DOI: 10.13376/j.cbls/2021031
文章编号：1004-0374(2021)03-0267-14

杨建华，中山大学教授、博士生导师，基因功能与调控教育部重点实验室副主任。2019年获广东省科技创新青年拔尖人才。长期致力于开发高通量组学方法，研究RNA及其互作蛋白的结构、功能和作用机制。以通讯作者或共同通讯作者身份在*Nature*、*Nat Cell Biol*、*Eur Urol*、*Nucleic Acids Res*、*Cell Rep*等杂志发表20多篇研究论文，多篇论文被*Nat Rev Genet*、*Nat Cell Biol*等杂志亮点评述。开发了包括starBase等软件和平台，被引用超过6 000次。受邀在Springer出版社出版了多篇关于非编码RNA和RNA修饰研究方法的论著章节，是期刊*Non-Coding RNA*、*Front Cell Dev Biol*和*Front Genet*的编委。

基于高通量测序的RNA信息解析技术

黄钧鸿#，黄巧娟#，李　斌，杨建华*
(中山大学生命科学学院，广州 510275)

摘　要：非编码RNA (noncoding RNA, ncRNA)占据真核生物转录组的绝大部分，在各种生理和病理过程中发挥重要作用。随着高通量测序技术的发展，人们利用RNA信息学技术解析到越来越多的非编码RNA的信息，并逐渐揭示其功能和作用机制。该文主要介绍非编码RNA及其靶标鉴定、RNA功能网络、RNA与蛋白质互作、RNA修饰及RNA二级结构的信息解析技术。

关键词：测序技术；非编码RNA；靶标鉴定；RNA-蛋白质互作；RNA修饰；二级结构
中图分类号：Q752　　**文献标志码：**A

Bioinformatic methods for analyzing noncoding RNAs from high-throughput sequencing data

HUANG Jun-Hong#, HUANG Qiao-Juan#, LI Bin, YANG Jian-Hua*
(School of Life Sciences, Sun Yat-sen University, Guangzhou 510275, China)

Abstract: Non-coding RNAs (ncRNAs) occupy most of the eukaryotic transcriptome, and play an important role in physiological and pathological processes. With the development of high-throughput sequencing technology, people have developed various bioinformatic tools and databases to analyze ncRNAs, and gradually revealed their regulatory functions and mechanisms. This review mainly introduces how to identify ncRNAs and their biological targets, regulatory networks of RNAs, interactions between RNAs and proteins, RNA modification as well as RNA secondary structures.

Key words: sequencing technology; noncoding RNA; target identification; RNA-protein interaction; RNA modification; secondary structure

收稿日期：2021-01-31
基金项目：国家重点研发计划(2019YFA0802202)
#共同第一作者
*通信作者：E-mail: yangjh7@mail.sysu.edu.cn

1 非编码RNA的鉴定

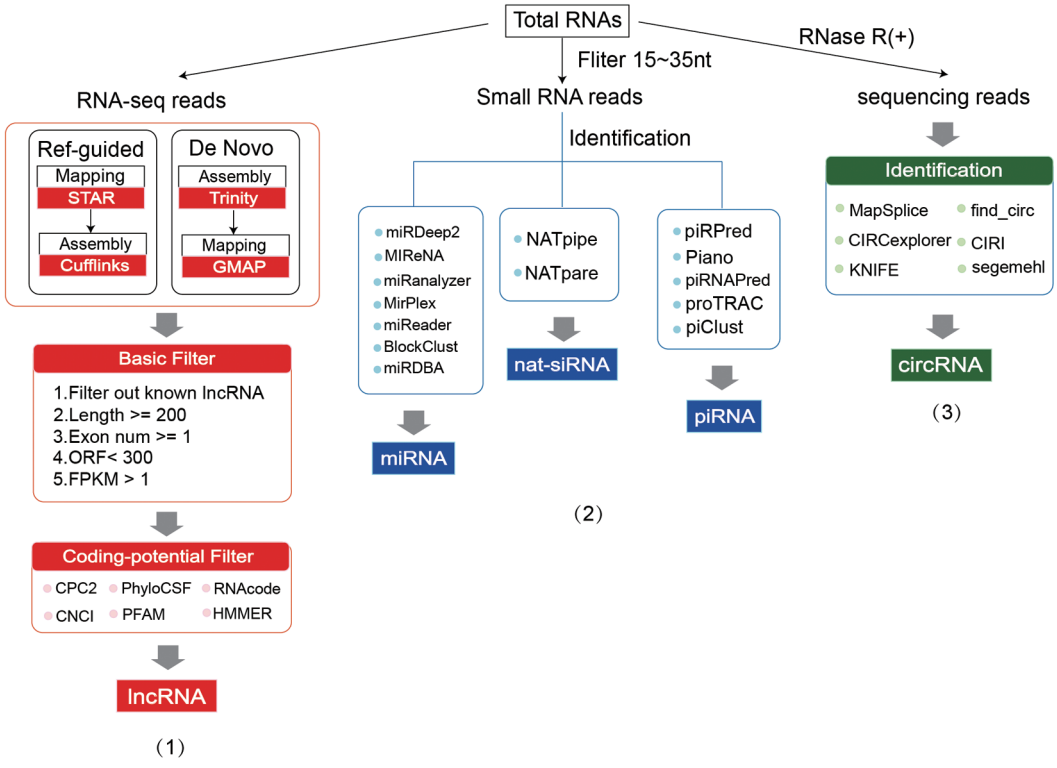
目前，基于高通量测序方法去鉴定非编码RNA的主要策略是根据不同的非编码RNA采取不同的富集方法，选择性地排除其他类型RNA的干扰。获得相应的非编码RNA测序数据后，利用计算方法对非编码RNA进行注释和预测(图1)。常规的RNA测序分三大类：转录组测序、小RNA (small RNA, sRNA) 测序、环状RNA (circular RNA, circRNA) 测序。其中转录组测序主要针对线性的mRNA以及长非编码RNA (long non-coding RNA, lncRNA)，小RNA测序主要针对15~35 nt的小分子RNA，而环状RNA测序则往往需要先降解掉线性的RNA分子，再进行建库测序[1]。

1.1 长非编码RNA的鉴定

lncRNA是指一类长度大于200 nt、没有蛋白质编码潜能的大分子RNA。lncRNA的鉴定主要基于测序技术，它包括早期的cDNA文库测序[2]、EST片段测序[3]以及目前主流的RNA-seq[4]；而计算机预测依据的特征主要包括序列特征、保守性和表观遗传修饰图谱。

由于一大部分lncRNA不含polyA结构，因此基于RNA-seq的测序策略通常采用去除核糖体RNA (rRNA)的方法，把编码mRNA与长非编码RNA一并测序并进行相关分析[5]。若有参考基因组，可使用STAR[6]等软件先将测序数据比对到参考基因组，然后使用Cufflinks[7]等软件组装得到转录本序列；若无参考基因组，则需要先使用Trinity[8]等软件从头组装获得转录本的序列。随后可以依据转录本长度、外显子数量、开放阅读框长度以及表达量等特征作进一步筛选。

lncRNA鉴定过程中的一个重要问题是区分编码与非编码转录本序列，目前已经有许多基于机



(1) 长非编码RNA (lncRNA)的鉴定。从总RNA中去除核糖体RNA后建立文库测序，获得测序数据后，对于有参考基因组的物种，先用STAR进行比对，然后使用Cufflinks等软件组装得到转录本序列；对于无参考基因组的物种，则先用Trinity进行从头组装，获得转录本序列。随后可以依据转录本长度、外显子数量、开放阅读框长度以及FPKM等特征作进一步筛选，最后通过相应的软件预测转录本的编码能力，获取候选的lncRNA。(2) 小RNA (sRNA)的鉴定。从总RNA中通过割胶分离获取长度约为15~35 nt的RNA片段，构建文库后进行测序，随后分别通过相应的软件对miRNA、nat-siRNA、piRNA进行鉴定。(3) 环状RNA (circRNA)的鉴定。通过RNase R去除线性RNA等方法对circRNA进行富集并测序，随后通过软件获取候选的circRNA。

图1 非编码RNA的鉴定流程

器学习的工具被开发，这些工具都使用序列的内部特征和结构特点预测 lncRNA，一般从已知的蛋白序列以及确定的非编码 RNA 中提取一些特征，然后使用支持向量机或逻辑回归的方法训练出分类算法。其中，基于开放阅读框及核苷酸 / 氨基酸频率特征的软件有 CNCI[9]、CONC[10]、CPAT[11]、CPC2[12]、iSeeRNA[13]、PLEK[14]；基于核苷酸替换模式的软件有 PhyloCSF[15]、RNAcode[16]；基于蛋白质结构域特征的软件有 HMMER[17]。

1.2 小非编码RNA的鉴定

近年来，许多小的非编码 RNA (small nocoding RNA, sncRNA) 被认为是植物和动物的重要基因和基因组的调控因子。目前研究较多的主要包括 microRNA (miRNA)、small interfering RNA (siRNA)、piwi-interacting RNA (piRNA) 三类。小 RNA 测序采用割胶分离技术，收集样品中 15~35 nt 的 RNA 片段进行高通量测序，进而鉴定已知的或新的小非编码 RNA。一些软件或数据库可完成对 sncRNA 的注释：SeqCluster[18] 和 DARIO[19] 可用于对整个 sRNA-seq 数据进行无偏好的注释和分类；ncPRO-seq[20] 提供所有类型的 sncRNA 的详细信息，并识别匹配 sncRNA 中显著富集的未注释区域；deepBase[21] 整合了现有的转录本测序数据，具备高通量和深度注释的小 RNA 数据。

