PIBB 生物化学与生物物理进展
Progress in Biochemistry and Biophysics
2025,52(1):32~49
www.pibb.ac.cn

哺乳动物可变多聚腺苷酸化*

张 宇[1,2] 池红霞[1,2] 杨乌日吐[3] 左永春[4,5]** 邢永强[1,2]**

([1] 内蒙古科技大学生命科学与技术学院，包头 014010；
[2] 内蒙古科技大学，内蒙古自治区生命健康与生物信息学重点实验室，包头 014010；
[3] 呼和浩特职业学院计算机系，呼和浩特 010070；[4] 内蒙古大学生命科学学院，呼和浩特 010020；
[5] 内蒙古大学省部共建草原家畜生殖调控与繁育国家重点实验室，呼和浩特 010020）

摘要 随着测序技术的快速发展，哺乳动物可变多聚腺苷酸化（alternative polyadenylation，APA）得到了更加精准的检测。APA通过改变poly(A)尾巴的长度和位置，精细地调控基因的表达，参与疾病发生、胚胎发育等生命过程。哺乳动物APA的研究主要聚焦于以下三个方面：第一，基于转录组数据识别APA事件，并阐明其特征；第二，研究APA与基因表达调控的相互关系，揭示其在生命调控中的重要作用；第三，挖掘APA与疾病发生、胚胎发育、分化等生命过程的内在联系，为疾病的诊治、胚胎发育调控机制的解析等提供新的视角和方法。本文详细阐述了APA的分类、发生机制及功能，系统总结了基于多种转录组数据的APA识别方法和APA数据资源，对APA研究进行了总结与展望，强调了测序技术对哺乳动物APA研究的推动作用。未来，随着测序技术的进一步发展，哺乳动物APA的调控机制将更加明晰。

关键词 可变多聚腺苷酸化，调控机制，转录组，识别方法，数据资源
中图分类号 Q-31，Q95，Q975 **DOI**：10.16476/j.pibb.2024.0298 **CSTR**：32369.14.pibb.20240298

可变多聚腺苷酸化（alternative polyadenylation，APA）是指mRNA在转录过程中，其poly(A)尾巴的长度、位置或结合的蛋白质发生改变的现象。这种改变不仅影响mRNA的稳定性，还可能影响其翻译过程，从而影响蛋白质的表达水平[1]。在哺乳动物基因组中，存在大量的APA事件，且参与发育、分化、疾病发生等过程[2]。随着测序技术的不断进步和生物信息学的快速发展，对APA事件的识别、发生和调控机制等研究正不断深入，但仍面临着一些亟需解决的问题。首先，需要结合转录组数据开发更加精准的APA检测方法；其次，需要结合表观组等数据对APA的调控机制进行更加全面深入探讨；最后，也需要进一步阐明APA在疾病发生、胚胎发育等生命过程中的调控作用，为疾病诊断和治疗、胚胎发育机制的解析等提供新的线索和方法。本文聚焦测序技术对哺乳动物APA研究的推动作用，详细综述了APA的分类、发生机制及其在哺乳动物发育和疾病等过程中的作用，总结了基于转录组测序数据的APA识别方法和数据资源，并对APA研究进行了总结与展望。

1 APA的类型和发生机制

1.1 APA的类型

APA是指具有多个多聚腺苷酸化位点（polyadenylation site，PAS）的基因在其mRNA 3'端成熟过程中，由于选择不同的PAS，导致产生多个3'非编码区（UTR）长度和序列组成不同的转录本异构体[3]。APA在所有真核生物中广泛存在，参与基因表达调控，且在控制mRNA的稳定性、定位、翻译、蛋白质编码和定位等方面也至关重要[4]。APA的近端和远端PAS独立发生，且在许

* 国家自然科学基金（62371265，62161039），内蒙古自治区自然科学基金（2024JQ10）和内蒙古自治区直属高校基本科研业务费（2023RCTD023）资助项目。
** 通讯联系人。
邢永强 Tel：0472-5955917，E-mail：xingyongqiang1984@163.com
左永春 Tel：0471-5227683，E-mail：yczuo@imu.edu.cn
收稿日期：2024-07-06，接受日期：2024-08-25

多情况下以顺序方式识别[5]。根据不同的剪切和拼接方式，APA可概括为4种基本类型（图1）[6-7]：a. 串联3' UTR APA（tandem 3' UTR APA），发生在同一个末端外显子内，通过选择不同的PAS，产生不同长度3' UTR转录本，不影响DNA编码区结构，不改变蛋白质的氨基酸序列，PAS位于3' UTR；b. 可变末端外显子APA（alternative terminal exon APA，也称为skipped terminal exon），导致原本通过剪接跳跃的外显子成为末端外显子，影响DNA编码区结构，改变蛋白质的氨基酸序列，PAS位于这些末端外显子的相邻内含子中；c. 内含子APA（intronic APA，也称为composite alternative terminal exon），导致内部外显子延长并成为末端外显子，影响DNA编码区结构，会改变蛋白质的氨基酸序列，PAS位于该外显子的下游内含子区；d. 内部外显子APA（internal exon APA），发生在编码区的内部外显子区，影响DNA编码区结构，改变蛋白质的氨基酸序列，PAS位于外显子区。

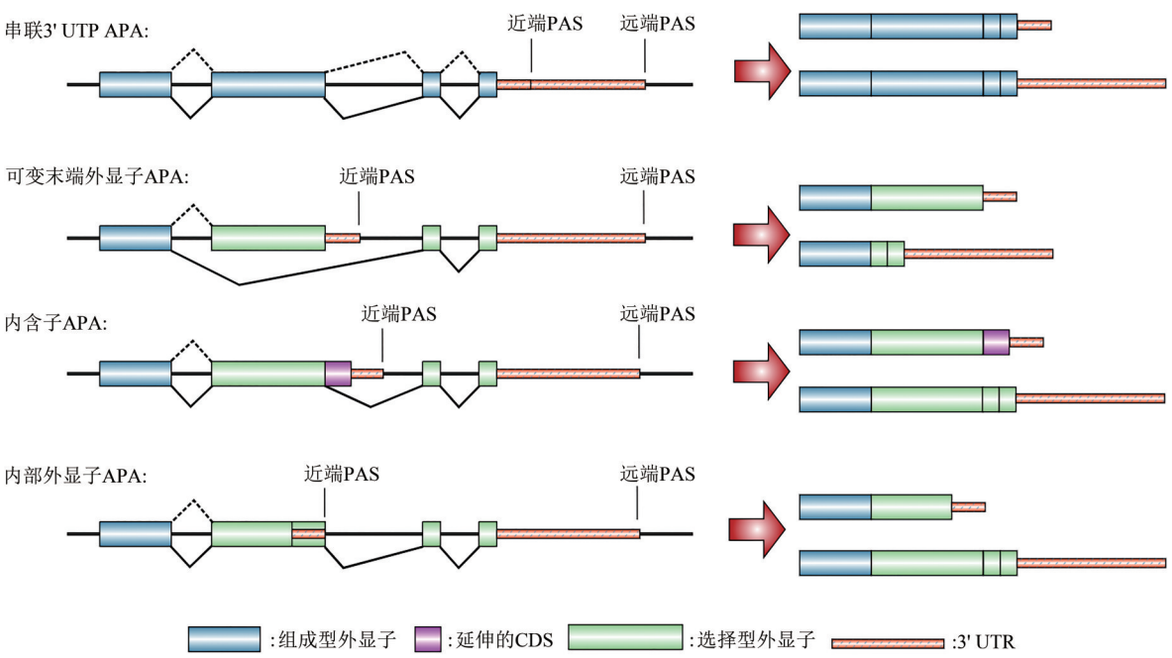


Fig. 1　Classification of APA

图1　APA的分类

虚线表示包含近端poly(A)转录本的剪接方式；实线表示包含远端poly(A)转录本的剪接方式。

1.2　APA的发生机制

APA的发生受大量顺式序列元件和反式结合蛋白的调控（表1）[8-9]。在pre-mRNA 3' UTR区包含大量调控APA发生的序列元件（图2），如上游序列元件（UGUA、PAS hexamer（AAUAAA）、胞质多聚腺苷酸化元件（cytoplasmic polyadenylation elements，CPE）、切割激活位点（cleavage activation site，CA））、下游序列元件（GU/U rich），以及多聚腺苷酸化结合元件（Musashi-binding element，MBE）、多聚腺苷酸化控制序列（translational control sequence，TCS）和多聚腺苷酸化增强子（Pumilio-binding element，PBE）等其他元件。其中，上游元件UGUA结合切割因子（CFI），切割和多聚腺酰化特异性因子（CPSF）结合PAS hexamer基序或其变体，CPE元件由PEB1蛋白识别。下游元件GU/U rich结合切割刺激因子（CstF）；MBE通过与多聚腺苷酸结合蛋白相互作用，影响mRNA的稳定性和翻译效率；TCS是一类位于基因启动子区域的元件，通过影响基因的转录效率，进而间接影响mRNA的多聚腺苷酸化；PBE是一类可以增强多聚腺苷酸化效率的元件，通常位于PAS附近或更远的基因组区域，通过与特定的转录因子结合来增强PAS的活性[10]。此外，还有一些辅助因子也参与促进poly(A)尾巴的形成。

Table 1　The binding of APA sequence regulatory elements with trans-acting proteins

表1　APA序列调控元件与反式蛋白的结合情况

序列元件名称	对应元件模体	反式蛋白类型	结合蛋白的组成
UGUA [6]	UGUA	CFI	CFI（CFIm）：CFIm25、CFIm68、CFIm59、CPSF5（NUDT21）、CPSF6、CPSF7
PAS六聚体 [6]	常见的PAS序列包括AAUAAA、AUUAAA、AGUAAA等	CPSF	CPSF：CPSF1、CPSF2、CPSF3、CPSF4、WDR33、FIP1（FIP1F1）
CPE [10]	UUUUA	CPEB1蛋白，且无需其他因子辅助	CPEB1
CA [6]	AAUAAA	CPSF、CstF、PAP，由CFII引发切割	CFII（CFIIm）：PCF11、CLP1
GU/U rich [6]	GU/U rich	CstF	CstF：CstF1、CstF2（CstF2T）、CstF3
MBE，TCS，PBE [10]	—	Symplekin（SYMPK）、PAP（及其不同异构体）、PAF1复合物、PABPN1、RBBP6、SCAF4、SCAF8等	—

