PIBB 生物化学与生物物理进展
Progress in Biochemistry and Biophysics
2012, 39(12): 1156~1166
www.pibb.ac.cn

基于 RNA-Seq 的长非编码 RNA 预测 *

孙 磊[1, 2] 张 林[1] 刘 辉[1]**
([1] 中国矿业大学信息与电气工程学院，徐州 221008；[2] 中国科学院北京基因组研究所计算生物学研究中心，北京 100029)

摘要 随着新一代生物技术和生物信息学的发展，研究发现，在真核生物转录组中存在大量长非编码 RNA(long non-coding RNA，lncRNA)，而这些 lncRNA 可能在基因表达调控过程中起到关键性的功能作用．当前 lncRNA 研究主要采用高通量 RNA-Seq 测序技术，并通过生物信息学方法对测序数据进行处理和分析，以挖掘其中 lncRNA 的序列、结构、表达及功能等信息．本文将对基于 RNA-Seq 的 lncRNA 预测流程进行介绍，对其中涉及的生物信息学方法进行较为全面的综述，就相关问题和挑战展开讨论，并对研究进行展望．

关键词 长非编码 RNA，RNA-Seq，lncRNA 预测，生物信息学
学科分类号 Q7，Q81，TP391
DOI: 10.3724/SP.J.1206.2012.00287

真核生物基因通过转录形成蛋白质编码 RNA (protein-coding RNA，mRNA) 与非编码 RNA (non-coding RNA，ncRNA)．以哺乳动物为例，预计有超过 50%的基因组能被表达，但迄今只发现了约 2%的基因组负责编码蛋白质，而其余的表达基因很可能转录成为 ncRNA[1]．其中，包括 snoRNA、microRNA、siRNA 和 piRNA 等在内的小 ncRNA(small ncRNA)已经得到广泛研究[2-4]．然而，对长(>200 nt)非编码 RNA (long non-coding RNA，lncRNA)的研究才刚刚进入发展阶段[5-10]．在过去很长一段时间里，由于生物技术的限制，研究者仅确定了个别功能性的 lncRNA，如 Xist 和 HOTAIR 等[11-12]．其中，Xist 位于哺乳动物的 X 染色体上，它是 X 染色体灭活(X inactivation)过程的主导因素．HOTAIR，又被称作 HOX 反义基因间 RNA，它能够控制哺乳动物 2 号染色体上的基因表达．研究也发现 HOTAIR 在转移性乳腺癌中具有高表达的特征[13]．从已知的 lncRNA 功能出发，生物学家推断 lncRNA 可能具有多种功能，它们或通过支架蛋白质以形成新的复合体，或引导转录组因子以抑制特定基因的表达[14]，或通过其他机制对生命体活动产生重要的功能作用[8, 15-22]，如图 1 所示．随着生物技术的不断发展和应用，陆续有研究发现了与胚胎干细胞全能性、细胞周期调控、免疫响应，以及前列腺癌、神经精神障碍等疾病有密切关联的 lncRNA[20, 23-26]．一些与疾病显著相关的 lncRNA 甚至可以作为生物学标记物，用于疾病的诊断和预测[24]．近年来，研究者借助二代测序在脊椎动物组织和细胞中发现了许多新型的 lncRNA[27]，但这些 lncRNA 的功能机制尚未清楚．正是由于 lncRNA 不断被发现且具有潜在的重要功能，lncRNA 研究已成为当前转录组研究领域的前沿问题．

早期研究者通过覆瓦式微阵列(tiling microarray)获取 lncRNA 信息[28-29]，但此方法存在交叉杂交(cross hybridization)以及不能准确定义 lncRNA 的结构等问题[30]．近年来，基于二代测序的 RNA-Seq[31]以其高通量、高灵敏度、低噪声，且能发现已知基因的新可变剪接及新基因等优点，已

* 中国博士后科学基金资助项目(2012M511335, 2012M511336)，中央高校基本科研业务费专项基金资助项目(2010QNA47, 2010QNA50)，霍英东教育基金会青年教师基金资助项目(121066).
** 通讯联系人.
Tel: 15262048535, E-mail: lhcumt@hotmail.com
收稿日期：2012-06-12，接受日期：2012-08-13

成为转录组研究的重要技术. 同时，RNA-Seq 也已取代覆瓦式微阵列成为当前 lncRNA 研究的主要方法.

尽管现有的 RNA-Seq 技术能够捕捉到 lncRNA，但要准确预测 lncRNA 的基因结构和功能仍是一个难题[32]. 在 lncRNA 的基因结构预测方面，主要存在两个难点. 首先，由于当前的测序深度、测序偏好[33]和测序错误[34]等问题，导致装配阶段产生部分(partial)转录本和人工转录本(artefact)，影响 lncRNA 识别. 其次，对于如何区分蛋白质编码基因和非编码基因，一直都是生物学与生物信息学的难点. 针对以上问题，相关研究基于 RNA-Seq 实验，采用特定的生物信息学方法，在脊椎动物组织和细胞中确定了数以千计的 lncRNA，并对其功能进行了分析和推断[18, 30, 35–36]. 综合这些研究，本文将主要介绍基于 RNA-Seq 的 lncRNA 预测流程，并对其中涉及的生物信息学方法进行综述. 在此基础上，还将对 lncRNA 的功能研究进行讨论.

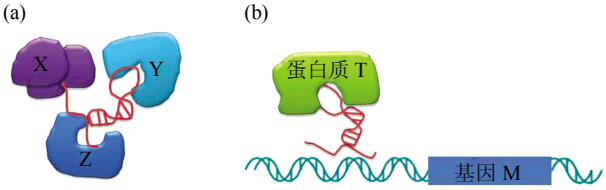

Fig. 1　Potential functions of lncRNA[5]

图 1　lncRNA 的潜在功能[5]

(a) lncRNA(红色)起支架(scaffold)作用，负责连接 X、Y、Z 三个蛋白质以形成新的复合体(complex). (b) lncRNA(红色)一端绑定在基因 M 附近，另一端引导蛋白质 T 与之结合，最终对基因 M 的表达产生抑制作用.

1　基于 RNA-Seq 的长非编码 RNA 预测流程

RNA-Seq 是二代 DNA 测序应用于转录组研究的技术. 随着 RNA-Seq 的日益成熟，相应的生物信息学处理与分析流程呈现出规范化和协议化的趋势[37]. 在此基础上，相关文献对流程进行了扩展，用以研究 lncRNA[35–36]. 基于 RNA-Seq 的 lncRNA 预测流程主要包括文库制备与测序、转录组重建及 lncRNA 识别与分析 3 个阶段，如图 2 所示.

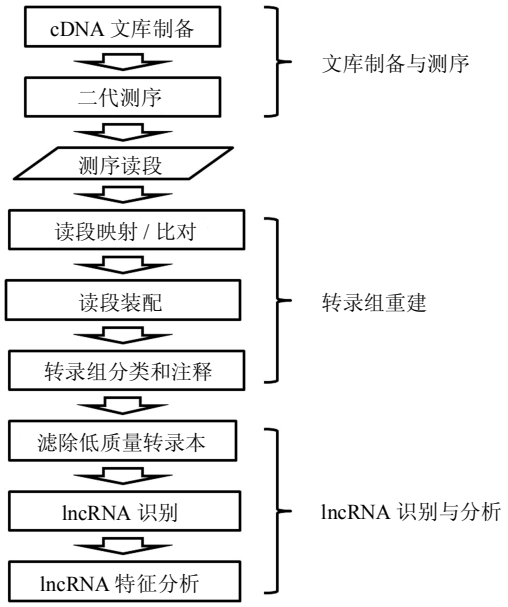

Fig. 2　Pipeline for lncRNA predcition based on RNA-Seq

图 2　基于 RNA-Seq 的 lncRNA 预测流程

整个流程顺序执行. 图右侧标示了 lncRNA 预测流程的 3 个主要阶段：文库制备与测序、转录组重建及 lncRNA 识别与分析；左侧分别对应各个子步骤.

