科学通报 2025 年 第 70 卷 第 15 期: 2356 ~ 2374
复旦大学生物医学研究院建院20周年专辑 评 述
《中国科学》杂志社 SCIENCE CHINA PRESS

转录组生物信息学: 从数据生成到分析框架

南芳[1,2*], 马旭凯[1,2], 杨力[1,2*]

1. 复旦大学生物医学研究院, 上海 200032
2. 复旦大学附属儿科医院, 上海 201102
* 联系人, E-mail: fangnan@fudan.edu.cn; liyang_fudan@fudan.edu.cn

2025-02-14 收稿, 2025-04-01 修回, 2025-04-24 接受, 2025-04-24 网络版发表
国家重点研发计划(2024YFC3405902, 2021YFA1300503)、国家自然科学基金(32430018)和上海市科学技术委员会(23JS14003000, 23DX1900102)资助

摘要 随着人类基因组计划的顺利完成和高通量测序技术的快速发展, 研究人员能够以前所未有的精度和深度对转录组进行全面探索, 揭示基因表达在转录组层面的复杂性及其在细胞和生理过程中的动态变化. 这些技术的突破大大提高了转录组数据的获取速度和准确性, 使研究人员能够对不同生理状态、发育阶段及疾病模型的基因表达模式进行精细的比较分析. 本综述归纳了转录组研究中的多种高通量测序数据获取及相关计算分析的核心思路, 在基于技术手段和分析目标差异对转录组测序技术进行系统分类的基础上, 介绍了不同转录组数据分析策略在多个研究方向的应用. 同时, 本文介绍了人工智能方法在转录组分析研究中的应用, 包括利用前沿深度学习技术建立的多种预测模型等, 期望为深入开展转录组信息挖掘及其应用提供新思路.

关键词 转录组, 高通量测序, 生物信息, 人工智能

高通量测序技术的进步与高效计算分析方法的广泛应用, 推动转录组研究进入了一个全新的时代. 相比于传统的微阵列芯片技术(microarray), 针对转录组核糖核酸(RNA)的高通量测序(RNA-seq)可以提供全面、准确的基因表达定量信息, 不仅能够有效检测低丰度的转录本, 而且能够解析复杂的可变剪接、基因融合等现象. 多种转录组测序技术和分析方法的不断发展和完善, 使得基因表达调控的研究变得更加系统和全面, 为转录组功能研究提供了强大的支持. 更为重要的是, 这些研究不仅证实RNA是遗传信息传递的中间载体, 更揭示其在表观遗传调控、细胞命运决定等生物学过程中的重要作用. 近年来, 随着人工智能技术的不断进步及其在生物数据分析中的广泛应用, 为转录组数据的解读提供了新的视角. 从传统的基因表达定量方法到目前基于机器学习和深度学习的前沿分析方法, 研究者不仅在数据处理效率上得到了显著提升, 还能够从大规模的数据中探索转录组与表观基因学、蛋白组学、代谢组学等其他“组学”之间的复杂关系[1].

因此, 基于高通量测序技术的发展和先进的分析方法的应用, 转录组研究已经进入了一个快速发展的新阶段. 人工智能的融入使得转录组学的研究不仅在技术上取得突破, 也在生物学理解和临床应用上带来了前所未有的机遇. 对转录组调控机制的认识因此更加全面和深入, 为精准医学、疾病防治以及生命科学的各个领域提供了强大的理论支持和实践指导. 本综述论文从RNA表达调控的复杂性入手, 深入探讨了不同转录组数据获取和分析的原理及其局限性, 强调了机器学习技术,尤其是深度学习模型在转录组研究中的重要推进作用, 为深入开展转录组信息挖掘及其潜在应用研究提供新思路.

引用格式: 南芳, 马旭凯, 杨力. 转录组生物信息学: 从数据生成到分析框架. 科学通报, 2025, 70: 2356–2374
Nan F, Ma X-K, Yang L. Bioinformatics in transcriptome: from sequencing strategies to analyzing pipelines (in Chinese). Chin Sci Bull, 2025, 70: 2356–2374, doi: 10.1360/TB-2025-0160

1 转录组复杂性

生物体内存在丰富的RNA类型，除了众所周知的mRNA(蛋白质表达模板)、tRNA(负责转运氨基酸)和rRNA(参与蛋白质合成)，还有大量的非编码RNA[2]. 目前研究人员已经对许多不同来源的非编码RNA进行深入研究，包括参与mRNA剪接的snRNA，参与rRNA化学修饰的snoRNA，长度介于21~23 nt可调控RNA降解的microRNA，主要存在于生殖细胞的piRNA，以共价闭合环形式存在的环形RNA，以及其他不同特征的长非编码RNA[1].

除了RNA种类的多样性，RNA还经过不同的加工发挥多种生物功能，包括但不限于选择性加尾，可变(反向)剪接，RNA编辑，RNA修饰等(图1). RNA的3′端可选择不同的多腺苷酸(即poly(A)尾)位置，导致RNA的长度和/或编码序列发生变化，并改变(m)RNA的稳定性、翻译效率、定位和翻译产物多样性[3]. 例如，某些选择性加尾位点可能导致RNA在细胞内的半衰期延长，并与免疫响应、细胞周期、应激反应以及癌症的发生密切相关，是调控基因表达的一个重要步骤[3]. 可变剪接是转录后调控的一个重要机制，指在mRNA前体剪接过程中，多个内含子和外显子的拼接方式可以发生变化，产生不同的mRNA异构体. 通过可变剪接，单一基因可以生成多个不同的mRNA产物进而编码不同的蛋白质产物，不仅增加了基因表达的多样性，还使得细胞能够通过调节剪接模式应对不同的生理需求，如在发育、免疫反应中或在癌症等疾病发生中的作用[4,5]. 除了经典剪接，mRNA前体中的外显子还可以通过反向剪接将下游外显子的5′剪接位点反向与上游外显子的3′剪接位点相连并形成3′,5′-磷酸二酯键，最终产生了共价闭合的环形RNA(circRNA)[6,7]. 此外，一些剪接后的内含子也可以逃逸脱分支酶的降解，形成稳定的内含子环形RNA(ciRNA)并在体内发挥调控功能[8]. 目前已经发现环形RNA在基因表达、转录调控、癌症发生、神经发育、免疫调控等多方面发挥功能[9,10].

单碱基水平的RNA编辑和修饰也进一步丰富了RNA的多样性. RNA分子生成后其转录自模板DNA的序列发生改变的过程被称为RNA编辑. 最常见的RNA

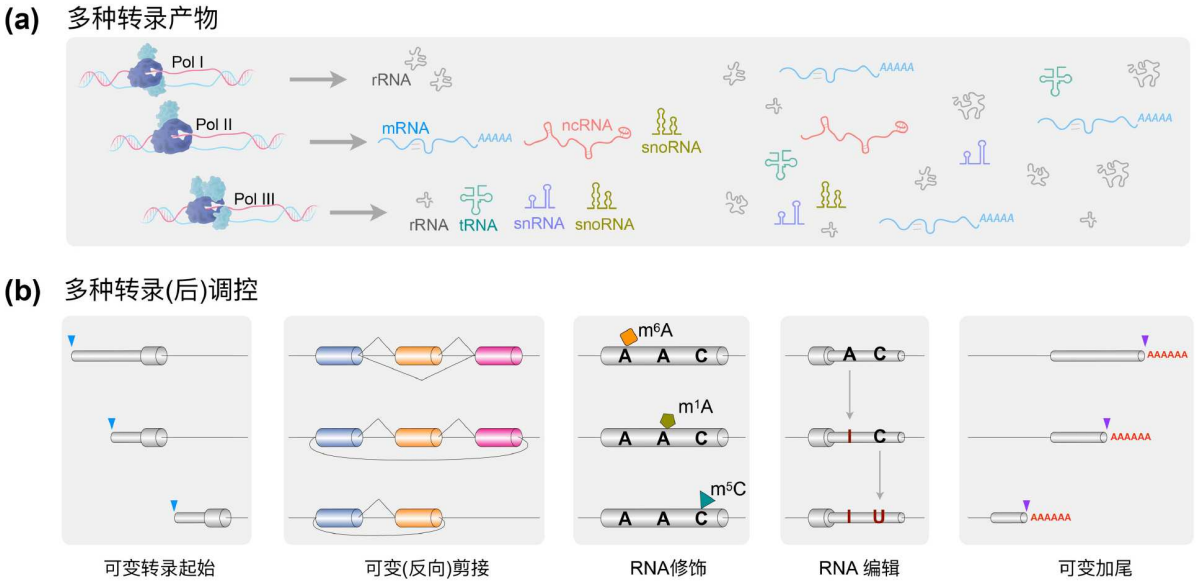

图 1 转录组复杂性. 转录组复杂性体现在转录本种类丰富性(a)和转录调控方式多样性(b). (a) 真核生物不同的转录产物: pol I主要生成28S、18S、5.8S rRNA等; pol II主要生成mRNA、ncRNA、snoRNA、miRNA等; pol III主要生成5S rRNA、tRNA、snRNA、snoRNA等. (b) 真核生物的多种转录调控方式，包括可变转录起始、可变(反向)剪接、RNA修饰(m^6A修饰、m^1A修饰、m^5C修饰等)、RNA编辑(A-to-I编辑，C-to-U编辑等)、可变加尾等

Figure 1 Transcriptome complexity. Transcriptome complexity is reflected in both the diversity of transcripts (a) and the variety of transcriptional regulation (b). (a) Diverse transcriptional products in eukaryotes: Pol I mainly transcribes 28S, 18S, 5.8S rRNAs, etc; Pol II mainly transcribes mRNAs, ncRNAs, snoRNAs, miRNAs, etc; Pol III mainly transcribes 5S rRNAs, tRNAs, snRNAs, snoRNAs, etc. (b) Multiple transcriptional regulatory mechanisms in eukaryotes: alternative transcription initiation, alternative (back) splicing, RNA modifications (m^6A modification, m^1A modification, m^5C modification, etc), RNA editing (A-to-I editing and C-to-U editing), and alternative polyadenylation, etc

编辑形式是腺苷(A)的脱氨基作用(A-to-I编辑)和胞苷(C)的脱氨基作用(C-to-U编辑), 其中以A-to-I编辑最为普遍[11]. RNA编辑能够在不改变基因组信息的情况下产生不同的mRNA异构体(可能产生不同的蛋白质产物)、改变mRNA的稳定性或翻译起始位点, 以及影响RNA的剪接模式[12]. 其广泛存在于神经系统[13]和免疫系统[14]中, 发挥重要的调控功能. RNA修饰是指不改变序列的情况下, 核苷酸发生特定的化学变化, 改变其功能的过程. RNA修饰通过改变RNA的结构或稳定性, 影响其翻译效率以及与其他分子的相互作用[15]. 例如, m^6A甲基化是目前研究最多的RNA修饰之一, 它通过在mRNA上加上甲基团调节RNA的稳定性、剪接和翻译. 除了m^6A外目前发现的还有m^5C、m^1A、5hmC等超过100种RNA修饰类型[16,17].