1.2.1 miRNA的鉴定

miRNA 是一种约为 22 nt 的小非编码 RNA，能够抑制蛋白质的翻译和影响 RNA 的稳定性。近年来，基于小 RNA 测序大规模预测 miRNA 的方法逐渐增多，为发现 miRNA 开辟了新途径。这些预测方法大致可以分为三类：基于 miRNA 前体、基于 miRNA 双链体、基于 read 聚类和注释。

基于 miRNA 前体的方法：首先将 read 比对到基因组上，将比对区域附近的基因组序列进行扫描和分析，生成一组假定的 miRNA 前体，随后分析 miRNA 前体最可能形成的二级结构。此类工具包括 miRDeep2[22]、miRSeqNovel[23]、miRdentify[24]、MIReNA[25] 和 miRDeep*[26]。此外，基于机器学习也可以预测 miRNA 前体，如 CoRAL[27] 和 miRanalyzer[28]，然而，对于大多数物种来说，暂没有足够的注释来生成足够的训练集。

基于 miRNA 双链体的方法：首先选择 10~30 nt 长的片段，并生成所有可能的配对，从而产生假定的 miRNA 双链，随后根据长度、未配对碱基的数量等特征，选择最可能真实的 miRNA 双链体。此类工具包括 MirPlex、miReader 等，其可以应用于缺乏参考基因组序列的物种，但只能鉴定到成熟的 miRNA；此外，需要同时存在 miRNA 和 miRNA*，read 才能被检出。

基于 read 聚类和注释的方法：将比对到基因组同一链上的临近 read 通过相似性进行聚类，从而获得较连续的区域，随后分析这些连续区域与哪些已知的 ncRNA 重合并进行注释，此类工具包括 BlockClust[29]、deepBlockAlign[30]、DARIO[19] 和 miRDBA[31]。

1.2.2 nat-siRNA的鉴定

天然反义转录物 (natural antisense transcript, NAT) 是由植物或动物的内源基因编码的成对互补转录物，其依靠退火区域的高互补性维持热力学稳定 [32]。天然反义转录物起源的小干扰 RNA (natural antisense transcript originated small interfering RNA, nat-siRNA) 鉴定的关键是先鉴定出天然反义转录物。

天然反义转录物有顺式和反式之分，如果两个转录物位于相同基因组位点的相反链，并且重叠区域长于 23 nt 以支持至少一个 siRNA 序列的产生，则它们有可能形成一对顺式天然反义转录物。而反式天然反义转录物则起源于两个遥远的基因位点，部分转录本可以形成完美的互补配对，可以通过搜索成对的转录单位来识别它们。它们一般具有长于 100 nt 的连续互补配对区域，以支持可能的 RNA-RNA 退火 [33]。可以通过软件 DINAMelt[34] 分析配对为 RNA-RNA 双链体的可能性。最后，将测序数据比对到已鉴定的天然反义转录物上 [35]。

目前，有一些计算分析流程可进行 nat-siRNA 的鉴定。NATpipe 通过使用 sRNA 测序数据可系统地发现 NAT 和 nat-siRNA[36]。NATpipe 针对阶段分布的 nat-siRNA，但 nat-siRNA 的产生也可能遵循位点特异性模式，因此 NATpipe 会遗漏此类 nat-siRNA；NATpare 对此进行了优化，其将 sRNA 数据、转录组数据和可选的降解组数据作为输入，能够识别顺式和反式 nat-siRNA 并预测其靶标 [37]。

1.2.3 piRNA的鉴定

piRNA 缺乏保守的结构基序，在不同物种间序列相似性相对较低，这使得对 piRNA 的精确计算预测非常具有挑战性。2011 年，Zhang 等 [38] 提出了一种基于小 RNA 测序且不依赖基因组数据来鉴定非模式生物 piRNA 的新方法 piRNA predictor，该方法基于 *k*-mer 串频率的 Fisher 判别式来预测 piRNA，精确度达 90% 以上。2014 年，Brayet 等 [39] 提出了一

种名为 piRPred 的新方法，该方法基于多核和支持向量机分类器，可随着训练集大小的增加而增加构建分类器的内存和时间。同年，Wang 等 [40] 提出了 piRNA 注释方法 Piano，其能够在特异性和敏感性之间取得很好的平衡，然而该方法仅局限于转座子的 piRNA 序列。2019 年，Monga 等 [41] 开发了一个综合的预测 piRNA 的框架 piRNAPred，其利用 *k*-mer 核苷酸组成、二级结构、热力学和物理化学性质等杂交特征，与其他最先进的 piRNA 预测方法相比，其在准确预测 piRNA 方面取得了最高的性能。

由于大多数 piRNA 来源于基因组 piRNA 簇 [42]，因此可以利用聚类位点信息进行 piRNA 鉴定。在基于 piRNA 聚类位点的方法中，proTRAC[43] 可以通过对映射序列 reads 的概率分析，从小 RNA-seq 数据中识别出 piRNA 簇和 piRNAs。此外，piClust[44] 使用了一种基于密度的聚类方法来识别 piRNA 簇，而无需假设任何参数分布模型。

1.3 环状RNA的鉴定

circRNA 是一类具有多样生物学功能的闭环 RNA，与 mRNA 相比，其缺乏 polyA 的尾部，且不易被 RNase R 消化。高通量测序和相关生物信息学工具的发展为深入研究 circRNA 提供了新的机会。目前能用于鉴定 circRNA 的测序数据，其建库策略主要包括：(1) rRNA(−)；(2) rRNA(−)、polyA(−)；(3) rRNA(−)、RNase R(+)；(4) RPAD。其中，方法 (3) 和 (4) 的 circRNA 富集程度较高，RPAD 实验方法 [45] 使用 RNase R 消化线性 RNA，随后除去含有 polyA 的 RNA，富集到的 circRNA 进行高通量测序，大多数 circRNA 鉴别工具倾向于使用 circRNA 富集后的 RNA-seq 数据集作为输入，能有效排除假阳性，同时提高检出率。

目前从高通量测序数据中大规模鉴定 circRNA 的工具主要分为两类：基于反向剪接位点和基于机器学习。基于反向剪接位点的算法大多基于 read 的拆分，或基于预先定义的反向剪接位点和 circRNA 的侧翼序列。MapSplice[46]、CIRCexplorer[47]、和 KNIFE[48] 需要依赖注释信息，find_circ[49]、segemehl[50] 和 CIRI[51] 能从头预测 circRNA，而不需要基因注释或外显子 - 内含子结构，这对于预测具有近端剪接位点的 circRNA 是有利的。Ularcirc[52] 和 UROBORUS[53] 可以检测总 RNA-seq 数据集中低表达水平的 circRNA，而无需 RNase R 处理。综合各个指标来看，CIRI、CIR-Cexplorer、KNIFE 这三款软件的性能更佳 [54]。但是单个软件往往因为算法的差异存在着一定的局限性，建议同时使用 2 个及以上的软件进行 circRNA 的预测 [55]。

近年来，机器学习方法越来越多地应用于生物信息学研究。已有研究分析 circRNA 形成过程中的影响因素，通过训练传统的机器学习算法 (支持向量机、随机森林和多核学习等) 来鉴别 circRNA，取得了较高的识别正确率。基于传统机器学习方法的工具主要包括 PredcircRNA[56]、H-ELM[57]、CirRNAPL[58] 等，但是这些方法需要先进行特征分析，而且这些选取的特征不能全面充分地表征反向剪接过程。深度学习算法能够处理大规模数据并自动提取有效特征，可以弥补传统机器学习模型的不足。基于深度学习方法的工具主要包括 DeepCirCode[59]、circDeep[60]、CRC[61] 等。

1.4 非编码RNA的鉴定小结

大规模测序技术为 ncRNA 预测提供了良好的数据基础，基于这些数据，结合 ncRNA 的生物学特征和加工机制开发出的算法可以达到高效准确的预测效果。近年来采用该方法鉴定出大量的新的 ncRNA 数据，使得 ncRNA 的研究获得进一步的发展。

目前，小 RNA 测序主要用于 miRNA 的分析与挖掘，而一些工具可以用于同时发掘多类小非编码 RNA，如 FlaiMapper[62]、DARIO[19] 或 miRDBA[32] 可以识别出包括 tRNA (transfer RNA)、scRNA (small cytoplasmic RNA)、snoRNA (small nucleolar RNA) 和 snRNA (small nuclear RNA) 在内的几种 ncRNA。此外，CoRAL[27] 允许预测其他五类 ncRNA，包括 lincRNA (long intergenic non-coding RNA)、scRNA、C/D box snoRNA、snRNA 和转座子衍生的 snRNA 以及 miRNA。