1.1　文库制备与测序

当前 lncRNA 研究主要以类 mRNA(mRNA-like)的 lncRNA 作为研究对象，采用 mRNA-Seq 的文库制备与测序方法[30, 35–36]，其过程如下. 首先分离出样品细胞 / 组织中的 RNA，用 Poly(T)寡聚核苷酸提取 Poly(A)+ RNA. 然后对 Poly(A)+ RNA 进行片段化(fragmentation)处理，接着利用随机引物和反转录酶合成 cDNA，再合成双链 cDNA，之后对 cDNA 进行末端修复并添加测序接头(adaptor). 为了提高测序效率，通常采用电泳切胶法提取一定长度范围的 cDNA，再对其进行 PCR 扩增，最终得到 cDNA 文库[38]. 若文库制备遵循 strand-specific 协议[39]，可为后续数据处理阶段的转录组重建及表达水平估计提供转录本的方向信息.

在测序阶段，将制备完成的 cDNA 文库放入测序平台(如 Illumina GA-analyzer、GA IIx 或 HiSeq 2000 等)的通道(lane). 测序仪器先对 cDNA 进行 PCR 扩增，之后开始测序. 测序主要通过检测荧光信号来确定测得的碱基及顺序. 按照是否从 cDNA 两端测序，RNA-Seq 可分为单端(single-end)和双端(paired-end)两种方式. 其中双端测序能提供

更准确的转录本信息，这有助于基因表达水平的估计．测序产生的核苷酸序列被称为“读段”(read/fragment)，其长度范围为35~500碱基．一般来说，测序深度越高，产生的读段数据量也越大．二代测序已能一次产生超过500千兆碱基(gigabases)的读段序列[37]，并以FASTQ格式的文件发布．FASTQ文件包含了读段的全部信息，包括测序仪器的基本信息、碱基排列，以及读段中碱基的质量得分(phred quality score)等．针对测序的质量控制(quality control)，可使用FastQC、RNA-SeQC[40]等软件提供的测试方法来评估读段质量[41]．也可对读段集合进行预处理，即通过Quake[42]、SeqTrim[43]、FASTX[44]、TagDust[45]等方法滤除低质量的读段或对读段进行剪裁．

1.2 转录组重建

通过预处理原始测序读段，可得到一个质量相对较高的读段集合．利用此读段集合恢复转录组结构的过程被称作转录组重建(transcriptome reconstruction)[46]．按照是否参考基因组序列，重建方法可分为基因组引导法(genome-guided)和基因组独立法(genome-independent)两大类[46]．基因组引导法也被称作基于参考的转录组装配(reference-based transcriptome assembly)[47]，即先将样本的所有读段映射至参考基因组序列，然后再装配转录组．基因组独立法也被称为 *de novo* 装配，它无需参考基因组，直接通过读段间的比对来完成转录组装配，它适用于尚无参考基因组序列的物种．相对于基因组引导法，基因组独立法需要较高的测序深度[47]，以保证装配质量．当前，lncRNA研究主要采用基因组引导法进行转录组重建，它包括读段映射/比对与读段装配两步．此外，为了协助下游的lncRNA识别与分析，一般还会对重建得到的转录组进行分类和注释．

1.2.1 读段映射/比对．读段映射/比对(read mapping/alignment)是基因组引导法重建转录组的第一步，即首先将测序读段映射至参考基因组．由于读段测自随机截取的RNA片段所对应的cDNA，因此读段既可能完整取自外显子内部，也可能来自外显子与内含子的剪接区域(splice junction)，分别被称作外显子读段(exonic read)和剪接读段(spliced read)．当前以lncRNA研究为目标的RNA-Seq数据处理流程主要采用能够识别剪接读段的映射方法，可分为两类．第一类采用外显子优先(exon-first)[46]策略，如MapSplice[48]、SpliceMap[49]和Tophat[34]等．以Tophat为例，其内部调用基于Burrows-Wheeler转换(BWT)的Bowtie[50]比对方法，采用外显子优先(exon-first)[46]策略来映射读段．在exon-first算法中，外显子读段优先被映射到基因组上，而后再将剪接读段映射至剪接位置．第二类被称为种子扩展(seed-extend)[46, 51]的映射算法，如GSNAP[52]和QPALMA[53]等．在seed-extend算法中，先为所有读段建立k-mers片段的查找表，然后将k-mers映射到基因组，最后根据查找表将k-mers扩展至完整测序读段．相对于seed-extend，exon-first算法占用的计算资源更少、速度更快，但seed-extend在处理剪接读段时比exon-first更准确[46]．基于以上两类映射算法，相关软件也提供了多种策略来提高映射效率．例如，使用Tophat时可通过-G选项提供现有注释基因(如RefSeq[54]、Ensembl[55]等)的GTF(gene transfer format)文件，则读段映射便可按照基因注释有序进行，即读段优先映射到注释基因的区域，其余不能成功映射的读段再与基因组其他区域进行比对．根据读段映射位置的唯一性与否，映射成功的读段(aligned reads)可分为唯一映射(unique-mapped)和多点映射(multiple-mapped)两种[31]．导致读段多点映射的可能原因包括读段源自基因组的重复序列或同源相似序列．解决方案是，可根据某一映射区域内唯一映射读段的数量来分派多点映射读段，或采用能产生长读段的454测序技术以减少多点映射的读段[31, 56]．映射完成后，即可得到以BAM/SAM格式文件发布的映射读段．这些BAM/SAM文件记录了读段在基因组上的位置信息及每个碱基的映射质量等．常用的BAM/SAM处理软件有SAMTools[57]、BEDTools[58]、IGVTools[59]等．

1.2.2 读段装配．

读段装配是基因组引导法的第二步，即根据映射读段在基因组上的覆盖和剪接信息，通过计算和统计模型预测转录本的结构．常用方法有Cufflinks[60]和Scripture[30]．尽管两者都采用了构建图(graph)的思路，但它们分别基于不同的图模型和优化算法．其中，Cufflinks定义了有向无环图(directed acyclic graph，DAG)，再根据Dilworth理论[61]确定构成图拓扑的最小转录本集合．而Scripture先根据映射读段的连接信息定义了连通图(connectivity graph)，再沿着图拓扑设定滑动窗口，继而进行分段统计测试，最终确定转录本结构．相对于Cufflinks，Scripture能预测到更多的转录本[46]．有研究将已注

释内含子作为评估对象，发现在恢复这些内含子的能力方面，Cufflinks 比 Scripture 具有更高的敏感性和特异性[62]. 也有研究将 Cufflinks 和 Scripture 的装配结果取并集，再从中识别 lncRNA[35-36].