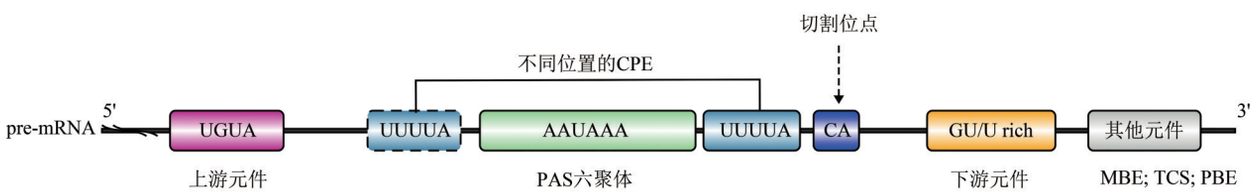


Fig. 2　The sequence regulatory elements of APA in the 3' UTR

图2　3' UTR的APA的序列调控元件

APA的发生分为切割和多聚腺苷酸化（cleavage and polyadenylation，CPA）两个步骤[6, 11-13]（图3）。首先，CPSF和CstF蛋白复合物结合在RNA聚合酶II（RNAP II）的羧基端结构域（CTD）。当RNAP II通过PAS时，CstF会转移到新的pre-mRNA上，CPSF会被释放并与PAS结合，由上一个pre-mRNA转移而来的CstF将与PAS下游元件结合。同时，CFIm复合物与PAS上游20~30 nt的元件结合，CFIIm复合物的PCF11亚基与RNAP II的CTD结合。然后，CPSF和CstF在PAS后约35 nt处（CA区）开始切割。聚腺苷酸化聚合酶（poly(A) polymerase，PAP）立即开始合成poly(A)尾部。几乎同时，细胞核中的聚腺苷酸化结合蛋白（PABPN1）与新形成的poly(A)序列结合，poly(A)合成过程受CPE影响，如CPE与PAS的相对位置及两者之间的距离均能线性地影响APA的程度，而且3' UTR内CPE的数量与多聚腺苷酸化的幅度呈正相关[10]。最后，CPSF开始解离，而PAP继续催化聚腺苷酸化并合成poly(A)尾，直至聚腺苷酸化停止并开始解离，而PABPN1继续保持结合状态出核。这个过程中，纯化的3'加工复合体含有约85种蛋白质，包括已知的核心3'加工因子和50多种可能与其他过程串扰的蛋白质，但这些蛋白质仍需进一步研究。

然而，3' UTR中近端和远端的PAS是如何进行选择？Tang等[5]提出了poly(A)渐进式聚腺苷酸化模型[5, 14]：在转录过程中，RNAP II会穿越上游的弱PAS，直至遇到下游的强PAS。这个强PAS能被CPA复合物精准识别并高效处理。在此过程中，如果近端PAS未与任何CPA成分结合，那么长3' UTR异构体将会被释放。相反，若受到CPA组分或其他潜在的核保留因子的约束，长3' UTR异构体将在核基质（nuclear matrix，NM）中受到限制，并进一步加工产生短3' UTR异构体。该模型表明，3'端加工动力学与PAS强度之间存在密切关联，且距离转录起始位点（transcription start site，TSS）较远的PAS通常比近端PAS信号更强，而相对较弱的近端PAS可以增强对远端强PAS的转录识别和处理。

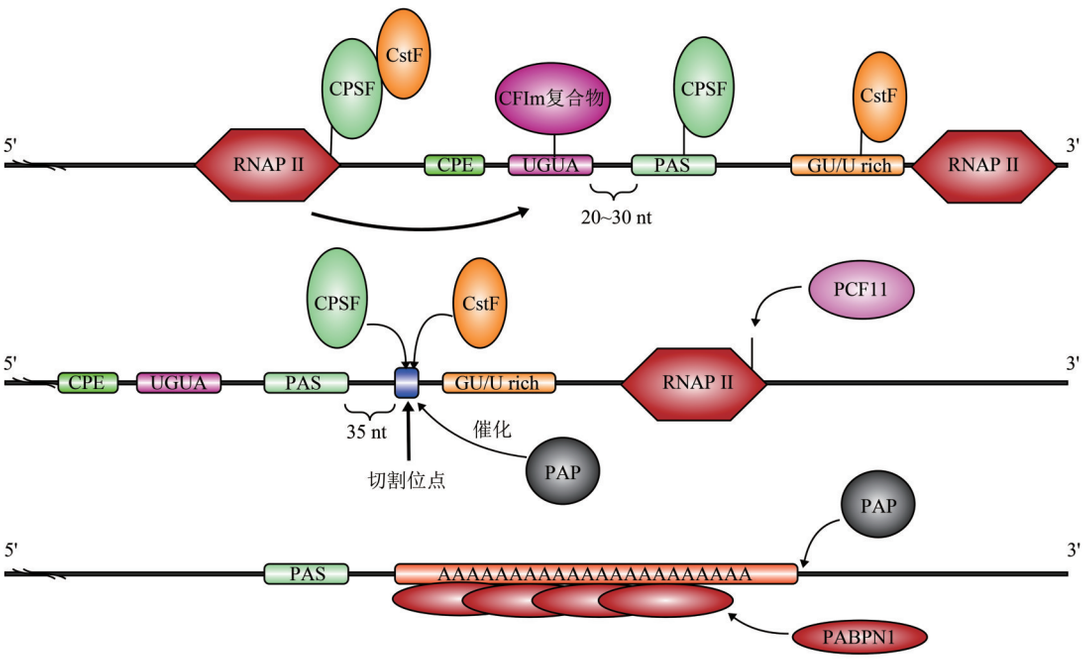


Fig. 3　The process of APA

图3　APA的发生过程

2　APA的功能

2.1　发育过程中的APA

APA与细胞增殖和分化状态密切相关，在胚胎发育中发挥着关键作用[15]。Li等[16]通过分析人类和小鼠着床前胚胎的单细胞RNA测序数据，研究了3' UTR长度的变化，并探讨了这些变化背后的RNA代谢途径。尽管人类和小鼠的3' UTR变化模式存在差异，但都与RNA代谢有很大关联。Agarwal等[17]提供了哺乳动物器官发生过程中APA的动态变化图谱，并发现在不同发育阶段的神经元细胞类型中，较长的3' UTR更为常见。针对小鼠胚胎和神经干细胞发育过程，Velten等[18]使用一种新的单细胞转录组学方法BATSeq对APA的使用进行全基因组定量，发现细胞的发育状态全局决定了亚型的使用，而且相同状态的单个细胞在APA亚型的选择上有所不同。因此，APA作为一种重要的转录后调控机制，对于哺乳动物的正常发育和生理功能至关重要。

除哺乳动物，其他物种的APA研究同样揭示了其在生物发育过程中的重要作用。在斑马鱼胚胎发育中，3' UTR的介导作用导致了显著的翻译抑制现象，这一发现突显了基因特异性的3' UTR在脊椎动物胚胎发育中的关键功能和重要性[19]。此外，Zhang等[20]探讨了果蝇神经系统发育过程中可变剪接和APA的协调作用。该研究指出，APA事件能够影响神经元的分化和突起生长，进而对神经网络的形成和功能产生影响。APA作为一种精细调控基因表达的转录后机制，在不同物种的胚胎发育和器官形成中起着核心作用，对理解生物体的发育过程和调控机制具有重要的科学价值。

2.2　疾病发生过程中的APA

APA在多种疾病的发生发展中扮演着关键角色。研究表明，APA的变化与癌症、神经退行性疾病以及心血管疾病等多种疾病具有密切的关联[21-22]。Zingone等[23]发现，APA介导肺癌中lncRNA和miRNA等非编码RNA的产生，指出APA的选择是调控非编码RNA表达的一种重要机制。通过该机制，APA能够调节基因表达水平和转录本的结构，进而影响细胞的生长、分化以及凋亡等关键生物学过程，最终导致疾病的发生和发展。

APA的异常表达在某些癌细胞中与肿瘤的发展和恶化密切相关[24]。为了深入探究APA事件在癌症中的调控作用，Wang等[25]对APA的不同亚型进行深入分析，特别关注了具有临床应用潜力的APA事件。通过一系列实验，他们发现CPSF1和

PABPN1是调控APA事件及三阴性乳腺癌细胞增殖的关键因子。当CPSF1或PABPN1的表达水平降低时，观察到细胞的增殖能力显著减弱，细胞凋亡率增加，细胞周期分布发生改变。此外，与乳腺癌肿瘤的发生、增殖、转移以及对化疗的敏感性相关的基因的APA模式也发生了逆转。这些发现不仅揭示了APA事件在癌症发展中的调控机制，而且为癌症的诊断和治疗提供了新的视角。特别是，针对CPSF1和PABPN1等关键因子的干预可能成为治疗癌症的潜在策略。未来的研究可以进一步探索APA事件在不同类型癌症中的作用，并开发针对APA事件的新型治疗手段，以提高癌症治疗的针对性和有效性。

在心血管疾病领域，APA对心脏发育和疾病过程中的基因表达调控具有显著影响。Cao等[26]综述了APA在心脏发育和心血管疾病中的生物学作用及其潜在的治疗应用。在心脏发育的不同阶段，特定基因的APA异构体可能具有不同的表达模式，这些模式对于心脏的正常发育至关重要。在心血管疾病的发展过程中，APA事件的改变可能导致心肌细胞的异常生长和分化，从而引发病理性心脏重塑。为了应对这一挑战，研究者们正在探索pre-mRNA靶向策略，以纠正心血管疾病中基因的异常APA模式。这些策略利用了RNA干扰和CRISPR-Cas9等技术，为调节APA提供了新的工具。通过这些技术，可以精确地调控特定基因的表达，恢复或改变其APA模式，从而有望治疗或预防心血管疾病。此外，基于RNA的疫苗治疗法，如mRNA疫苗，已经证明了其在治疗和预防人类疾病中的潜力。这些技术同样适用于心血管疾病的治疗，为开发新的治疗手段提供了新的思路。通过深入理解APA在心脏发育和疾病中的作用，可以为心血管疾病的治疗提供更为精准的分子靶点，推动基于RNA治疗策略的发展。

