文章编号：1004-0374(2010)07-0598-10

动物 snmRNA 的系统识别与鉴定

邵 鹏，屈良鹄*
（中山大学基因工程教育部重点实验室，
有害生物控制与资源利用国家重点实验室， 广州510275）

摘 要：小分子非编码RNA（snmRNA）在调控基因的转录和转录后加工、细胞分化和个体发育、遗传和表观遗传等几乎所有的重要生命活动中发挥关键作用。建立和发展snmRNA研究技术，系统地发现和注释基因组中的snmRNA基因并阐明其生物学意义是当前RNA组学的首要任务。围绕snmRNA的系统识别与鉴定等问题，该文对近年来采用实验技术和计算机预测方法发掘snmRNA所取得的主要研究成果进行综述。

关键词：小分子非编码RNA；RNA组学；大规模测序
中图分类号：Q522；Q752 **文献标识码**：A

Systematic identification of animal snmRNAs

SHAO Peng, QU Liang-hu*
(Key Laboratory of Gene Engineering of the Ministry of Education, State Key Laboratory for Biocontrol and Resource Utilization, Sun Yat-Sen University, Guangzhou 510275, China)

Abstract: Small non-messenger RNAs (snmRNAs) play important roles in the regulation of gene expression at the transcriptional and post-transcriptional level, cell differentiation, development, inheritance and epigenetic inheritance. Developing new approaches to identify the increasing number of novel snmRNAs is critical to our ability to characterize the molecular details of these RNAs and understand their biological function. In this review, we summarize the recent advances in the methodology used to discover animal snmRNAs.

Key words: small non-messenger RNAs; RNomics; high-throughput sequencing

自人类基因组序列测定完成之后，科学家们惊奇地发现，在庞大的基因组中，原来绝大部分的区域(约占95%)并不编码蛋白质。一些DNA片段转录为信使RNA (mRNA)前体后，可以被加工成其他类型的RNA，比如调控性的RNA，从而调节其他基因的合成与表达。这类不编码蛋白质的具有功能性的RNA分子被称为非信使RNA (non-messenger RNA，nmRNA)或者是非编码RNA (non-coding RNA，ncRNA)[1]，而编码这类型RNA的基因称为非编码RNA基因。与其他大片段RNA分子相比，大部分已知的nmRNA长度在20~500个核苷酸(nt)，这类nmRNA称为小分子非编码RNA (small non-messenger RNA，snmRNA)，或者是sRNA (small RNA)、tncRNA(tiny non-coding RNA)等。snmRNA主要包括微RNA (microRNA/miRNA)、小干扰RNA (small interfering RNA,siRNA)、与piwi蛋白相关的RNA (piwi-associated RNA，piRNA)、tRNA、核小RNA (small nuclear RNA，snRNA)和核仁小分子RNA (small nucleolar RNA，snoRNA)。研究表明，snmRNA在调控基因的转录和转录后加工、细胞分化和个体发育、遗传和表观遗传等几乎所有的

收稿日期：2010-03-23；**修回日期**：2010-05-19
基金项目：国家重点基础研究发展计划(“973”项目)(2005CB724600)； 国家自然科学基金重点项目(30830066)；中山大学青年教师培养项目(09lgpy37)
***通讯作者**：E-mail:lssqlh@mail.sysu.edu.cn

重要生命活动中发挥关键作用。建立和发展相应的研究技术，系统地发现和注释基因组中的ncRNA基因并阐明其生物学意义是当前RNA组学的首要任务。围绕snmRNA的系统识别与鉴定方法，本文拟对近年来实验发掘和计算机分析snmRNA方面所取得的研究进展进行总结。