多种多样的RNA生成加工产生了丰富的转录组RNA类型, 为生命活动提供了不同的调控途径和灵活的调节方式, 而高通量测序技术的发展极大地推动了对转录组的系统研究, 加深了人们对生命活动过程中RNA所发挥的重要功能的理解.

2 高通量测序技术在转录组研究中的应用

高通量测序技术的进步, 使得转录组研究进入了一个全新的时代. 相比于传统的微阵列芯片技术, 高通量测序提供了更宽的检测范围、更高的灵敏度, 以及更准确的基因表达定量信息, 能够有效检测低丰度的转录本, 并且能够解析复杂的剪接变异、可变加尾、动态编辑/修饰和基因融合等现象, 使得基因表达调控的研究更加系统和全面, 为转录组功能研究提供了强大的数据支持[18]. 转录组测序技术可以根据其测序/分析技术特点进行系统性分类(图2和表1). 从测序技术平台的角度, 可将转录组测序技术简单地分为二代短读长测序技术(测序长度50~500 bp)和三代长读长测序技术(单分子全长, 最长可达4 Mb[19])两大类. 根据富集目标和分析目标的不同, 二代测序技术还可细分为以下三类: (1) 转录本全序列富集; (2) 转录本目标片段富集; (3) 引入突变的转录本片段富集. 此外, 在二代测序技术框架下, 可以根据细胞分辨率进一步细分: 一种是针对细胞群体水平(bulk RNA-Seq)的技术, 另一种是能够解析单个细胞转录组特征的单细胞转录组测序(single-cell RNA-Seq)技术和空间转录组测序(spatial RNA-seq)技术. 这些分类主要基于测序技术、富集目标和分析目标的差异, 适用于不同的转录组研究场景.

2.1 二代测序技术

2.1.1 转录本全序列富集

随着高通量测序技术的进步, 研究人员不再局限于利用微阵列芯片检测特定转录本表达谱, 而是利用oligo(dT)捕获含有poly(A)尾的转录本(以mRNA为主), 对富集的RNA序列建库并测序, 实现对转录本表达水平的检测[20]. 该建库测序方法(poly(A)+ RNA-seq)在早期转录组研究中被广泛使用, 实现比较不同条件下基因的差异表达情况.

除了含poly(A)尾的mRNA, 细胞内还有大量无poly(A)尾的转录本, 研究人员建立了去除rRNA并选择不含poly(A)尾RNA的富集方法(poly(A)– RNA-seq), 对mRNA以外的其他不具有poly(A)尾的转录本进行建库测序, 识别了一系列新型长非编码RNA如sno-lncRNA[21], 并发现其中一个转录本*SLERT*具有调控pol I转录的功能[22,23]. 另外, 利用核酸外切酶RNase R降解线性RNA后, 对剩余的RNA测序并鉴定出大量环形RNA, 包括ciRNA[8]和circRNA[6]. 同时, 通过仅去除rRNA的富集建库方法(Ribo– RNA-seq), 研究人员也可以直接对转录组中的大部分RNA(无论有无poly(A)尾)统一分析, 相对于poly(A)+ RNA-seq和poly(A)– RNA-seq获得更为全面的转录组动态信息[24]. 这些RNA-seq数据除了用于表达差异定量和新型转录本鉴定, 还可以利用针对性分析流程研究可变(反向)剪接、RNA编辑等调控过程.

研究人员还利用特异性富集建库测序, 分离具有不同特征的转录本并用于鉴定新型RNA分子. 例如Fib-RIP-seq(Fibrillarin RIP-seq)通过抗体富集含有snoRNA的转录本, 从中鉴定到两端均为snoRNA的sno-lncRNA以及3′端具有poly(A)尾但5′端为snoRNA的一类新型RNA分子*SPA*[25]; NAP-seq通过3′-OH末端筛选和添加特定接头进行扩增, 可以富集并鉴定多种napRNA(noncapped RNAs)[26]; RIP-PEN-seq/PEN-seq/sub-PEN-seq等测序技术则可用于具有后向K-turn结构和序列特征的bktRNA(backward K-turn RNA)的鉴定[27].

除了富集成熟的转录本, 研究人员还开发了4sUDRB-seq用于研究新生转录本[28,29]. 该方法利用转录抑制剂DRB(5,6-dichloro-1-β-D-ribofuranosylbenzimidazole)抑制RNA聚合酶II的活性使转录暂停, 随后加入4sU(4-硫代尿苷)并重启转录使其替代尿苷(U)掺入新合成的RNA分子, 然后通过4sU抗体富集特定时间间隔

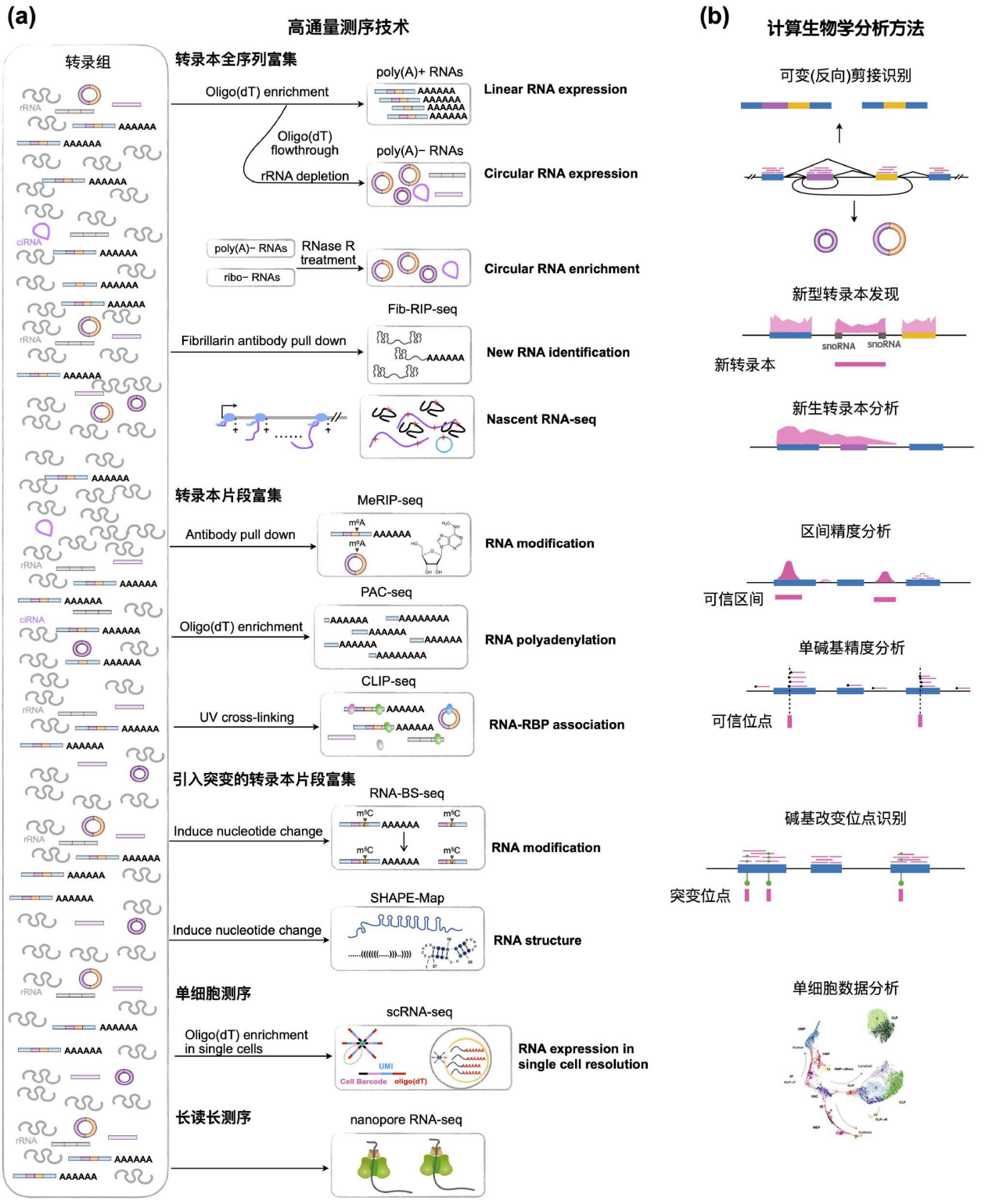

图 2 转录组研究相关测序技术及分析方法. (a) 转录组研究不同的建库测序方法原理示意图. (b) 根据不同建库测序方法的识别目标建立的相应分析方法

Figure 2 High-throughput sequencing technologies and analysis methods for transcriptome research. (a) Schematic illustration of different library preparation and sequencing approaches for transcriptome research. (b) Analytical methods based on detection objectives from different library sequencing approaches

表 1 转录组测序技术及分析方法

Table 1 High-throughput sequencing technologies and analysis methods for transcriptome research

测序平台	细胞分辨率	富集目标	测序技术名称	RNA富集方法	分析目标	应用场景	局限性
二代测序(短读长)	细胞群	转录本全序列	poly(A)+ RNA-seq	oligo(dT)富集有poly(A)尾的RNA	转录本全序列	mRNA表达差异/可变剪接/RNA编辑	仅可检测mRNA
			Ribo– RNA-seq	去除rRNA	转录本全序列	表达差异/可变剪接/RNA编辑位点	低表达转录本难以检测
			poly(A)– RNA-seq	去除rRNA和含poly(A)的RNA	转录本全序列	非编码RNA/环形RNA	无法检测mRNA
			poly(A)–/RNase R RNA-seq	去除rRNA和含poly(A)的RNA, 并用RNase R消化线性RNA	转录本全序列	环形RNA	仅可检测环形RNA或其他不被RNase R降解的RNA
			4sUDRB-seq	4sU富集新生转录本	转录本全序列	新生转录本	实验难度大, 存在假阳性
			Fib-RIP-seq	与蛋白交联后富集	转录本全序列	鉴定与Fibrillarin蛋白互作的RNA	分辨率低
		转录本目标片段	Poly(A)-ClickSeq	富集RNA加尾位点上游序列	加尾位点	识别RNA加尾位点	定量准确性受限
			CLIP-seq	与蛋白交联固定、打断后, 抗体富集的RNA片段	交联序列	蛋白与RNA的结合位点	分辨率低
			iCLIP-seq	交联后反转录停止于交联位点	反转录停止位点	蛋白与RNA的结合位点	实验难度大
			eCLIP-seq	交联后RNA片段化, 反转录停止于交联位点	反转录停止位点	蛋白与RNA的结合位点	实验难度大
			SPLASH	RNA互作区域交联后打断	交联序列	RNA互作片段	交联效率低, 假阳性难以排除
			PARIS	psoralen交联RNA双链区域	交联序列	RNA互作片段/RNA结构双链区域	交联效率低, 假阳性难以排除
			CLASH	交联后RNA片段连接与测序	交联序列	RNA互作片段	交联效率低, 假阳性难以排除
			m^6A-seq	抗体富集含有m^6A修饰片段	抗体富集片段	m^6A修饰位点鉴定	分辨率低
			MeRIP-seq	抗体富集含有m^6A修饰片段	抗体富集片段	m^6A修饰位点鉴定	分辨率低
			m^6A-CLIP-seq	抗体结合后交联, 反转录停止于结合位点	反转录停止位点	m^6A修饰位点鉴定	实验难度大
			icSHAPE	NAI-N3标记后反转录停止于标记位点	反转录停止位点	RNA二级结构	对测序深度要求较高
		引入突变的转录本片段	PAR-CLIP	细胞中掺入光敏核苷酸并进行交联, 反转录时引入碱基改变	突变位点	蛋白与RNA的结合位点	碱基改变位置和实际结合位置存在偏差
			RNA-BS-seq	将无修饰的C转化为U	突变位点	m^5C修饰位点鉴定	C转化不完全, 存在假阳性
			SHAPE-MaP	含有突变的反转录产物	突变位点	RNA二级结构	难以实现全转录组检测
	单细胞	磁珠探针富集序列	10 × Genomics Chromium	5′ cap富集/3′-poly(A)富集	转录本全序列	单细胞基因表达谱	基因覆盖不全, 无法用于可变剪接分析
		转录本全序列	SMART-seq2	全序列	转录本全序列	单细胞基因表达谱	单次捕获单细胞数量少
		磁珠探针富集序列	Visium	3′-poly(A)富集	转录本全序列	带空间信息的单细胞基因表达谱	检测灵敏度受限
三代测序(长读长)	细胞群/单细胞	转录本全长	Nanopore	直接测序	转录本全长	全长转录本检测	准确率略低
			PacBio	直接测序	转录本全长	全长转录本检测	成本较高

内新生成的RNA并进行建库测序，推动了对转录动态过程的研究.