Singh等[27]发现，内含子多聚腺苷酸化（intronic polyadenylation，IPA）在正常人类免疫细胞中普遍表达，并且其表达模式在免疫细胞亚型中通常保持稳定。然而，在癌症状态下，IPA的表达模式出现异常，揭示了其在肿瘤免疫微环境中的重要作用。Devany等[28]发现，IPA位点的激活导致细胞周期抑制蛋白p21（CDKN1A）截短型异构体的产生。这种截短型异构体的产生影响其正常的细胞周期调控功能。同时，较短mRNA异构体的积累为CDKN1A在UVC（紫外线C波长）诱导的DNA损伤响应期间的调控机制提供了一种新的视角。此外，IPA与肿瘤的发生发展也密切相关，Zhao等[29]发现，一些IPA动态变化在肿瘤样本中表现出上调趋势，这可能影响抑癌基因的功能，如IPA类型的*TSC1*基因产生截断蛋白，从而失去抑制mTOR信号通路的能力。随着对IPA类型功能和调控机制的深入了解，迫切需要开发和完善新的IPA识别工具，以促进该领域的研究进展。

3 APA的识别

在现代生命科学研究中，测序技术的重要性日益凸显，其发展极大地促进了APA的识别。下面系统介绍基于表达序列标签（expressed sequence tags，EST）、混池转录组（bulk RNA-seq）、单细胞转录组（scRNA-seq）、全长转录组（full-length transcriptome）、空间转录组（spatial transcriptome）等技术识别APA事件的研究进展（表2）。

Table 2 Identification of APA

表2 APA的识别

实验技术	识别
EST	利用EST序列与基因组序列的差异，检测poly(A)尾巴的长度和位置的变化[30-33]
Bulk RNA-seq	使用PolyAMiner-Bulk、APATrap和QAPA等方法全面地检测基因的表达水平和转录组的结构，从而揭示APA事件的发生和调控[34-49]
scRNA-seq	在揭示细胞异质性的同时，可利用scDapars、TCEtool等方法学鉴定单细胞水平的APA事件[50-62]
Full-length transcriptome	能够提供全长转录本信息，从而更准确、全面地鉴定APA事件的类型和位置[63-68]
Spatial transcriptome	能够在空间分辨率下识别组织或器官中的转录本和APA事件，为研究细胞发育和疾病发生机制提供了关键技术[69]

3.1 基于表达序列标签识别APA事件

EST是一种通过测序和分析cDNA库中的短序列标签识别转录本的方法。高通量测序技术出现之前经常基于EST技术识别APA，其基本原理是利用EST序列与基因组序列的差异识别APA事件，这种方法的关键在于准确地对EST序列进行聚类

和拼接，以获得完整的mRNA序列[30-31]。其优势包括：a. 覆盖范围大。EST库中包含了大量的cDNA序列标签，覆盖了全基因组的转录本信息，因此可以用于在全基因组范围识别基因的APA事件。b. 发现转录本多样性，APA可导致一个基因产生多个不同的mRNA转录本。通过EST数据比对，可以发现这些基因的不同PAS，进而揭示APA调控下不同组织、发育阶段或疾病状态下基因多样性转录。c. 辅助基因结构预测，利用EST可以更准确地预测基因的结构，包括UTR的长度和编码区域的组成。d. 应用广，EST库中的转录本信息可以用于不同物种和组织类型的APA事件分析，具有较广泛的适用性。e. 可扩展，EST库中的转录本信息可以作为其他高通量测序数据的验证和补充，有助于验证APA事件的真实性。

然而，基于EST的APA识别方法也存在一些局限性[32]。首先，EST测序技术本身在建库效率、测序深度和测序质量等方面存在一些问题，这些因素可能影响APA识别的准确性。其次，EST数据集的大小和质量可能影响APA事件的覆盖度和分辨率。为了克服这些问题，需要采用更为先进的高通量测序技术和方法。例如，基于3'端的测序技术，通过特异性捕获和分析mRNA的3'端，为探索转录后调控提供了直接的分子证据。Shepard等[33]开发了一种基于深度测序的方法PAS-Seq，用于在转录组水平上定量分析RNA多聚腺苷酸化，不仅鉴定出比基于整个小鼠EST数据库方法更多的PAS，而且还检测到全局APA谱的动态变化。

3.2　基于混池转录组识别APA

基于bulk RNA-seq鉴定APA为人们提供了一个全面深入了解APA事件的方法[34-35]。在bulk RNA-seq鉴定APA的上游实验阶段主要包括如下步骤：a. 收集样本；b. 测序文库构建，包括典型的二代转录本测序建库方法和3'端建库方法；c. 测序，通过Illumina等高通量测序平台完成，获得FASTQ格式的原始测序数据。近年来，A-seq[36]、3P-seq[37]、3'READS[38]、PAS-seq[39]和PolyA-seq[40]等直接用于测定RNA分子3'端序列的高通量测序技术已被报道。A-seq技术是一种通过连接适配器序列到RNA分子3'端，并利用反转录和PCR技术来确定转录本末端序列的方法；3P-Seq技术通过化学标记RNA分子3'端的磷酸基团，富集并测序，研究RNA末端修饰；3'READS技术通过富集RNA分子3'端进行高通量测序；PAS-seq技术专注于分析mRNA 3'端poly(A)信号和长度；PolyA-seq技术通过富集poly(A)尾RNA分子，分析转录本3'端特征和poly(A)尾变化。这些技术为实验上测定APA序列提供了多种策略。在下游生物信息学分析阶段，首先使用HTSeq、featureCounts等工具对基因和转录本进行表达定量，为后续的APA事件分析提供基础数据。接着，利用PolyAMiner-Bulk、APATrap、QAPA等算法准确鉴定APA事件（表3）。

PolyAMiner-Bulk、APATrap、QAPA、Aptardi、IPAFinder和TAPAS等方法利用不同的算法基于bulk RNA-seq数据识别和量化PAS。PolyAMiner-Bulk通过深度学习模型C/PAS-BERT提供了一种强大的分析bulk RNA-seq数据的方法，尤其擅长处理大规模数据集，并能够揭示APA的动态变化。它的设计旨在克服传统方法在处理大量数据时的局限性，并且在实际应用中已经显示出了高效性和准确性。APATrap、QAPA和Aptardi则更侧重于预测和定量APA事件，可以通过分析RNA-seq数据来识别PAS。这些工具在样本特异性分析和APA预测方面中表现出了良好的性能。IPAFinder作为一个较新的工具，虽然在比较研究中提及较少，但它的开发表明了APA分析工具的快速发展和迭代趋势。TAPAS因其在基准测试中的表现而受到关注，它能够准确地从bulk RNA-seq数据中预测PAS。与其他方法相比，TAPAS在某些情况下（如弱监督学习、异步训练环境、关注难以分类的样本以及端到端训练方法等方面）显示出更高的准确性和可靠性。

Jonnakuti等[48]使用PolyAMiner-Bulk揭示了硬皮病病理学中全新的APA动力学，并识别出该疾病基因中的差异性APA。该算法不仅处理数据通量高，而且识别APA准确性高。但也存在一些缺点，如计算资源需求高、参数复杂性高、包或工具依赖性强等。Ye等[43]开发了APATrap方法，并成功应用于模拟数据以及人类和拟南芥的RNA-seq数据集的APA识别。通过与ChangePoint和DaPars两种方法的比较，证明了APAtrap适用于任何具有注释基因组的生物。APATrap在识别新的PAS和改进基因组注释的同时，还可统计所有潜在的PAS。此外，还可以从基因组序列中识别和分析启动子区域，找出其中可能存在的转录调控元件，如转录因子结合位点、增强子等，并对潜在调控元件序列特征预测。QAPA基于RNA-seq数据准确定量APA事

Table 3 The methods for identifying APA based on bulk RNA-seq

表3 基于bulk RNA-seq识别APA的方法

方法	原理	识别APA类型	输入文件格式	分析对象	发表时间
Dapars [41]	一个从头开始设计的工具，可以识别基因模型中的近端PAS以及评估长短3' UTR的表达水平	3' UTR APA	Sample.wig（sorted.bam转化）	比较不同样本之间的APA差异性，每组样本可有多个wig文件输入	2014.11
QAPA [42]	QAPA通过分析RNA-seq数据中的读段覆盖情况，推断出PAS的使用情况。使用PAU指数（3' UTR亚型的相对使用率）衡量PAS使用频率	3' UTR APA	quant.sf（Salmon生成）	分析两组样本之间的APA差异性；每组样本单个文件输入	2018.03
APATrap [43]	基于均方误差模型进行APA的识别和量化	新3' UTRs、3' UTR延伸的APA	bedgraph文件（sorted.bam转化）	分析两组样本之间PAS使用率的百分比差异以及两组之间远、近端位点的使用相关性；每组样本单个文件输入	2018.06
TAPAS [44]	基于时间序列分析的方法识别数据中的关键变化点，并引入一系列过滤技术排除那些可能被误判为APA的假阳性点	3' UTR APA	sqa_data（序列质量评估：原始的DNA、RNA或蛋白质序列数据）	多组序列数据文件	2018.08
MoutainClimber [45]	通过查找读取覆盖率中的显着变化点来识别RNA-Seq中的APA	远端poly(A)、串联APA、IPA、内部外显子APA	bedgraph文件	单组APA/两组样本差异APA均可识别；每组样本单个文件输入	2019.10
APAlyzer [46]	通过比较在高质量PAS划定的区域进行测序读数，检查3' UTR APA、IPA和基因表达变化	3' UTR APA、IPA	sorted.bam	单组APA/两组样本差异APA均可识别；每组样本单个文件输入	2020.06
Aptardi [47]	利用机器学习范式中的DNA序列和RNA测序来预测表达的PAS	3' UTR APA	aptardi.gtf（sorted.bam转化）	分析每组转录本的3'末端外显子并注释3'；单个文件输入	2021.03
IPAFinder [29]	通过分析RNA-seq数据中的读取覆盖率来推断内含子PAS的使用，并基于“change point”模型进行*de novo*定量IPA事件	IPA	sorted.bam	分析两组样本之间的IPA位点使用情况差异性；每组多个文件输入	2021.11
PolyAMiner-Bulk [48]	基于深度学习模型C/PAS-BERT准确推断和解码APA动力学	3' UTR APA	sorted.bam	比较不同样本之间的APA差异性，每组样本可有多个文件输入	2023.06
InPACT [49]	通过检查上下文序列模式和RNA-seq读数比对来识别和量化IPA位点	IPA	sorted.bam（需bai文件）	单组样本单个文件输入	2024.03