1　克隆测序技术鉴定snmRNA

采用克隆测序的方法系统地鉴定snmRNA的研究可追溯到1996年Kiss-Laszlo等[2]通过在小RNA的5’和3’末端加接头构建cDNA文库的方法来鉴定人snoRNA的工作。到了2001年，Huttenhofer等[3]采用分段割胶获取50~500 nt的小RNA，构建了小鼠脑snmRNA的cDNA文库。通过序列测定，意外地发现细胞中存在着一个隐蔽的snoRNA世界，打开了实验RNA组学(experimental RNomics)[4]的大门。可以说，目前很多snmRNA的克隆方法都源自早期克隆snoRNA的策略。随着snmRNA研究的推动，实验RNA组学方法的发展日新月异，特别是第二代或下一代DNA测序(next generation sequencing)技术[5]，又称为深度测序(deep-sequencing)的广泛应用，极大地促进本领域的快速发展。

1.1　常规cDNA克隆测序技术鉴定snmRNA

snmRNA长度短，缺乏诸如mRNA多聚腺苷酸，即poly(A)尾结构。在构建snmRNA的cDNA文库前需要将预先合成的RNA接头连接到经过富集处理的RNA样品上，或利用poly(A)聚合酶在其3’末端加上寡聚核苷酸尾，随后采用与接头相对应的特异引物或者是寡聚核苷酸poly(dT)作为引物进行反转录反应获得cDNA(图1)。经过以上的处理后，就可以