相比于小非编码 RNA，由于 lncRNA 的多样性，目前暂没有鉴定 lncRNA 的标准流程，其鉴定工作仍然面临挑战，未来可以基于 RNA-seq，结合更加多样的鉴定方法，如结合 ORF 长度、外显子数量、表达水平等，发掘更多 lncRNA 区别于其他非编码 RNA 的特点，创建更优的模型；同时，根据不同的情况，可以针对不同类型的 lncRNA 分别进行鉴定。

越来越多的证据表明，circRNA 在疾病发生过程中发挥重要作用，可在人体体液和外泌体中检测到，这使得 circRNA 被持续关注，也意味着需要开发更专业化的 circRNA 工具来满足这些需求。一些 circRNA 对 RNase R 敏感 [63]，RNase R(+) 文库的制备方法将导致此类 circRNA 丰度相对较低，因此文库制备方法的类型可能会对下游结果产生很大影响。随着 circRNA 鉴定工具数量的增加，对于 circRNA

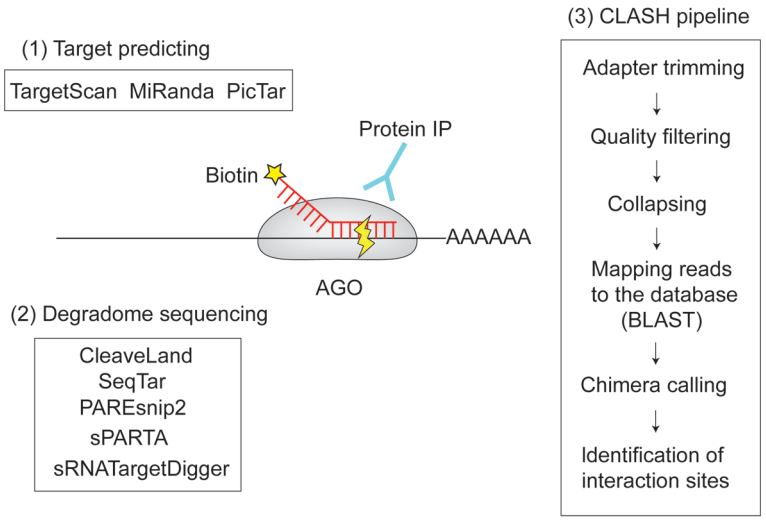
数据集的置信度仍然有很多要求，未来需要更好地统计模型来模拟 circRNA 数据集。此外，目前基于机器学习的工具依然数量有限，如何将 circRNA 的鉴定问题转化为分类问题，依然是未来重要的发展方向。

总的来说，目前鉴定到的主要是保守的和普遍表达的 ncRNA，而还有许多组织或物种特异低表达的 ncRNA 有待发掘，因此，开发出更精确的 ncRNA 高通量测序技术及 RNA 信息解析技术，更好地排除其他 RNA 的“噪音”，仍然是未来发展的重要方向。

2　非编码RNA靶标的鉴定

目前，非编码 RNA 靶标中研究得比较深入的主要是 miRNA 的靶标。确定 miRNA 靶基因的最常用方法是依赖计算机算法，如 TargetScan[64]、MiRanda[65] 和 PicTar[66](图 2)，它们预测 miRNA 种子区的结合主要遵循种子区匹配、保守性、可接近性、AU 含量和结合能量等特征。预测 miRNA 靶基因的算法虽然在不断升级，但计算机模拟仍有一定的局限性，如可能会给出假阳性结果，并需要进行额外的实验验证，若一些靶标不具有典型特征，更难以直接通过算法进行预测。

基于高通量测序来更直观地寻找 miRNA 靶基因，能够明显降低 miRNA 结合位点的假阳性预测率，其主要包括以下几种技术。(1) 针对 RNA 互作蛋白进行免疫共沉淀：AGO-CLIP[66] 通过 RNA 诱导沉默复合物 (RNA-induced silencing complex, RISC) 中 Argonaute (AGO) 蛋白的抗体进行 miRNA 及其靶标的免疫共沉淀；CRAC[67] 技术给 AGO 蛋白加上组氨酸标签，通过免疫共沉淀捕获 miRNA 后进行亲和纯化，具有更高的特异性；而 CLASH[68] 则增加了 mRNA 与 miRNA 的连接，通过生物信息学流程 hyb[69] 可以进行 CLASH 数据的解析，其将连接的序列分别比对到基因组和 miRNA 库，可以准确找到 miRNA 的靶位点。此外，TarPmiR[70] 基于机器学习，利用从 CLASH 数据中学习到的 13 个特征来预测 miRNA 靶点，其性能优于传统的基于 RNA-seq 的预测程序。(2) 针对 RNA 进行下拉富集：生物素下拉富集 [71] 利用生物素标记 miRNA，然后用磁珠将 miRNA 连同其靶 mRNA 下拉，这种方法能确定一个特定的 miRNA 靶基因，具有高度特异性。(3) 针对靶标的变化：在植物中，miRNA 通常与靶标序列紧密结合，mRNA 会被直接剪切为两段，其中之一是含有 3'-polyA 尾巴且 5' 不含 cap 的片段，降解组测序 [72] 针对这种裂解片段进行捕获，将测序片段比对到转录本上，能寻找特定 sRNA 引导的裂解信号；生物信息学工具如 CleaveLand[73]、SeqTar[74]、PAREsnip2[75]、sPARTA[76]、sRNA-TargetDigger[77] 充分利用降解数据进行靶标预测，其根据序列互补，筛选潜在的 sRNA 靶基因，通过降解信号的匹配分析，获得潜在靶基因的特异性裂解位点，最后根据靶标的裂解位点和结合位点之间的相关性来确定调



(1)非编码RNA靶标直接预测软件；(2)基于降解组测序的分析软件；(3) CLASH数据的分析流程。CLASH数据的分析流程主要包括：去除接头、过滤低质量碱基、去除PCR冗余、通过BLAST将测序的read比对到相应的数据库(如miRNA)、嵌合体的鉴定、互作位点的鉴定。

图2　非编码RNA靶标的鉴定软件及流程

节关系 (图 2)。

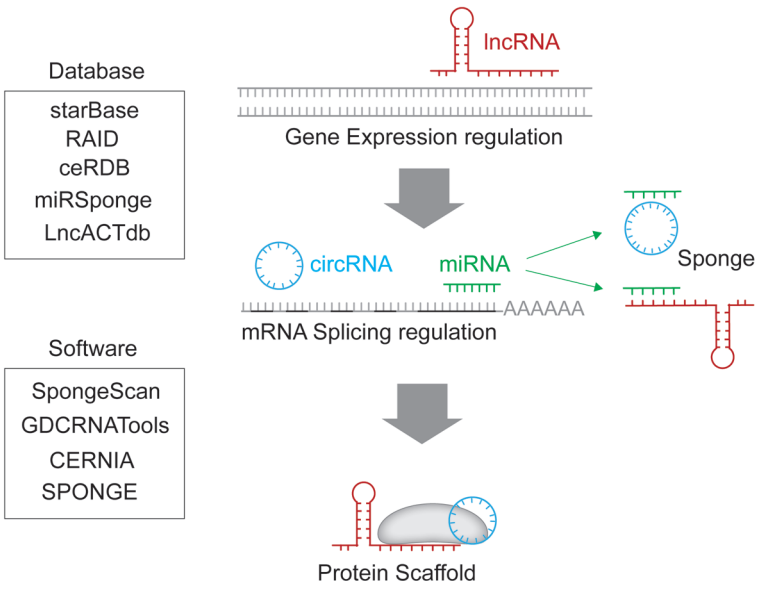
这些技术同样可以应用于其他非编码 RNA 靶标的预测，如 Yuan 等 [78] 使用抗 Miwi 蛋白的抗体进行 CLIP-seq，用于获得 piRNAs 和靶片段，最后鉴定出 3 781 个 mRNA 作为可能的 piRNA 靶标。Shen 等 [79] 则通过 CLASH 鉴定到秀丽隐杆线虫中的 piRNAs 和相关靶标 RNA 的结合位点。Chu 等 [80] 提出了 ChIRP-seq (chromatin isolation by RNA purification) 技术：通过设计生物素偶联的 RNA 探针，捕获 lnc-RNA 及与其结合的 DNA 与蛋白质，分离出 DNA 后进行高通量测序，从而获取 lncRNA 在染色质上的靶标位点。

目前，鉴定非编码 RNA 靶标的手段非常多样化，总的来说，由于互作往往涉及蛋白质、RNA、靶标这三者，因此策略主要包括对 RNA 互作蛋白的免疫共沉淀、RNA 本身的下拉、RNA 作用靶标后靶标的变化 (图 2)，每种策略都有其固有的优势和局限性，与基因表达谱法相比，免疫沉淀法排除了 RISC 外的假阳性靶基因，准确率大大提高。然而，依赖特异性抗体分离靶基因进一步降低了靶基因分离鉴定的效率，后续也需要生物信息学分析来揭示 miRNA-mRNA 的相互作用。目前最有希望的策略是下拉方法，其直接分离 RNA 相关的靶标。但有相关研究结果显示，3′- 生物素化极大地阻碍了 miRNA 与其靶点在 RISC 中的关联 [81]，为了解决这个问题，未来有待开发更优的化学修饰为 RNA 添加标记。