件，它通过提取和注释3' UTR序列，并基于转录本水平的丰度计算替代3' UTR异构体的相对使用率。Ha等[42]使用QAPA方法，基于RNA-seq数据估计差异3' UTR异构体表达，并准确推断APA的变化，成功应用它分析了小鼠胚胎干细胞分化为谷氨酸能神经元过程中的APA。然而，QAPA在设计时只针对特定的物种和注释数据库，这可能限制了其在其他物种或缺乏高质量注释的物种上的应用，而且在区分不同的APA异构体时也存在一定的局限性，尤其是在APA事件导致的异构体差异较小或者APA信号较弱的情况。

这些工具基于不同的原理从测序数据中识别APA位点和类型。通过深入分析这些数据，研究人员可以揭示APA在不同组织或发育阶段中的动态变化和调控规律。然而，bulk RNA-seq在检测APA事件时也存在一些挑战。首先，bulk RNA-seq的分辨率可能受到测序深度和建库质量的影响，导致一些差异性较小的APA事件难以被检测到。其次，bulk RNA-seq数据通常来源于大量细胞的混合样本，这可能掩盖了特定细胞群体中存在的APA事件的差异。基于单细胞测序技术识别APA可以有效解决上述问题。

3.3　基于单细胞转录组识别APA

单细胞转录组测序技术的快速发展进一步促进了APA的研究[50]。scRNA-seq能够全面地检测单个细胞的基因表达水平和转录组结构，从而揭示APA事件的细胞特异性和不同细胞群间的差异APA事件。目前已有SCAPE、MAAPER、scAPAtrap、scDAPA、scDaPars、TECtool等一系列方法用于分析scRNA-seq数据中APA事件[51]（表4）。SCAPE基于贝叶斯方法，利用DNA片段的插入大小信息从头识别和量化PAS；scAPAtrap专注于识别和量化单细胞中的PAS，特别适合处理3'端富集的scRNA-seq数据，并且能够精确定位低覆盖率的PAS；TECtool使用mRNA和3'端测序数据识别新型末端外显子，能可视化bam文件里所有的IPA，并标注基因及其所在位置。MAAPER、scDAPA和scDaPars在识别scRNA-seq数据中的APA事件的同时，可进行降维分析，更直观的展示了细胞间的APA差异性。

Table 4　The methods for identifying APA based on scRNA−seq

表4　基于scRNA−seq识别APA的方法

方法	原理	识别APA类型	输入文件格式	分析/输入对象	发表时间
scDAPA[52]	以bam/sam文件和细胞簇标签为输入，基于直方图的方法和Wilcoxon秩和检验检测APA动态	3' UTR	序列比对结果（bam/sam文件）和细胞注释信息（csv文件）	识别单组样本APA事件；每组单个文件输入	2020.02
scAPAtrap[53]	基于3'端建库生成的scRNA-seq数据（10x、CEL-seq、Drop-seq），通过峰值识别和PAS序列的读取锚定，有效识别并精确定位PAS	3' UTR	sorted.bam	识别单组样本APA事件；每组单个文件输入	2021.07
MAAPER[54]	基于概率模型的计算方法，利用nearSite（接近于PAS的测序读段）读段进行APA分析	3' UTRs和IPA	sorted.bam（需bai文件）	识别两组样本之间差异APA事件；每组可有多个文件输入	2021.08
scDaPars[55]	输入基因组覆盖率数据，形成线性回归模型，共同推断近端PAS的确切位置，进而生成稀疏APA矩阵构建最近邻图，再使用非负最小二乘（NNLS）回归模型来细化相邻细胞基因的PDUI	3' UTR	PDUI.txt（Dapars v2.0生成）	稀疏APA矩阵	2021.10
TECtool[56]	根据转录终点的数量、位置和表达水平等信息，对APA事件进行定量和定性分析，从而识别基因的不同APA形式	IPA	sorted.bam（STAR比对）	识别单组样本APA事件；每组单个文件输入	2021.10
scUTRquant[57]	基于Snakemake管道，使用3'端标签定量3' UTR亚型	3' UTR	sorted.bam（CellRanger输出）	识别单组样本APA事件；每组可有多个文件输入	2021.11

续表4

方法	原理	识别APA类型	输入文件格式	分析/输入对象	发表时间
scAPA[58]	通过识别细胞中不同基因的3' UTR区的APA，从而揭示细胞间APA的差异性，适用于带3'标签scRNA-seq数据	3' UTR	sample.name.bam（CellRanger输出）	识别单组样本APA事件；每组可有多个文件输入	2022.01
SCAPE[59]	将每个PAS的分布特性建模为高斯分布，并为每个亚型分配相应的权重，权重的分配反映了不同亚型在总体分布中的贡献程度	3' UTR	sorted.bam	识别单组样本APA事件；每组单个文件输入	2022.06
SCINPAS[60]	通过提取含有poly(A)的读段来识别PAS，并且修改了读取去重步骤	3' UTR	3'_sorted.bam（CellRanger输出）	识别单组样本APA事件	2023.09

在单细胞APA的研究中，Shulman等[61]通过对不同组织来源的数据集进行综合分析，揭示了APA在不同细胞类型中的广泛调控作用，该调控作用导致内含子PAS的全局性3' UTR长度变化及切割效率的增强，为理解APA的生物学意义提供了重要视角。Zhou等[59]在单细胞水平上提出了一种基于贝叶斯统计的分析方法SCAPE，通过对36个小鼠器官样本中的31 558个位点进行鉴定，验证了该方法的准确性与鲁棒性。此外，作者使用SCAPE揭示了APA亚型与miRNA结合位点的相关性，并指出APA事件的调控具有组织特异性、细胞类型特异性及肿瘤特异性。SCAPE的优势在于其结构感知能力，能够捕捉文本结构信息，从而深入理解基因的多聚腺苷酸化选择性及其调控机制，能够考虑文本上下文信息，全面理解APA选择性的影响因素，以及通过预训练学习丰富的语言模型，提高对基因组数据的分析能力。SCAPE面临的挑战包括对大量标注数据的需求、预训练过程中的计算资源和时间成本，以及深度学习模型内部机制的复杂性导致的解释性问题。Li等[54]提出的MAAPER方法，基于概率模型，以高精度和高灵敏度预测PAS，并通过统计学方法验证了不同APA事件的存在。MAAPER在单细胞转录组数据中展示了其性能，并适用于未配对或配对的实验设计。该方法的优势在于其多方面注意力机制，能够同时关注基因序列特征和上下游序列信息，全面理解APA的调控机制，具有在不同基因组数据中的鲁棒性，适用不同细胞类型和物种的数据特征，以及在PAS识别和预测方面的高准确性。尽管如此，MAAPER在训练过程中需要大量标注数据，参数调优和模型调整的复杂性，以及深度学习模型的解释性限制，都是该方法面临的挑战。

scAPAtrap方法由Wu等[53]开发，该工具结合了峰值鉴定和poly(A)读长锚定技术，能够精确识别那些读长覆盖率较低的PAS。scAPAtrap的一个显著特点是其能够在无需先验基因组注释的情况下识别PAS，这有助于在以前未被充分研究的区域中发现新的PAS，并改进基因组注释。scAPAtrap的优势在于其专注于单细胞分析，能够整合不同来源的scRNA-seq数据，以及在APA事件识别和定位方面表现出的高灵敏度和准确性，使其成为探索细胞间APA动态变化和单个细胞中APA亚型异质性表达的有力工具。然而，该工具也存在数据处理复杂性、对输入数据质量的依赖性，以及结果解释性限制等挑战。Gao等[55]开发了scDaPars（图4）。scDaPars能够利用3'端（10×Chromium）或全长（Smart-seq2）数据，在单细胞和单基因分辨率下精确定量APA事件，展现了高分辨率分析的能力。通过真实和模拟数据的验证，scDaPars显示出了对由少量mRNA测序引起的单细胞中缺失APA事件的稳健恢复能力，即在鲁棒性方面表现优异。在癌症和人类内胚层分化数据的应用中，scDaPars不仅揭示了细胞类型特异性的APA调控，还发现了常规基因表达分析难以观察到的细胞亚群，从而有助于深入理解转录后水平上的细胞异质性。尽管如此，scDaPars的性能受限于高质量的scRNA-seq数据，且在处理大规模数据集时需要较大的计算资源。

scRNA-seq在检测APA事件时也存在一些挑战。首先，scRNA-seq的测序深度和建库质量可能影响APA事件的检测准确性。其次，由于单细胞测序技术的限制，可能存在一些技术偏差和误差，