以上技术利用不同的分离富集方法捕获不同特征的转录本, 并实现序列特征、表达差异等分析比较, 极大地推动了转录本类型丰富性和表达特异性的研究.

2.1.2 转录本目标片段富集

除了上述转录本全序列富集研究，研究人员还开发了多种针对性的富集测序方法，可以富集目标RNA的特定片段，用于RNA可变加尾、RNA修饰、RNA与蛋白互作、RNA与RNA互作、RNA二级结构等研究. 根据实验原理和对应分析方法，不同方法还可以分为区间精度和单碱基精度.

通过oligo(dT)可以富集mRNA转录本全序列并检测表达水平，对RNA加尾位点的识别可以使用相同的富集方法. 首先通过oligo(dT)富集mRNA，随后对反转录体系进行改进，使建库的片段多集中在RNA的poly(A)尾部[30]. 例如poly(A)-ClickSeq[31]在利用oligo(dT)富集后的反转录过程使用一定比例被叠氮修饰的核酸(azido-nucleotide), 使反转录过程提前终止, 然后对产物进行建库, 实现对poly(A)尾附近序列的富集和测序. 在分析过程中根据mRNA的序列信息确定具体加尾位点, 并进行进一步的分析.

RNA与蛋白的互作区域常利用RNA-蛋白质交联后免疫沉淀进行富集测序，根据其分辨率差异分为CLIP-seq[32,33]/iCLIP-seq[34]/eCLIP-seq[35]等方法. 它们的关键步骤都是通过紫外交联使蛋白质和其结合的RNA形成稳定复合物，然后通过免疫沉淀富集与蛋白结合的RNA片段, 并进行建库测序. 仅通过片段富集识别的RNA-蛋白结合位点分辨率有限，因此研究者改进了实验技术(iCLIP-seq/eCLIP-seq)，在交联之后增加了反转录的步骤. 由于与RNA结合的蛋白质的阻碍, 反转录会停止在蛋白质结合的位置, 对反转录得到的cDNA片段进行建库测序, 通过测序读序(reads)比对分析识别反转录停止位点可以实现单碱基精度的蛋白结合位点鉴定.

另外交联固定的策略还可以用于检测RNA分子之间的相互作用序列. 例如CLASH[36]、SPLASH[37]、PARIS[38]等方法均通过不同的方式将RNA-RNA的互作序列交联固定并连接，通过高通量测序及序列比对确定哪些序列之间存在相互作用. 不同的技术实验细节和适用范围略有差异，如CLASH技术通过抗体富集特定RNA结合蛋白，适用研究特定蛋白介导的RNA-RNA互作; PARIS增加了双向电泳的步骤, 适用于解析RNA长距离作用，同时还可以用于RNA二级结构的解析.

相似的思路也被应用于RNA修饰位点的鉴定. RNA修饰的丰度通常较低，为了全面深入地对RNA修饰进行研究，目前已有多种不同原理的RNA修饰位点鉴定方法被开发. 其中一种经典的方法是利用特异性抗体富集具有修饰的RNA片段并进行测序，例如用于检测m^6A修饰的m^6A-seq[39]和MeRIP-seq[40]，但该方法分辨率较低, 约100~200 nt, 且可能存在非特异性结合. 为了提高对m^6A修饰位点的识别精度, 研究人员参考了CLIP-seq的实验方法，在m^6A与抗体结合后进行交联, 随后进行反转录并停止于交联位点，建立了单碱基精度鉴定m^6A修饰位点的m^6A-CLIP-seq实验技术[41].

另一个利用富集RNA片段进行研究的领域是RNA二级结构解析. 与RNA-蛋白结合位点识别技术相似, RNA二级结构同样可以通过识别反转录停止位点进行解析: 通过化合物对RNA的单链区域标记, 随后利用随机引物进行反转录, 转录延伸在标记位点停止, 可以在单碱基精度上确定被标记的未配对碱基[42]，但受限于细胞内RNA的丰度，该方法更适用于RNA二级结构的体外检测，随后icSHAPE技术利用生物素富集改善了这一缺点[43]. 该技术利用NAI-N3标记RNA中的单链核苷酸并进行反转录，然后通过生物素富集标记片段的RNA/cDNA杂交链并进行建库测序，实现鉴定细胞内全转录组的RNA二级结构状态.

2.1.3 引入突变的转录本片段富集

得益于序列比对算法的进步，可以通过测序读序准确识别转录本序列发生改变的位点. 除了识别生物体自然发生的碱基改变, 如RNA编辑位点, 还可以人为引入突变, 通过对突变位点的识别实现特定研究目标.

一个典型的应用是m^5C修饰位点识别方法RNA-BS-seq[44]，通过亚硫酸氢盐处理将未甲基化的胞嘧啶(C)转化为尿嘧啶(U)，而甲基化的胞嘧啶(m^5C)不会转化. 对转化后的RNA进行建库测序并识别其中发生转化的胞嘧啶, 可以鉴定RNA中的m^5C修饰位点.

研究人员还尝试利用引入突变的策略研究RNA与蛋白的结合位点. 例如PAR-CLIP[45]技术通过在活细胞中引入光活性核糖核苷类似物(如4sU)，使其掺入RNA分子中, 随后交联固定RNA与蛋白结合位点, 在反转录过程中交联位点会产生特异性突变(T>C)，通过高通量测序可以实现单碱基精度的结合位点识别，但转换信

号与真实发生结合的位点可能存在一定偏差.

另外, 相似的思路还可以用在RNA二级结构解析, 通过改造反转录扩增体系引入随机突变, 建立了RNA二级结构的解析方法SHAPE-MaP[46]. 与icSHAPE等方法类似, SHAPE-MaP同样需要利用SHAPE化合物对RNA的未配对碱基进行标记, 然后在反转录体系加入Mn^{2+}离子使RNA反转录酶可以跨过被标记碱基持续延伸, 但是由于无法正确识别被标记碱基, 延伸过程会随机连接核苷酸从而在合成的序列中引入突变. 通过对反转录产物的建库测序及并识别突变位点, 可以确定哪些位点的碱基处于单链状态, 从而解析RNA二级结构.

2.2 单细胞/空间转录组

常见的细胞群体水平的RNA-seq测序技术反映的是基因在细胞群的平均表达水平, 难以解析细胞亚群间的表达异质性, 限制了胚胎发育谱系分化、免疫细胞动态分群、肿瘤微环境细胞互作等复杂生理过程的研究. 为突破这一分辨率限制, 单细胞RNA测序技术(scRNA-seq)应运而生, 通过微流控系统[47](如10x Genomics Chromium)或微孔板(如Smart-seq2[48])实现单细胞分离与独立建库, 使同时开展数百至数万个细胞的并行转录组分析得以实现.

不同的单细胞捕获技术决定了它们的数据特点和适用范围. 微流控液滴技术通过油水两相系统生成液滴, 将单个细胞与携带独特细胞条形码(cell barcode)的磁珠共同包裹, 实现对RNA来源的确定[47]. 磁珠表面连接的寡核苷酸探针除了细胞条形码, 还有唯一分子标识符(unique molecular identifier, UMI, 用于校正PCR扩增偏差以实现分子绝对定量)以及oligo(dT)序列(特异性捕获mRNA的poly(A)尾)构成完整的单细胞分子标签, 实现对单个细胞内RNA的捕获和标记[47]. 该技术实现了高通量捕获细胞, 细胞捕获量约10^3~10^4, 但仅能检测mRNA的部分片段, 且基因覆盖度受限[47]. 微孔板分离单细胞(Smart-seq2)的技术通过显微操作或微孔芯片物理隔离获取单细胞, 可以实现转录本全序列测序, 且基因覆盖度优于基于微流控液滴技术的单细胞测序方法, 但是其细胞捕获量受限[48]. 因此, 利用微流控液滴技术搭建的单细胞测序平台(10x Genomics为代表), 通常用于大规模细胞群体的测序, 有利于构建复杂细胞群体表达图谱(如免疫细胞), 解释复杂细胞群体中不同细胞类型的表达差异, 值得注意的是10x Genomics平台的测序数据由于不同的建库策略会导致读序在基因区域有5′或3′的分布偏好性, 可以用于单细胞精度鉴定加尾位点[49]. 而Smart-seq2等利用微孔板分离的单细胞测序技术, 虽然通量较低, 但是其读序一般可覆盖整个基因区域, 多用于单细胞水平可变剪接、融合基因、点突变等复杂结构变异的研究.

同时, 单细胞测序技术的发展也极大地降低了空间转录组(spatial transcriptomics)测序的技术难度和实验成本. 早期空间转录组通过RNA原位测序实现, 操作复杂且检测范围有限. 随着单细胞测序技术的出现, 可以利用预先设计有空间条形码的芯片(如10x Genomics Visium平台), 将组织切片铺在载玻片上捕获释放的mRNA, 利用反转录和扩增获得转录组表达信息. 每个捕获区域(spot)具有唯一条形码, 能够将测序数据和样品的空间位置对应起来. 以芯片和单细胞测序技术为基础的空间转录组可以有效揭示区域性基因表达和细胞间互作, 在研究具有复杂功能结构的器官时具有独特优势. 在大脑中, 空间转录组可以解析不同脑区的基因表达差异, 揭示神经元、胶质细胞等细胞类型的空间分布及其功能网络[50]; 在发育生物学中, 空间转录组技术能够追踪器官发育过程中基因表达的空间动态, 揭示导致器官发育异常的分子机制[51], 为精准医疗和基础研究提供关键数据支持.

单细胞测序/空间转录组技术推动转录组研究从“群体模糊侧写”迈向“单细胞精准解析”, 极大地推动了对复杂生物系统转录组的认识, 为胚胎发育、免疫细胞分化、肿瘤微环境、神经科学等领域的研究提供了有效技术支持.