3 非编码RNA的功能网络

非编码 RNA 具有多方面的调控功能 (图 3)，目前研究较多的是竞争性内源 RNA (competing endogenous RNA, ceRNA)：多种类型的 RNA 能成为海绵分子，比如 circRNA、lncRNA、mRNA 和假基因 [82]，间接抑制 miRNA 对靶基因的影响。目前已有一些数据库记录并整合了基于高通量测序的 RNA 互作数据，专门针对 RNA 互作网络进行分析 (图 3)：starBase v2.0[23] 收集整理了超过 700 个 CLIP 数据集，提供实验支持的 RNA 互作网络，同时提供了 miRFunction 和 ceRNAFunction 两个程序，可通过超几何检验的方法计算两个 RNA 分子是否能形成 ceRNA 对；RAID[83] 整合了计算预测的和实验证明的 RNA 互作；ceRDB[84] 基于 TargetScan 预测 miRNA-mRNA 相互作用，提供假定的 ceRNA 互作；此外，miRSponge[85] 及 LncACTdb 2.0[86] 数据库提供实验支持的 ceRNA 数据。

进行 ceRNA 网络分析一般先获得差异表达的 mRNA、miRNA、ceRNA，基于上述数据进行共表达分析，即 mRNA-ceRNA 之间的表达水平应为显著的正相关，miRNA 与 mRNA/ceRNA 的表达水平应为显著的负相关 [87]，相关系数由皮尔森相关系数进行计算，最后筛选出被同一个 miRNA 靶向的 mRNA-



中心法则过程中可能涉及的非编码RNA调控作用包括：调控DNA的转录、调控mRNA前体的剪接、调控mRNA的翻译、作为miRNA海绵、作为蛋白质支架。图中左侧展示的是常用的ceRNA数据库和分析软件。

图3 非编码RNA的功能网络及分析工具

ceRNA 对即可构建 ceRNA 网络。2016 年，Furió-Tarí 等 [88] 开发了 SpongeScan，与传统的 miRNA 靶标预测方法不同，其专注于更能显示 ceRNA 作用的特征；2018 年，Li 等 [89] 开发了 R 包 GDCRNA-Tools，并定义了 ceRNA 的鉴定指标：(1) ceRNA-mRNA 共享大量 miRNAs；(2) ceRNA 与 mRNA 表达正相关；(3) ceRNA 和 mRNA 共有的 miRNA 应以相似的方式调节其表达 [90]。

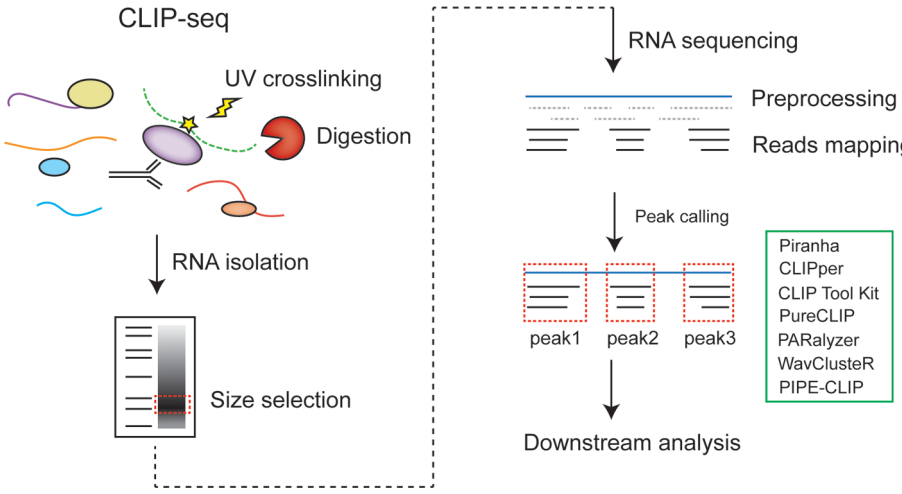
识别 ceRNA-ceRNA 互作或 miRNA 海绵互作的计算方法主要包括：成对相关方法、部分关联方法和数学建模方法。Sardina 等 [91] 提出了一种新的 ceRNA 预测算法 CERNIA，其基于 DT-Hybrid 算法，可用于研究不同组织类型和不同类别基因之间的 ceRNA 竞争。目前较为有效的一种 ceRNA 预测工具是 SPONGE (Sparse Partial correlation ON Gene Expression)[92]，其使用多敏感度相关性的量度来大规模推断 ceRNA 互作。

尽管 ceRNA 网络是一种研究非常广泛的 ncRNA 网络，但是 ncRNA 种类及功能的多样化将增加这些网络的复杂性，使得全面了解 ncRNA 在多级网络中的功能变得更加困难，而 RNA 与其他分子互作的检测技术的进步将有利于更好地阐明 RNA 调控网络。此外，合并大规模的非编码 RNA 数据需要制定普适的标准和命名规范。

4　RNA-蛋白质互作研究的生物信息学方法

RNA 结合蛋白在 RNA 调节的各种机制中发挥作用，影响 mRNA 前体的剪接、3' 末端加工，以及 RNA 修饰、翻译、稳定性和定位等，近年来发展了许多基于高通量测序的 RNA- 蛋白质互作研究方法。RIP-seq 通过使用目标蛋白的抗体，对蛋白质及靶 RNA 进行特异性捕获，捕获后经过纯化与建库，对 RNA 进行高通量测序 [93]。RIP 并不非常适合于研究与蛋白质直接接触的 RNA，因为它保留了多个蛋白质之间的相互作用。因此，需要具有更高特异性的方法，能够保留内源蛋白质与 RNA 的接触，同时确保仅纯化单个特异性 RBP。为此，研究人员开发了 CLIP-seq[94]，方法是利用 254 nm 紫外光共价交联蛋白质与 RNA 片段，这使 CLIP 能够在足够严格的条件下纯化与特定 RBP 结合的 RNA，以防止共纯化其他 RBP 或游离 RNA。随后，更多的 CLIP-seq 改进方法的出现大大提高了蛋白质交联位点的分辨率，如 PAR-CLIP[95] 引入了光激活核苷，并改用 365 nm 紫外线照射；iCLIP[96] 采用环化的建库策略；ir-CLIP[97] 引入了一种红外荧光染料来可视化免疫沉淀的质量；BrdU-CLIP[98] 使用 BrdUTP 并在逆转录过程整合到 cDNA 中，使得 cDNA 可被 BrdU 抗体严格纯化；eCLIP[99] 使用两步独立的接头连接步骤等。

各类 CILP-seq 数据的分析流程主要包括质控、比对到参考基因组、峰鉴定、模体分析、下游分析 (图 4)。其中，有集合多个功能的分析管道可以使用，如 CLIP Tool Kit (CTK)[100] 可用于质控、比对与峰鉴定，在此基础上 CLIPSeqTools[101] 和 CLIPZ[102] 还



产生和分析CLIP-seq数据的主要步骤：首先通过紫外交联蛋白质与RNA，通过酶消化RNA片段，分离并纯化RNA后，建立文库进行测序；测序数据先通过一定的预处理后比对到基因组上，随后使用一到多款软件进行峰的鉴定，获得峰的具体位置后可进行一系列的下游分析。

图4　CILP-seq数据的分析流程

能用于模体分析。由于这些高通量技术基本都使用了紫外交联 (除了 RIP-seq)，因此会导致交联位点处产生截断、突变、插入或缺失等诊断事件，这种诊断事件反过来可用于定位交联位置。目前，一些工具已经被开发用于 CLIP 数据的计算分析，但其假阳性率往往较高，不同的峰鉴定工具可能会鉴定到不同的结合区域。此外，它们往往具有不同的缺点，例如两款最常用的软件，CLIPper[99] 和 Piranha[103]，不能检测单个的交联位点。CLIPper 即使使用多个线程，其鉴定峰也需要很长的时间。Piranha 不能利用基因组信息，其通过设置恒定的步长对序列进行扫描和分析，不同的步长可能会导致不同的结论。CTK 包含太多的步骤，同时需要 Perl 环境，复杂的工作流程对新手来说并不友好。PIPE-CLIP[104] 是一个在线工具，难以实现更大规模的数据分析和整合。更重要的是，许多工具只能应用于特定的某种 CLIP-seq 数据。例如，PARalyzer[105]、WavClusteR[106] 只能用于 PAR-CLIP 数据，PureCLIP[107] 只适用于 iCLIP 或 eCLIP 数据，这使得不同的结果难以比较和整合。因此，未来需要开发能克服这些缺点的工具，更准确、快速地鉴定峰区域和结合位点，有利于更好地整合大规模的测序数据。