对来自不同通道的测序数据来说，可采用不同的组合策略来重建转录组. 可对各个样本的读段集合(FASTQ 格式)分别映射和装配，最后合并；也可在全部或部分数据集合并后，再进行映射和装配；或在映射完各个样本后，合并映射文件，再进行装配等. 为了节省计算资源及避免合并后由于剪接形式的过于复杂而导致的装配错误，Cufflinks 的设计者在最近发表的文章中建议先单独装配各个样本的读段(第一种策略)，再利用 cuffmerge 合并各样本的预测转录本[37].

1.2.3 转录组分类和注释. 重建得到的转录组不仅会包含已注释的基因，还可能含有新型的转录本、部分转录本、人工转录本、非 Poly(A)+ RNA 分子或污染物片段. 因此，需要对重建后的转录组进行分类和注释. Cuffcompare[37]提供了一种有效的分类和注释方法，即将重建转录组与现有基因注释进行比较，以获取重建转录组的分类，并用类别代码(classcode)加以标示. 例如，“=”代码表示此预测转录本与注释基因的所有内含子完全吻合，但它们在第一外显子(first exon)的起始端或最后外显子(last exon)的末端可能有差别. 然而，这并不影响将“=”类重建转录本判定为已注释转录本. 又如，有些转录本标有“j”类别代码，表明此转录本至少有一个内含子与已注释基因的内含子相同，而其他位置可能不同. 据此可推断此类转录本可能是注释基因的一个新异构体(novel isoform). 转录组类别代码不仅可以为研究者提供预测转录本的分类信息，还可用于 lncRNA 的识别过程.

1.3 lncRNA 识别与分析

1.3.1 滤除低质量转录组. 理想情况下，若测序深度足够，转录组重建便能从原始测序数据中恢复出完整的转录组. 但由于当前测序技术存在诸多问题，如测序深度低、测序偏好[33]和测序错误[34]等，严重影响了转录组重建，导致低质量转录组(主要是部分转录本和人造转录本)的产生[63]. 对于低表达的转录本，重建方法仅能将其部分恢复，而这些部分转录本很可能被误判为 lncRNA，导致假阳性结果. 此外，映射错误也会导致人工转录本的产生，这些原本不存在的转录本也可能被误判为 lncRNA. 其后果是，低质量的转录本会直接影响到下游的数据处理和分析[37, 60]. 针对预测转录本的质量问题，有些转录组重建方法会提供一些解决方案[60]. 例如，Cufflinks 中的 RABT[64]即是利用现有基因注释，通过合成读段(synthetic reads)补全转录本，从而提高重建质量. 然而，在与注释不能交叠的区域，无法利用注释信息，因此难以识别这些位置的低质量转录本. 为此，相关研究提供了不同类型的阈值用以滤除低质量转录本. 例如，Cufflinks 提供了一些过滤方法. 其中，-F 选项依据基因转录本的最高表达量来设置阈值，-I 选项用来设置内含子的最大长度，--min-frags-per-transfrag 选项可设置覆盖转录本的读段的最小数量. 除了利用软件提供的过滤选项之外，也可通过统计学习方法从数据中获得阈值[35]. 有研究发现，装配出的部分转录本与完整转录本呈现出差异性的读段覆盖分布. 因此，可通过学习数据中的完整转录本与部分转录本的读段覆盖(read coverage)来获取最优的覆盖阈值，用以滤除部分转录本和人造转录本[35]. 利用以上方法将低质量转录本滤除之后，便可得到一个质量相对较高的转录本集合.

1.3.2 lncRNA 识别.

lncRNA 识别即是从高质量转录本集合中识别 lncRNA 的过程. Guttman[30]、Cabili[35]、Pauli[36]等都采用了图 3 所示的流程从高质量转录本中识别 lncRNA.

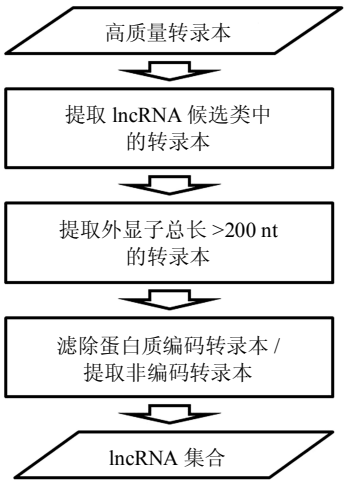

Fig. 3 lncRNA identification

图 3 lncRNA 识别

lncRNA 识别过程分为 3 步：a. 提取 lncRNA 候选类中的转录本；b. 提取外显子总长 > 200 nt 的转录本；c. 滤除蛋白质编码转录本 / 提取非编码转录本. 详细内容参见 **1.3.2** 节.

a．提取 lncRNA 候选类中的转录本可利用前面 **1.2.3** 节得到的预测转录组分类信息．当前 lncRNA 的分类一般是根据它们与相邻蛋白质编码基因的相对位置来划分，如反义 lncRNA(antisense lncRNA)、内含子 lncRNA(intronic lncRNA)、双向 lncRNA (bidirectional lncRNA) 及基因间 lncRNA (intergenic lncRNA)等[17]．为了避免已注释基因对 lncRNA 研究的干扰，Guttman 和 Cabili 等仅提取基因间 RNA(类别代码为“u”)，将其作为候选 lncRNA[30, 35]．也有研究由于在测序阶段采用了 strand-specific 协议，转录本预测准确性得到提高，故将 lncRNA 候选类扩展至其他转录本类别[36]．在此基础上，Pauli 等[36]还分析了反义 lncRNA(类别代码为“o”)等其他类型的 lncRNA．综上，候选类的确定可根据 RNA-Seq 的实验情况合理选择，但一般会包括“u”类转录本．一旦确定 lncRNA 的候选类，即可通过简单的脚本程序从高质量转录本集合中提取相应的转录本．

b．提取外显子总长度大于 200 碱基的转录本[35]．此阈值是由 lncRNA 的定义所决定，本质上是用来区分 lncRNA 与小 ncRNA(如 miRNA 等)[6, 65]．尽管有些已知的小 ncRNA 长度大于 200 碱基，但这并不影响将此阈值用于判别新的 lncRNA[17]．