2.3 长读长测序技术

传统二代测序技术在建库过程中需要打断, 测序读长通常为50~500 bp, 难以准确组装基因组中的重复区域和结构变异区域, 也无法通过转录本全序列测序获得相关区域的转录本异构体表达、复杂剪接位点的情况, 因此研究人员开发出长读长测序技术, 实现对单个核酸分子的全长测序. 不同的长读长测序平台与测序样品的组合展现出不同的数据特点, 数据长度中位数从0.6~2.1 kb不等, 最长可达4 Mb[19,52].

目前较为成熟的长读长测序平台主要有Pacific Biosciences(PacBio)公司的单分子实时测序技术(Single-molecule Real-time, SMRT)[53]和Oxford Nanopore公司(Oxford Nanopore Technologies, ONT)的纳米孔测序

技术[54]. PacBio公司的实时测序技术[53]使用的信号识别系统与二代技术类似，通过合成过程中不同碱基的荧光信号识别碱基类型. 另外在建库过程中将目标分子的双链两端加接头形成环状分子实现滚环测序(circular consensus sequencing, CCS), 通过对同一目标分子的多次测序降低测序错误率(<0.1%)[55]. Nanopore公司的纳米孔测序则利用核苷酸分子通过纳米孔时产生的不同电流变化识别序列中不同的碱基类型，并且还可以利用电流信号的差异直接检测RNA修饰位点[56]. 但由于没有类似SMRT技术的CCS测序矫正机制，其测序准确度低，且依赖碱基识别(Base calling)算法[57]的效果, 目前准确度可达到99%以上[58].

三代测序克服了二代测序读序短、无法区分重复序列的缺点，更加适合于复杂基因组组装、转录本异构体鉴定比较、基因组/转录本结构变异等分析，实现了高质量的基因组组装，推动了基因结构变异相关疾病的临床诊断[59]. 在转录调控研究领域, 三代长读长测序突破了二代测序依赖读段拼接的局限，直接获得完整转录本信息，成功用于研究新生转录本、选择性加尾、可变(反向)剪接[60]、RNA修饰[61]、RNA结构[62]等转录调控机制，实现了对转录组更加精细全面的解析[52]. 研究人员还利用三代测序探索等位基因特异性表达和转录本结构变异与疾病的关系，并结合长读长基因组、甲基化、表观组等数据解析复杂表型疾病[63,64]. 另外研究人员也推动三代测序技术应用于单细胞/空间转录组测序，解析癌症中的单细胞水平的结构变异, 在单细胞水平实现对转录本更加全面地检测[65].

总的来说, 针对不同的研究目的, 可以利用不同的RNA富集手段和建库测序技术，实现对转录组的表达水平、RNA剪接、RNA编辑、RNA修饰、RNA互作、RNA结构等不同层面的研究.

3 转录组分析中的生物信息学方法

针对不同的转录组测序数据，研究人员开发了不同的分析流程，为转录组研究提供可靠的分析结果(图2(b)). 不同类型转录数据分析的核心均是测序读序(reads)数据的比对和定量，根据参考基因组信息将测序读序比对到正确的基因组位置，目前已有许多成熟的比对工具，如: BWA[66]、Bowtie1/2[67,68]、TopHat2[69]、Hisat2[70~72]、STAR[73]等，可以根据不同数据特性实现测序数据准确快速地比对，并应用于下游分析.

3.1 针对转录本全序列测序数据的分析

转录本全序列测序数据根据比对后的读序坐标和基因注释信息，即可利用featureCounts、DESeq2、edgeR等工具完成基本的基因表达定量和差异分析[20,74]. 对于特定方法富集的新型转录本，则需要从头组装转录本, 并与不同富集方法的测序数据交叉比较, 确定新的转录本注释信息. 例如Fib-RIP-seq富集snoRNA进行测序并鉴定到的sno-lncRNA和*SPA*，它们均是含有snoRNA的新型lncRNA分子. 其中sno-lncRNA的两端均为snoRNA，通过Fib-RIP-seq测序数据可以检测到snoRNA之间区域的表达，而*SPA*的5′端是snoRNA其3′端是poly(A)尾，通过Fib-RIP-seq和poly(A)+ RNA-seq可以观察到除已有的基因注释区域外，存在连续的转录活跃区间, 帮助鉴定新型转录本[25].

转录本全序列测序数据还可用于可变剪接的分析定量，研究人员已经建立cuffdiff2、JunctionSeq、rMATS、LeafCutter、SUPPA2、MAJIQ等十余种工具, 实现基于转录本异构体或外显子/剪接事件数据覆盖度差异的可变剪接事件识别及定量. 以rMATS为例[75~77], 通过RNA-Seq数据的比对结果与基因组注释信息，精确识别各个外显子是否参与剪接事件，进而确定不同剪接事件的类型. 具体来说，rMATS通过分析不同RNA-Seq数据中跨剪接位点读序的比对情况，能够准确地识别出哪些外显子参与了剪接，并标记剪接的具体位置. 结合注释信息, rMATS进一步确认多种剪接事件类型, 主要包括跳跃外显子、内含子保留、选择性5′剪接和选择性3′剪接. 这些剪接事件可以通过计算特定剪接事件在不同样本中的读序数量进行定量，进而评估剪接模式的偏好和强度变化. 另外有多项研究系统比较了用于研究可变剪接事件的多种工具的性能，显示基于外显子覆盖度的方法以及MAJIQ和rMATS在多个衡量指标均表现良好[78~80].

除了经典的可变剪接，大量通过外显子反向剪接形成的环形RNA可以在poly(A)– RNA-seq或Ribo–/RNase R RNA-seq中检测到[24]. 对外显子反向剪接形成的环形RNA(circRNA)，已建立了代表性鉴别定量流程CIRCexplorer系列[6,7,81]，该方法巧妙地利用TopHat-fusion识别融合基因的能力，结合转录组注释信息构建潜在的反向剪接转录本注释，识别测序数据中跨越反向剪接位点的读序，并根据该读序来源的外显子位置确定环形RNA信息并进行定量，为环形RNA的全面研

究提供了有力工具. 其他的一些环形RNA鉴定分析流程还有CIRI2[82]、DCC[83]和MapSplice[84]等, 均基于相似的原理实现环形RNA的鉴别与定量. 研究人员对这些鉴定工具也进行过系统的比较和验证, 为环形RNA的全面研究提供参考[24,85].

RNA编辑(A-to-I editing)位点的鉴定同样可以在常规转录组RNA-seq样本中实现. 例如, 可一次性鉴定12种不同类型RNA突变(主要是A-to-I 编辑)的新型计算分析流程RADAR (RNA-editing analysis-pipeline to decode all-twelve-types of RNA-editing)[86], 同时还可准确高效地检测碱基编辑中存在的RNA水平脱靶效应. 分析流程中除了测序读序的精确匹配, 还需要识别存在多个错配(mismatch)的读序, 并通过已有数据库排除SNP位点, 根据覆盖度、编辑效率等阈值综合筛选, 确定可信的RNA编辑位点.

3.2 针对转录本片段测序数据的分析

片段富集的建库测序数据可根据分辨率差异分为区间精度和单碱基精度, 对可变加尾、RNA修饰、与蛋白质互作、RNA二级结构解析等不同研究目标, 在对富集片段进行序列比对后, 根据不同的分析目标和数据特性进行下游分析.

用于鉴定RNA加尾位点的Poly(A)-ClickSeq[31]技术, 通过测序读序中连续A碱基上游的读序进行序列比对, 比对到基因组单一位置的读序末端(转录方向下游)视为加尾位点. 为排除基因组中A富集区域造成的假阳性, 基因转录方向下游含有连续A的位点被剔除, 然后将临近区间内的读序定义为源于相同的加尾位点.

利用抗体富集RNA修饰位点、识别RNA与蛋白质互作区域的CLIP-seq[32,33]等技术均通过测序读序比对到的区间鉴定互作区域并根据相应区间的读序丰度进行进一步筛选. 例如Piranha[87]是针对CLIP-seq数据建立的分析方法, 在比对过程中允许读序含有3个错配并保留具有单一比对位置的读序. 随后利用零截断负二项分布(zero-truncated negative binomial)拟合读序分布情况, 根据拟合情况计算p值, 区分背景和可信结合位点. 针对单碱基精度识别蛋白互作位点的iCLIP-seq[34]/eCLIP-seq[35]数据, 研究人员利用隐马尔可夫模型建立分析方法PureCLIP[88]. 该方法认为iCLIP-seq/eCLIP-seq数据中相邻碱基覆盖度非独立分布, 碱基可以分为四种隐藏状态: (1) 非富集+非交联, (2) 非富集+交联, (3) 富集+非交联, (4) 富集+交联. PureCLIP使用读序分布情况和读序开始位点在基因组上的覆盖数用于建模, 并利用空白对照排除假阳性, 同时通过交联相关的特征序列排除非特异性交联序列, 获得可靠的RNA蛋白结合位点.

对于利用反转录停止位置解析RNA二级结构的icSHAPE[43]等方法, 通常通过序列比对识别反转录停止位点, 并统计单碱基覆盖度, 获得单碱基精度的SHAPE化合物标记强度. icSHAPE-pipe[89]流程首先去除rRNA等高峰度的非目标RNA, 然后比对到基因组上, 对每个位点的覆盖度进行统计, 利用标记组和对照组计算标记强度并进行标准化处理, 实现细胞内全转录组的二级结构解析.

分析用于识别RNA-RNA相互作用序列的测序数据需要进行允许缺口(gap)的读序比对, 以便确定互作RNA区域的位置. 例如用于分析PARIS[90]的流程利用比对工具STAR实现了含有拼接序列的测序数据的比对, 成功比对到基因组参考序列之后去除由正常剪接产生的测序数据, 其他拼接方式的测序数据可以显示不同RNA分子间的相互作用区域. 另外将测序数据比对到RNA参考序列还可以显示同一RNA内部的相互作用区域, 用于解析RNA二级结构.

3.3 针对含有引入突变的转录本片段测序数据的分析

识别测序数据中突变位点最典型的方法是RNA编辑位点识别. 利用该方法可以人为引入碱基改变并识别该位点用于分析, 如RNA修饰位点可以利用修饰位点碱基转换测序数据的读序比对结果进行鉴定. 例如, 利用亚硫酸氢盐转换C到U, 但保留发生m^5C修饰的C. 由于转化效率的差异, 存在C不完全转化到U的情况, 对序列比对的准确性造成影响. 为了获得可靠的m^5C修饰位点, 研究人员建立针对性的分析流程如meRanTK[91], 将测序数据的C转变为T(U)然后将其比对到C转变为T(U)的转录组参考序列, 保留可以比对到唯一确定位置的读序进行进一步的分析, 确定发生修饰的m^5C位点及其可信度.

针对PAR-CLIP数据, 研究者建立了PARalyzer[92]通过识别测序数据中可靠的T>C转换位点识别RNA与蛋白的结合位置. 首先对测序读序进行允许两个错配的比对, 根据参考序列筛选可能由T>C转换造成错配的读序, 将这些读序的错配碱基转换回T后再次进行比

对，保留具有唯一比对位置的读序用于结合位点的鉴定, 并进行下游统计分析.

