriptional profiling of single cells [J]. Nat Commun, 2017, 8: 14049.

[4] Sun Y, Kong L, Huang J, et al. A comprehensive survey of dimensionality reduction and clustering methods for single–cell and spatial transcriptomics data [J]. Brief Funct Genomics, 2024, 23(6): 733–744.

[5] Butler A, Hoffman P, Smibert P, et al. Integrating single–cell transcriptomic data across different conditions, technologies, and species [J]. Nat Biotechnol, 2018, 36(5): 411–420.

[6] Ståhl P L, Salmén F, Vickovic S, et al. Visualization and analysis of gene expression in tissue sections by spatial transcriptomics [J]. Science, 2016, 353(6294): 78–82.

[7] Ramsköld D, Luo S, Wang Y C, et al. Full–length mRNA–Seq from single–cell levels of RNA and individual circulating tumor cells [J]. Nat Biotechnol, 2012, 30(8): 777–782.

[8] Picelli S, Faridani O R, Björklund A K, et al. Full–length RNA–seq from single cells using Smart–seq2 [J]. Nat Protoc, 2014, 9(1): 171–181.

[9] Ziegenhain C, Vieth B, Parekh S, et al. Comparative analysis of single–cell RNA sequencing methods [J]. Mol Cell, 2017, 65(4): 631–643.e4.

[10] De Jonghe J, Kaminski T S, Morse D B, et al. spinDrop: a droplet microfluidic platform to maximise single–cell sequencing information content [J]. Nat Commun, 2023, 14(1): 4788.

[11] Deslattes Mays A, Schmidt M, Graham G, et al. Single–molecule real–time (SMRT) full–length RNA–sequencing reveals novel and distinct mRNA isoforms in human bone marrow cell subpopulations [J]. Genes, 2019, 10(4): 253.

[12] Wang Y, Zhao Y, Bollas A, et al. Nanopore sequencing technology, bioinformatics and applications [J]. Nat Biotechnol, 2021, 39(11): 1348–1365.

[13] Wenger A M, Peluso P, Rowell W J, et al. Accurate circular consensus long–read sequencing improves variant detection and assembly of a human genome [J]. Nat Biotechnol, 2019, 37(10): 1155–1162.

[14] Torma G, Tombácz D, Csabai Z, et al. Combined short and long–read sequencing reveals a complex transcriptomic architecture of african swine fever virus [J]. Viruses, 2021, 13(4): 579.

[15] Wu S Z, Al–Eryani G, Roden D L, et al. A single–cell and spatially resolved atlas of human breast cancers [J]. Nat Genet, 2021, 53(9): 1334–1347.

[16] Rodriques S G, Stickels R R, Goeva A, et al. Slide–seq: a scalable technology for measuring genome–wide expression at high spatial resolution [J]. Science, 2019, 363(6434): 1463–1467.

[17] Chen K H, Boettiger A N, Moffitt J R, et al. RNA imaging. Spatially resolved, highly multiplexed RNA profiling in single cells [J]. Science, 2015, 348(6233): aaa6090.

[18] Shi W, Zhang Z, Xu X, et al. Single–cell and spatial transcriptomics integration: new frontiers in tumor microenvironment and cellular communication [J]. Front Immunol, 2025, 16: 1649468.

[19] 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.

[20] Kivioja T, Vähärautio A, Karlsson K, et al. Counting absolute numbers of molecules using unique molecular identifiers [J]. Nat Methods, 2011, 9(1): 72–74.

[21] van Dijk D, Sharma R, Nainys J, et al. Recovering gene interactions from single–cell data using data diffusion [J]. Cell, 2018, 174(3): 716–729.e27.

[22] Risso D, Perraudeau F, Gribkova S, et al. A general and flexible method for signal extraction from single–cell RNA–seq data [J]. Nat Commun, 2018, 9(1): 284.

[23] Chockalingam S P, Aluru M, Aluru S. SCEMENT: scalable and memory efficient integration of large–scale single–cell RNA–sequencing data [J]. Bioinformatics, 2025, 41(2): btaf057.

[24] Lopez R, Regier J, Cole M B, et al. Deep generative modeling for single–cell transcriptomics [J]. Nat Methods, 2018, 15(12): 1053–1058.

[25] Yiu T, Chen B, Wang H, et al. Transformative advances in single–cell omics: a comprehensive review of foundation models, multimodal integration and computational ecosystems [J]. J Transl Med, 2025, 23(1): 1176.