c．滤除 mRNA/ 提取 ncRNA 是关于如何区分 mRNA 与 ncRNA 的经典问题，方法主要分为以下 4 类：1)通过 ORF 长度判别．对于编码蛋白质的 mRNA 来说，其开放阅读框(ORF)长度一般大于 300 碱基或 100 氨基酸．因此，若 RNA 序列的 ORF 小于 300 碱基，其编码蛋白质的可能性会非常小．例如，早期的 FANTOM 便是通过此 ORF 阈值(300 碱基)来判别 mRNA[66]．相应地，若 RNA 序列的假定 ORF(putative ORF)长度小于 300 碱基，则会被判定为 ncRNA．然而，这种武断的判定方法存在一定问题．例如，有些 lncRNA，如 H19、Xist、Mirg、Gtl2、KcnqOT1 等[67]，由于其假定 ORF 长度大于 300 碱基，因此，在此 ORF 长度标准下它们会被错误划分为 mRNA．类似地，有些 ORF 长度小于此阈值的 mRNA 也会被误判为 ncRNA．2)根据 ORF 保守性，采用比较基因组学的方法进行判别．mRNA 的 ORF 具有保守性，即可编码蛋白质的转录本序列与已注释的蛋白质或蛋白质结构域有同源相似性．因此，可采用 BLASTX[68]、rsCDS[69]、Pfam[70]、SUPERFAMILY[71]等方法，将预测转录本的氨基酸序列放入蛋白质库进行搜索，最后根据比对得到的同源相似性得分来判别该转录本是否可能编码蛋白质．以上直接比对方法的缺点在于它们依赖于现有蛋白质库的准确性[1]．换句话说，若现有蛋白质库错误地包含了非编码 RNA，那么通过比对得到的结果也会存在偏差．此外，有些从 mRNA 演变出的 ncRNA 也会表现出与蛋白质序列类似的同源相似性[72-73]，从而被错判为 mRNA．另外，可采用 CSTminer[69]、CRITICA[74]、PhyloCSF[75]等方法．这些方法的基本假设是氨基酸序列倾向于发生同义碱基变化(synonymous base change)[1]．此类方法通过学习多序列比对，建立蛋白质编码和非编码序列的密码子替换模型，用以判别查询序列是否编码蛋白质．其中，PhyloCSF 不仅能对完整转录本的氨基酸序列进行判别，还能对局部肽链[76]进行预测．3)通过 RNA 二级结构保守性预测．常用的根据二级结构保守性来识别 ncRNA 的方法有 QRNA[77]、RNAz[78]、EvoFOLD[79]等．其缺点在于不能区分同样具有二级结构保守性的 mRNA．4)综合性方法．有研究通过整合以上方法来判别 mRNA 和 ncRNA，主要分为两种．一种采用监督机器学习(supervised machine learning)方法，如 CPC[80]、CONC[81]、incRNA[82]等．此类方法通过学习肽链长度、氨基酸构成、蛋白质同源性、二级结构、蛋白质比对或表达等多种特征，建立分类模型．以 CPC 为例，其分类模型主要基于序列 ORF 长度和蛋白质同源性等特征；另一种综合性方法是将以上方法串联，形成一个过滤流程，用以区分 mRNA 与 ncRNA．例如，有文献[35]整合了 PhyloCSF[75]和 Pfam[70]库搜索，用于识别 lncRNA．串联方法的局限性在于每增加一种方法，对预测的要求越严格，从而产生的预测数量也会越少，因此很可能排除了真实的 ncRNA．

1.3.3 lncRNA 特征分析．

对于预测得到的 lncRNA 集合，研究一般会分析其中 lncRNA 的基本特征．这些特征包括转录本长度、外显子个数、表达水平、可变剪接等．

转录本长度和外显子个数可通过解析 GTF 文件获得．由于 lncRNA 表达属于转录本水平，因此可采用 rSeq[83], Cufflinks[60]、MISO[84]、Alexa-seq[85]、RSEM[86]、IsoformEx[87]等方法估计 lncRNA 的表达．这些方法大多选取 RPKM/FPKM 作为量化指标，即根据转录本长度和映射读段的数量对读段计数进行量化．对于转录本重叠区域的读段来说，由

于其转录本来源难以确定，导致表达水平估计的困难．为此，rSeq、Cufflinks、RSEM 和 MISO 将表达水平设为参数，建立似然函数，再通过最大似然估计(MLE)法预测表达水平．IsoformEx 则采用非负最小二乘法来估计表达水平[87]．Alexa-seq 则通过对唯一映射到某一转录本的读段进行计数，用以量化转录本的表达水平．然而，若转录本没有独特的外显子，Alexa-seq 便不能对读段进行计数，因而失效．对样本间 lncRNA 进行差异表达分析的方法有 Cuffdiff[60]、DESeq[88]、DEGSeq[89]、EdgeR[90]、DEXSeq[91]、GENE-counter[92]等．早期的差异表达分析方法普遍采用泊松分布(Poisson distribution)模型[83, 93]，但泊松分布并不能适应(fit)RNA-Seq 的生物学偏差(biological variability)[50, 90]．为此，EdgeR、Cuffdiff、DESeq 和 DEXSeq 等方法引入负二项分布(negative bionormial distribution)模型，此模型能够很好地适应生物学偏差．其中，Cuffdiff 能够分析转录本水平的差异表达．DEXSeq 可分析外显子水平的差异表达．对于 lncRNA 的可变剪接分析，可采用 SpliceGrapher[94]、Cuffdiff[37]、MATS[95]等方法．

通过以上方法，相关研究对脊椎动物的 lncRNA 进行分析后发现，相对于 mRNA，lncRNA 具有更短的转录本长度、更少的外显子个数及更低的表达水平[35]，同时也发现 lncRNA 具有进化保守性(evolutionary conservation)[35-36]、表达水平呈现出与胚胎发育过程相关[36]等特点．这些分析方法和结果均可为今后关于 lncRNA 的研究提供参考．

Table 1 Bioinformatic methods and packages for lncRNA prediction

表 1 lncRNA 预测相关的生物信息学方法及软件包

处理阶段	基本模型 / 方法	软件包
质量控制	**	FastQC[41], RNA-SeQC[40]
读段剪裁和过滤	**	Quake[42], SeqTrim[43], FASTX[44], TagDust[45]
数据格式转换和处理	**	SAMtools[57], IGVTools[59], BEDtools[58]
转录组装配	基因组引导法	Cufflinks[60], Scripture[30]
转录组分类和注释	与注释基因比较	Cuffcompare[37]
滤除低可靠性转录本	定义表达量阈值	Cufflinks[60], Scripture[30]
	学习读段覆盖	*
mRNA 与 ncRNA 的分类	ORF 长度阈值	*
	比较基因组学	BLASTX[68], rsCDS[69], Pfam[70], SUPERFAMILY[71], CSTminer[69], CRITICA[74], PhyloCSF[75]
	二级结构保守性	QRNA[77], RNAz[78], EvoFOLD[79]
	综合性方法	CPC[80], CONC[81], incRNA[82]
表达水平估计	最大似然估计	rSeq[83], Cufflinks[60], RSEM[86], MISO[84]
	对转录本的确定读段计数	Alexa-seq[85]
	非负最小二乘法	IsoformEx[87]
差异表达分析	负二项分布	Cuffdiff[60], DESeq[88], EdgeR[90], DEXSeq[91], GENE-counter[92]
	泊松分布	DEGSeq[89]
可变剪接分析	**	Cuffdiff[37], SpliceGrapher[94], MATS[95]

* 尚无软件发布; ** 采用多种模型 / 方法.

2 lncRNA 研究的问题和挑战

RNA-Seq 能够捕捉到生命体组织或细胞中的 lncRNA，但要从海量测序数据中挖掘 lncRNA 的基因结构和功能等信息需要上一节所介绍的预测流程及相关的生物信息学方法，如表 1 所示．尽管 lncRNA 研究在现有 RNA-Seq 技术及数据处理分析方法的协助下取得了很大进展，但仍然面临诸多问题和挑战，这里重点在以下三方面进行讨论．

2.1 提高 lncRNA 预测的准确度

lncRNA 预测的准确度不仅依赖于可靠的 RNA-Seq 技术, 也需要其他相关技术的辅助与支持.