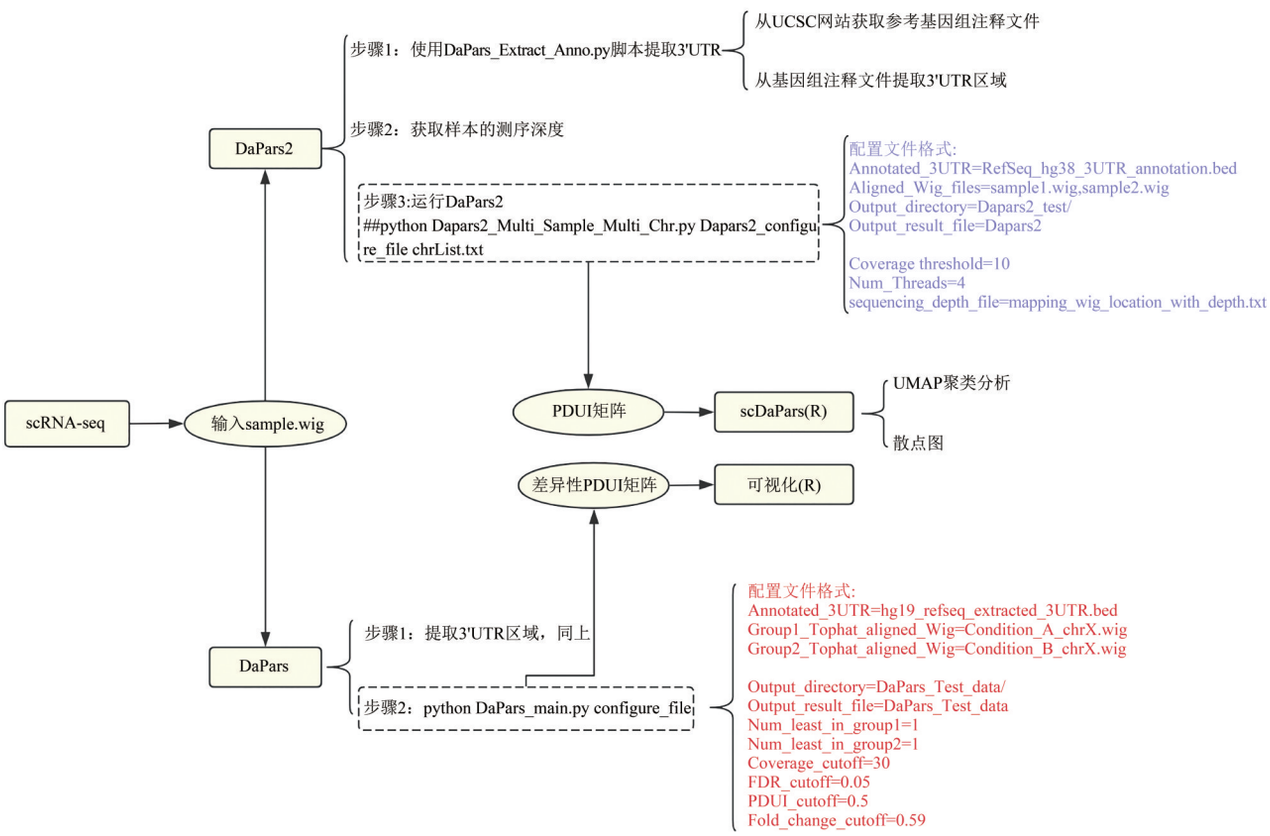


Fig. 4　The workflow of identifying APA with DaPars

图4　DaPars识别APA的流程

需要采用适当的校正方法和标准品进行质量控制。为了克服这些问题，需要进一步优化单细胞测序技术和数据处理方法，提高APA事件检测的准确性和可靠性。因此，Wang等[62]开发了单细胞多聚腺苷酸化测序（single-cell polyadenylation sequencing，scPolyA-seq）来识别单细胞分辨率下的APA，scPolyA-seq技术通过富集和测序带有poly(A)尾巴的mRNA分子，从而获得与poly(A)相关的信息，相比于传统的scRNA-seq测序技术，scPolyA-seq在测序读长中包含了较高比例的PAS相关序列，显著增强了对基因组中特定PAS精确定位的能力，进而使得scPolyA-seq技术在APA分析中更为适用。

综上，scRNA-seq为APA事件的检测提供了更为精确的手段，有助于更深入地研究APA的调控机制和生物学功能。随着单细胞测序技术的不断发展，有望更全面地揭示APA事件的探索，为疾病诊断和治疗提供新的思路和方法。

3.4　基于全长转录组识别APA

全长转录组测序技术（如PacBio和Oxford Nanopore测序平台）提供了较高的准确度和较长的读长，使得全长转录本的测序结果更加准确和完整[63]。与传统的基于短读段的测序方法不同，全长转录组测序能够直接测定包含5' UTR、编码区域、3' UTR以及poly(A)尾部的完整mRNA分子，所以无需将短读段拼接成长读段，从而减少了拼接错误和遗漏的可能性。为了更加全面系统的鉴定基因3' UTR中的APA事件，可以基于全长转录组数据识别APA事件[64]。

全长转录组分析APA通常包括以下步骤。a. 数据预处理：对全长转录组数据进行质量控制和预处理，包括去除低质量序列、去除接头序列和去除重复序列等步骤。b. 3' UTR定位：通过比对全长转录组数据到参考基因组，确定基因的3' UTR。c. APA事件鉴定：通过分析3' UTR的PAS，鉴定基因的不同APA事件。d. 功能分析：对鉴定出的APA事件进行功能富集分析、通路分析和基因调控

网络分析等，了解APA事件对基因功能和调控的影响。

目前基于全长转录组数据鉴定APA的方法有PRAPI、tappAS、TAPIS、ASAPA等（表5）。PRAPI由Gao等[65]提出，是一个多功能的Iso-Seq分析工具，它不仅可以处理ATI、AS、APA和circRNA的差异表达分析，还能整合Iso-Seq全长异构体与短读长数据，具备新基因注释和错误注释校正的功能。PRAPI的可视化工具能够生成高品质的矢量图形，但其主要限制在于它主要针对Iso-Seq数据设计，可能在其他类型的全长转录组数据上的应用受到限制。tappAS由de la Fuente等[66]开发，专注于功能性异构体的全长转录组分析，特别强调对APA事件的功能影响进行评估。tappAS的优势在于其能够对编码和非编码功能结构域、motifs和位点进行异构体解析注释，并且集成了多种生物信息学工具和数据库，提供全面的数据分析支持。此外，tappAS可能需要特定全长转录组类型的数据输入，并且对计算资源有较高需求。TAPIS由Abdel-Ghany等[67]开发，用于鉴定全长剪接异构体和APA。TAPIS的优势在于其高度准确性和利用全长信息的能力，但同样面临着数据需求高、计算资源消耗大等挑战。通过Wang等[68]在拟南芥数据上的测试，ASAPA展示了其在揭示基因表达调控现象方面的潜力，如剪接过程中内含子的同步性和可变剪接与可变转录起始或APA之间的耦合现象。综合来看，PRAPI、tappAS、TAPIS和ASAPA四种工具各有其优势和应用场景。tappAS专注于功能性异构体分析；PRAPI适合进行Iso-Seq数据的全面分析，包括新基因注释和错误校正；TAPIS作为一个鉴定全长剪接异构体和APA的工具，它的准确性和利用全长信息的能力使其在某些研究中非常有用；ASAPA则通过分析揭示了基因表达调控中的复杂相互作用，尤其是在探索APA事件间的关联性方面表现出色。

Table 5　The methods for identifying APA based on full−length transcriptomes

表5　基于全长转录组识别APA的方法

方法	原理	实验平台	发表时间
PRAPI[65]	Iso-Seq分析的一站式解决方案，可分析选择性转录起始（ATI）、可变剪接（AS）、APA、天然反义转录本（NAT）和环状RNA（circRNA）	PacBio	2018.05
tappAS[66]	使用编码和非编码功能域、基序和位点的亚型注释3' UTR，并引入新的量化方法来评估APA事件的功能影响	PacBio	2020.05
TAPIS[67]	用于分析全长剪接异构体和PAS的转录组分析方法，通过比对和拼接的方式，对转录组数据进行重建，识别出基因组中的包含APA在内的各种剪接亚型	PacBio或Oxford Nanopore	2016.06
ASAPA[68]	通过分析长读长序列，识别Iso-Seq数据中的AS、ATI和APA事件	PacBio或Illumina HiSeq	2024.03

基于全长转录组测序鉴定APA，在揭示基因表达调控机制、发现新的基因调控事件以及疾病诊断和发育机制解析等方面具有重要价值。然而，该技术仍面临一些挑战。a. 数据处理复杂：全长转录组数据的处理需要大量的计算资源和时间。b. 误差率高：由于全长转录组数据的复杂性，分析过程中可能会引入一定程度的误差，特别是在对低丰度基因的检测上容易出现假阳性或假阴性结果。c. 生物学解释的复杂性：APA的鉴定只是第一步，解释其生物学意义需要进一步的功能实验和验证，这可能需要大量的时间和资源。d. 数据分析的标准化和统一性：由于全长转录组数据的分析方法和流程并不统一，不同实验室或研究者的结果可能存在一定程度的差异，因此需要更多的标准化和统一化的方法来进行比较和验证。尽管存在这些挑战，随着技术的进步和方法的改进，基于全长转录组的APA鉴定在转录组学研究中已展现出广阔的应用前景。

3.5　基于空间转录组识别APA

随着生物技术的不断发展，对APA的研究已经进入了新的阶段。空间转录组学作为一种新兴技术，能够在单细胞水平上解析组织或器官中的基因表达情况，为研究细胞发育和疾病发生机制提供了有力工具。然而，如何利用空间转录组学数据鉴定APA是一个具有挑战性的问题。例如，如何精确检测低丰度mRNA的PAS、如何解析空间异质性以及如何解析长距离的基因表达调控等。为了克服这些难题，需要进一步发展相关技术和方法。

Ji等[69]开发了stAPAminer工具包，用于从空间条形码的ST数据中挖掘APA的空间模式，进而鉴定和量化APA。stAPAminer的优点包括：a. 空

间模式挖掘，专门设计用于分析空间分辨的转录组数据，能够揭示APA事件在不同空间位置的分布和模式；b. 有效的数据填补，该工具包采用基于*k*近邻算法的插补模型，恢复APA信号，有助于在复杂的空间数据中准确识别APA；c. 减少数据稀疏性，通过与基因表达剖面结合，能够插补PAS的使用情况，从而减少数据的稀疏性，提高分析的准确性；d. 可重复性验证，识别的空间APA模式在不同重复实验中具有可重复性，验证了其结果的可靠性。目前，stAPAminer受限于输入数据的质量和覆盖度，PAS识别的准确性和完整性还有待提高。基于空间转录组数据，通过识别不同组织或细胞类型间表达差异的基因，进一步分析这些基因的PAS来鉴定APA事件。

实验上可以利用CLIP-seq或polyA-seq等分子标记技术，通过特异性标记PAS，在单分子水平上解析mRNA的长度和稳定性，从而在不同条件下比较标记结果以鉴定差异APA事件。还可利用smFISH或MERFISH等单分子荧光原位杂交技术，通过直接观察mRNA在细胞中的分布和定位来鉴定APA事件。这些方法共同为深入理解APA的生物学功能和作用机制提供了强有力的工具。

