高通量测序在可变剪接中的研究进展

张黎娅[1] 牛宝华[1] 宋宁[2]*

([1]昆明理工大学灵长类转化医学研究院, 省部共建非人灵长类生物医学国家重点实验室, 昆明 650500;
[2]深圳大学, 高等研究院, 深圳 518000)

摘要 在真核生物基因表达的过程中, mRNA的可变剪接(alternative splicing, AS)导致同一基因蛋白质亚型多样性的产生, 同时也增加了基因表达调控的多样性。高达95%的人类基因可以通过AS来产生具有不同功能的蛋白质。除此之外, 约15%的人类遗传疾病和癌症与AS相关。作为一种精密的基因表达调控方式, AS协助完成重要的生物过程, 如细胞发育和分化等。近年来, 高通量测序的发展推动了AS在分析组织特异性基因表达领域的研究。然而, 两者的有机结合应用仍然具有挑战性。该文总结了高通量测序在AS研究中的应用, 进一步分析了其中存在的问题, 并提出了解决方法, 为推动该领域的发展提供了新的策略与思路。

关键词 可变剪接; 高通量测序; 转录组测序; 单细胞转录组测序

Advances in High-Throughput Sequencing for the Study of Alternative Splicing

ZHANG Liya[1], NIU Baohua[1], SONG Ning[2]*

([1]*State Key Laboratory of Primate Biomedical Research, Institute of Primate Translational Medicine, Kunming University of Science and Technology, Kunming 650500, China*; [2]*Institute for Advanced Study, Shenzhen University, Shenzhen 518000, China*)

Abstract AS (alternative splicing) achieves the diversity of most proteins in eukaryotes by producing multiple different protein isoforms from a single gene, adding an additional regulatory layer for gene expression. Up to 95% of human genes can be spliced to produce multiple proteins with different functions. In addition, about 15% of human genetic diseases and cancers are associated with alternative splicing. The regulation of alternative splicing is a delicated set of interactive mechanisms that assist important biological processes such as cell development and differentiation. In recent years, development of high-throughput sequencing drives research on alternative splicing in the field of tissue-specific gene expression. However, it is still challenging to combine both technologies. The application of high-throughput sequencing in alternative splicing is summerized in this review, moreover this study analyze the problems existing in this technology for giving solutions, which provide new strategies and ideas to promote the development of this field.

Keywords alternative splicing; high-throughput sequencing; RNA-sequencing; single cell RNA-sequencing

人类基因组含有约两万个蛋白编码基因。然而, 这两万个蛋白编码基因却可能产生不少于20万种蛋白质[1]。基因数量与蛋白质数量有如此巨大差异的原因主要在于生物体存在可变剪接(alternative splicing, AS)。在人类不同的细胞类型中, 绝大多数多外显子基因通过AS来转录翻译产生具有不同功能的

收稿日期: 2022-05-08 接受日期: 2022-09-07
国家自然科学基金(批准号: 32260176)和中国博士后科学基金(批准号: 2022MD723788)资助的课题
*通讯作者。Tel: 13662619927, E-mail: ningsong@szu.edu.cn
Received: May 8, 2022 Accepted: September 7, 2022
This work was supported by the National Natural Science Foundation of China (Grant No.32260176) and the China Postdoctoral Science Foundation (Grant No.2022MD723788)
*Corresponding author. Tel: +86-13662619927, E-mail: ningsong@szu.edu.cn

蛋白质[1]。近些年的研究表明, AS与胚胎发育[2-3]和细胞干性[4-5]息息相关。对AS的深入研究, 可以加强我们对mRNA复杂性及其调控机制的理解, 有助于解析疾病致病机理和开发剪接相关疾病的治疗干预方法。同时也有利于进一步阐明影响胚胎发育的关键因素, 加深对细胞干性维持及分化机制的理解, 对建立不同的干细胞系并将其应用到临床具有巨大价值。

1 可变剪接

1.1 AS 的发生

基因由内含子与外显子组成, 在转录过程中, 内含子与外显子会根据基因的表达需要而被保留或者删除(表1)。一个基因的成熟转录本上所有内含子都被切掉而保留所有外显子, 这种剪接方式属于组成型剪接(constitutive splicing)(图1)。AS与组成型剪接的主要区别在于最终所形成的成熟mRNA结构。AS存在一定程度上的灵活性, 最终mRNA的结构可选择性地包含前体RNA的部分外显子和/或内含子片段, 从而使蛋白质在序列组成上产生变化。因此, AS为蛋白质多样性奠定了基础[7]。现利用高通量测序数据, 根据剪接位点的差异, 可鉴别出7种AS的类型(图1)[6]。参与AS这一基本生物学过程的各因素间的协同作用对细胞正常运作是至关重要的, 而这一过程的任何异常都可能导致正常细胞功能的紊乱和疾病的发生。

此外, 高等真核生物中一些蛋白质编码基因也可通过外显子和/或内含子的反向剪接产生环状RNA(circRNA)(图1)。研究发现, AS事件在circRNA的产生过程中也普遍存在, 且具有明显的核内倾向的空间定位, 同时circRNA的AS表现出组织和发育阶段特异的表达模式, 参与的剪接因子不同于mRNA的剪接因子, 这表明circRNA的AS可能受到与已知机制不同的调控作用。目前, circRNA被发现可以作为miRNA和RNA结合蛋白(RNA-binding protein, RBP)等的缓冲结合物从而参与基因表达调控, 作为物理支架介导某些酶和底物之间的相互作用, 甚至可以通过不依赖于帽子结构的翻译(cap-independent translation)产生多肽或蛋白产物。