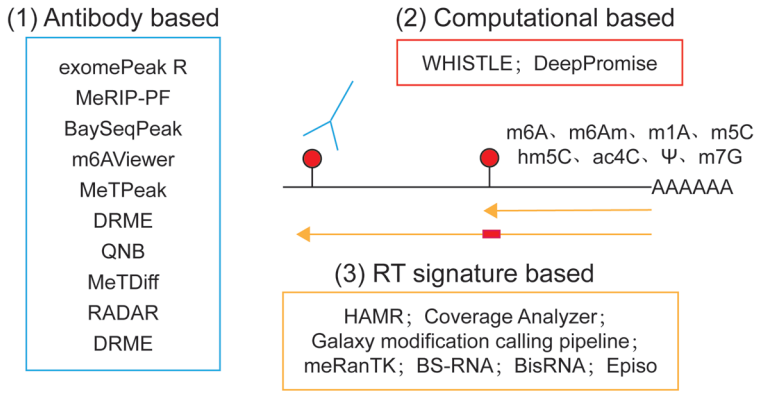
5 RNA修饰的计算机检测方法

在 RNA 中发现了至少 170 种化学修饰 [108]，其主要包括 m6A、m6Am、m1A、m5C、hm5C、ac4C、Ψ、m7G 等。随着测序技术的不断进步，出现了越来越多的在全转录组范围鉴定 RNA 修饰的技术，结合相应的生信分析工具，极大地加快了表观转录组的功能研究。基于不同的检测策略，目前使用较多的生信分析工具可以分为：分析抗体富集数据的工具、分析反转录信号的工具、基于计算预测的工具 (图 5)。

甲基化 RNA 免疫沉淀测序 (MeRIP-Seq 或 m6A-Seq)[109] 是迄今为止使用最广泛的实验方法。通过将此类测序数据的 read 比对到基因组，能在 RNA 修饰位点附近获得富集的峰信号，通过富集组和非富集对照组的统计学分析，能够获得全转录组的 RNA 修饰分布。目前已有一些程序能用于 MeRIP-Seq 的富集峰鉴定，如 exomePeak[110]、MeRIP-PF[111]、BaySeqPeak[112]、m6AViewer[113]、MeTPeak[114] 等。鉴定甲基化位点之后，可以进行甲基化的差异分析，从而解析一些生物学问题，此类程序包括 QNB[115]、MeTDiff[116]、RADAR[117]、DRME[118] 等。

一些 RNA 修饰的检测策略基于反转录信号的变化，一般为诱导反转录终止或引入突变。如假尿嘧啶修饰的检测是通过化学试剂 CMC 在修饰处留下一个庞大的基团，并导致逆转提前终止 [119]。m5C 可以通过亚硫酸氢盐处理的 RNA 高通量测序 (RNA bisulfite sequencing, RNA-BisSeq) 进行检测 [120]，该测序将所有未修饰的胞嘧啶转化为尿嘧啶，留下修饰的胞嘧啶不受影响。而某些 RNA 修饰本身可以影响反转录的进行，如 m1A。基于此，一些程序能对反转录的变化进行分析，如 HAMR[121]、Coverage Analyzer[122]、Galaxy modification calling pipeline[123] 等。而 meRanTK[124]、BS-RNA[125]、BisRNA[126]、Episo[127] 等软件能针对 RNA-BisSeq 进行分析。

实际上，大部分的 RNA 修饰位点都是通过计算预测得到的，即通过提取预测特征并利用机器学习或深度学习分类器来预测假定的 RNA 修饰位点。针对不同的 RNA 修饰，目前有很多相关的预测程序，基于支持向量机算法的 WHISTLE 程序 [128] 和



(1)基于抗体富集数据的工具；(2)基于算法直接预测的工具；(3)基于反转录信号检测的工具。

图5 RNA修饰检测的分析工具

基于深度学习的 DeepPromise 程序[129]的预测性能更佳。

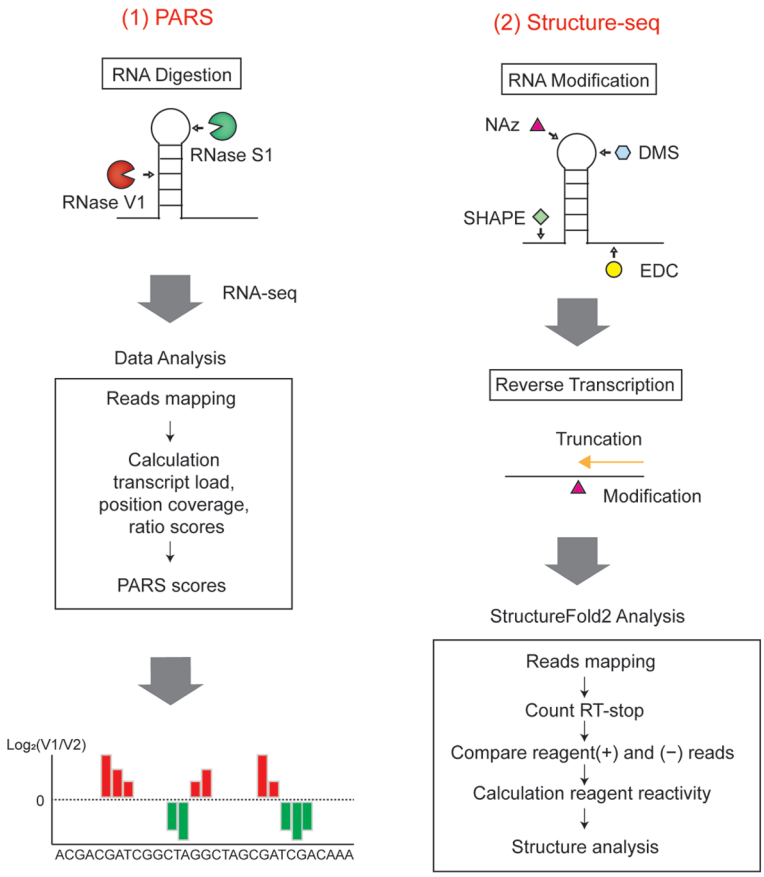
尽管基于高通量测序的 RNA 修饰检测技术发展迅速，但目前生物信息学分析工具和分析流程仍然较为局限，未来需要进一步扩展到其他 RNA 修饰中。此外，由于 RNA 修饰位点往往是动态的且具有组织特异性，这些实验技术往往无法检测到所有修饰位点，同时，这些实验往往忽略 RNA 高级结构的影响，因此未来需要进一步提高实验和计算机检测的准确性和灵敏度。

6　RNA二级结构的检测

RNA 结构是生物学中许多重要功能的基础。近年来，多种方法将结构探测技术与高通量测序相结合，从而在全基因组范围内检测结构信息，其主要包括基于核酸酶的方法和基于化学修饰的方法（图 6）。

最早应用的是基于核酸酶的检测方法。PARS (parallel analysis of RNA structures)[130]利用 RNase V1 和 RNase S1 的结构特异性，分别用于双链 RNA 和单链 RNA 的切割，获得高通量数据后，回帖到转录组，然后将每个 read 的 5' 端上游 1 nt 确定为切割位点，最后使用 RNase V1 与 RNase S1 切割位点数量的比值来计算 PARS 评分。FragSeq (fragmentation sequencing)[131]与 PARS 相似，只是它仅依赖于 ssRNase 核酸酶 P1 来识别未配对区域。与 PARS 和 FragSeq 相比，ds/ssRNA-seq[132]并不寻找直接的切割位点，而是旨在对 RNA 被 RNase 彻底消化后剩余的 RNA 进行测序，这种方法缺乏一定的分辨率，但可以对 RNA 结构进行更局部的观察，且可以同时对 RNA 二级结构和 RNA- 蛋白质相互作用位点进行观察。

目前更为主流的是基于化学修饰的方法，即使



(1) PARS技术及分析步骤。通过RNase V1与RNase S1分别特异性切割RNA双链和单链，随后构建文库进行常规的RNA-seq；获得测序数据后，将read比对到基因组上计算出转录本的覆盖度、位点覆盖度及比例等信息；最后，使用RNase V1与RNase S1切割位点数量的比值来计算PARS评分。(2) Structure-seq技术及分析步骤。通过不同的化学物质特异性地对RNA单链进行修饰，形成的空间位阻导致随后的反转录停顿，使用软件StructureFold2进行停顿位点与试剂可及性的计算，最终获得可能的二级结构。

图6　RNA二级结构检测的技术及分析步骤

用化学物质渗透进入活细胞，并在无保护的单链位点对 RNA 进行共价修饰，这些修饰会导致随后的逆转录停顿，从而产生截断的 cDNA 片段，然后通过检测随后的逆转录终止位点读出碱基修饰，Structure-seq 是此类方法中的一种 [133]。突变测序分析 (mutational profiling with sequencing, MaPseq)[134] 是一种最近发展起来的替代逆转录终止的分析方法，其利用一些高亲和逆转录酶的偏好，读取 RNA 碱基修饰，并在修饰位点对面插入 cDNA 突变，该方法的优点是可以读取单个分子上的多个修饰，增加给定修饰碱基的测序深度。随着技术的不断进步，用于检测 RNA 二级结构的化学物质越来越多，使用二甲基硫酸酯 (DMS) 能够使 Watson-Crick 中无保护的腺嘌呤和胞嘧啶残基甲基化 [135]，使用乙二醇可以修饰鸟嘌呤 [136]，使用 1- 乙基 -3- 碳二亚胺 (EDC) 可以修饰尿嘧啶和鸟嘌呤 [136]，而使用 SHAPE 试剂 (selective 2'-hydroxyl acylation and profiling experiment) 则可以修饰所有碱基的核糖部分 [137]，而烟碱酰叠氮化合物 (NAz) 可以在紫外光激发下快速探测非 Watson-Crick 的碱基互作 [138]。基于逆转录停顿的测序数据，可以使用软件 StructureFold2 进行二级结构分析 [139]。SeqFold 则结合了高通量 RNA 结构分析数据和计算预测，其能够整合各种高通量的 RNA 结构数据，并适用于分析转录组中的 RNA 结构 [140]。