4　APA的数据资源

目前基于高通量测序技术产生的转录组测序数据，已发布了一些APA公共数据资源，如scAPAdb、polyA_DB和Animal-APAdb等（表6）。这些数据库主要涵盖了真核生物poly（A）尾巴基因、可变剪接、转录多样性、多聚腺苷酸化等信息。其中包括了存储和查询poly(A)尾巴基因信息的PolyA_DB和PolyA_DB 3，专注于APA事件和可变剪接的ASTD和APASdb，以及评估遗传变异对APA影响的SNP2APA等数据库。此外，还有针对不同细胞类型、组织的APA事件的APAatlas和Animal-APAdb，以及单细胞水平的可变剪接和APA事件的scAPAdb和scAPAatlas。最近发布了PolyASite 2.0、3'aQTL-atlas和ipaQTL-atlas等3' UTR APA数据库，以及涉及人类癌症中APA事件功能的CAFuncAPA。此外，APADB数据库提供了关于脊椎动物PAS和单细胞转录组信息。这些数据库为研究人员提供了丰富的数据资源和工具，有助于深入理解基因调控和转录组学领域的相关信息。

这些数据库大多数基于bulk RNA-seq和scRNA-seq测序数据生成，不仅提供了基因组区域和PAS的信息，还提供了针对这些数据的分析和工具，如评估生物信息学方法鉴定的PAS的真实性、遗传变异对APA影响等。这些工具有助于研究人员进一步理解APA的基因表达调控机制和生物学功能及致病机制。此外，这些数据库中还提供了可视化工具和API接口，使得研究人员可以更方便地获取和处理数据。在获取和使用APA数据资源时，需要注意数据的可重复性和可验证性。不同的测序技术和实验条件可能导致数据的质量和可比性存在差异。因此，在使用这些数据资源时，需要进行适当的质控和标准化处理，确保分析结果的可靠性和准确性。

Table 6　The list of APA databases

表6　APA数据库列表

名称	数据库简介	网址	发布时间
polyA_DB [70]	一个专门用于存储和查询真核生物poly(A)尾巴基因的数据库，收集了大量来自不同物种和细胞类型的poly(A)尾巴测序数据，并提供了丰富的工具和资源来进行数据分析和挖掘	https://exon.apps.wistar.org/PolyA_DB/	2005.01
ASTD [71]	一个专注于可变剪接和转录多样性的数据库，涵盖了APA事件的数据	http://www.ebi.ac.uk/astd	2009.03
APASdb [72]	具有可视化所有基因在全基因组范围内不同APA亚型的精确图谱和使用量化的功能，现已更新到2.0版本	http://genome.bucm.edu.cn/utr/	2015.01
PolyA_DB 3 [73]	数据库增加了几个新功能，包括其他七种脊椎动物中人类PAS的同源基因组区域以及与PAS相邻的顺式元素信息，更新后的数据库旨在扩大脊椎动物基因组中PAS的覆盖范围，并提供评估生物信息学鉴定的PAS真实性的方法	https://exon.apps.wistar.org/polya_db/v3/	2018.01

续表6

名称	数据库简介	网址	发布时间
SNP2APA [74]	用于评估遗传变异对人类癌症APA影响的数据库，使用来自癌症基因组图谱（TCGA）和癌症3' UTR Altas（TC3A）的9 082个样本的基因型和APA数据，系统地鉴定了影响32种癌症类型中APA事件的 SNP，并将其定义为APA数量性状位点（apaQTLs）	http://gong_lab.hzau.edu.cn/SNP2APA/#!/	2020.01
PolyASite 2.0 [75]	来自3'端测序的APA的综合图谱，通过对所有数据的综合处理，识别并聚类了间隔很近并共享APA信号的位点	https://polyasite.unibas.ch	2020.01
APAatlas [76]	系统地从53个人体组织的9 475个样本中鉴定了APA事件，并检查了它们与跨组织的多种性状和基因表达的关联	https://hanlaboratory.com/apa/	2020.01
Animal-APAdb [77]	收集了多种动物的APA事件信息，包括了PAS的位置信息、调控因子信息、组织特异性表达信息等	http://gong_lab.hzau.edu.cn/Animal-APAdb/	2021.01
scAPAdb [58]	专门用于储存和分析单细胞水平的AS和APA事件的数据库。scAPAdb收集了来自scRNA-seq数据的APA事件信息，并提供了丰富的工具和资源进行数据分析和挖掘。	http://www.bmibig.cn/scAPAdb	2022.01
scAPAatlas [78]	用于探索不同细胞类型的APA，并解释潜在的生物学功能，基于来自24个人类和25个小鼠正常组织的精选scRNA-seq数据，探索人类和小鼠不同细胞类型的细胞类型特异性APA事件	http://www.bioailab.com:3838/scAPAatlas	2022.01
3' aQTL-atlas [79]	一个专注于人类正常组织中3' UTR APA的遗传影响的数据库，旨在帮助解释全基因组关联分析（GWAS）中识别的非编码单核苷酸多态性（SNPs）的功能	https://wlcb.oit.uci.edu/3aQTLatlas	2022.01
ipaQTL-atlas [80]	一个专注于IPA事件的数据库，提供了一个全面的资源来研究这些事件在人类不同组织中的分布及其与疾病风险的关联	http://bioinfo.szbl.ac.cn/ipaQTL	2023.01
CAFuncAPA [81]	该数据库包含了15 478例人类泛癌，旨在帮助研究人员系统地注释人类癌症中APA事件的功能	https://relab.xidian.edu.cn/CAFuncAPA/	2023.01
APADB [82]	一个通过3'末端测序确定的脊椎动物APA数据库，使用对互补DNA末端的大量分析，存储的信息为数以千计的PAS和APA事件提供了实验证据	http://tools.genxpro.net/apadb/	2023.09
scTEA-db [83]	通过分析53 069个公开可用的单细胞转录组，提供了12 063个迄今为止尚未注释的末端外显子和相关转录亚型	www.scTEA-db.org	2024.01

5　总结与展望

本综述对APA的概念、调控机制以及其在疾病和发育过程中的作用进行了阐述，系统总结了针对 EST、bulk RNA-seq、scRNA-seq、full-length transcriptome 以及 spatial transcriptome 等组学数据的APA识别方法和已有数据资源。随着测序技术的不断进步和数据资源的日益丰富，APA的研究将更加深入和全面。首先，随着技术的进步，能够更加深入地研究不同组织、细胞类型和疾病状态下APA事件的发生和精细调控机制。有助于更好地理解APA在生命过程中的调控作用，也为相关疾病的诊疗提供新的思路和方法。其次，随着APA数据资源的大量积累，更多的APA事件被发现，APA的研究也将更加全面和系统。此外，随着人工智能技术的发展，结合机器学习算法对APA进行自动挖掘和分析也是必然趋势。APA的调控机制极为复杂，仍有许多参与调控因素尚待深入探究；APA事件与发育和疾病之间的关联同样错综复杂。依赖多种精确的检测技术、综合考量多种调控因素，有望在APA研究领域取得突破性进展。

参　考　文　献

[1] Xu C, Zhang J. Alternative polyadenylation of mammalian transcripts is generally deleterious, not adaptive. Cell Syst, 2018, **6**(6): 734-742.e4

[2] Han T, Kim J K. Driving glioblastoma growth by alternative polyadenylation. Cell Res, 2014, **24**(9): 1023-1024

[3] Yuan F, Hankey W, Wagner E J, *et al*. Alternative polyadenylation of mRNA and its role in cancer. Genes Dis, 2019, **8**(1): 61-72

[4] Chen W, Jia Q, Song Y, *et al*. Alternative polyadenylation: methods, findings, and impacts. Genomics Proteomics Bioinformatics, 2017, **15**(5): 287-300

[5] Tang P, Yang Y, Li G, *et al*. Alternative polyadenylation by sequential activation of distal and proximal PolyA sites. Nat Struct Mol Biol, 2022, **29**(1): 21-31

[6] Zhang Y, Liu L, Qiu Q, *et al*. Alternative polyadenylation: methods, mechanism, function, and role in cancer. J Exp Clin Cancer Res, 2021, **40**(1): 51

[7] Elkon R, Ugalde A P, Agami R. Alternative cleavage and polyadenylation: extent, regulation and function. Nat Rev Genet, 2013, **14**(7): 496-506

[8] Lianoglou S, Garg V, Yang J L, *et al*. Ubiquitously transcribed genes use alternative polyadenylation to achieve tissue-specific expression. Genes Dev, 2013, **27**(21): 2380-2396

[9] Tian B, Manley J L. Alternative polyadenylation of mRNA precursors. Nat Rev Mol Cell Biol, 2017, **18**(1): 18-30

[10] Xiang K, Ly J, Bartel D P. Control of poly(A) -tail length and translation in vertebrate oocytes and early embryos. Dev Cell, 2024, **59**(8): 1058-1074.e11

[11] Gallicchio L, Olivares G H, Berry C W, *et al*. Regulation and function of alternative polyadenylation in development and differentiation. RNA Biol, 2023, **20**(1): 908-925

[12] Sun Y, Fu Y, Li Y, *et al*. Genome-wide alternative polyadenylation in animals: insights from high-throughput technologies. J Mol Cell Biol, 2012, **4**(6): 352-361

[13] Guo S, Lin S. mRNA alternative polyadenylation (APA) in regulation of gene expression and diseases. Genes Dis, 2021, **10**(1): 165-174

[14] Mitschka S, Mayr C. Context-specific regulation and function of mRNA alternative polyadenylation. Nat Rev Mol Cell Biol, 2022, **23**(12): 779-796

[15] Zhou X, Zhang Y, Michal J J, *et al*. Alternative polyadenylation coordinates embryonic development, sexual dimorphism and longitudinal growth in *Xenopus tropicalis*. Cell Mol Life Sci, 2019, **76**(11): 2185-2198

[16] Li N, Cai Y, Zou M, *et al*. CFIm-mediated alternative polyadenylation safeguards the development of mammalian pre-implantation embryos. Stem Cell Reports, 2023, **18**(1): 81-96