如上文所述，RNA二级结构同样可以通过引入突变的方式进行解析，但与RNA编辑和RNA修饰中的点突变不同，通过反转录引入的变异要比上述测序方法中的突变更为复杂，它包括碱基的突变、缺失、插入以及复杂的混合情况[46,93]. 研究人员建立的分析流程ShapeMapper2[94]利用Bowtie2中的局部比对模式识别测序读序中这些复杂的碱基突变情况，但对序列中的连续相同碱基的标记位置确认仍存在偏差. 同时通过提高目标RNA整体的测序数据覆盖度，可以一定程度上提高对标记位点鉴定的准确度[94]. 通过对含有突变的读序的识别和统计，可以得到每个碱基位点的突变率，结合空白对照组突变率可以获得单碱基精度的标记强度, 用于RNA二级结构建模.

3.4 针对单细胞/空间转录组测序数据的分析

单细胞数据的分析除了将读序比对到基因组上正确位置，还需要根据细胞条形码(Cell Barcode)区分读序来源的细胞, 并根据UMI去除PCR扩增偏差, 最终生成细胞-基因表达矩阵. 10x Genomics公司发布的Cell-Ranger是用于单细胞测序数据初步比对分析的主流工具，可以直接完成原始测序数据到细胞-基因表达矩阵的分析. 在生成矩阵后可以利用Seurat[95]或Scanpy[96]进行下游分析, 通常包括: (1) 质控: 对单细胞数据进行基本的质控, 筛选可用于分析的细胞, 包括线粒体基因占比、UMI计数量、表达基因的数量等指标; (2) 筛选高变异系数基因: 筛选出可用于分析的细胞后, 需要对表达矩阵进行标准化，如果有多批单细胞数据还需要对整合的数据进行批次效应校正，然后筛选高变异系数的基因用于后续分析; (3) 降维、聚类、细胞注释: 在获得高变异系数基因后对其进行主成分分析(PCA)并利用该结果进行细胞聚类和可视化展示细胞分群情况，然后根据不同细胞群特异表达的基因确定细胞类型. 目前也有一系列自动注释细胞类群的工具如SingleR[97]、singleCellNet[98]、CellTypist[99]等，对不同的工具进行比较分析显示不同细胞类型注释方法对不同来源的数据和不同的细胞类型准确性各异[100]，需要研究者谨慎对待标注结果.

除了以上的常规分析步骤，还有多种下游分析方向可应用于单细胞数据进行进一步挖掘. 如比较细胞数量或基因表达的差异剖析不同细胞类型或生理病理样本变化; 拟时序分析可用于推测干细胞或免疫细胞分化的过程[101]; RNA速率分析(RNA Velocity)可实现预测细胞短期状态转变[102]; 细胞通讯分析工具(如CellPhoneDB[103])通过统计不同细胞类型中受体和配体的表达及配对情况，推断不同细胞之间的相互作用. 同时单细胞还可以与细胞群体RNA测序(bulk RNA-seq)数据或其他组学数据进行联合分析, 获得更丰富的信息.

空间转录组数据的比对与单细胞类似，但含有空间条形码的芯片可以提供单细胞之间的位置关系，可以对比不同位置的细胞类型和基因表达差异，对肿瘤发生、神经科学、发育生物学等众多领域的研究都有重要意义. 由于测序芯片的分辨率和捕获效率较低, 空间转录组的测序数据覆盖度较低且每个spot难以达到单细胞精度，分析过程中的一大挑战是正确地还原每个spot的表达谱. 目前已经发展了多种方法用于减弱邻近spots之间的交叉污染，矫正空间数据并增强数据分辨率[104~107]. 获得可靠细胞的空间表达谱后, 可以进行空间细胞通讯、发育过程结构变化等下游分析，为转录组研究提供组织/器官在空间维度的知识.

3.5 针对三代长读长测序数据的分析

不同于二代测序技术的高准确性，三代测序数据存在多种测序错误类型, 包括替换、插入、缺失等, 导致现有的二代测序比对方法不适合用于长读长测序数据的处理[108]. 目前已经有一些方法可以用于长读长测序数据的比对，如minimap2[109]基于稀疏哈希索引实现长读长数据比对和剪接比对，GraphMap2[110]基于图比对(graph-based)思路可处理重复区域和复杂变异产生的读序. 随着长读长测序技术的发展, 配套分析工具已逐步完善并形成多维度应用实现不同分析目标，包括基因组组装(Hifiasm、Flye、Canu等)[111]，转录本异构体的鉴定(Cupcake、TAMA、TALON、FLAIR等)，剪接事件分析(SUPPA)，转录本定量(Salmon、FLAIR、TALON等)，转录本差异分析(DRIMSeq)，RNA修饰位点识别(m6Anet、EpiNano、Tombo等)，结构变异鉴定(Sniffles2、MAJIQ)等[52,56]. 也有研究对长读长测序数据分析工具进行分类总结，包含了质检、比对、组装、定量等的分析工具并收录于long-read-tools.org[112]. 长读长测序技术和相关分析工具的开发仍在不断发展中，随着建库测序方法的优化和算法的改进，长读长测序技术将会得到更加广泛的应用.

4 人工智能技术促进转录组数据挖掘

人工智能技术对高通量测序数据的解析有巨大的帮助, 其在生物学研究中的应用主要分为两个阶段: 在技术发展初期(2010年前后), 研究者主要基于传统机器学习方法构建分析框架: (1) 非监督学习方法如k-means聚类、主成分分析(PCA)和非负矩阵分解(NMF)等被广泛应用于基因表达模式识别; (2) 监督学习算法包括决策树、随机森林、支持向量机(SVM)和K近邻(KNN)算法等在样本分类预测中展现优势; (3) 半监督策略如自训练算法和生成式模型则有效缓解了标注数据稀缺的困境. 随着神经网络架构的突破性进展, 深度学习技术已成为领域的研究焦点.

近年来, 深度学习模型被应用于生物研究的多个领域, 例如利用CNN或RNN模型通过DNA序列预测基因表达量(DeepBind[113]和DeepSEA[114]); 基于Transformer或LSTM的模型预测RNA剪接位点(SpliceAI[115]); 预测不同细胞系中RNA与蛋白结合情况(PrismNet[116]); 预测蛋白结构和RNA结构以及互作情况(AlphaFold3等); 单细胞数据细胞类型注释工具(scANVI); 基于结构的蛋白或RNA生成模型(RFAA[117]、RfamGen[118]). 深度学习技术已经进入生物学研究的各个领域, 加速了生命现象机制的解析进程.

当前的深度学习模型有多种架构应用于不同的生物学问题, 但高质量的训练数据是人工智能模型性能的决定性因素[119]. 高通量测序技术产生的海量组学数据为深度学习模型提供了丰富的训练数据, 但由于生物数据的复杂性, 训练数据的选择需要综合考虑数据的规模、多样性和准确性. 其次, 输入数据的编码方式直接影响着人工智能模型对转录组数据的理解和处理效率.

4.1 训练数据的选择

训练数据的选择对模型的效果起到关键作用, 合适的训练数据可以帮助模型更好地学习数据特征, 进而提高模型泛化能力. 通常可以从数据量、准确性、平衡性、多样性等方面衡量数据质量. 目前产生的大量转录组数据可以提供丰富的训练数据, 利用传统生信分析方法进行准确分析整合是数据准确性的保障. 在训练数据的初步筛选中, 通常先行去除低质量数据、矫正技术偏差(如测序批次和平台差异), 并进行标准化等预处理. 值得注意的是, 虽然高通量测序数据和分析流程可以为模型提供大量准确的训练数据, 但绝大多数数据为正例(positive data), 在训练过程中容易发生数据不平衡, 可以利用过采样、欠采样或损失函数加权等方法解决类别不平衡问题, 或者利用生成对抗网络(GANs)生成合成数据. 在筛选相应负例(negative data)过程要注意避免引入未知的正例造成模型性能的下降, 这一过程很大程度上依赖研究人员对相应生物数据的理解, 另外可以利用半监督学习剔除负例中的潜在正例, 提高负例的准确性.

同时根据训练数据和实际输入数据的差异, 对训练数据进行恰当的处理可以提高模型的准确性. 例如, 用于鉴定3′-tag单细胞转录组数据中PAS的工具SCAPTURE中内置的深度学习模型DeepPASS(图3(a))[49]可以通过序列特征鉴定该位置是否含有加尾位点. 为了实现对3′-tag单细胞转录组数据中PAS位点的准确识别, 基于3′-tag单细胞转录组的数据分布特征, 选取训练数据时对PAS位点上下游序列进行符合正态分布的偏移采样, 避免DeepPASS模型仅可识别位于序列中间位置的PAS位点. 相较于训练数据未进行偏移采样的模型DeepPAS-fixed, DeepPASS模型的识别准确率明显提高, 实现了对单细胞转录组数据中PAS位点的准确鉴定. 这一例子说明了恰当的训练数据构建对模型效果的影响, 在利用转录组数据建立模型时需要考虑不同数据集的特征, 充分利用数据的信息.

4.2 数据的编码方式

在深度学习中, 编码是将输入数据(通常是离散的, 如文本或类别特征)转换为连续向量表示的过程. 根据建立人工智能模型的不同任务目标, 输入的训练数据有所差异, 转录组数据用于模型训练的输入形式通常有两种: 序列和表达量.

利用序列信息构建的转录组模型通常使用RNA序列进行编码并根据已知分类标签进行学习, 实现对未知序列的定位、功能等信息的预测. 在序列功能预测的模型建立过程中, 最常用的方法是独热编码(one-hot encoding), 该方法将每一个类别用一个向量表示, RNA序列由AUCG四种核糖核苷酸组成, 其表示方式简单但由于不包含顺序信息, 可能会遗漏相应特征. 为了弥补独热编码的这一缺陷, 研究人员通过将RNA序列分解为长度为k的子序列(称为k-mer), 然后对每个k-mer进行编码, 称为k-mer编码(k-mer encoding), 相较于独热编码它可以捕捉局部的碱基序列特征, 一定程度上为模型提供序列上下游信息.