目前，基于高通量测序研究 RNA 的高级结构尚在发展阶段，细胞中的缓冲环境和某些功能机制都有可能影响 RNA 的天然结构。同时，RNA 本身的某些化学修饰同样可以导致逆转录终止。这些都是未来有待考虑与改进的方面。重要的是，RNA 结构的大规模信息解析技术的发展将有助于深入了解 RNA 的调控功能。

7 结语与展望

非编码 RNA 及其作用靶标的类型、RNA 的功能网络、RNA 与蛋白质的互作、RNA 修饰及 RNA 二级结构皆是 RNA 重要的功能信息，如何排除假阳性和假阴性获得更准确的 RNA 信息，获得海量 RNA 信息后如何阐明其复杂的功能网络，以及如何应用到临床或生产，仍需要大量的研究。随着实验技术 (如文库构建策略) 的发展以及测序技术 (如单分子测序、空间转录组) 的不断革新，将会有更多关于 RNA 的不同层面的数据产生，为生物信息技术的发展提供前所未有的机遇和挑战。现有的信息学工具的完善与更新，以及全新的信息分析方法 (如结合多组学的分析) 的开发，将有助于加深对 RNA 在生物体内的重要作用的理解。

[参 考 文 献]

[1] Sekar S, Geiger P, Cuyugan L, et al. Identification of circular RNAs using RNA sequencing. J Vis Exp, 2019, doi: 10.3791/59981

[2] Imanishi T, Suzuki Y, O'Donovan C, et al. Integrative annotation of 21,037 human genes validated by full-length cDNA clones. PLoS Biol, 2004, 2: e162

[3] Wen J, Parker BJ, Weiller GF. *In silico* identification and characterization of mRNA-like noncoding transcripts in *Medicago truncatula*. In Silico Biol, 2007, 7: 485-505

[4] Prensner JR, Iyer MK, Balbin OA, et al. Transcriptome sequencing across a prostate cancer cohort identifies PCAT-1, an unannotated lincRNA implicated in disease progression. Nat Biotechnol, 2011, 29: 742-9

[5] Necsulea A, Soumillon M, Warnefors M, et al. The evolution of lncRNA repertoires and expression patterns in tetrapods. Nature, 2014, 505: 635-40

[6] Dobin A, Davis CA, Schlesinger F, et al. STAR: ultrafast universal RNA-seq aligner. Bioinformatics, 2013, 29: 15-21

[7] Trapnell C, Williams BA, Pertea G, et al. Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. Nat Biotechnol, 2010, 28: 511-5

[8] Grabherr MG, Haas BJ, Yassour M, et al. Full-length transcriptome assembly from RNA-Seq data without a reference genome. Nat Biotechnol, 2011, 29: 644-52

[9] Sun L, Luo H, Bu D, et al. Utilizing sequence intrinsic composition to classify protein-coding and long non-coding transcripts. Nucleic Acids Res, 2013, 41: e166

[10] Liu J, Gough J, Rost B. Distinguishing protein-coding from non-coding RNAs through support vector machines. PLoS Genet, 2006, 2: e29

[11] Wang L, Park HJ, Dasari S, et al. CPAT: Coding-Potential Assessment Tool using an alignment-free logistic regression model. Nucleic Acids Res, 2013, 41: e74

[12] Kang YJ, Yang DC, Kong L, et al. CPC2: a fast and accurate coding potential calculator based on sequence intrinsic features. Nucleic Acids Res, 2017, 45: W12-16

[13] Sun K, Chen X, Jiang P, et al. iSeeRNA: identification of long intergenic non-coding RNA transcripts from transcriptome sequencing data. BMC Genomics, 2013, 14 Suppl 2: S7

[14] Li A, Zhang J, Zhou Z. PLEK: a tool for predicting long non-coding RNAs and messenger RNAs based on an improved *k*-mer scheme. BMC Bioinformatics, 2014, 15: 311

[15] Lin MF, Jungreis I, Kellis M. PhyloCSF: a comparative genomics method to distinguish protein coding and non-coding regions. Bioinformatics, 2011, 27: i275-82

[16] Washietl S, Findeiss S, Müller SA, et al. RNAcode: robust discrimination of coding and noncoding regions in comparative sequence data. RNA, 2011, 17: 578-94

[17] Finn RD, Clements J, Eddy SR. HMMER web server: interactive sequence similarity searching. Nucleic Acids Res, 2011, 39: W29-37
[18] Pantano L, Estivill X, Martí E. A non-biased framework for the annotation and classification of the non-miRNA small RNA transcriptome. Bioinformatics, 2011, 27: 3202-3
[19] Fasold M, Langenberger D, Binder H, et al. DARIO: a ncRNA detection and analysis tool for next-generation sequencing experiments. Nucleic Acids Res, 2011, 39: W112-7
[20] Chen CJ, Servant N, Toedling J, et al. ncPRO-seq: a tool for annotation and profiling of ncRNAs in sRNA-seq data. Bioinformatics, 2012, 28: 3147-9
[21] Yang JH, Shao P, Zhou H, et al. deepBase: a database for deeply annotating and mining deep sequencing data. Nucleic Acids Res, 2010, 38: D123-30
[22] Friedländer MR, Mackowiak SD, Li N, et al. miRDeep2 accurately identifies known and hundreds of novel microRNA genes in seven animal clades. Nucleic Acids Res, 2012, 40: 37-52
[23] Li JH, Liu S, Zhou H, et al. starBase v2.0: decoding miRNA-ceRNA, miRNA-ncRNA and protein-RNA interaction networks from large-scale CLIP-Seq data. Nucleic Acids Res, 2014, 42: D92-7
[24] Hansen TB, Venø MT, Kjems J, et al. miRdentify: high stringency miRNA predictor identifies several novel animal miRNAs. Nucleic Acids Res, 2014, 42: e124
[25] Mathelier A, Carbone A. MIReNA: finding microRNAs with high accuracy and no learning at genome scale and from deep sequencing data. Bioinformatics, 2010, 26: 2226-34
[26] An J, Lai J, Lehman ML, et al. miRDeep*: an integrated application tool for miRNA identification from RNA sequencing data. Nucleic Acids Res, 2013, 41: 727-37
[27] Yee LY, Paul R, Ungar LH, et al. CoRAL: predicting non-coding RNAs from small RNA-sequencing data. Nucleic Acids Res, 2013, 41: e137
[28] Michael H, Naiara RE, Aransay AM. miRanalyzer: an update on the detection and analysis of microRNAs in high-throughput sequencing experiments. Nucleic Acids Res, 2011, 39: W132-8
[29] Videm P, Rose D, Costa F, et al. BlockClust: efficient clustering and classification of non-coding RNAs from short read RNA-seq profiles. Bioinformatics, 2014, 30: i274-82
[30] Langenberger D, Pundhir S, Ekstrm CT, et al. DeepBlock-Align: a tool for aligning RNA-seq profiles of read block patterns. Bioinformatics, 2011, 28: 17-24
[31] Pundhir S, Gorodkin J. MicroRNA discovery by similarity search to a database of RNA-seq profiles. Front Genet, 2013, 4: 133
[32] Wight M, Werner A. The functions of natural antisense transcripts. Essays Biochem, 2013, 54: 91-101
[33] Zhang W, Zhou X, Xia J, et al. Identification of microRNAs and natural antisense transcript-originated endogenous siRNAs from small-RNA deep sequencing data. Methods Mol Biol, 2021, 883: 221-7
[34] Markham NR, Zuker M. DINAMelt web server for nucleic acid melting prediction. Nucleic Acids Res, 2005, 33: W577-81
[35] Dimitrov RA, Zuker M. Prediction of hybridization and melting for double-stranded nucleic acids. Biophys J, 2004, 87: 215-26
[36] Yu D, Meng Y, Zuo Z, et al. NATpipe: an integrative pipeline for systematical discovery of natural antisense transcripts (NATs) and phase-distributed nat-siRNAs from *de novo* assembled transcriptomes. Sci Rep, 2016, 6: 21666
[37] Thody J, Folkes L, Moulton V. NATpare: a pipeline for high-throughput prediction and functional analysis of nat-siRNAs. Nucleic Acids Res, 2020, 48: 6481-90
[38] Zhang Y, Wang X, Kang L. A *k*-mer scheme to predict piRNAs and characterize locust piRNAs. Bioinformatics, 2011, 27: 771-6
[39] Brayet J, Zehraoui F, Jeanson-Leh L, et al. Towards a piRNA prediction using multiple kernel fusion and support vector machine. Bioinformatics, 2014, 30: i364-70
[40] Wang K, Liang C, Liu J, et al. Prediction of piRNAs using transposon interaction and a support vector machine. BMC Bioinformatics, 2014, 15: 419
[41] Monga I, Banerjee I. Computational identification of piRNAs using features based on RNA sequence, structure, thermodynamic and physicochemical properties. Curr Genomics, 2019, 20: 508-18
[42] Erwin AA, Galdos MA, Wickersheim ML, et al. piRNAs are associated with diverse transgenerational effects on gene and transposon expression in a hybrid dysgenic syndrome of *D. virilis*. PLoS Genet, 2015, 11: e1005332
[43] Rosenkranz D, Zischler H. proTRAC--a software for probabilistic piRNA cluster detection, visualization and analysis. BMC Bioinformatics, 2012, 13: 5
[44] Jung I, Park JC, Kim S. piClust: A density based piRNA clustering algorithm. Comput Biol Chem, 2014, 50: 60-7
[45] Panda AC, Supriyo D, Ioannis G, et al. High-purity circular RNA isolation method (RPAD) reveals vast collection of intronic circRNAs. Nucleic Acids Res, 2017, 45: e116
[46] 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
[47] Gao Y, Zhang J, Zhao F. Circular RNA identification based on multiple seed matching. Brief Bioinform, 2017, 19: 803-10
[48] Szabo L, Morey R, Palpant NJ, et al. Statistically based splicing detection reveals neural enrichment and tissue-specific induction of circular RNA during human fetal development. Genome Biol, 2016, 16: 126
[49] Memczak S, Jens M, Elefsinioti A, et al. Circular RNAs are a large class of animal RNAs with regulatory potency. Nature, 2013, 495: 333-8
[50] Hoffmann S, Otto C, Doose G, et al. A multi-split mapping algorithm for circular RNA, splicing, trans-splicing and fusion detection. Genome Biol, 2014, 15: R34