随着测序技术的发展，RNA-Seq 的读段质量会越来越高，读段长度也会变得更长. 这些变化都将有助于提高后期读段映射及装配的效率和质量. 相对于 single-end 测序，paired-end 测序能提供更加准确的转录组信息，特别是有利于与 mRNA 发生重叠的 lncRNA 表达水平的估计. 类似地，采用 strand-specific 协议的测序可为后续数据处理提供 RNA 的方向信息，有助于识别反义 lncRNA. 由于二代测序技术采用 PCR 扩增，引入的测序偏好会影响数据处理阶段的转录组重建及表达水平估计. 值得称赞的是，新一代单分子测序技术[96]无需 PCR 扩增，因此不会产生二代测序中出现的测序偏好问题. 但这项新技术尚未成熟，还存在测序读段可靠性不高等问题. 现有的解决方法是根据高质量二代测序数据对其进行校正，再从原始单分子测序读段中截取高质量的片段用于进一步的处理和分析.

除了采用 RNA-Seq，还可使用加帽端测序(CAGE-Seq)和聚腺苷酸端测序(3P-Seq)，以提高 lncRNA 的 5′端和 3′端的预测准确度[17]. 另外，可整合来自表观遗传学的信息，利用染色质签名[97](H3K4me3 和 H3K36me3，K4-K36 域)的 ChIP-Seq 信息[28, 98-100]，以帮助确定 lncRNA 在基因组上的转录起始位置和转录区域[17].

2.2 系统分析 lncRNA 的功能

lncRNA 研究的目的是确定 lncRNA 在生命体活动中起到的功能作用及其与相关疾病之间的关系. 具体来说，即研究 lncRNA 的调控靶标和调控机制. 最近有研究描述了 lncRNA 与 *p53* 基因之间相互作用的机制[28, 101-102]. 然而，大部分 lncRNA 的功能机制以及它们与表现型之间的关系都尚未确定[37, 103].

在对 lncRNA 进行功能分析时，可借助多种技术和信息. 比如可通过 RNA 免疫沉淀及测序(RIP-Seq)来研究 lncRNA 与蛋白质之间的结合关系[27, 29, 104]. 也可采用最新的关联推定法(guilt by association)[17]和基因功能获得 / 丧失研究(gain- or loss-of-function studies)[15, 17]，从而根据基因表达的动态信息来分析 lncRNA 的功能. 其中，关联推定法是通过差异表达分析来识别与蛋白质基因作用通路相关的 lncRNA 家族(families of lncRNAs)[17]. 通过基因功能获得 / 丧失研究，可得到与获得 / 丧失 lncRNA 相关的其他基因的表达图谱[15].

对于以上来源不同的 lncRNA 信息，需要建立有效的基因模型[17]，进而利用系统生物学方法分析 lncRNA 在全局范围内的功能作用. 此外，在辅助疾病诊断方面，研究者可通过分析特定 lncRNA 于相关疾病(如乳腺癌)不同样本中的表达信息，进而建立起 lncRNA 表达与临床信息之间的关系图，以供诊断使用.

2.3 lncRNA 数据库及基因注释

随着 RNA-Seq 等生物技术的广泛应用，生物体内不同细胞和组织中将有更多的 lncRNA 会被发现. 因此, 当前亟需记录有详细 lncRNA 注释信息[105]的数据库，如 NONCODE v3.0[106]、lncRNAdb[107]、RNAcentral[108]、Rfam[109]、HGNC[110]等，以供进一步的信息挖掘及辅助相关疾病的诊断等[111]. 然而，由于现在对 lncRNA 的研究尚处在开始阶段，以上数据库关于 lncRNA 的信息也都不够全面，并且缺乏统一和系统的注释[111]. 为了统一各个数据库的注释，NRDR[112]提出了一种分类法用以划分现有数据库，以帮助研究者快速找到所需的 lncRNA 信息，包括 RNA 家族、信息源、信息内容和可行的搜索机制等. 这一方法值得其他数据库借鉴. 同时，数据库还应提供优良的可视化方法以方便生物学家分析 lncRNA，和对不同实验结果的交叉参考等[107].

3 总结与展望

高通量 RNA-Seq 测序技术正在对转录组研究产生革命性影响，它同时也是对传统微阵列方法的有力补充. RNA-Seq 的优势源于它是一种开放式的生物技术，只需利用有效的生物信息学方法，即可从测序数据中获取转录组信息，包括已知和未知的信息. lncRNA 研究正是在此背景下迅速发展起来的. 本文围绕基于 RNA-Seq 的 lncRNA 预测流程，对各阶段所涉及的代表性的生物信息学方法进行了归纳总结. 在此基础上，我们的研究发现了一些与小鼠红血细胞分化相关的 lncRNA，并采用实时定量 PCR(qRT-PCR)[113]对其中 3 个 lncRNA 进行了生物学验证[114]，它们的基因结构也已上传到 GenBank[115](查询号: JQ173108、JQ173110).

lncRNA 研究尚处于发展阶段，因此还面临着诸多问题和挑战. 针对这些问题和挑战，我们有如下建议. 首先，相关研究在 RNA-Seq 预测的基础上，应采用多层次和多角度的信息，以提高

lncRNA 预测的准确度；其次，在进行 lncRNA 功能研究时，应采用 RIP-Seq、关联推定法和基因功能获得 / 丧失实验等方法，以建立 lncRNA 的调控模型；最后，在 lncRNA 研究过程中应充分利用现有数据库的注释信息.

随着未来生物技术和生物信息学的发展，越来越多的 lncRNA 会被准确识别. 相关领域将在 lncRNA 的序列、结构及功能等各个层面开展更加广泛和系统化的研究. 其中，对 lncRNA 的功能分析将是未来研究的重点，但也是一项复杂的系统性研究. 此外，还可开展相关方向的探索，如研究 lncRNA 与 mRNA 在进化方面的差别、lncRNA 与染色质之间的交互机制、lncRNA 与 DNA 交互时的序列和 / 或结构偏好[17]等，这些都将有助于增进对 lncRNA 的科学认识. 同时，lncRNA 与人类疾病之间的关系也会逐渐明晰. 希望本文能对正在或即将进行 lncRNA 研究的学者提供参考和帮助.

致谢 感谢中国科学院北京基因组研究所张治华博士、昆士兰大学 Tim Bailey 博士的讨论和帮助.

参 考 文 献

[1] Dinger M E, Pang K C, Mercer T R, *et al*. Differentiating protein-coding and noncoding RNA: Challenges and ambiguities. PLoS Comput Biol, 2008, **4**(11): e1000176

[2] Landgraf P, Rusu M, Sheridan R, *et al*. A mammalian microRNA expression atlas based on small RNA library sequencing. Cell, 2007, **129**(7): 1401−1414

[3] Dunoyer P, Schott G, Himber C, *et al*. Small RNA duplexes function as mobile silencing signals between plant cells. Science, 2010, **328**(5980): 912−916

[4] Lakatos L, Csorba T, Pantaleo V, *et al*. Small RNA binding is a common strategy to suppress RNA silencing by several viral suppressors. EMBO J, 2006, **25**(12): 2768−2780

[5] Baker M. Long noncoding RNAs: the search for function. Nat Meth, 2011, **8**(5): 379−383

[6] Kapranov P, Cheng J, Dike S, *et al*. RNA maps reveal new RNA classes and a possible function for pervasive transcription. Science, 2007, **316**(5830): 1484−1488

[7] Bertone P, Stolc V, Royce T E, *et al*. Global identification of human transcribed sequences with genome tiling arrays. Science, 2004, **306**(5705): 2242−2246