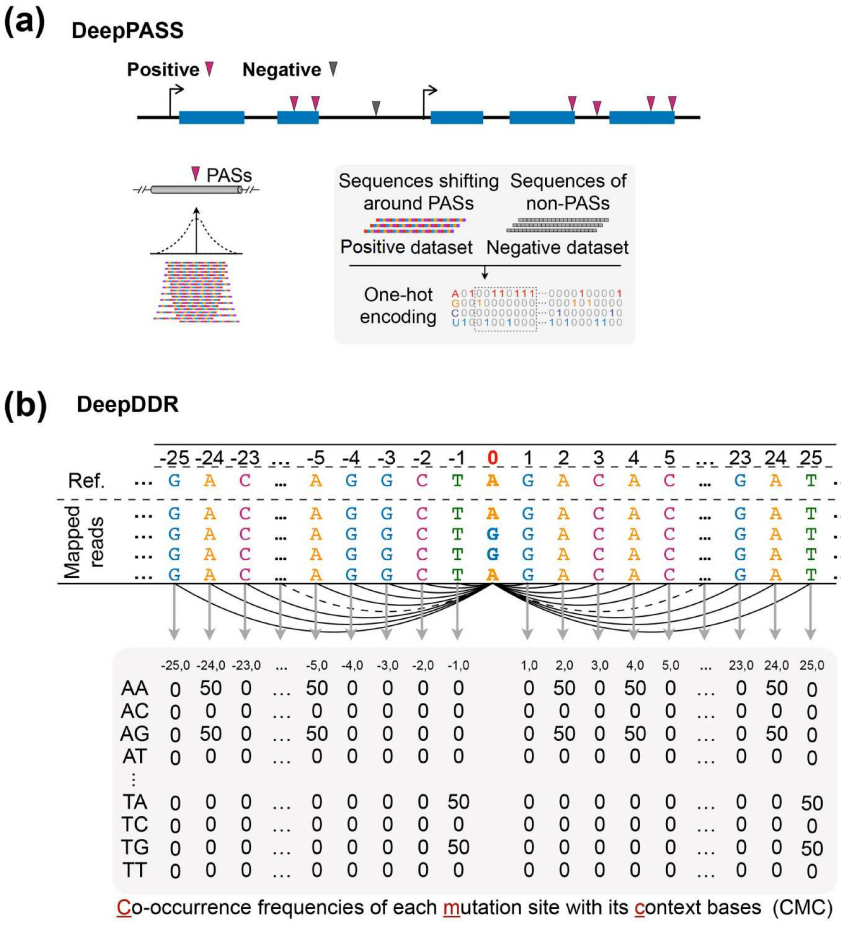

图 3 转录组人工智能模型中的数据筛选与编码方式. (a) 单细胞加尾位点识别模型SCAPTURE的训练数据构建方法. SCAPTURE使用数据库的已知加尾位点上下游序列(200 nt)作为正例, 从基因间区随机筛选200 nt序列作为负例. 正例中的加尾位点选取上下游序列时遵循正态分布进行偏移, 使其符合单细胞数据读序的分布特征, 提升模型性能. (b) 转录组突变位点鉴定模型DeepDDR使用共变频率矩阵(CMC)编码突变位点及其上下游序列. CMC以突变位点为中心, 统计测序数据中突变位点与上下游各25 nt形成不同碱基对(16种)出现的频率, 每个突变位点的16×50 CMC矩阵为模型的输入数据. (c) 单细胞预训练模型Geneformer的训练数据编码方式. Geneformer使用秩编码, 即基因在不同细胞中表达水平的排序, 而非实际表达值, 作为编码方式

Figure 3 Training dataset construction and encoding ways for AI model of transcriptome. (a) Training sets construction for DeepPASS. Positive training set were 200-nt sequences extracted surrounding known PASs, which shifted around these PASs with a normal distribution. These shifting sequences accorded with single-cell sequencing reads distribution on gene loci. Negative sequences were constructed with 200 bp sequences randomly extracted in intergenic regions without overlapping with any of the annotated PASs. (b) Training data of DeepDDR were encoded by co-occurrence frequencies of each mutation site with its context bases (CMC). The CMC was determined by 51-nucleotide sequences of aligned RNA sequencing reads containing the mutation site in the middle. As shown in the figure, all sites excluding the mutation site were calculated frequencies of 16 dinucleotides (paired with mutation site) by aligned reads from RNA-seq data. The 16×50 CMC matrix of all mutation sites was input training sets for DeepDDR model. (c) Rank value encoding is used in Geneformer model. This method represents genes by rank of their expression

除了对序列直接编码，还可以通过嵌入(embedding)将高维的、离散的对象(如序列等)映射到低维的连续向量空间. 对输入序列进行嵌入的过程通常是模型的一部分，通过浅层神经网络学习能够使得相似的对象在向量空间中靠得更近. 例如Word2Vec是一种基于神经网络的词嵌入方法，它通过学习大量文本中的词语之间的共现关系，将每个词映射为一个低维的、稠密的向量. 通过训练，Word2Vec能够捕捉到词语的语义信息，使得相似语义的词在向量空间中距离较近. 广受关注的Transformer[120]模型将输入信息(input embedding)和位置编码(positional encoding)相结合，一同输入到后续多头自注意力(Multi-Head Self-Attention)架构中, 捕捉序列信息和位置的关系, 使模型可以同时捕捉输入元素的语义和顺序信息，提高了语言翻译、文本分类等任务的完成效果.

在实际应用过程中，研究人员还尝试改进编码方式为模型提供更多的信息，以提升转录组模型的性能. 例如DEMINING中的DNA/RNA突变位点分类模型DeepDDR(图3(b))，创新性地将突变位点上下游序列和测序读序编码，构建带注意力的双碱基上下文共轭同频矩阵(matrix of the co-occurrence frequencies of each mutation site with its context bases, CMC)，作为DeepDDR模型的编码输入. 该编码方式创新性地将突变位点与其上下游碱基信息进行组合，使DeepDDR模型不仅能够识别突变位点，还能够捕捉到这些突变在更大范围内的上下文信息，这可能是有效区分RNA编辑和DNA突变的关键所在[121].

此外研究人员尝试将序列信息转换为图像信息，用于模型训练. 例如SVision-pro算法，将样本间结构变异差异检测问题从序列层面转化为图像空间的变异实例分割问题，直接比较图像化的样本测序差异[122]. 该算法的编码方式实现了测序信息图像化表征，对图像内变异的类型、位置和等位基因频率进行高分辨率识别，保证了对全类型结构变异和复杂结构变异的全面检测.

除了将序列作为训练数据，还有一些模型将转录表达谱作为训练数据，学习不同细胞类型或生理病理状态下的表达特征. 如Geneformer[123]利用海量单细胞测序数据建立预训练模型，模型经过数据微调后可用于下游关键网络调节因子和候选治疗靶点的发现. 该模型使用秩编码(rank value encoding)处理单细胞表达谱数据后进行训练，即对单个细胞的基因表达谱进行排序，使用排序而非实际的基因表达值(图3(c))进行模型的训练. 相较于独热编码, 秩编码可以保留数据之间的顺序信息. 这种方法可以将普遍高表达的管家基因标准化到较低的等级来降低其优先级，但是转录因子等具有较高细胞表达异质性的基因则在编码中移动到更高的位置, 有利于模型识别不同的细胞状态.

5 展望

高通量测序技术和分析方法的交替发展引领转录组研究迈入新纪元，实现了从单一基因研究向多维度、多层次研究的重要转变. 这一变革不仅体现在对全转录组的系统性解析能力上，更突破了传统研究在时间、空间以及生理病理状态等方面的局限性. 从最初专注于少量特定转录本的研究，发展为能够全面揭示新型转录本、深入解析复杂的转录调控网络，并精准捕捉不同时间和空间维度下转录组的动态变化. 海量多维度数据的积累为转录组研究提供了丰富的资源库，跨组学整合分析进一步拓展了对生命系统复杂性的理解, 但同时也带来了前所未有的数据分析挑战[1]. 当前转录组研究正沿着多个前沿方向同步推进. 在技术层面, 从二代短读长测序到单分子长读长测序的突破使研究者能够更准确地解析转录异构体、重复序列以及复杂结构变异；从细胞群体到单细胞分辨率的提升正重塑我们对细胞异质性的理解，而空间转录组学的兴起则为揭示组织微环境中的转录调控提供了新视角；表观转录组学研究通过整合RNA修饰、RNA-蛋白互作及RNA三维结构数据, 正深入解析转录后调控机制, 这一领域与传统转录组学的融合将极大丰富对RNA功能多样性及调控机制的认知. 在临床转化层面, 基于液体活检的RNA标志物研究和单细胞转录组诊断技术展现出广阔前景，尤其在肿瘤异质性和免疫微环境评估方面具有独特优势. 循环核酸和外泌体RNA作为非侵入性生物标志物的研究正在快速发展，有望成为肿瘤动态变化的实时监测工具.

深度学习技术的迅猛发展为应对这些挑战提供了强有力的技术支撑. 基于深度学习的模型和方法不断涌现, 不仅能够为解析单组学数据提供新的见解, 更能实现多组学数据的整合分析与联合建模，同时在生物分子的预测与设计中展现出巨大潜力. 但深度学习技术在转录组学中的应用仍面临显著挑战: 首先, 现有模型多基于“中心法则”已包含所有生命活动信息的假设, 但实际难以全面捕捉转录组的动态复杂性; 其次, 训练数据的偏好性(如现存RNA结构数据多数是短RNA片

段)严重限制了模型的适用范围; 此外, 基于统计相关性的黑箱模型难以提供可解释的生物学机制. 未来深度学习在转录组领域的突破点可能在于多模态学习框架的建立, 将转录组数据与表观组、蛋白质组等多维信息有机整合, 并引入因果推断机制增强可解释性. 除计算方法外, 转录组研究的标准化也是亟待解决的问题, 尤其是在单细胞和空间转录组数据的批次效应消除、质量控制和整合分析方面. 此外, 临床样本的非侵入性采集技术和现场即时测序分析能力的发展将进一步推动转录组学向精准医学转化.

综合这些技术与方法的协同发展, 未来转录组学研究有望实现从静态观测向动态预测的转变, 建立对生命系统中转录组调控的时空精细网络模型, 为理解复杂疾病发生机制和开发靶向治疗策略提供全新视角.