[51] Gao Y, Wang J, Zhao F. CIRI: an efficient and unbiased algorithm for *de novo* circular RNA identification. Genome Biol, 2015, 16: 4
[52] Humphreys DT, Fossat N, Demuth M, et al. Ularcirc: visualization and enhanced analysis of circular RNAs via back and canonical forward splicing. Nucleic Acids Res, 2019, 47: e123
[53] Song X, Zhang N, Han P, et al. Circular RNA profile in gliomas revealed by identification tool UROBORUS. Nucleic Acids Res, 2016, 44: e87
[54] Zeng X, Lin W, Guo M, et al. A comprehensive overview and evaluation of circular RNA detection tools. PLoS Comput Biol, 2017, 13: e1005420
[55] Hansen TB. Improved circRNA identification by combining prediction algorithms. Front Cell Dev Biol, 2018, 6: 20
[56] Pan X, Xiong K. PredcircRNA: computational classification of circular RNA from other long non-coding RNA using hybrid features. Mol Biosyst, 2015, 11: 2219-26
[57] Chen L, Zhang YH, Huang G, et al. Discriminating cirRNAs from other lncRNAs using a hierarchical extreme learning machine (H-ELM) algorithm with feature selection. Mol Genet Genomics, 2017, 293: 137-49
[58] Niu M, Zhang J, Li Y, et al. CirRNAPL: A web server for the identification of circRNA based on extreme learning machine. Comput Struct Biotechnol J, 2020, 18: 834-42
[59] Wang J, Wang L. Deep learning of the back-splicing code for circular RNA formation. Bioinformatics, 2019, 35: 5235-42
[60] Chaabane M, Williams RM, Stephens AT, et al. circDeep: deep learning approach for circular RNA classification from other long non-coding RNA. Bioinformatics, 2020, 36: 73-80
[61] Liu C, Liu YC, Huang HD, et al. Biogenesis mechanisms of circular RNA can be categorized through feature extraction of a machine learning model. Bioinformatics, 2019, 35: 4867-70
[62] Hoogstrate Y, Jenster G, Martens-Uzunova ES. FlaiMapper: computational annotation of small ncRNA-derived fragments using RNA-seq high-throughput data. Bioinformatics, 2015, 31: 665-73
[63] You X, Conrad TO. Acfs: accurate circRNA identification and quantification from RNA-Seq data. Sci Rep, 2016, 6: 38820
[64] Lewis BP, Burge CB, Bartel DP. Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are microRNA targets. Cell, 2005, 120: 15-20
[65] Betel D, Koppal A, Agius P, et al. Comprehensive modeling of microRNA targets predicts functional non-conserved and non-canonical sites. Genome Biol, 2010, 11: R90
[66] Krek A, Grün D, Poy MN, et al. Combinatorial microRNA target predictions. Nat Genet, 2005, 37: 495-500
[67] Granneman S, Kudla G, Petfalski E, et al. Identification of protein binding sites on U3 snoRNA and pre-rRNA by UV cross-linking and high-throughput analysis of cDNAs. Proc Natl Acad Sci U S A, 2009, 106: 9613-8
[68] Helwak A, Kudla G, Dudnakova T, et al. Mapping the human miRNA interactome by CLASH reveals frequent noncanonical binding. Cell, 2013, 153: 654-65
[69] Travis AJ, Moody J, Helwak A, et al. Hyb: a bioinformatics pipeline for the analysis of CLASH (crosslinking, ligation and sequencing of hybrids) data. Methods, 2014, 65: 263-73
[70] Ding J, Li X, Hu H. TarPmiR: a new approach for microRNA target site prediction. Bioinformatics, 2016, 32: 2768-75
[71] Tan Shen M, Kirchner R, Jin J, et al. Sequencing of captive target transcripts identifies the network of regulated genes and functions of primate-specific miR-522. Cell Rep, 2014, 8: 1225-39
[72] German MA, Pillay M, Jeong DH, et al. Global identification of microRNA-target RNA pairs by parallel analysis of RNA ends. Nat Biotechnol, 2008, 26: 941-6
[73] Addo-Quaye C, Miller W, Axtell M, et al. CleaveLand: a pipeline for using degradome data to find cleaved small RNA targets. Bioinformatics, 2008, 25: 130-1
[74] Zheng Y, Li YF, Sunkar R, et al. SeqTar: an effective method for identifying microRNA guided cleavage sites from degradome of polyadenylated transcripts in plants. Nucleic Acids Res, 2011, 40: e28
[75] Joshua T, Leighton F, Zahara MC, et al. PAREsnip2: A tool for high-throughput prediction of small RNA targets from degradome sequencing data using configurable targeting rules. Nucleic Acids Res, 2018, 46: 8730-9
[76] Kakrana A, Hammond R, Patel P, et al. sPARTA: a parallelized pipeline for integrated analysis of plant miRNA and cleaved mRNA data sets, including new miRNA target-identification software. Nucleic Acids Res, 2014, 42: e139
[77] Ye X, Yang Z, Jiang Y, et al. sRNATargetDigger: A bioinformatics software for bidirectional identification of sRNA-target pairs with co-regulatory sRNAs information. PLoS One, 2020, 15: e0244480
[78] Yuan J, Zhang P, Cui Y, et al. Computational identification of piRNA targets on mouse mRNAs. Bioinformatics, 2015, 32: 1170-7
[79] Shen EZ, Chen H, Ozturk AR, et al. Identification of piRNA binding sites reveals the Argonaute regulatory landscape of the *C. elegans* germline. Cell, 2018, 172: 937-51.e18
[80] Chu C, Qu K, Zhong FL, et al. Genomic maps of long noncoding RNA occupancy reveal principles of RNA-chromatin interactions. Mol cell, 2011, 44: 667-78
[81] Imig J, Brunschweiger A, Brümmer A, et al. miR-CLIP capture of a miRNA targetome uncovers a lincRNA H19–miR-106a interaction. Nat Chem Biol, 2014, 11: 107-14
[82] Hansen TB, Jensen TI, Clausen BH, et al. Natural RNA circles function as efficient microRNA sponges. Nature, 2013, 495: 384-8
[83] Ying Y, Yue Z, Chunhua L, et al. RAID v2.0: an updated resource of RNA-associated interactions across organisms. Nucleic Acids Res, 2017, 45: D115-8
[84] Sarver AL, Subramanian S. Competing endogenous RNA