1.2 AS 的功能调控

很多顺式剪接元件存在于DNA序列上调控AS, 它们分别界定内含子与上下游外显子边界的5′和3′剪接位点及3′剪接位点上游的分支位点和聚嘧啶束。此外, 外显子或侧翼内含子中的辅助顺式元件也可作为剪接增强子或沉默子, 与反式剪接调节子作用来促进或抑制外显子剪接。在不同组织或发育的不同阶段, AS通过广泛的RNA-蛋白相互作用进行调控, 展现出组织特异性和时间上的变化差异, 从

表 1 AS的类型

Table 1 Types of alternative splicing

剪接名称 Name of splicing	AS的类型 Types of alternative splicing	定义 Definition
Constitutive splicing	Constitutive splicing	Removing introns from the pre-mRNA, joining the exons together to form a mature mRNA
Alternative splicing	Exon skip	Exon is cleaved from the original transcript
	Retained intron	Intron is retained from the original transcript
	Alternate donor site	The same 3′-end splicing site but different 5′-end splicing site. The 5′-end exon is extended
	Alternate acceptor site	The same 5′-end splicing site but different 3′-end splicing site. The 3′-end exon is extended
	Alternate promoter	The transcripts differ in the first exon
	Alternate terminator	The transcripts differ in the last exon
	Mutually exclusive exons	The same exons between transcripts are called constitutive exon, and the different exons are called inclusive exon. Inclusive exon cannot exist in the same transcript, but can only exist in different transcripts. Such a variable splicing event is called mutually exclusive exon
CircRNA backsplicing	Exonic circRNA	All circRNAs derive from maternal exons
	Circular intronic RNA	Lasso circRNAs derive entirely from introns
	Exon-intron circRNA	CircRNAs derive from exons and introns
	Sense overlapping circRNA	Circular RNAs that do not belong to exon and intron circular RNAs
	Antisense circular RNA	Antisense circRNA derive from antisense chain transcripts
	Intergenic circular RNA	CircRNAs derive from intergenic sequences or other unannotated genomic sequences