[17] Agarwal V, Lopez-Darwin S, Kelley D R, *et al*. The landscape of alternative polyadenylation in single cells of the developing mouse embryo. Nat Commun, 2021, **12**(1): 5101

[18] Velten L, Anders S, Pekowska A, *et al*. Single-cell polyadenylation site mapping reveals 3' isoform choice variability. Mol Syst Biol, 2015, **11**(6): 812

[19] Fernandes S F, Fior R, Pinto F, *et al*. Fine-tuning of fgf8a expression through alternative polyadenylation has a selective impact on Fgf-associated developmental processes. Biochim Biophys Acta Gene Regul Mech, 2018, **1861**(9): 783-793

[20] Zhang Z, Bae B, Cuddleston W H, *et al*. Coordination of alternative splicing and alternative polyadenylation revealed by targeted long read sequencing. Nat Commun, 2023, **14**(1): 5506

[21] Yang Y, Paul A, Bach T N, *et al*. Single-cell alternative polyadenylation analysis delineates GABAergic neuron types. BMC Biol, 2021, **19**(1): 144

[22] Lee S, Aubee J I, Lai E C. Regulation of alternative splicing and polyadenylation in neurons. Life Sci Alliance, 2023, **6**(12): e202302000

[23] Zingone A, Sinha S, Ante M, *et al*. A comprehensive map of alternative polyadenylation in African American and European American lung cancer patients. Nat Commun, 2021, **12**(1): 5605

[24] Venkat S, Tisdale A A, Schwarz J R, *et al*. Alternative polyadenylation drives oncogenic gene expression in pancreatic ductal adenocarcinoma. Genome Res, 2020, **30**(3): 347-360

[25] Wang L, Lang G T, Xue M Z, *et al*. Dissecting the heterogeneity of the alternative polyadenylation profiles in triple-negative breast cancers. Theranostics, 2020, **10**(23): 10531-10547

[26] Cao J, Kuyumcu-Martinez M N. Alternative polyadenylation regulation in cardiac development and cardiovascular disease. Cardiovasc Res, 2023, **119**(6): 1324-1335

[27] Singh I, Lee S H, Sperling A S, *et al*. Widespread intronic polyadenylation diversifies immune cell transcriptomes. Nat Commun, 2018, **9**(1): 1716

[28] Devany E, Park J Y, Murphy M R, *et al*. Intronic cleavage and polyadenylation regulates gene expression during DNA damage response through U1 snRNA. Cell Discov, 2016, **2**: 16013

[29] Zhao Z, Xu Q, Wei R, *et al*. Cancer-associated dynamics and potential regulators of intronic polyadenylation revealed by IPAFinder using standard RNA-seq data. Genome Res, 2021, **31**(11): 2095-2106

[30] Beaudoing E, Gautheret D. Identification of alternate polyadenylation sites and analysis of their tissue distribution using EST data. Genome Res, 2001, **11**(9): 1520-1526

[31] Muro E M, Herrington R, Janmohamed S, *et al*. Identification of gene 3' ends by automated EST cluster analysis. Proc Natl Acad Sci USA, 2008, **105**(51): 20286-20290

[32] Gilat R, Goncharov S, Esterman N, *et al*. Under-representation of PolyA/PolyT tailed ESTs in human ESTdb: an obstacle to alternative polyadenylation inference. Bioinformation, 2006, **1**(6): 220-224

[33] Shepard P J, Choi E A, Lu J, *et al*. Complex and dynamic landscape of RNA polyadenylation revealed by PAS-Seq. RNA, 2011, **17**(4): 761-772

[34] Chen M, Ji G, Fu H, *et al*. A survey on identification and quantification of alternative polyadenylation sites from RNA-seq data. Brief Bioinform, 2020, **21**(4): 1261-1276

[35] Yang Y, Wu X, Yang W, *et al*. Dynamic alternative polyadenylation during iPSC differentiation into cardiomyocytes. Comput Struct Biotechnol J, 2022, **20**: 5859-5869

[36] Martin G, Gruber A R, Keller W, *et al*. Genome-wide analysis of pre-mRNA 3' end processing reveals a decisive role of human cleavage factor I in the regulation of 3' UTR length. Cell Rep, 2012, **1**(6): 753-763

[37] Jan C H, Friedman R C, Ruby J G, *et al*. Formation, regulation and evolution of *Caenorhabditis elegans* 3'UTRs. Nature, 2011, **469**(7328): 97-101

[38] Hoque M, Ji Z, Zheng D, *et al*. Analysis of alternative cleavage and polyadenylation by 3' region extraction and deep sequencing. Nat Methods, 2013, **10**(2): 133-139

[39] Derti A, Garrett-Engele P, MacIsaac K D, *et al*. A quantitative atlas

of polyadenylation in five mammals. Genome Res, 2012, **22**(6): 1173-1183

[40] Yao C, Shi Y. Global and quantitative profiling of polyadenylated RNAs using PAS-seq. Methods Mol Biol, 2014, **1125**: 179-185

[41] 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

[42] Ha K C H, Blencowe B J, Morris Q. QAPA: a new method for the systematic analysis of alternative polyadenylation from RNA-seq data. Genome Biol, 2018, **19**(1): 45

[43] Ye C, Long Y, Ji G, *et al*. APAtrap: identification and quantification of alternative polyadenylation sites from RNA-seq data. Bioinformatics, 2018, **34**(11): 1841-1849

[44] Arefeen A, Liu J, Xiao X, *et al*. TAPAS: tool for alternative polyadenylation site analysis. Bioinformatics, 2018, **34**(15): 2521-2529

[45] Cass A A, Xiao X. mountainClimber identifies alternative transcription start and polyadenylation sites in RNA-seq. Cell Syst, 2019, **9**(4): 393-400.e6

[46] Wang R, Tian B. APAlyzer: a bioinformatics package for analysis of alternative polyadenylation isoforms. Bioinformatics, 2020, **36**(12): 3907-3909

[47] Lusk R, Stene E, Banaei-Kashani F, *et al*. Aptardi predicts polyadenylation sites in sample-specific transcriptomes using high-throughput RNA sequencing and DNA sequence. Nat Commun, 2021, **12**(1): 1652

[48] Jonnakuti V S, Wagner E J, Maletić-Savatić M, *et al*. PolyAMiner-Bulk: a machine learning based bioinformatics algorithm to infer and decode alternative polyadenylation dynamics from bulk RNA-seq data. Cell Rep Methods, 2024, **4**(2): 100707

[49] Liu X, Chen H, Li Z, *et al*. InPACT: a computational method for accurate characterization of intronic polyadenylation from RNA sequencing data. Nat Commun, 2024, **15**(1): 2583

[50] Li L, Dong J, Yan L, *et al*. Single-cell RNA-seq analysis maps development of human germline cells and gonadal niche interactions. Cell Stem Cell, 2017, **20**(6): 891-892

[51] Ji G, Xuan W, Zhuang Y, *et al*. Learning association for single-cell transcriptomics by integrating profiling of gene expression and alternative polyadenylation. bioRxiv, 2021. DOI: 10.1101/2021.01.04.425335

[52] Ye C, Zhou Q, Wu X, *et al*. scDAPA: detection and visualization of dynamic alternative polyadenylation from single cell RNA-seq data. Bioinformatics, 2020, **36**(4): 1262-1264

[53] Wu X, Liu T, Ye C, *et al*. scAPAtrap: identification and quantification of alternative polyadenylation sites from single-cell RNA-seq data. Brief Bioinform, 2021, **22**(4): bbaa273

[54] Li W V, Zheng D, Wang R, *et al*. MAAPER: model-based analysis of alternative polyadenylation using 3' end-linked reads. Genome Biol, 2021, **22**(1): 222

[55] Gao Y, Li L, Amos C I, *et al*. Analysis of alternative polyadenylation from single-cell RNA-seq using scDaPars reveals cell subpopulations invisible to gene expression. Genome Res, 2021, **31**(10): 1856-1866

[56] Gruber A J, Gypas F, Riba A, *et al*. Terminal exon characterization with TECtool reveals an abundance of cell-specific isoforms. Nat Methods, 2018, **15**(10): 832-836

[57] Fansler M M, Mitschka S, Mayr C. Quantifying 3'UTR length from scRNA-seq data reveals changes independent of gene expression. Nat Commun, 2024, **15**(1): 4050

[58] Zhu S, Lian Q, Ye W, *et al*. scAPAdb: a comprehensive database of alternative polyadenylation at single-cell resolution. Nucleic Acids Res, 2022, **50**(D1): D365-D370

[59] Zhou R, Xiao X, He P, *et al*. SCAPE: a mixture model revealing single-cell polyadenylation diversity and cellular dynamics during cell differentiation and reprogramming. Nucleic Acids Res, 2022, **50**(11): e66

[60] Moon Y, Burri D, Zavolan M. Identification of experimentally-supported poly(A) sites in single-cell RNA-seq data with SCINPAS. NAR Genom Bioinform, 2023, **5**(3): lqad079

[61] Shulman E D, Elkon R. Cell-type-specific analysis of alternative polyadenylation using single-cell transcriptomics data. Nucleic Acids Res, 2019, **47**(19): 10027-10039

[62] Wang J, Chen W, Yue W, *et al*. Comprehensive mapping of alternative polyadenylation site usage and its dynamics at single-cell resolution. Proc Natl Acad Sci USA, 2022, **119**(49): e2113504119

[63] Legnini I, Alles J, Karaiskos N, *et al*. FLAM-seq: full-length mRNA sequencing reveals principles of poly(A) tail length control. Nat Methods, 2019, **16**(9): 879-886

[64] Chang T, An B, Liang M, *et al*. PacBio single-molecule long-read sequencing provides new light on the complexity of full-length transcripts in cattle. Front Genet, 2021, **12**: 664974

[65] Gao Y, Wang H, Zhang H, *et al*. PRAPI: post-transcriptional regulation analysis pipeline for Iso-Seq. Bioinformatics, 2018, **34**(9): 1580-1582

[66] de la Fuente L, Arzalluz-Luque Á, Tardáguila M, *et al*. tappAS: a comprehensive computational framework for the analysis of the functional impact of differential splicing. Genome Biol, 2020, **21**(1): 119