database. Bioinformation, 2012, 8: 731-3
[85] Wang P, Zhi H, Zhang Y, et al. miRSponge: a manually curated database for experimentally supported miRNA sponges and ceRNAs. Database (Oxford), 2015, 2015: bav098
[86] Wang P, Li X, Gao Y, et al. LncACTdb 2.0: an updated database of experimentally supported ceRNA interactions curated from low- and high-throughput experiments. Nucleic Acids Res, 2019, 47: D121-7
[87] Le TD, Zhang J, Liu L, et al. Computational methods for identifying miRNA sponge interactions. Brief Bioinform, 2017, 18: 577-90
[88] Furió-Tarí P, Tarazona S, Gabaldón T, et al. spongeScan: a web for detecting microRNA binding elements in lncRNA sequences. Nucleic Acids Res, 2016, 44: W176-80
[89] Li R, Qu H, Wang S, et al. GDCRNATools: an R/Bioconductor package for integrative analysis of lncRNA, miRNA and mRNA data in GDC. Bioinformatics, 2018, 34: 2515-7
[90] Paci P, Colombo T, Farina L. Computational analysis identifies a sponge interaction network between long non-coding RNAs and messenger RNAs in human breast cancer. BMC Syst Biol, 2014, 8: 1-15
[91] Sardina DS, Alaimo S, Ferro A, et al. A novel computational method for inferring competing endogenous interactions. Brief Bioinform, 2016, 18: 1071-81
[92] List M, Dehghani Amirabad A, Kostka D, et al. Large-scale inference of competing endogenous RNA networks with sparse partial correlation. Bioinformatics, 2019, 35: i596-604
[93] Zhao J, Ohsumi TK, Kung JT, et al. Genome-wide identification of polycomb-associated RNAs by RIP-seq. Mol Cell, 2010, 40: 939-53
[94] Licatalosi DD, Mele A, Fak JJ, et al. HITS-CLIP yields genome-wide insights into brain alternative RNA processing. Nature, 2008, 456: 464-9
[95] Hafner M, Landthaler M, Burger L, et al. Transcriptome-wide identification of RNA-binding protein and microRNA target sites by PAR-CLIP. Cell, 2010, 141: 129-41
[96] 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: 909-15
[97] Zarnegar BJ, Flynn RA, Shen Y, et al. irCLIP platform for efficient characterization of protein-RNA interactions. Nat Methods, 2016, 13: 489-92
[98] Weyn-Vanhentenryck SM, Mele A, Yan Q, et al. HITS-CLIP and integrative modeling define the Rbfox splicing-regulatory network linked to brain development and autism. Cell Rep, 2014, 6: 1139-52
[99] Van Nostrand EL, Pratt GA, Shishkin AA, et al. Robust transcriptome-wide discovery of RNA-binding protein binding sites with enhanced CLIP (eCLIP). Nat Methods, 2016, 13: 508-14
[100] Shah A, Qian Y, Weyn-Vanhentenryck SM, et al. CLIP Tool Kit (CTK): a flexible and robust pipeline to analyze CLIP sequencing data. Bioinformatics, 2017, 33: 566-7
[101] Maragkakis M, Alexiou P, Nakaya T, et al. CLIPSeqTools--a novel bioinformatics CLIP-seq analysis suite. RNA, 2016, 22: 1-9
[102] Khorshid M, Rodak C, Zavolan M. CLIPZ: a database and analysis environment for experimentally determined binding sites of RNA-binding proteins. Nucleic Acids Res, 2011, 39: D245-52
[103] Uren PJ, Bahrami-Samani E, Burns SC, et al. Site identification in high-throughput RNA-protein interaction data. Bioinformatics, 2012, 28: 3013-20
[104] Chen B, Yun J, Kim MS, et al. PIPE-CLIP: a comprehensive online tool for CLIP-seq data analysis. Genome Biol, 2014, 15: R18
[105] Corcoran DL, Georgiev S, Mukherjee N, et al. PARalyzer: definition of RNA binding sites from PAR-CLIP short-read sequence data. Genome Biol, 2011, 12: R79
[106] Comoglio F, Sievers C, Paro R. Sensitive and highly resolved identification of RNA-protein interaction sites in PAR-CLIP data. BMC Bioinformatics, 2015, 16: 32
[107] 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
[108] Zhao LY, Song J, Liu Y, et al. Mapping the epigenetic modifications of DNA and RNA. Protein Cell, 2020, 11: 792-808
[109] Meyer KD, Saletore Y, Zumbo P, et al. Comprehensive analysis of mRNA methylation reveals enrichment in 3' UTRs and near stop codons. Cell, 2012, 149: 1635-46
[110] Meng J, Lu Z, Liu H, et al. A protocol for RNA methylation differential analysis with MeRIP-Seq data and exomePeak R/Bioconductor package. Methods, 2014, 69: 274-81
[111] Li Y, Song S, Li C, et al. MeRIP-PF: an easy-to-use pipeline for high-resolution peak-finding in MeRIP-Seq data. Genomics Proteomics Bioinformatics, 2013, 11: 72-5
[112] Zhang M, Li Q, Xie Y. A Bayesian hierarchical model for analyzing methylated RNA immunoprecipitation sequencing data. Quant Biol, 2018, 6: 275-86
[113] Antanaviciute A, Baquero-Perez B, Watson CM, et al. m6aViewer: software for the detection, analysis and visualization of N^6-methyl-adenosine peaks from m^6A-seq/ME-RIP sequencing data. RNA, 2017, 23: 1493
[114] Cui X, Jia M, Zhang S, et al. A novel algorithm for calling mRNA m^6A peaks by modeling biological variances in MeRIP-seq data. Bioinformatics, 2016, 32: i378-85
[115] Liu L, Zhang SW, Huang Y, et al. QNB: differential RNA methylation analysis for count-based small-sample sequencing data with a quad-negative binomial model. BMC Bioinformatics, 2017, 18: 1-12
[116] Cui X, Zhang L, Meng J, et al. Metdiff: a novel differential RNA methylation analysis for MeRIP-Seq data. IEEE/ACM Trans Comput Biol Bioinform, 2018, 15: 526-34
[117] Zhang Z, Zhan Q, Eckert M, et al. RADAR: differential analysis of MeRIP-seq data with a random effect model. Genome Biol, 2019, 20: 1-17
[118] Liu L, Zhang SW, Gao F, et al. DRME: count-based

differential RNA methylation analysis at small sample size scenario. Anal Biochem, 2016, 499: 15-23

[119] Carlile TM, Rojas-Duran MF, Zinshteyn B, et al. Pseudouridine profiling reveals regulated mRNA pseudouridylation in yeast and human cells. Nature, 2014, 515: 143-6

[120] Schaefer M, Pollex T, Hanna K, et al. RNA methylation by Dnmt2 protects transfer RNAs against stress-induced cleavage. Genes Dev, 2010, 24:1590-5

[121] Ryvkin P, Leung YY, Silverman IM, et al. HAMR: high-throughput annotation of modified ribonucleotides. RNA, 2013, 19: 1684-92

[122] Hauenschild R, Werner S, Tserovski L, et al. CoverageAnalyzer (CAn): a tool for inspection of modification signatures in RNA sequencing profiles. Biomolecules, 2016, 6: 42

[123] Schmidt L, Werner S, Kemmer T, et al. Graphical workflow system for modification calling by machine learning of reverse transcription signatures. Front Genet, 2019, 10: 876

[124] Rieder D, Amort T, Kugler E, et al. meRanTK: methylated RNA analysis ToolKit. Bioinformatics, 2015, 32: 782-5

[125] Liang F, Hao L, Wang J, et al. BS-RNA: an efficient mapping and annotation tool for RNA bisulfite sequencing data. Comput Biol Chem, 2016, 65: 173

[126] Legrand C, Tuorto F, Hartmann M, et al. Statistically robust methylation calling for whole-transcriptome bisulfite sequencing reveals distinct methylation patterns for mouse RNAs. Genome Res, 2017, 27: 1589-96

[127] Liu J, An Z, Luo J, et al. Episo: quantitative estimation of RNA 5-methylcytosine at isoform level by high-throughput sequencing of RNA treated with bisulfite. Bioinformatics, 2019, 36: 2033-9

[128] Kunqi C, Zhen W, Qing Z, et al. WHISTLE: a high-accuracy map of the human N^6-methyladenosine (m^6A) epitranscriptome predicted using a machine learning approach. Nucleic Acids Res, 2019, 23: e41

[129] Chen Z, Zhao P, Li F, et al. Comprehensive review and assessment of computational methods for predicting RNA post-transcriptional modification sites from RNA sequences. Brief Bioinform, 2020, 21: 1676-96

[130] Kertesz M, Yue W, Mazor E, et al. Genome-wide measurement of RNA secondary structure in yeast. Nature, 2010, 467: 103-7

[131] Underwood JG, Uzilov AV, Katzman S, et al. FragSeq: transcriptome-wide RNA structure probing using high-throughput sequencing. Nat Methods, 2010, 7: 995-1001

[132] Li F, Zheng Q, Ryvkin P, et al. Global analysis of RNA secondary structure in two metazoans. Cell Rep, 2012, 1: 69-82

[133] Ding Y, Kwok KC, Tang Y, et al. Genome-wide profiling of *in vivo* RNA structure at single-nucleotide resolution using structure-seq. Nat Protoc, 2015, 10: 1050-66

[134] Zubradt M, Gupta P, Persad S, et al. DMS-MaPseq for genome-wide or targeted RNA structure probing *in vivo*. Nat Methods, 2016, 14: 75-82

[135] Ritchey LE, Su Z, Tang Y, et al. Structure-seq2: sensitive and accurate genome-wide profiling of RNA structure *in vivo*. Nucleic Acids Res, 2017, 45: e135

[136] Mitchell D 3rd, Renda AJ, Douds CA, et al. *In vivo* RNA structural probing of uracil and guanine base-pairing by 1-ethyl-3-(3-dimethylaminopropyl)carbodiimide (EDC). RNA, 2019, 25: 147-57

[137] Spitale RC, Crisalli P, Flynn RA, et al. RNA SHAPE analysis in living cells. Nat Chem Biol, 2013, 9: 18-20

[138] Feng C, Chan D, Joseph J, et al. Light-activated chemical probing of nucleobase solvent accessibility inside cells. Nat Chem Biol, 2018, 14: 276-83

[139] Tack DC, Tang Y, Ritchey LE, et al. StructureFold2: bringing chemical probing data into the computational fold of RNA structural analysis. Methods, 2018, 143: 12-5

[140] Ouyang Z, Snyder MP, Chang HY. SeqFold: genome-scale reconstruction of RNA secondary structure integrating high-throughput sequencing data. Genome Res, 2013 , 23: 377-87