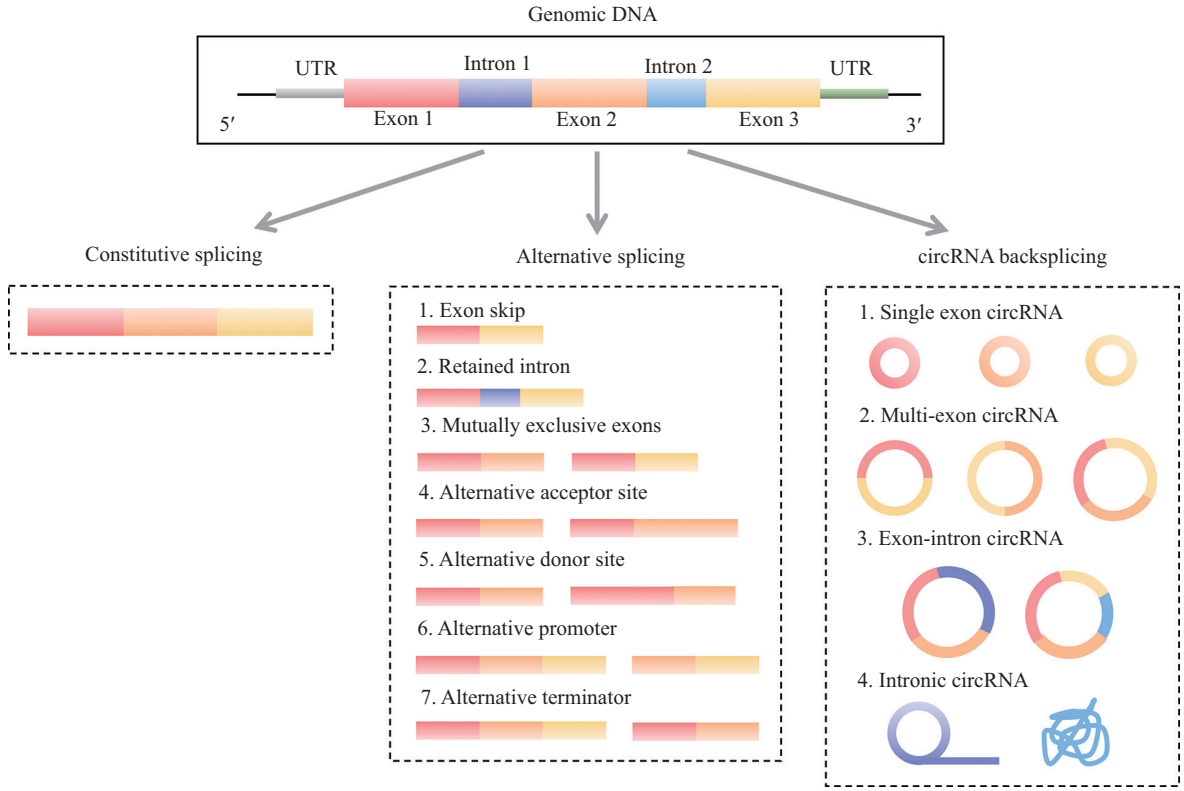

图1 AS产生多个不同的转录本

Fig.1 Alternative splicing produces multiple different transcripts

而与正常的生命活动或疾病相关联。在早衰(premature)中, 核纤层蛋白基因*LMNA*的AS失衡是导致早衰最重要的原因[7]。*LMNA*正常剪接时有三个外显子, 翻译出含C-端结构域的蛋白, 但突变发生会导致剪接方式改变, 5′-AS位点发生变化、可翻译区域变短、失去定位信号, 使得起支撑作用的核纤层蛋白在细胞核的分布变得紊乱, 进而影响下游基因的表达, 最终导致产生早衰的表型。另有研究表明, AS参与RNA剪接调控蛋白的表达。RBM4是一个与RNA剪接相关的RNA结合蛋白, 该蛋白在很多癌症中下调, 从而改变很多基因的剪接方式, 进而影响细胞的增殖、死亡、凋亡等过程, 最终导致癌症的发生和发展[8]。上述两个例子说明, AS依赖于顺式作用元件与反式作用因子两方面的调控, 它们的突变都可导致剪接方式的改变, 最终引起疾病。因此理解AS的变化对于疾病的治疗及预防具有重要价值。

2 高通量测序技术在AS研究中的应用

2.1 AS研究的技术发展

AS的研究方法随着技术的突破而不断发展。传统的方法是通过RT-PCR(reverse transcription-polymerase chain reaction)来量化AS[9]。在20世纪90年代, 通过对全长mRNA片段的表达序列标签(expressed sequence tag, EST)测序, 揭示了真核生物中AS的广泛存在[10]。在21世纪初期, 剪接敏感的微阵列(microarray)的发展使研究人员能够研究不同细胞状态、跨组织和物种的剪接调控机制[11]。在2008年, 三项具有里程碑意义的研究证明使用转录组测序(RNA-sequencing, RNA-seq)能够表征哺乳动物组织中AS的特性[12-14]。此后, RNA-seq迅速超越微阵列, 成为转录组水平分析AS的标准方法[15]。

2.2 高通量测序技术概况

在传统的二代测序中, RNA-seq是最为人所熟知的。RNA-seq常用于分析两组稳定状态下的组织、器官样本或细胞群的差异表达基因(differentially expressed gene, DEG)。然而通过RNA-seq得到的只是基因表达的均值, 因此细胞群体中单个细胞的特异性信息往往会被掩盖(比如特异表达的基因或RNA不同的亚型)。scRNA-seq的出现更新了以往人们对细胞研究的认知。scRNA-seq是近年来生物学领域的热门技术, 它有着极高的分辨率, 能够精准地剖析样本中的细胞组成信息, 进而大规模地揭示单个细

胞的基因结构和基因表达状态, 从而反映细胞间的异质性[12]。这项技术开辟了超越细胞状态描述研究的新领域。

2.3 早期高通量测序技术在AS研究中的应用

RNA-seq是转录组学分析的强大工具。随着深度测序成本的大幅下降, 基因表达和AS的大规模研究成为热点问题。RNA-seq数据可以直接显示外显子部分的AS: 在高通量测序平台进行RNA转录本的分割测序、读数与基因组比对, 从而显示外显子和剪接接头的覆盖情况。剪接百分比(percent splicing index, PSI), 指的是包含的读数(inclusion reads, IR)即重叠读数和排除读数(exclusion reads, ER)即重叠分割排列这二者的比率, 表明外显子部分被剪接到不同亚型中的效率。PSI的提出向人们展示了一个定量的、全局评估外显子的指标, 结合其他识别不同同源亚型处理的工具, 甚至可以将一个新的基因位点的、复杂的剪接事件以外显子为中心的方式进行可视化, 并在不同条件下进行比较。

RNA-seq主要通过高通量、短读长的测序方法对得到的cDNA进行扩增和测序, 进而获得组织样本的测序数据。在理想情况下, 利用一些比对软件可以将转录组测序得到的读数组装以重建被转录的基因组区域, 然后利用AS分析软件对AS事件进行检出和定量。尽管如此, 因存在有限的读段长度, AS的精确量化仍受到技术限制。二代测序RNA-seq的读长范围一般为50 bp~150 bp, 导致相同基因的不同转录本亚型通常难以区分, 因此使用短读长发现新的转录本是RNA-seq中最具挑战性的任务之一。对于低表达转录本, 几乎从未跨越剪接位点的短读段可能会使全长转录本的推断复杂化。为了解决上述问题, 研究人员开发了各种软件, 包括rMATS、MAJIQ、LeafCutter、SUPPA2、SplAdder和Whippet等(表2)。