参考文献

1 Yang L, Ulitsky I, Gilbert W V, et al. The challenges of investigating RNA function. Mol Cell, 2024, 84: 3567–3571
2 Chen L L, Kim V N. Small and long non-coding RNAs: past, present, and future. Cell, 2024, 187: 6451–6485
3 Tian B, Manley J L. Alternative polyadenylation of mRNA precursors. Nat Rev Mol Cell Biol, 2017, 18: 18–30
4 Nilsen T W, Graveley B R. Expansion of the eukaryotic proteome by alternative splicing. Nature, 2010, 463: 457–463
5 Kalsotra A, Cooper T A. Functional consequences of developmentally regulated alternative splicing. Nat Rev Genet, 2011, 12: 715–729
6 Zhang X O, Wang H B, Zhang Y, et al. Complementary sequence-mediated exon circularization. Cell, 2014, 159: 134–147
7 Zhang X O, Dong R, Zhang Y, et al. Diverse alternative back-splicing and alternative splicing landscape of circular RNAs. Genome Res, 2016, 26: 1277–1287
8 Zhang Y, Zhang X O, Chen T, et al. Circular intronic long noncoding RNAs. Mol Cell, 2013, 51: 792–806
9 Li X, Yang L, Chen L L. The biogenesis, functions, and challenges of circular RNAs. Mol Cell, 2018, 71: 428–442
10 Yang L, Wilusz J E, Chen L L. Biogenesis and regulatory roles of circular RNAs. Annu Rev Cell Dev Biol, 2022, 38: 263–289
11 Nishikura K. A-to-I editing of coding and non-coding RNAs by ADARs. Nat Rev Mol Cell Biol, 2016, 17: 83–96
12 Deffit S N, Hundley H A. To edit or not to edit: regulation of ADAR editing specificity and efficiency. WIREs RNA, 2016, 7: 113–127
13 Chen T, Xiang J F, Zhu S, et al. ADAR1 is required for differentiation and neural induction by regulating microRNA processing in a catalytically independent manner. Cell Res, 2015, 25: 459–476
14 Yuan J, Xu L, Bao H J, et al. Biological roles of A-to-I editing: implications in innate immunity, cell death, and cancer immunotherapy. J Exp Clin Cancer Res, 2023, 42: 149
15 Li S, Mason C E. The pivotal regulatory landscape of RNA modifications. Annu Rev Genom Hum Genet, 2014, 15: 127–150
16 Li K, Peng J, Yi C. Sequencing methods and functional decoding of mRNA modifications. Fundamental Res, 2023, 3: 738–748
17 Sun W J, Li J H, Liu S, et al. RMBase: a resource for decoding the landscape of RNA modifications from high-throughput sequencing data. Nucleic Acids Res, 2016, 44: D259–D265
18 Goodwin S, McPherson J D, McCombie W R. Coming of age: ten years of next-generation sequencing technologies. Nat Rev Genet, 2016, 17: 333–351
19 Scarano C, Veneruso I, De Simone R R, et al. The third-generation sequencing challenge: novel insights for the omic sciences. Biomolecules, 2024, 14: 568
20 Mortazavi A, Williams B A, McCue K, et al. Mapping and quantifying mammalian transcriptomes by RNA-Seq. Nat Methods, 2008, 5: 621–628
21 Yin Q F, Yang L, Zhang Y, et al. Long noncoding RNAs with snoRNA ends. Mol Cell, 2012, 48: 219–230
22 Xing Y H, Yao R W, Zhang Y, et al. SLERT regulates DDX21 rings associated with pol I transcription. Cell, 2017, 169: 664–678.e16
23 Wu M, Xu G, Han C, et al. lncRNA *SLERT* controls phase separation of FC/DFCs to facilitate Pol I transcription. Science, 2021, 373: 547–555
24 Ma X K, Zhai S N, Yang L. Approaches and challenges in genome-wide circular RNA identification and quantification. Trends Genet, 2023, 39: 897–907
25 Wu H, Yin Q F, Luo Z, et al. Unusual processing generates SPA lncRNAs that sequester multiple RNA binding proteins. Mol Cell, 2016, 64: 534–548
26 Liu S, Huang J, Zhou J, et al. NAP-seq reveals multiple classes of structured noncoding RNAs with regulatory functions. Nat Commun, 2024, 15: 2425
27 Li B, Liu S, Zheng W, et al. RIP-PEN-seq identifies a class of kink-turn RNAs as splicing regulators. Nat Biotechnol, 2024, 42: 119–131
28 Fuchs G, Voichek Y, Benjamin S, et al. 4sUDRB-seq: measuring genomewide transcriptional elongation rates and initiation frequencies within

cells. Genome Biol, 2014, 15: R69
29 Fuchs G, Voichek Y, Rabani M, et al. Simultaneous measurement of genome-wide transcription elongation speeds and rates of RNA polymerase II transition into active elongation with 4sUDRB-seq. Nat Protoc, 2015, 10: 605–618
30 Xia Z, Donehower L A, Cooper T A, et al. Dynamic analyses of alternative polyadenylation from RNA-seq reveal a 3′-UTR landscape across seven tumour types. Nat Commun, 2014, 5: 5274
31 Routh A, Ji P, Jaworski E, et al. Poly(A)-ClickSeq: click-chemistry for next-generation 3′-end sequencing without RNA enrichment or fragmentation. Nucleic Acids Res, 2017, 45: e112
32 Licatalosi D D, Mele A, Fak J J, et al. HITS-CLIP yields genome-wide insights into brain alternative RNA processing. Nature, 2008, 456: 464–469
33 Yeo G W, Coufal N G, Liang T Y, et al. An RNA code for the FOX2 splicing regulator revealed by mapping RNA-protein interactions in stem cells. Nat Struct Mol Biol, 2009, 16: 130–137
34 König J, Zarnack K, Rot G, et al. iCLIP reveals the function of hnRNP particles in splicing at individual nucleotide resolution. Nat Struct Mol Biol, 2010, 17: 909–915
35 Van Nostrand E L, Pratt G A, Shishkin A A, et al. Robust transcriptome-wide discovery of RNA-binding protein binding sites with enhanced CLIP (eCLIP). Nat Methods, 2016, 13: 508–514
36 Helwak A, Tollervey D. Mapping the miRNA interactome by cross-linking ligation and sequencing of hybrids (CLASH). Nat Protoc, 2014, 9: 711–728
37 Aw J G A, Shen Y, Wilm A, et al. *In vivo* mapping of eukaryotic RNA interactomes reveals principles of higher-order organization and regulation. Mol Cell, 2016, 62: 603–617
38 Lu Z, Zhang Q C, Lee B, et al. RNA duplex map in living cells reveals higher-order transcriptome structure. Cell, 2016, 165: 1267–1279
39 Dominissini D, Moshitch-Moshkovitz S, Schwartz S, et al. Topology of the human and mouse m6A RNA methylomes revealed by m6A-seq. Nature, 2012, 485: 201–206
40 Meyer K D, Saletore Y, Zumbo P, et al. Comprehensive analysis of mRNA methylation reveals enrichment in 3′ UTRs and near stop codons. Cell, 2012, 149: 1635–1646
41 Hsu P J, He C. High-resolution mapping of N(6)-methyladenosine using m(6)A crosslinking immunoprecipitation sequencing (m(6)A-CLIP-Seq). Methods Mol Biol, 2019, 1870: 69–79
42 Lucks J B, Mortimer S A, Trapnell C, et al. Multiplexed RNA structure characterization with selective 2′-hydroxyl acylation analyzed by primer extension sequencing (SHAPE-Seq). Proc Natl Acad Sci USA, 2011, 108: 11063–11068
43 Flynn R A, Zhang Q C, Spitale R C, et al. Transcriptome-wide interrogation of RNA secondary structure in living cells with icSHAPE. Nat Protoc, 2016, 11: 273–290
44 Motorin Y, Lyko F, Helm M. 5-methylcytosine in RNA: detection, enzymatic formation and biological functions. Nucleic Acids Res, 2010, 38: 1415–1430
45 Danan C, Manickavel S, Hafner M. PAR-CLIP: a method for transcriptome-wide identification of RNA binding protein interaction sites. Methods Mol Biol, 2016, 1358: 153–173
46 Smola M J, Rice G M, Busan S, et al. Selective 2′-hydroxyl acylation analyzed by primer extension and mutational profiling (SHAPE-MaP) for direct, versatile and accurate RNA structure analysis. Nat Protoc, 2015, 10: 1643–1669
47 Zheng G X Y, Terry J M, Belgrader P, et al. Massively parallel digital transcriptional profiling of single cells. Nat Commun, 2017, 8: 14049
48 Picelli S, Faridani O R, Björklund Å K, et al. Full-length RNA-seq from single cells using Smart-seq2. Nat Protoc, 2014, 9: 171–181
49 Li G W, Nan F, Yuan G H, et al. SCAPTURE: a deep learning-embedded pipeline that captures polyadenylation information from 3′ tag-based RNA-seq of single cells. Genome Biol, 2021, 22: 221
50 Li Y, Li Z, Wang C, et al. Spatiotemporal transcriptome atlas reveals the regional specification of the developing human brain. Cell, 2023, 186: 5892–5909.e22
51 Sanchez-Ferras O, Pacis A, Sotiropoulou M, et al. A coordinated progression of progenitor cell states initiates urinary tract development. Nat Commun, 2021, 12: 2627
52 Monzó C, Liu T, Conesa A. Transcriptomics in the era of long-read sequencing. Nat Rev Genet, 2025, 37: 1155–1162
53 Eid J, Fehr A, Gray J, et al. Real-time DNA sequencing from single polymerase molecules. Science, 2009, 323: 133–138
54 Deamer D, Akeson M, Branton D. Three decades of nanopore sequencing. Nat Biotechnol, 2016, 34: 518–524

55 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. Nat Biotechnol, 2019, 37: 1155–1162

56 Amarasinghe S L, Su S, Dong X, et al. Opportunities and challenges in long-read sequencing data analysis. Genome Biol, 2020, 21: 30

57 Rang F J, Kloosterman W P, de Ridder J. From squiggle to basepair: computational approaches for improving nanopore sequencing read accuracy. Genome Biol, 2018, 19: 90

58 Wang Y, Zhao Y, Bollas A, et al. Nanopore sequencing technology, bioinformatics and applications. Nat Biotechnol, 2021, 39: 1348–1365

59 Owusu R, Savarese M. Long-read sequencing improves diagnostic rate in neuromuscular disorders. Acta Myol, 2023, 42: 123–128

60 Qin Y, Long Y, Zhai J. Genome-wide characterization of nascent RNA processing in plants. Curr Opin Plant Biol, 2022, 69: 102294

61 Wongsurawat T, Jenjaroenpun P, Nookaew I. Direct sequencing of RNA and RNA modification identification using nanopore. Methods Mol Biol, 2022, 2477: 71–77

62 Aw J G A, Lim S W, Wang J X, et al. Determination of isoform-specific RNA structure with nanopore long reads. Nat Biotechnol, 2021, 39: 336–346

63 Glinos D A, Garborcauskas G, Hoffman P, et al. Transcriptome variation in human tissues revealed by long-read sequencing. Nature, 2022, 608: 353–359

64 Vollger M R, Korlach J, Eldred K C, et al. Synchronized long-read genome, methylome, epigenome and transcriptome profiling resolve a Mendelian condition. Nat Genet, 2025, 57: 469–479

65 Gupta P, O'neill H, Wolvetang E J, et al. Advances in single-cell long-read sequencing technologies. NAR Genomics BioInf, 2024, 6: lqae047

66 Li H, Durbin R. Fast and accurate short read alignment with Burrows–Wheeler transform. Bioinformatics, 2009, 25: 1754–1760

67 Langmead B, Trapnell C, Pop M, et al. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol, 2009, 10: R25

68 Langmead B, Salzberg S L. Fast gapped-read alignment with Bowtie 2. Nat Methods, 2012, 9: 357–359

69 Kim D, Pertea G, Trapnell C, et al. TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. Genome Biol, 2013, 14: R36

70 Kim D, Langmead B, Salzberg S L. HISAT: a fast spliced aligner with low memory requirements. Nat Methods, 2015, 12: 357–360

71 Pertea M, Kim D, Pertea G M, et al. Transcript-level expression analysis of RNA-seq experiments with HISAT, StringTie and Ballgown. Nat Protoc, 2016, 11: 1650–1667

72 Kim D, Paggi J M, Park C, et al. Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. Nat Biotechnol, 2019, 37: 907–915

73 Dobin A, Davis C A, Schlesinger F, et al. STAR: ultrafast universal RNA-seq aligner. Bioinformatics, 2013, 29: 15–21

74 Stark R, Grzelak M, Hadfield J. RNA sequencing: the teenage years. Nat Rev Genet, 2019, 20: 631–656

75 Shen S, Park J W, Huang J, et al. MATS: a Bayesian framework for flexible detection of differential alternative splicing from RNA-Seq data. Nucleic Acids Res, 2012, 40: e61

76 Shen S, Park J W, Lu Z, et al. rMATS: robust and flexible detection of differential alternative splicing from replicate RNA-Seq data. Proc Natl Acad Sci USA, 2014, 111: E5593–E5601