[8] Rinn J L, Kertesz M, Wang J K, *et al*. Functional demarcation of active and silent chromatin domains in human HOX loci by noncoding RNAs. Cell, 2007, **129**(7): 1311−1323

[9] Novikova I V, Hennelly S P, Sanbonmatsu K Y. Structural architecture of the human long non-coding RNA, steroid receptor RNA activator. Nucleic Acids Research, 2012: **40**(11): 5034−5051

[10] Atkinson S R, Marguerat S, Bähler J. Exploring long non-coding RNAs through sequencing. Seminars in Cell & Developmental Biology, 2012, **23**(2): 200−205

[11] Nagano T, Fraser P. No-nonsense functions for long noncoding RNAs. Cell, 2011, **145**(2): 178−181

[12] Bernstein E, Allis C D. RNA meets chromatin. Genes & Development, 2005, **19**(14): 1635−1655

[13] Gupta R A, Shah N, Wang K C, *et al*. Long non-coding RNA HOTAIR reprograms chromatin state to promote cancer metastasis. Nature, 2010, **464**(7291): 1071−1076

[14] Spitale R C, Tsai M C, Chang H Y. RNA templating the epigenome: Long noncoding RNAs as molecular scaffolds. Epigenetics, 2011, **6**(5): 539−543

[15] Guttman M, Donaghey J, Carey B W, *et al*. lincRNAs act in the circuitry controlling pluripotency and differentiation. Nature, 2011, **477**: 295−300

[16] Kretz M, Webster D E, Flockhart R J, *et al*. Suppression of progenitor differentiation requires the long noncoding RNA ANCR. Genes & Development, 2012, **26**(4): 338−343

[17] Rinn J L, Chang H Y. Genome regulation by long noncoding RNAs. Annual Review of Biochemistry, 2012, **81**(1): 145−166

[18] Guttman M, Rinn J L. Modular regulatory principles of large non-coding RNAs. Nature, 2012, **482**(7385): 339−346

[19] Geisler S, Lojek L, Khalil A M, *et al*. Decapping of long noncoding RNAs regulates inducible genes. Molecular Cell, 2012, **45** (3): 279−291

[20] Cui Z, Ren S, Lu J, *et al*. The prostate cancer-up-regulated long noncoding RNA PlncRNA-1 modulates apoptosis and proliferation through reciprocal regulation of androgen receptor [J/OL]. Urologic Oncology: Seminars and Original Investigations, 2012 [2012-12-13]. http://download.journals.elsevierhealth.com/pdfs/journals/1078-1439/PIIS1078143911004340.pdf (DOI: 10.1016/j.urolonc.2011.11.030)

[21] Ng S Y, Johnson R, Stanton L W. Human long non-coding RNAs promote pluripotency and neuronal differentiation by association with chromatin modifiers and transcription factors. EMBO J, 2012, **31**(3): 522−533

[22] Rackham O, Shearwood A M J, Mercer T R, *et al*. Long noncoding RNAs are generated from the mitochondrial genome and regulated by nuclear-encoded proteins. RNA, 2011, **17**(12): 2085−2093

[23] Ren S, Peng Z, Mao J H, *et al*. RNA-seq analysis of prostate cancer in the Chinese population identifies recurrent gene fusions, cancer-associated long noncoding RNAs and aberrant alternative splicings. Cell Res, 2012, **22**(5): 806−821

[24] Mitra S A, Mitra A P, Triche T J. A central role for long non-coding RNA in cancer. Frontiers in Genetics, 2012(3): 17

[25] Lin M, Pedrosa E, Shah A, *et al*. RNA-Seq of human neurons derived from iPS cells reveals candidate long non-coding RNAs involved in neurogenesis and neuropsychiatric disorders. PLoS ONE, 2011, **6**(9): e23356

[26] Ellatif S K A, Gutschner T, Diederichs S. Long Noncoding RNA Function and Expression in Cancer Regulatory RNAs//Mallick B, Ghosh Z. Regulatory RNAs. Berlin, Heidelberg: Springer, 2012: 197−226

[27] Zhao J, Ohsumi T K, Kung J T, *et al*. Genome-wide Identification of Polycomb-Associated RNAs by RIP-seq. Mol Cell, 2010, **40**(6):

939-953

[28] Guttman M, Amit I, Garber M, *et al*. Chromatin signature reveals over a thousand highly conserved large non-coding RNAs in mammals. Nature, 2009, **458**(7235): 223-227

[29] Khalil A M, Guttman M, Huarte M, *et al*. Many human large intergenic noncoding RNAs associate with chromatin-modifying complexes and affect gene expression. Proc Natl Acad Sci USA, 2009, **106**(28): 11667-11672

[30] Guttman M, Garber M, Levin J Z, *et al*. Ab initio reconstruction of cell type-specific transcriptomes in mouse reveals the conserved multi-exonic structure of lincRNAs. Nat Biotech, 2010, **28**(5): 503-510

[31] Mortazavi A, Williams B A, McCue K, *et al*. Mapping and quantifying mammalian transcriptomes by RNA-seq. Nat Methods, 2008, **5**(7): 621-628

[32] Schorderet P, Duboule D. Structural and functional differences in the long non-coding RNA hotair in mouse and human. PLoS Genet, 2011, **7**(5): e1002071

[33] Roberts A, Trapnell C, Donaghey J, *et al*. Improving RNA-Seq expression estimates by correcting for fragment bias. Genome Biology, 2011, **12**(3): R22

[34] Trapnell C, Pachter L, Salzberg S L. TopHat: discovering splice junctions with RNA-Seq. Bioinformatics, 2009, **25**(9): 1105-1111

[35] Cabili M N, Trapnell C, Goff L, *et al*. Integrative annotation of human large intergenic noncoding RNAs reveals global properties and specific subclasses. Genes & Development, 2011, **25** (18): 1915-1927

[36] Pauli A, Valen E, Lin M F, *et al*. Systematic identification of long noncoding RNAs expressed during zebrafish embryogenesis. Genome Research, 2012, **22**(3): 577-591

[37] Trapnell C, Roberts A, Goff L, *et al*. Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. Nat Protocols, 2012, **7**(3): 562-578

[38] 王 曦, 汪小我, 王立坤, 等. 新一代高通量 RNA 测序数据的处理与分析. 生物化学与生物物理进展, 2010, **37**(8): 834-846
Wang X, Wang X W, Wang L K, *et al*. Prog Biochem Biophys, 2010, **37**(8): 834-846

[39] Levin J Z, Yassour M, Adiconis X, *et al*. Comprehensive comparative analysis of strand-specific RNA sequencing methods. Nat Meth, 2010, **7**(9): 709-715

[40] DeLuca D S, Levin J Z, Sivachenko A, *et al*. RNA-SeQC: RNA-seq metrics for quality control and process optimization. Bioinformatics, 2012, **28**(11): 1530-1532

[41] Andrews S. FastQC: A quality control tool for high throughput sequence data. http://www.bioinformatics.babraham.ac.uk/projects/fastqc/. 2012. Version 0.10.1

[42] Kelley D, Schatz M, Salzberg S. Quake: quality-aware detection and correction of sequencing errors. Genome Biology, 2010, **11**(11): 1-13

[43] Falgueras J, Lara A, Fernandez-Pozo N, *et al*. SeqTrim: a high-throughput pipeline for pre-processing any type of sequence read. BMC Bioinformatics, 2010, **11**(1): 38