利用不同的高通量RNA-seq数据, 同时结合不同的分析软件, 可以阐释很多AS在疾病发生和发育调控中的作用[12-13,16-17]。通过高通量RNA-seq技术, 发现在儿童B淋巴细胞白血病(B-cell acute lymphoblastic leukemia, B-ALL)中编码表面抗原*CD22*的mRNA存在异常剪接, 导致该蛋白水平下调, 使恶性肿瘤细胞对靶向CD22的免疫疗法产生抵抗性[18]; BLENCOWE团队[19]报道了一种调节胚胎干细胞的多能性和重编程的剪接开关, 他们经过对高通量RNA-seq数据的分析, 发现了一个改变转录因子FOXP1的DNA结合偏好性的胚胎干细胞特异性AS事件。FOXP1特异性AS刺激了胚胎干细胞多能性转录因子基因包括*OCT4*、*NANOG*、*SOX2*和*DPPA3*等的表达, 同时抑制了胚胎干细胞分化相关基因的表达[19]。此外, 北京大学杜鹏团队[20]也结合高通量测序与实验验证, 通过对特异亚型进行抑制, 成功将小鼠胚胎干细胞(embryonic stem cells, ESCs)逆转到全能性状态。因此, 通过高通量RNA-seq测序, 能深入地了解AS在多种生物学过程如细胞增殖和存活、细胞分化、内环境稳态、机体对压力的反应以及疾病的产生等中的作用[21-22]。

2.4 高通量测序技术在可变剪接研究中的限制因素

虽然RNA-seq使得转录组水平的AS分析成为可能, 但测序得到的是一群细胞的RNA表达水平的平均值, 无法检测到细胞间的差异表达基因, 从而掩盖了细胞异质性表达[23-24]。然而单细胞AS研究可以评估AS在细胞群体中的分布和相关细胞亚群特征, 以及在动态过程中AS的变化, 例如在细胞水平上评估一个基因不同AS的丰度。scRNA-seq表明, 一个细胞通常不会表达所有的AS类型, 但是却能显著地表达或不表达某个亚型[25]。从2009年开始, 二代单细胞转录组至今已发展成为科学研究中强有力的工

表 2 六种AS软件在亚型研究上的小结

Table 2 Summary of six alternative splicing software for isoform study

软件 Software	原理 Principle	检测到的AS的类型数量 The number of types of alternative splicing detected	能否检测未注释转录本 Detect unannotated transcripts
rMATS	Event-based	5	Yes
MAJIQ	Event-based	5	No
SUPPA2	Event-based	7	No
SplAdder	Event-based	5	Yes
Whippet	Event-based	9	No
LeafCutter	Intron excision	Unknown	Yes

具, 使得研究者们在探索单细胞分辨率上的各类群细胞表达以及拟时序分析成为可能。但由于二代单细胞转录组测序平台的局限性, 双端150 bp的测序长度并不能检测到平均长度在1.5 Kb左右的全长转录本。下文将讨论高通量测序技术在可变剪接研究中最主要的三个限制因素。

2.4.1 全长转录本的捕获 基于转录本方法存在的不足之处: 从短读长中推断出全长的mRNA不同亚型并不容易, 且结果对转录本注释的选择很敏感。此外, 对于具有多个AS区的基因, 将mRNA亚型丰度的变化归因于特定外显子或不同剪接位点的剪接调控并非易事。不同类型的剪接变异可以发生在pre-mRNA分子的不同位点。因此, 对转录本进行不完全测序就可能会忽略部分剪接事件。在Illumina测序文库制备中, 一种短的、随机的寡核苷酸单分子标签——独特的分子标识符(unique molecular identifiers, UMI)在PCR之前的逆转录步骤中被加入到cDNA上, 因此由比对到同一基因不同read生成的cDNA分子将具有不同的UMI序列[26]。由于对UMI的统计是在PCR扩增之后, 因此可以通过匹配的UMI和映射位点的reads来纠正阻碍准确表达量化的非线性放大。比起高通量RNA-seq, 高通量scRNA-seq由于需要更多轮的PCR来获得足够的cDNA以进行文库构建和测序, 因此同时引入了更多的偏差。目前基于UMI的单细胞转录组分析主要是基于从3′端扩增或者5′端扩增的短读长测序研究, 例如CEL-seq2[27]、inDrop[28]、Drop-seq[29]、MARS-seq[30]、SCRB-seq[31]和STRT-seq[26], 绝大多数仅能进行基因定量, 不能进行转录本定量以及转录本结构等全貌研究。UMIs分子条码能够使嵌合cDNA被错误地注释为新转录本的风险降至最低, 但利用UMI的方法只能对转录本的一个片段进行测序, 当差异不位于该片段时就无法对AS进行鉴别。因此, 在亚型的研究中若要获得全长转录本信息就需要放弃UMI的使用。由Rickard SANDBERG实验室开发的Smart-seq[32](switching mechanism at 5′ end of the RNA transcript)能在全转录组范围内进行scRNA-seq, 以全长mRNA建库, 进而提升转录本的测序覆盖度。在2014年升级为Smart-seq2[33]后, 它已成为几乎每个单细胞测序公司都采用的方法。虽然高通量Illumina测序从相同的转录本中生成许多短序列进行标记, 再通过计算方法对转录组进行组装, 但这种通过计算的组装不能恢复不同AS的结构, 只能将定量限制在外显子的表达水平。而低通量Smart-seq或Smart-seq2测序通过增强逆转录作用来捕获整个转录本, 因此Smart-seq或Smart-seq2较高通量的scRNA-seq更适用于AS研究[34](表3)。研究人员利用SMRT(single molecule real time)测序技术可以生成高准确度和长读长的HiFi reads, 数十Kb的读长在一次测序中足以读取大多数转录本, 并且无需下游的生物信息学组装, 在没有参考基因组的条件下, 完全能够进行转录本亚型的全长测序和后续分析。PacBio和Nanopore(nanopore平台可以直接测序RNA, 而PacBio平台则必须先合成cDNA), 这些长读长技术能够克服短读长技术自身的缺陷, 如: ambiguous reads更少, 可以捕捉到更多的亚型, 减少单细胞转录组工具带来的splice-junction错误。