77 Wang Y, Xie Z, Kutschera E, et al. rMATS-turbo: an efficient and flexible computational tool for alternative splicing analysis of large-scale RNA-seq data. Nat Protoc, 2024, 19: 1083–1104

78 Mehmood A, Laiho A, Venäläinen M S, et al. Systematic evaluation of differential splicing tools for RNA-seq studies. Brief BioInf, 2020, 21: 2052–2065

79 Vaquero-Garcia J, Aicher J K, Jewell S, et al. RNA splicing analysis using heterogeneous and large RNA-seq datasets. Nat Commun, 2023, 14: 1230

80 Kubota N, Chen L, Zheng S. Shiba: a versatile computational method for systematic identification of differential RNA splicing across platforms. Nucleic Acids Res, 2025, 53: gkaf098

81 Ma X K, Wang M R, Liu C X, et al. CIRCexplorer3: a CLEAR pipeline for direct comparison of circular and linear RNA expression. Genomics Proteomics BioInf, 2019, 17: 511–521

82 Gao Y, Zhang J, Zhao F. Circular RNA identification based on multiple seed matching. Brief BioInf, 2018, 19: 803–810

83 Cheng J, Metge F, Dieterich C. Specific identification and quantification of circular RNAs from sequencing data. Bioinformatics, 2016, 32: 1094–1096

84 Wang K, Singh D, Zeng Z, et al. MapSplice: accurate mapping of RNA-seq reads for splice junction discovery. Nucleic Acids Res, 2010, 38: e178
85 Vromman M, Anckaert J, Bortoluzzi S, et al. Large-scale benchmarking of circRNA detection tools reveals large differences in sensitivity but not in precision. Nat Methods, 2023, 20: 1159–1169
86 Wang X, Ding C, Yu W, et al. Cas12a base editors induce efficient and specific editing with low DNA damage response. Cell Rep, 2020, 31: 107723
87 Uren P J, Bahrami-Samani E, Burns S C, et al. Site identification in high-throughput RNA–protein interaction data. Bioinformatics, 2012, 28: 3013–3020
88 Krakau S, Richard H, Marsico A. PureCLIP: capturing target-specific protein–RNA interaction footprints from single-nucleotide CLIP-seq data. Genome Biol, 2017, 18: 240
89 Li P, Shi R, Zhang Q C. icSHAPE-pipe: a comprehensive toolkit for icSHAPE data analysis and evaluation. Methods, 2020, 178: 96–103
90 Lu Z, Gong J, Zhang Q C. PARIS: psoralen analysis of RNA interactions and structures with high throughput and resolution. Methods Mol Biol, 2018, 1649: 59–84
91 Rieder D, Amort T, Kugler E, et al. meRanTK: methylated RNA analysis ToolKit. Bioinformatics, 2016, 32: 782–785
92 Corcoran D L, Georgiev S, Mukherjee N, et al. PARalyzer: definition of RNA binding sites from PAR-CLIP short-read sequence data. Genome Biol, 2011, 12: R79
93 Siegfried N A, Busan S, Rice G M, et al. RNA motif discovery by SHAPE and mutational profiling (SHAPE-MaP). Nat Methods, 2014, 11: 959–965
94 Busan S, Weeks K M. Accurate detection of chemical modifications in RNA by mutational profiling (MaP) with ShapeMapper 2. RNA, 2018, 24: 143–148
95 Satija R, Farrell J A, Gennert D, et al. Spatial reconstruction of single-cell gene expression data. Nat Biotechnol, 2015, 33: 495–502
96 Hao Y, Stuart T, Kowalski M H, et al. Dictionary learning for integrative, multimodal and scalable single-cell analysis. Nat Biotechnol, 2024, 42: 293–304
97 Aran D, Looney A P, Liu L, et al. Reference-based analysis of lung single-cell sequencing reveals a transitional profibrotic macrophage. Nat Immunol, 2019, 20: 163–172
98 Tan Y, Cahan P. SingleCellNet: a computational tool to classify single cell RNA-Seq data across platforms and across species. Cell Syst, 2019, 9: 207–213.e2
99 Domínguez Conde C, Xu C, Jarvis L B, et al. Cross-tissue immune cell analysis reveals tissue-specific features in humans. Science, 2022, 376: eabl5197
100 Abdelaal T, Michielsen L, Cats D, et al. A comparison of automatic cell identification methods for single-cell RNA sequencing data. Genome Biol, 2019, 20: 194
101 Cao J, Spielmann M, Qiu X, et al. The single-cell transcriptional landscape of mammalian organogenesis. Nature, 2019, 566: 496–502
102 La Manno G, Soldatov R, Zeisel A, et al. RNA velocity of single cells. Nature, 2018, 560: 494–498
103 Efremova M, Vento-Tormo M, Teichmann S A, et al. CellPhoneDB: inferring cell–cell communication from combined expression of multi-subunit ligand–receptor complexes. Nat Protoc, 2020, 15: 1484–1506
104 Huang S, Ouyang L, Tang J, et al. Spatial transcriptomics: a new frontier in cancer research. Clin Cancer Bull, 2024, 3: 13
105 Ni Z, Prasad A, Chen S, et al. SpotClean adjusts for spot swapping in spatial transcriptomics data. Nat Commun, 2022, 13: 2971
106 Wang Y, Song B, Wang S, et al. Sprod for de-noising spatially resolved transcriptomics data based on position and image information. Nat Methods, 2022, 19: 950–958
107 Zhao E, Stone M R, Ren X, et al. Spatial transcriptomics at subspot resolution with BayesSpace. Nat Biotechnol, 2021, 39: 1375–1384
108 Weirather J L, de Cesare M, Wang Y, et al. Comprehensive comparison of Pacific Biosciences and Oxford Nanopore Technologies and their applications to transcriptome analysis. F1000Res, 2017, 6: 100
109 Li H, Alkan C. New strategies to improve minimap2 alignment accuracy. Bioinformatics, 2021, 37: 4572–4574
110 Marić J, Sović I, Križanović K, et al. Graphmap2 - splice-aware RNA-seq mapper for long reads. bioRxiv, 2019, 720458
111 Espinosa E, Bautista R, Larrosa R, et al. Advancements in long-read genome sequencing technologies and algorithms. Genomics, 2024, 116: 110842
112 Amarasinghe S L, Ritchie M E, Gouil Q. long-read-tools.org: an interactive catalogue of analysis methods for long-read sequencing data. GigaScience, 2021, 10: giab003

113 Alipanahi B, Delong A, Weirauch M T, et al. Predicting the sequence specificities of DNA- and RNA-binding proteins by deep learning. Nat Biotechnol, 2015, 33: 831–838
114 Zhou J, Troyanskaya O G. Predicting effects of noncoding variants with deep learning–based sequence model. Nat Methods, 2015, 12: 931–934
115 Jaganathan K, Kyriazopoulou Panagiotopoulou S, McRae J F, et al. Predicting splicing from primary sequence with deep learning. Cell, 2019, 176: 535–548.e24
116 Sun L, Xu K, Huang W, et al. Predicting dynamic cellular protein–RNA interactions by deep learning using *in vivo* RNA structures. Cell Res, 2021, 31: 495–516
117 Krishna R, Wang J, Ahern W, et al. Generalized biomolecular modeling and design with RoseTTAFold All-Atom. Science, 2024, 384: eadl2528
118 Sumi S, Hamada M, Saito H. Deep generative design of RNA family sequences. Nat Methods, 2024, 21: 435–443
119 Ma X K, Yu Y, Huang T, et al. Bioinformatics software development: principles and future directions. Innov Life, 2024, 2: 100083
120 Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need. arXiv, 1706.03762
121 Fu Z C, Gao B Q, Nan F, et al. DEMINING: a deep learning model embedded framework to distinguish RNA editing from DNA mutations in RNA sequencing data. Genome Biol, 2024, 25: 258
122 Wang S, Lin J, Jia P, et al. *De novo* and somatic structural variant discovery with SVision-pro. Nat Biotechnol, 2025, 43: 181–185
123 Theodoris C V, Xiao L, Chopra A, et al. Transfer learning enables predictions in network biology. Nature, 2023, 618: 616–624

Summary for “转录组生物信息学: 从数据生成到分析框架”

Bioinformatics in transcriptome: from sequencing strategies to analyzing pipelines

Fang Nan[1,2*], Xu-Kai Ma[1,2] & Li Yang[1,2*]

[1] *Institutes of Biomedical Sciences, Fudan University, Shanghai 200032, China*
[2] *Children's Hospital of Fudan University, Shanghai 201102, China*
* Corresponding authors, E-mail: fangnan@fudan.edu.cn; liyang_fudan@fudan.edu.cn

The completion of the Human Genome Project (HGP) and the advent of affordable high-throughput sequencing technologies in the past decade have enabled researchers to explore the complex transcriptomes and their dynamics comprehensively. In addition, technological advancements in the last decades have significantly enhanced the speed and accuracy of transcriptomic data analyses, achieving in-depth comparative studies of gene expression patterns across various physiological states, varying developmental stages, and distinct pathological conditions.

Here, we provide an overview of different strategies for transcriptomic analyses, from library preparation to bioinformatic pipelines. In brief, transcriptomic profiling technologies can be classified into two main categories based on sequencing platforms: next-generation short-read sequencing or third-generation long-read sequencing. Based on data enrichment/analysis objectives, the next-generation sequencing can be further divided into three categories: (1) whole transcript enrichment, (2) transcript target fragment enrichment, and (3) artificial-mutagenesis transcript fragment enrichment. Alternatively, based on starting materials, sequencing methods can be categorized as bulk RNA-seq and single-cell/spatial transcriptome sequencing. Of note, recent transcriptomic analyses have been extended from bulk cell profiling to single-cell/spatial levels.

Each sequencing method employs specialized enrichment strategies and computational frameworks tailored to specific research goals. Whole-transcript enrichment data contain expression information of full-length transcripts, which supports comprehensive analyses including differential gene expression quantification, alternative (back-) splicing event identification, or *de novo* transcript discovery after alignment. Specific fragment enrichment data are ideal for exploring features of RNA modification, RNA-protein interaction, RNA-RNA interaction, or RNA secondary structure. Artificial mutagenesis strategies combined with advanced alignment algorithms allow precise identification of RNA modification sites and structural features at single-nucleotide resolution. Advances in alignment tools enable researchers to accurately and rapidly identify mutation sites embedded in transcriptomes.

Moreover, single-cell RNA-seq datasets based on droplet microfluidics/multi-well microplates provide unprecedented resolution for studying intricate biological processes such as embryogenesis, immune cell differentiation, and tumor microenvironment dynamics. In addition, long-read sequencing allows the recovery of full-length transcripts without assembly steps, improving the understanding of transcript isoforms, splicing, RNA modification, and RNA structure.

We further discuss recent trends in transcriptomic analyses with artificial intelligence, including machine learning and deep learning. Using published models as examples, we demonstrate how training data selection and sequence encoding methods can critically influence model development and performance. Finally, we propose future directions for AI-powered transcriptomic data mining, emphasizing its potential to unlock novel biological insights and to guide potential applications.

transcriptome, high-throughput sequencing, bioinformatics, artificial intelligence

doi: 10.1360/TB-2025-0160