[44] Gordon A. FASTX. http://hannonlab.cshl.edu/fastx_toolkit/. 2012. Version 0.0.13

[45] Lassmann T, Hayashizaki Y, Daub C O. TagDust—a program to eliminate artifacts from next generation sequencing data. Bioinformatics, 2009, **25**(21): 2839-2840

[46] Garber M, Grabherr M G, Guttman M, *et al*. Computational methods for transcriptome annotation and quantification using RNA-seq. Nat Meth, 2011, **8**(6): 469-477

[47] Martin J A, Wang Z. Next-generation transcriptome assembly. Nat Rev Genet, 2011, **12**(10): 671-682

[48] Wang K, Singh D, Zeng Z, *et al*. MapSplice: Accurate mapping of RNA-seq reads for splice junction discovery. Nucl Acid Res, 2010, **38**(18): e178

[49] Au K F, Jiang H, Lin L, *et al*. Detection of splice junctions from paired-end RNA-seq data by SpliceMap. Nucl Acid Res, 2010, **38**(14): 4570-4578

[50] Langmead B, Trapnell C, Pop M, *et al*. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol, 2009, **10**(3): R25

[51] Kent W. BLAT-the BLAST-like alignment tool. Genome Res, 2002, **12**(4): 656-664

[52] Wu T D, Nacu S. Fast and SNP-tolerant detection of complex variants and splicing in short reads. Bioinformatics, 2010, **26** (7): 873-881

[53] De Bona F, Ossowski S, Schneeberger K, *et al*. Optimal spliced alignments of short sequence reads. Bioinformatics, 2008, **24**: i175-i180

[54] Pruitt K D, Tatusova T, Maglott D R. NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. Nucl Acid Res, 2007, **35**(Suppl 1): D61-D65

[55] Hubbard T, Barker D, Birney E, *et al*. The Ensembl genome database project. Nucl Acid Res, 2002, **30**(1): 38-41

[56] Wang Z, Gerstein M, Snyder M. RNA-Seq: a revolutionary tool for transcriptomics. Nature Reviews Genetics, 2009, **10**(1): 57-63

[57] Li H, Handsaker B, Wysoker A, *et al*. The sequence alignment/map format and SAMtools. Bioinformatics, 2009, **25**(16): 2078-2079

[58] Quinlan A R, Hall I M. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics, 2010, **26**(6): 841-842

[59] Robinson J T, Thorvaldsdottir H, Winckler W, *et al*. Integrative genomics viewer. Nat Biotech, 2011, **29**(1): 24-26

[60] Trapnell C, Williams B A, Pertea G, *et al*. Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. Nat Biotech, 2010, **28**(5): 511-515

[61] Dilworth R P. A Decomposition Theorem for Partially Ordered Sets//Gessel I, Rota G C. Classic Papers in Combinatorics. Boston: Birkhäuser Boston, 1987: 139

[62] Robertson G, Schein J, Chiu R, *et al*. De novo assembly and analysis of RNA-seq data. Nat Meth, 2010, **7**(11): 909-912

[63] Kozarewa I, Ning Z, Quail M, *et al*. Amplification-free Illumina sequencing-library preparation facilitates improved mapping and assembly of (G+C)-biased genomes. Nature Methods, 2009, **6**(4): 291-295

[64] Roberts A, Pimentel H, Trapnell C, *et al*. Identification of novel transcripts in annotated genomes using RNA-Seq. Bioinformatics, 2011, **27**(17): 2325-2329

[65] Esteller M. Non-coding RNAs in human disease. Nat Rev Genet,

2011, **12**(12): 861−874

[66] Consortium T F, Carninci P, Kasukawa T, *et al*. The transcriptional landscape of the mammalian genome. Science, 2005, **309**(5740): 1559−1563

[67] Prasanth K V, Spector D L. Eukaryotic regulatory RNAs: an answer to the 'genome complexity' conundrum. Genes & Development, 2007, **21**(1): 11−42

[68] Gish W, States D J. Identification of protein coding regions by database similarity search. Nature Genetics, 1993, **3**(3): 266−272

[69] Furuno M, Kasukawa T, Saito R, *et al*. CDS annotation in full-length cDNA sequence. Genome Research, 2003, **13** (6b): 1478−1487

[70] Finn R D, Tate J, Mistry J, *et al*. The Pfam protein families database. Nucl Acid Res, 2008, **36**(suppl 1): D281−D288

[71] Gough J, Karplus K, Hughey R, *et al*. Assignment of homology to genome sequences using a library of hidden Markov models that represent all proteins of known structure. J Mol Biol, 2001, **313**(4): 903−919

[72] Allen E, Xie Z, Gustafson A M, *et al*. Evolution of microRNA genes by inverted duplication of target gene sequences in *Arabidopsis thaliana*. Nat Genet, 2004, **36**(12): 1282−1290

[73] Duret L, Chureau C, Samain S, *et al*. The xist RNA gene evolved in eutherians by pseudogenization of a protein-coding gene. Science, 2006, **312**(5780): 1653−1655

[74] Bolotin A, Quinquis B, Renault P, *et al*. Complete sequence and comparative genome analysis of the dairy bacterium Streptococcus thermophilus. Nat Biotech, 2004, **22**(12): 1554−1558

[75] Lin M F, Jungreis I, Kellis M. PhyloCSF: a comparative genomics method to distinguish protein coding and non-coding regions. Bioinformatics, 2011, **27**(13): i275−i282

[76] Ingolia Nicholas T, Lareau Liana F, Weissman Jonathan S. Ribosome profiling of mouse embryonic stem cells reveals the complexity and dynamics of mammalian proteomes. Cell, 2011, **147**(4): 789−802

[77] Rivas E, Eddy S. Noncoding RNA gene detection using comparative sequence analysis. BMC Bioinformatics, 2001, **2**(1): 8

[78] Washietl S, Hofacker I L, Stadler P F. Fast and reliable prediction of noncoding RNAs. Proc Natl Acad Sci USA, 2005, **102**(7): 2454−2459

[79] Pedersen J S, Bejerano G, Siepel A, *et al*. Identification and classification of conserved RNA secondary structures in the human genome. PLoS Comput Biol, 2006, **2**(4): e33

[80] Kong L, Zhang Y, Ye Z Q, *et al*. CPC: assess the protein-coding potential of transcripts using sequence features and support vector machine. Nucl Acid Res, 2007, **35**(suppl 2): W345−W349

[81] Liu J, Gough J, Rost B. Distinguishing protein-coding from non-coding RNAs through support vector machines. PLoS Genet, 2006, **2**(4): e29

[82] Lu Z J, Yip K Y, Wang G, *et al*. Prediction and characterization of noncoding RNAs in *C. elegans* by integrating conservation, secondary structure, and high-throughput sequencing and array data. Genome Research, 2011, **21**(2): 276−285

[83] Jiang H, Wong W H. Statistical inferences for isoform expression in RNA-seq. Bioinformatics, 2009, **25**(8): 1026−1032

[84] Li H, Ruan J, Durbin R. Mapping short DNA sequencing reads and calling variants using mapping quality scores. Genome Research, 2008, **18**(11): 1851−1858

[85] Griffith M, Griffith O L, Mwenifumbo J, *et al*. Alternative expression analysis by RNA sequencing. Nat Meth, 2010, **7**(10): 843−847