2.4.2 低表达转录本的定量 单个细胞中某些RNA含量低, 阻碍了捕获效率。在逆转录过程中, 常常会出现由于转录本未被捕获到而被认定为基因未表达的情况[35-36]。因此, 表达水平非常低的转录本受此影响很大, 并且零表达值也不能真正等同于生物学上的不表达[37]。因此, mRNA捕获效率限制了scRNA-seq中可以检测到的转录本总数, 但是对于超

表3 高、低通量测序技术在亚型研究上的限制因素

Table 3 Limitation factors of high-throughput/low-throughput sequencing methods for isoform study

限制因素 Limitation factors	高通量 High-throughput	低通量 Low-throughput	
	Illumina RNA-seq	Smart-seq	SMS
Sequencing depth	Medium	High	Low
Expression quantification	High	Medium	Low
Sequencing errors	Low	Low	High
Number of cells sequenced	High	Medium	Low
Number of isoforms detected	Low	Medium	High

过这个检测限度的表达转录本, 提高测序深度是提高灵敏度[38]的关键。高通量单细胞转录组领域关于达到饱和测序深度的普遍共识是, 当测序后的每个细胞得到超过100万个Illumina读数时, 几乎就不会再产生任何新信息了[39-40]。由于非显性AS通常表达水平较低, 未被测到的概率很高, 因此每个基因的AS多样性很容易被低估, 在亚型水平上达到饱和时需要每个细胞有超过100万次读取。除此之外, scRNA-seq中要求的质量过滤步骤也会对低表达转录本的定量产生影响, 包括: (1) 去除低质量的细胞(例如检测到少量特征的细胞[41]); (2) 过滤在大部分细胞中没有表达的基因。虽然Smart-seq2相较于高通量的scRNA-seq能检测到更多的表达基因, 具有更高的敏感度。但Smart-seq2中低表达基因在所有被检测到的细胞中占比很少。而在高通量单细胞数据中, 存在大量的低表达的基因, 说明数据中mRNA在很低的表达水平上随机捕获有较高的噪声。基于高通量单细胞数据, 发现更严重的丢弃(dropout)问题——dropout是scRNA-seq数据的一大特点, 指很多基因在某些细胞中根本就检测不到表达, 但是在其他的细胞中却被检测为高表达。通常认为dropout是由于在文库构建过程中低表达基因没有被成功反转录而导致的[33-38,42-43]。

2.4.3 长读长测序中的错误检测 错误检测主要分为测序错误和实验误差, 前者是由于测序过程中碱基的误测造成的, 后者通常出现在扩增和逆转录过程中[40], 这些都会影响对AS的分析结果。测序错误在长读长技术中很常见, 相比于传统的二代高通量测序, 三代测序错误率要高上1~2个数量级。PacBio的测序错误率为2%~5%[44-45], 如此高的错误率对于AS研究显然是不利的, 这与Illumina测序的高准确度(错误率约0.005%)形成鲜明对比。三代测序中常见的错误是插入−缺失, 但这对于高通量RNA-seq而言影响甚微。因为RNA-seq并不要求每个碱基都正确, 它只需比对到转录本的亚型上即可。测序错误可能会识别错误的供体/受体位点, 导致新AS的错误识别[46]。对于高错误率, PacBio可利用CCS(circular consensus sequence)来多次测序以进行纠错。但实际中考虑测序成本, 并不能无限制测序, 而且过多的纠错可能会导致最终得到的独特的转录本数减少。Illumina测序在很大程度上采用多通路使用的策略以最小化批次效应以及UMI计数来消除扩增偏差。这些流程产生的错误会给分析增加额外的困难。一些错误在计算中遗留下来, 导致错误的比对, 从而产生并非真实的新亚型; 一些错误在逆转录过程中出现, 因此不能使用UMIs来识别。例如基因内部的poly(A)序列会产生较短的cDNA伪基因, 可能被误认为具有上游TTS的AS[47]。此外, mRNA分子形成的二级结构可以阻止逆转录酶进入mRNA的某些区域, 导致这些区域以交替剪接的AS形式出现[48]。

2.5 现阶段高通量测序技术在 AS 研究中的应用

三代测序平台能够获得全长转录组数据, 将数据分析从基因分析水平提升至了转录本分析水平。然而三代测序的低通量会导致对低表达转录本检测的敏感度下降。不过随着通量的增加, 长读长策略就能达到跟高通量测序一样的敏感性, 而且特异性更高。虽然Smart-seq2测序结果有着更为均一的分布, 而高通量单细胞测序存在mRNA的3′端或5′端偏差性, 却能覆盖大量细胞, 更好地检测到稀有细胞类型。因此, 不同平台检测到的细胞簇之间的差异表达基因表明这些技术是具有互补性的(表4)。目前一般采用二、三代结合的方法来进行Iso-seq(isoform-seq), 这样既可以增加测序深度, 又可以提高低表达基因转录本定量的准确性。因此现阶段一般会把Illumina RNA-seq与三代测序结合[49-50], 这样既可以增加测序深度、敏感度和特异性, 同时也可以保证转录本定量的准确性。