[67] Abdel-Ghany S E, Hamilton M, Jacobi J L, *et al*. A survey of the sorghum transcriptome using single-molecule long reads. Nat Commun, 2016, **7**: 11706

[68] Wang F, Jin Z, Wang S, *et al*. ASAPA: a bioinformatic pipeline based on Iso-Seq that identifies the links among alternative splicing, alternative transcription initiation and alternative polyadenylation. Funct Integr Genomics, 2024, **24**(2): 67

[69] Ji G, Tang Q, Zhu S, *et al*. stAPAminer: mining spatial patterns of alternative polyadenylation for spatially resolved transcriptomic studies. Genomics Proteomics Bioinformatics, 2023, **21**(3): 601-618

[70] Zhang H, Hu J, Recce M, *et al*. PolyA_DB: a database for mammalian mRNA polyadenylation. Nucleic Acids Res, 2005, **33**(Database issue): D116-D120

[71] Koscielny G, Le Texier V, Gopalakrishnan C, *et al*. ASTD: The

Alternative Splicing and Transcript Diversity database. Genomics, 2009, **93**(3): 213-220

[72] You L, Wu J, Feng Y, *et al*. APASdb: a database describing alternative poly(A) sites and selection of heterogeneous cleavage sites downstream of poly(A) signals. Nucleic Acids Res, 2015, **43**(Database issue): D59-D67

[73] Wang R, Nambiar R, Zheng D, *et al*. PolyA_DB 3 catalogs cleavage and polyadenylation sites identified by deep sequencing in multiple genomes. Nucleic Acids Res, 2018, **46**(D1): D315-D319

[74] Yang Y, Zhang Q, Miao Y R, *et al*. SNP2APA: a database for evaluating effects of genetic variants on alternative polyadenylation in human cancers. Nucleic Acids Res, 2020, **48**(D1): D226-D232

[75] Herrmann C J, Schmidt R, Kanitz A, *et al*. PolyASite 2.0: a consolidated atlas of polyadenylation sites from 3' end sequencing. Nucleic Acids Res, 2020, **48**(D1): D174-D179

[76] Hong W, Ruan H, Zhang Z, *et al*. APAatlas: decoding alternative polyadenylation across human tissues. Nucleic Acids Res, 2020, **48**(D1): D34-D39

[77] Jin W, Zhu Q, Yang Y, *et al*. Animal-APAdb: a comprehensive animal alternative polyadenylation database. Nucleic Acids Res, 2021, **49**(D1): D47-D54

[78] Yang X, Tong Y, Liu G, *et al*. scAPAatlas: an atlas of alternative polyadenylation across cell types in human and mouse. Nucleic Acids Res, 2022, **50**(D1): D356-D364

[79] Cui Y, Peng F, Wang D, *et al*. 3'aQTL-atlas: an atlas of 3'UTR alternative polyadenylation quantitative trait loci across human normal tissues. Nucleic Acids Res, 2022, **50**(D1): D39-D45

[80] Ma X, Cheng S, Ding R, *et al*. ipaQTL-atlas: an atlas of intronic polyadenylation quantitative trait loci across human tissues. Nucleic Acids Res, 2023, **51**(D1): D1046-D1052

[81] Huang K, Wu S, Yang X, *et al*. CAFuncAPA: a knowledgebase for systematic functional annotations of APA events in human cancers. NAR Cancer, 2023, **5**(1): zcad004

[82] Müller S, Rycak L, Afonso-Grunz F, *et al*. APADB: a database for alternative polyadenylation and microRNA regulation events. Database, 2014, **2014**: bau076

[83] Barquin M, Kouzel I U, Ehrmann B, *et al*. scTEA-db: a comprehensive database of novel terminal exon isoforms identified from human single cell transcriptomes. Nucleic Acids Res, 2024, **52**(D1): D1018-D1023

Alternative Polyadenylation in Mammalian*

ZHANG Yu[1,2], CHI Hong-Xia[1,2], YANG Wu-Ri-Tu[3], ZUO Yong-Chun[4,5]**, XING Yong-Qiang[1,2]**

([1])*College of Life Sciences and Technology, Inner Mongolia University of Science and Technology, Baotou* 014010, China;

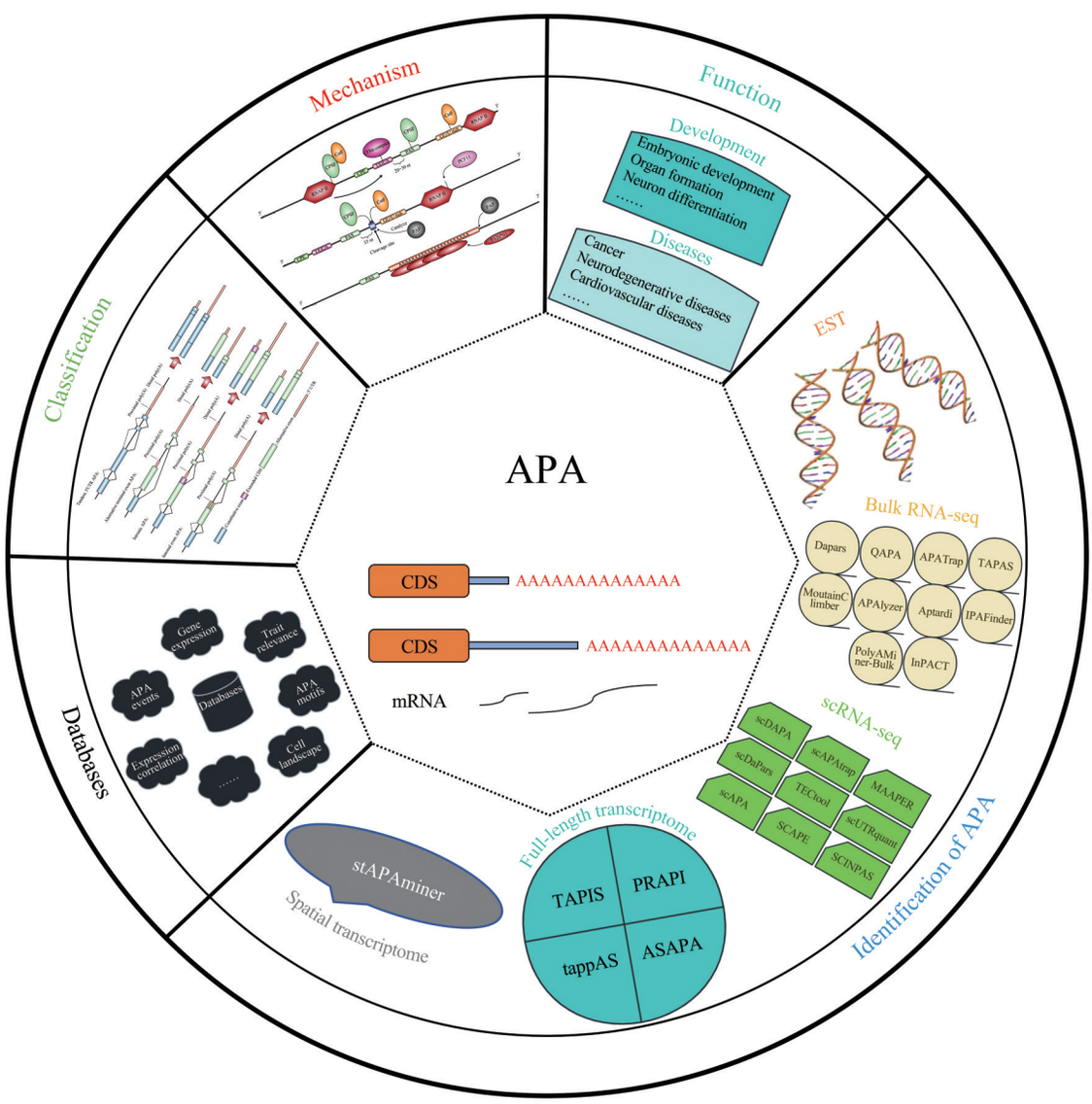
[2])*The Inner Mongolia Key Laboratory of Life Health and Bioinformatics, Inner Mongolia University of Science and Technology, Baotou* 014010, China;

[3])*Computer Department, Hohhot Vocational College, Hohhot* 010070, China;

[4])*College of Life Sciences, Inner Mongolia University, Hohhot* 010020, China;

[5])*State Key Laboratory of Reproductive Regulation and Breeding of Grassland Livestock, Inner Mongolia University, Hohhot* 010020, China)

Graphical abstract



* This work was supported by grants from The National Natural Science Foundation of China (62371265, 62161039), the Natural Science Foundation of Inner Mongolia (2024JQ10), and Basic Scientific Research Funding for Universities Directly Under Inner Mongolia Autonomous Region (2023RCTD023).

** Corresponding author.

XING Yong-Qiang. Tel：86-472-5955917，E-mail：xingyongqiang1984@163.com

ZUO Yong-Chun. Tel：86-471-5227683，E-mail：yczuo@imu.edu.cn

Received：July 6，2024　Accepted：August 25，2024

Abstract With the rapid development of sequencing technologies, the detection of alternative polyadenylation (APA) in mammals has become more precise. APA precisely regulates gene expression by altering the length and position of the poly(A) tail, and is involved in various biological processes such as disease occurrence and embryonic development. The research on APA in mammals mainly focuses on the following aspects: (1) identifying APA based on transcriptome data and elucidating their characteristics; (2) investigating the relationship between APA and gene expression regulation to reveal its important role in life regulation; (3) exploring the intrinsic connections between APA and disease occurrence, embryonic development, differentiation, and other life processes to provide new perspectives and methods for disease diagnosis and treatment, as well as uncovering embryonic development regulatory mechanisms. In this review, the classification, mechanisms and functions of APA were elaborated in detail and the methods for APA identifying and APA data resources based on various transcriptome data were systematically summarized. Moreover, we epitomized and provided an outlook on research on APA, emphasizing the role of sequencing technologies in driving studies on APA in mammals. In the future, with the further development of sequencing technology, the regulatory mechanisms of APA in mammals will become clearer.

Key words alternative polyadenylation, regulatory mechanism, transcriptome, recognition methods, data resources

DOI：10.16476/j.pibb.2024.0298 **CSTR**：32369.14.pibb.20240298