[86] Li B, Dewey C. RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. BMC Bioinformatics, 2011, **12**(1): 323

[87] Kim H, Bi Y, Pal S, *et al*. IsoformEx: isoform level gene expression estimation using weighted non-negative least squares from mRNA-Seq data. BMC Bioinformatics, 2011, **12**(1): 305

[88] Anders S, Huber W. Differential expression analysis for sequence count data. Genome Biology, 2010, **11**(10): R106

[89] Wang L, Feng Z, Wang X, *et al*. DEGseq: an R package for identifying differentially expressed genes from RNA-seq data. Bioinformatics, 2010, **26**(1): 136−138

[90] Robinson M D, McCarthy D J, Smyth G K. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics, 2010, **26**(1): 139−140

[91] Anders S, Huber W. Differential expression analysis for sequence count data. Genome Biology, 2010, **11**(10): R106

[92] Cumbie J S, Kimbrel J A, Di Y, *et al*. GENE-counter: A computational pipeline for the analysis of RNA-Seq data for gene expression differences. PLoS ONE, 2011, **6**(10): e25279

[93] Marioni J C, Mason C E, Mane S M, *et al*. RNA-seq: an assessment of technical reproducibility and comparison with gene expression arrays. Genome Res, 2008, **18**: 1509−1517

[94] Rogers M, Thomas J, Reddy A, *et al*. SpliceGrapher: detecting patterns of alternative splicing from RNA-Seq data in the context of gene models and EST data. Genome Biology, 2012, **13**(1): R4

[95] Shen S, Won Park J, Huang J, *et al*. MATS: a Bayesian framework for flexible detection of differential alternative splicing from RNA-Seq data. Nucl Acid Res, 2012, **40**(8): e61

[96] Sam L T, Lipson D, Raz T, *et al*. A comparison of single molecule and amplification based sequencing of cancer transcriptomes. PLoS ONE, 2011, **6**(3): e17305

[97] Ulitsky I, Shkumatava A, Jan Calvin H, *et al*. Conserved function of lincRNAs in vertebrate embryonic development despite rapid sequence evolution. Cell, 2011, **147**(7): 1537−1550

[98] Johnson D S, Mortazavi A, Myers R M, *et al*. Genome-wide mapping of *in vivo* potein-DNA interactions. Science, 2007, **316**(5830): 1497−1502

[99] Mikkelsen T S. Genome-wide maps of chromatin state in pluripotent and linearge-committed cells. Nature, 2007, **448**(7153): 553− 560

[100]Marson A, Levine S S, Cole M F, *et al*. Connecting microRNA genes to the core transcriptional regulatory circuitry of embryonic stem cells. Cell, 2008, **134**(3): 521−533

[101]Hung T, Wang Y, Lin M F, *et al*. Extensive and coordinated transcription of noncoding RNAs within cell-cycle promoters. Nat Genet, 2011, **43**(7): 621−629

[102]Huarte M, Guttman M, Feldser D, *et al*. A large intergenic noncoding RNA induced by p53 mediates global gene repression in the p53 response. Cell, 2010, **142**(3): 409−419

[103]Wang K C, Chang H Y. Molecular mechanisms of long noncoding

RNAs. Mol Cell, 2011, **43**(6): 904-914
[104]Gerber A P, Herschlag D, Brown P O. Extensive association of functionally and cytotopically related mRNAs with Puf family RNA-binding proteins in yeast. PLoS Biol, 2004, **2**(3): e79
[105]Liao Q, Liu C, Yuan X, *et al*. Large-scale prediction of long non-coding RNA functions in a coding-non-coding gene co-expression network. Nucl Acid Res, 2011, **39**(9): 3864-3878
[106]Bu D, Yu K, Sun S, *et al*. NONCODE v3.0: integrative annotation of long noncoding RNAs. Nucl Acid Res, 2012, **40**(D1): D210-D215
[107]Amaral P P, Clark M B, Gascoigne D K, *et al*. lncRNAdb: a reference database for long noncoding RNAs. Nucl Acid Res, 2011, **39**(Suppl 1): D146-D151
[108]Bateman A, Agrawal S, Birney E, *et al*. RNAcentral: A vision for an international database of RNA sequences. RNA, 2011, **17**(11): 1941-1946
[109]Gardner P P, Daub J, Tate J G, *et al*. Rfam: updates to the RNA families database. Nucl Acid Res, 2009, **37**(Suppl 1): D136-D140
[110]Wright M W, Bruford E A. Naming 'junk': human non-protein coding RNA (ncRNA) gene nomenclature. Human Genomics, 2011, **5**(2): 90-98
[111]Reis E M, Verjovski-Almeida S. Perspectives of long noncoding RNAs in cancer diagnostics and therapy. Frontiers in Genetics, 2012, **3**(464): 1071-1076
[112]Paschoal A R, Maracaja-Coutinho V, Setubal J C, *et al*. Non-coding transcription characterization and annotation: A guide and web resource for non-coding RNA databases. RNA Biology, 2012, **9**(3): 274-282
[113]Livak K J, Schmittgen T D. Analysis of relative gene expression data using real-time quantitative PCR and the $2^{-\Delta\Delta CT}$ method. Methods, 2001, **25**(4): 402-408
[114]Tallack M R, Magor G W, Dartigues B, *et al*. Novel roles for KLF1 in erythropoiesis revealed by mRNA-seq[J/OL]. Genome Research, 2012 [2012-12-13]. http://genome.cshlp.org/content/early/2012/07/26/gr.135707.111.full.pdf+html(DOI:10.1101/gr.135707.111)
[115]Benson D A, Karsch-Mizrachi I, Lipman D J, *et al*. GenBank: update. Nucl Acid Res, 2004, **32**(suppl 1): D23-D26

Prediction of Long Non-Coding RNAs Based on RNA-Seq*

SUN Lei[1,2], ZHANG Lin[1], LIU Hui[1]**

([1] *School of Information and Electrical Engineering, China University of Mining and Technology, Xuzhou* 221008, China;
[2] *Center for Computational Biology, Beijing Institute of Genomics, Chinese Academy of Sciences, Beijing* 100029, China)

Abstract With the development of the new generation of biotechnology and bioinformatics, studies on the transcriptome of eukaryotes have detected a number of long non-coding RNAs (lncRNAs) and the lncRNAs may play key functional roles in gene expression and regulation. Currently, high-throughput RNA-Seq has become the main technique for lncRNA study and several bioinformatic methods have been used to process and analyze the sequencing data for exploring lncRNAs' information including sequence, structure, expression, function and so on. This paper represents a pipeline for the lncRNA prediction based on RNA-Seq, and the relevant bioinformatic methods are reviewed comprehensively. We also discussed several challenges and future works related to the lncRNA study.

Key words long non-coding RNA, RNA-Seq, lncRNA prediction, bioinformatics
DOI: 10.3724/SP.J.1206.2012.00287

* This work was supported by grants from China Postdoctoral Science Foundation (2012M511335, 2012M511336), The Central Special Fund for Operating Expenses of College Basic Research (2010QNA47, 2010QNA50) and Fok Ying-Tung Education Foundation for Young Teachers (121066).
**Corresponding author.
Tel: 86-15262048535, E-mail: lhcumt@hotmail.com
Received: June 12, 2012 Accepted: August 13, 2012