相对于PacBio平台, Nanopore平台的三代全长转录组的主要优势在于单张芯片产量更高, 通过结合二代测序数据进行数据校正, 可同时在单细胞水平上进行基因及转录本分析。2020年RAINER等[51]开发了ScNaUmi-seq(single-cell nanopore sequencing with UMIs)技术, 将Nanopore测序技术与UMI以及10× Genomics单细胞分离系统相结合, 获得了经过纠错的全长序列信息, 使得在单细胞水平上检测差异RNA剪接成为可能。同一年, 北京大学汤富酬组[52]开发了一种基于三代单分子测序平台的高灵敏度单细胞转录组测序方法——SCAN-seq(single cell amplification and sequencing of full-length RNAs by nanopore platform), 能够以单细胞分辨率直接获取全长转录本序列信息, 表现出了高灵敏度和高稳定性, 与之前基于二代测序平台最灵敏的单细胞转录组测序方法相比不相上下。他们还首次利用单细胞转录组三代测序数据将一个单细胞中的父源和母源转录本准确

表4 近两年高通量测序结合三代测序的应用

Table 4 Application of high-throughput sequencing combined with third-generation sequencing in recent years

技术名称 Technology	方法 Method	特点 Feature	参考文献 Reference
ScNaUmi-seq	Single-cell transcriptome+Nanopore	The number of target samples is large, eg: tumor tissue sample, cell sample etc.	[51]
SCAN-seq	Manual+Nanopore	The number of target samples is small, eg: a small number of specific samples that are stream sorted from samples at various stages of embryonic development etc.	[52]
FlsnRNA-seq	Single-nucleus transcriptome	Nuclear cell suspension/break the cell diameter limit, eg: arabidopsis root tip maize root tip etc.	[53]

区分开, 分别进行精准定量分析。南方科技大学翟继先组[53]2021年开发了FlsnRNA-seq(full-length single-nucleus RNAseq)技术, 通过结合10× Genomics单细胞测序和三代Nanopore全长转录组测序, 不需经过原生质体就可直接检测植物单细胞核。同时利用Illumina二代测序数据捕获基因表达丰度信息, 并以Nanopore三代测序数据捕获同源亚型信息, 从而最大限度地保留单个细胞核的转录状态。在单细胞水平揭示AS以及与多聚腺苷酸化相关的RNA isoform信息有助于识别细胞表征。这些利用高通量测序对SMS数据进行校正的方法, 必须产生足够多的cDNA, 并分别使用Illumina和Nanopore进行测序。以SMS结果为重点, 利用Illumina reads数据对SMS的测序高错误率进行校正, 最终达到对AS分析的最优化[54]。

3 总结与展望

40多年前GILBERT[55]首次提出了AS的概念, 在分子实验、高通量测序和生物信息学的帮助下, 虽然对AS的理解不断加深, 但仍有很多方面需要在细胞水平上进行功能探索。如上所述, 全长转录本的获得、低表达基因的定量以及错误检测等方面的限制是基于高通量测序研究AS的三个主要障碍。在不久的将来, 以下两种技术的进步可能会极大推动对AS的研究。(1) 长读长测序。SMRT和Nanopore测序技术通过测通整个转录本, 能够避免通过采用计算方法来重建的问题, 目前已有报道证明长读长测序适用于揭示全长转录本和AS事件[56-57]。最近, STEINMETZ团队[58-59]利用长读直接进行RNA测序, 发现从合成酵母基因组中的DNA序列表达的全长RNA分子的起点、终点和数量的变化, 以及基因的重新定位会影响其RNA输出的长度和丰度; (2) 通过优化的scRNA-seq在细胞分辨率上检测基因表达。通过合适的实验流程和生物信息分析, scRNA-seq更有潜力探索低表达的转录本和各种剪接事件。虽然这两种技术都还存在由于测序深度有限导致scRNA-seq得到的表达基因较少、技术噪音和生物噪音难以区分等局限性, 但是随着这些新技术与传统技术的结合和优化, 以及越来越多的强大的计算工具的出现和应用, 高通量测序会在AS研究中发挥不可估量的作用。

参考文献 (References)

[1] HU Z, SCOTT H S, QIN G, et al. Revealing missing human protein isoforms based on ab initio prediction, RNA-seq and proteomics [J]. Sci Rep, 2015, 5(1): 1-15.

[2] DAVINA, TONDELEIR, DRIEKE, et al. Actin isoform expression patterns during mammalian development and in pathology: insights from mouse models [J]. Cell Motil Cytoskeleton, 2009: 798-815.

[3] ZHANG W, XIA W, WANG Q, et al. Isoform switch of TET1 regulates DNA demethylation and mouse development [J]. Mol Cell, 2016, 64(6): 1062-73.

[4] ARSIC N, GADEA G, LAGERQVIST E L, et al. The p53 isoform δ133p53β promotes cancer stem cell potential-sciencedirect [J]. Stem Cell Rep, 2015, 4(4): 531-40.

[5] PAPAMICHOS S I, KOTOULA V, TARLATZIS B C, et al. OCT4B1 isoform: the novel OCT4 alternative spliced variant as a putative marker of stemness [J]. Mol Hum Reprod, 2009, 15(5): 269-70.

[6] ZONG Z, LI H, YI C, et al. Genome-wide profiling of prognostic alternative splicing signature in colorectal cancer [J]. Front Oncol, 2018, 8: 1-9.

[7] ERIKSSON M, BROWN W T, GORDON L B, et al. Recurrent de novo point mutations in lamin a cause hutchinson-gilford progeria syndrome [J]. Nature, 2003, 423(6937): 293-8.

[8] WANG Y, CHEN D, QIAN H, et al. The splicing factor rbm4 controls apoptosis, proliferation, and migration to suppress tumor progression [J]. Cancer Cell, 2014, 26(3): 374-89.

[9] PERCIFIELD R, MURPHY D, STOILOV P. Medium throughput analysis of alternative splicing by fluorescently labeled RT-PCR [J]. Methods Mol Biol, 2014, 1126: 299-313.

[10] MODREK B, LEE C. A genomic view of alternative splicing [J]. Nat Genet, 2002, 30(1): 13-9.

[11] LEE C, ROY M. Analysis of alternative splicing with microarrays: successes and challenges [J]. Genome Biol, 2004, 5(7): 1-4.

[12] PAN Q, SHAI O, LEE L J, et al. Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing [J]. Nat Genet, 2008, 40(12): 1413-5.
[13] WANG E T, SANDBERG R, LUO S, et al. Alternative isoform regulation in human tissue transcriptomes [J]. Nature, 2008, 456(7221): 470-6.
[14] MORTAZAVI A, WILLIAMS B A, MCCUE K, et al. Mapping and quantifying mammalian transcriptomes by RNA-seq [J]. Nat Methods, 2008, 5(7): 621-8.
[15] WANG Z, GERSTEIN M, SNYDER M. RNA-seq: a revolutionary tool for transcriptomics [J]. Nat Rev Genet, 2009, 10(1): 57-63.
[16] RAJ B, BLENCOWE B J. Alternative splicing in the mammalian nervous system: recent insights into mechanisms and functional roles [J]. Neuron, 2015, 87(1): 14-27.
[17] TEICHROEB J H, KIM J, BETTS D H. The role of telomeres and telomerase reverse transcriptase isoforms in pluripotency induction and maintenance [J]. RNA Biol, 2016, 13(8): 707-19.
[18] ZHENG S, GILLESPIE E, NAQVI A S, et al. Modulation of cd22 protein expression in childhood leukemia by pervasive splicing aberrations: implications for cd22-directed immunotherapies [J]. Blood Cancer Discov, 2022, 3(2): 103-15.
[19] GABUT M, SAMAVARCHI-TEHRANI P, WANG X, et al. An alternative splicing switch regulates embryonic stem cell pluripotency and reprogramming [J]. Cell, 2011, 147(1): 132-46.
[20] SHEN H, YANG M, LI S, et al. Mouse totipotent stem cells captured and maintained through spliceosomal repression [J]. Cell, 2021, 184(11): 2843-59.
[21] CHEPELEV I, CHEN X. Alternative splicing switching in stem cell lineages [J]. Front Biol, 2013, 8(1): 50-9.
[22] PARK J W, FU S, HUANG B, et al. Alternative splicing in mesenchymal stem cell differentiation [J]. Stem Cells, 2020, 38(10): 1229-40.
[23] PELECHANO V, WEI W, JAKOB P, et al. Genome-wide identification of transcript start and end sites by transcript isoform sequencing [J]. Nat Protoc, 2014, 9(7): 1740-59.
[24] BUETTNER F, NATARAJAN K N, CASALE F P, et al. Computational analysis of cell-to-cell heterogeneity in single-cell RNA-sequencing data reveals hidden subpopulations of cells [J]. Nat Biotechnol, 2015, 33(2): 155-60.
[25] SHALEK A K, SATIJA R, ADICONIS X, et al. Single-cell transcriptomics reveals bimodality in expression and splicing in immune cells [J]. Nature, 2013, 498(7453): 236-40.
[26] ISLAM S, ZEISEL A, JOOST S, et al. Quantitative single-cell RNA-seq with unique molecular identifiers [J]. Nat Methods, 2014, 11(2): 163-6.
[27] HASHIMSHONY T, SENDEROVICH N, AVITAL G, et al. Cel-seq2: sensitive highly-multiplexed single-cell RNA-seq [J]. Genome Biol, 2016, 17: 1-7.
[28] KLEIN A M, MAZUTIS L, AKARTUNA I, et al. Droplet barcoding for single-cell transcriptomics applied to embryonic stem cells [J]. Cell, 2015, 161(5): 1187-201.
[29] 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): 1202-14.
[30] 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): 776-9.
[31] SOUMILLON M, CACCHIARELLI D, SEMRAU S, et al. Characterization of directed differentiation by high-throughput single-cell RNA-seq [J]. bioRxiv, 2014, doi:10.1101/003236.
[32] RAMSKOLD 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): 777-82.
[33] PICELLI S, FARIDANI O R, BJORKLUND A K, et al. Full-length RNA-seq from single cells using Smart-seq2 [J]. Nat Protoc, 2014, 9(1): 171-81.
[34] ZHU Y Y, MACHLEDER E M, CHENCHIK A, et al. Reverse transcriptase template switching: a smart approach for full-length cdna library construction [J]. Biotechniques, 2001, 30(4): 892-7.
[35] MARINOV G K, WILLIAMS B A, MCCUE K, et al. From single-cell to cell-pool transcriptomes: stochasticity in gene expression and RNA splicing [J]. Genome Res, 2014, 24(3): 496-510.
[36] STEGLE O, TEICHMANN S A, MARIONI J C. Computational and analytical challenges in single-cell transcriptomics [J]. Nat Rev Genet, 2015, 16(3): 133-45.
[37] BRENNECKE P, ANDERS S, KIM J K, et al. Accounting for technical noise in single-cell RNA-seq experiments [J]. Nat Methods, 2013, 10(11): 1093-5.
[38] ZIEGENHAIN C, VIETH B, PAREKH S, et al. Comparative analysis of single-cell RNA sequencing methods [J]. Mol Cell, 2017, 65(4): 631-43.
[39] BACHER R, KENDZIORSKI C. Design and computational analysis of single-cell RNA-sequencing experiments [J]. Genome Biol, 2016, 17: 1-14.
[40] TARDAGUILA M, DE LA FUENTE L, MARTI C, et al. Sqanti: extensive characterization of long-read transcript sequences for quality control in full-length transcriptome identification and quantification [J]. Genome Res, 2018, 28(3): 396-411.
[41] ILICIC T, KIM J K, KOLODZIEJCZYK A A, et al. Classification of low quality cells from single-cell RNA-seq data [J]. Genome Biol, 2016, 17: 1-15.
[42] WESTOBY J, HERRERA M S, FERGUSON-SMITH A C, et al. Simulation-based benchmarking of isoform quantification in single-cell RNA-seq [J]. Genome Biol, 2018, 19(1): 1-14.
[43] OLIVIERI J E, DEHGHANNASIRI R, WANG P L, et al. RNA splicing programs define tissue compartments and cell types at single-cell resolution [J]. eLife, 2021, doi: 10.7554/eLife.70692.
[44] JAIN M, FIDDES I T, MIGA K H, et al. Improved data analysis for the minion nanopore sequencer [J]. Nat Methods, 2015, 12(4): 351-6.
[45] JAWORSKI E, ROUTH A. Parallel clickseq and nanopore sequencing elucidates the rapid evolution of defective-interfering RNAs in flock house virus [J]. PLoS Pathog, 2017, 13(5): e1006365.
[46] HOUSELEY J, TOLLERVEY D. Apparent non-canonical trans-splicing is generated by reverse transcriptase *in vitro* [J]. PLoS One, 2010, 5(8): e12271.
[47] NAM D K, LEE S, ZHOU G, et al. Oligo(dt) primer generates a high frequency of truncated cdnas through internal poly(A) priming during reverse transcription [J]. Proc Natl Acad Sci USA, 2002, 99(9): 6152-6.
[48] COCQUET J, CHONG A, ZHANG G, et al. Reverse transcrip-

tase template switching and false alternative transcripts [J]. Genomics, 2006, 88(1): 127-31.

[49] ZHU F Y, CHEN M X, YE N H, et al. Proteogenomic analysis reveals alternative splicing and translation as part of the abscisic acid response in arabidopsis seedlings [J]. Plant J, 2017, 91(3): 518-33.

[50] WANG M, WANG P, LIANG F, et al. A global survey of alternative splicing in allopolyploid cotton: landscape, complexity and regulation [J]. New Phytol, 2018, 217(1): 163-78.

[51] LEBRIGAND K, MAGNONE V, BARBRY P, et al. High throughput error corrected nanopore single cell transcriptome sequencing [J]. Nat Commun, 2020, 11(1): 1-8.

[52] FAN X, TANG D, LIAO Y, et al. Single-cell RNA-seq analysis of mouse preimplantation embryos by third-generation sequencing [J]. PLoS Biol, 2020, 18(12): e3001017.

[53] LONG Y, LIU Z, JIA J, et al. FlsnRNA-seq: protoplasting-free full-length single-nucleus RNA profiling in plants [J]. Genome Biol, 2021, 22(1): 1-14.

[54] BYRNE A, BEAUDIN A E, OLSEN H E, et al. Nanopore long-read RNAseq reveals widespread transcriptional variation among the surface receptors of individual B cells [J]. Nat Commun, 2017, 8: 1-11.

[55] GILBERT W. Why genes in pieces [J]? Nature, 1978, 271(5645): 501.

[56] MA J, XIANG Y, XIONG Y, et al. SMRT sequencing analysis reveals the full-length transcripts and alternative splicing patterns in ananas comosus var. *Bracteatus* [J]. PeerJ, 2019, 7: e7062.

[57] DE JONG L C, CREE S, LATTIMORE V, et al. Nanopore sequencing of full-length brca1 mRNA transcripts reveals co-occurrence of known exon skipping events [J]. Breast Cancer Res, 2017, 19(1): 1-9.

[58] BRUNELLO L. Neighbourly modulation of transcript isoforms [J]. Nat Rev Genet, 2022, 23(5): 264.

[59] BROOKS A N, HUGHES A L, CLAUDER-MUNSTER S, et al. Transcriptional neighborhoods regulate transcript isoform lengths and expression levels [J]. Science, 2022, 375(6584): 1000-5.