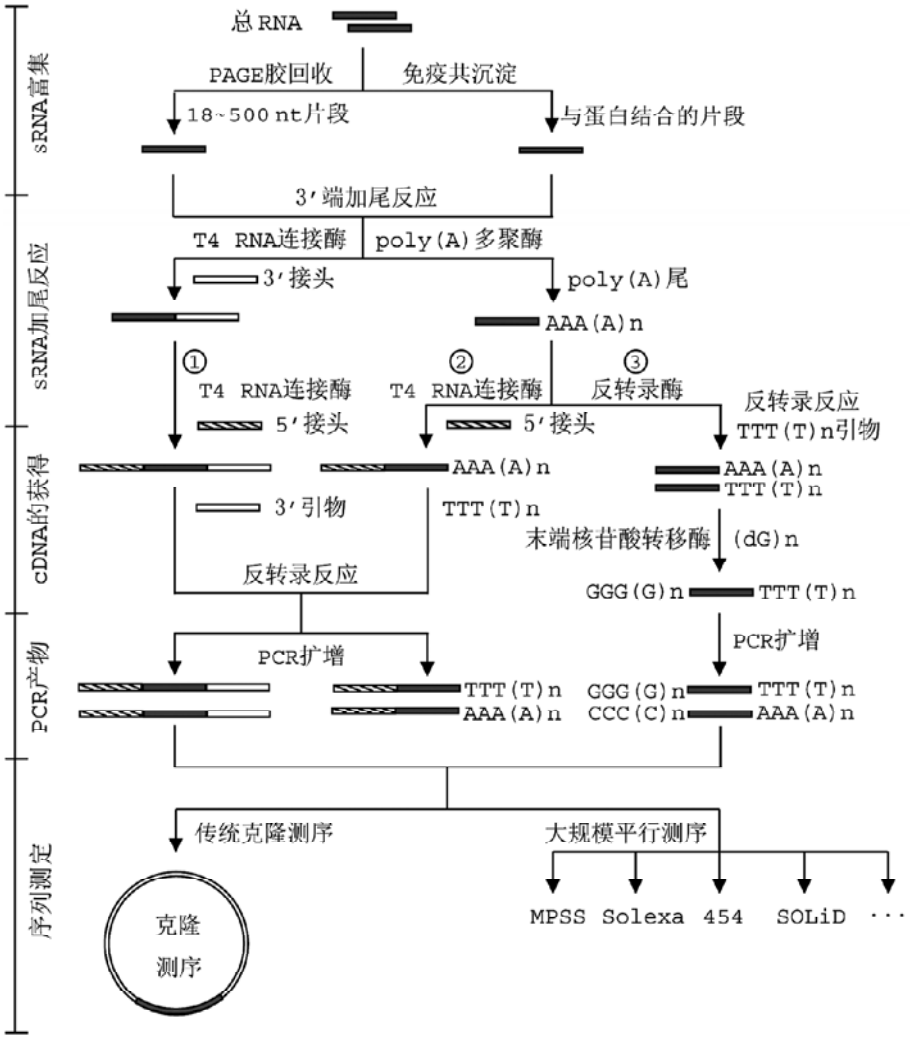

图1　采用不同cDNA文库构建策略测序鉴定snmRNA

采用图中①②③三种方法都能克隆不具有帽子结构的snmRNA；对于具有5’端帽子结构的小RNA，可采用方法③进行

方便地如同操作mRNA一样来克隆snmRNA，最后进行序列测定。根据所要克隆的snmRNA性质不同，一般可以采用两种方法来富集snmRNA：(1)对于具有一定长度范围的snmRNA，通常采用变性凝胶电泳(如变性聚丙烯酰胺凝胶电泳)的方法，切取所需片段大小的RNA组分，采用盐溶液从凝胶中回收RNA。Huttenhofer等[3]就是采用在割胶回收的RNA的3'末端直接加上poly(C)尾的方法，在小鼠组织中鉴定了201个新的snmRNA。由于poly(A)聚合酶在UTP、CTP等为底物时的掺入效率低，我们实验室改用ATP做底物，从而使其加尾效率极大地提高。该方法已应用于克隆snoRNA[6]及其他新类型的snmRNA，如miRNA等[7,8](图1-③)。(2)目前已知的snmRNA大部分都与特异性结合蛋白形成复合体，可以通过免疫共沉淀的方法富集snmRNA(图1)。对比前一种的方法，这种方法的突出优点就是所获得的snmRNA具有很高的特异性，但对操作技术要求较高，步骤繁琐。Kiss等[9]使用box H/ACA RNA蛋白复合体(RNP)的抗体anti-GAR1，通过免疫共沉淀的方法成功富集HeLa细胞中的box H/ACA snoRNA。部分研究也采用免疫共沉淀miRNP复合物和Argonaute蛋白复合体的方法来发现miRNA[10]和内源siRNA(endogenous short interfering RNAs，endo-siRNAs)[11]。目前这种方法也广泛用于富集piRNA。

尽管以上两种方法可以富集不同种类的snmRNA，但在实际操作中，往往还需要根据所要分离的RNA序列、结构特征或者是其亚细胞分布特点，同时采用多种不同的富集策略。比如，snoRNA和tRNA等基因序列在长度上比较接近，采用割胶回收snoRNA的方法并不能避免tRNA等其他小RNA的污染。利用snoRNA定位于核仁，而其他tRNA等富集于细胞质的特点，可以通过纯化细胞核仁的方法初步富集snoRNA[6]。本实验室在此基础上还建立了“锚定引物法”来构建特定类型snoRNA的cDNA文库[12]。该方法设计原理主要是基于snoRNA序列在3'末端具有保守的CUGA(box C/D)或ACA(box H/ACA)的元件，采用具有锚定CUGA或ACA结构的poly(dT)作为引物，对加poly(A)尾的RNA进行反转录反应，高效获得snoRNA的cDNA，而tRNA等其他小RNA没有此保守元件而无法得到cDNA产物。进行PCR反应后，根据不同类型的snoRNA长度大小切取一定长度范围的PCR产物，从而可以构建高度富集的snoRNA的cDNA文库。该方法目前已成功地应用于鸡等其他物种的snoRNA研究[13]。

值得一提的是，不同的snmRNA往往具有表达特异性或者阶段性，所以在富集snmRNA时，样本来源也是一个需要考虑的重要因素，比如最近发现piRNA，都是来源于动物的生殖器官和生殖细胞系，包括卵母细胞和精子细胞。Northern杂交进一步验证了这些piRNA在生殖细胞/器官中特异性表达。

cDNA克隆、测序分析的方法易于操作，而且即使在缺乏基因组信息和目标基因功能未知的情况下都可以使用，所以直接克隆测序的方法已经广泛应用于在秀丽隐杆线虫、果蝇、斑马鱼、爪蟾、鸡、哺乳动物等不同模式生物中发现各种新snmRNA类型。

1.2 第二代测序技术鉴定snmRNA

由于不同RNA在细胞中丰度不一，而传统的cDNA克隆技术倾向于克隆丰度较高的组分，因此对于表达水平较低的snmRNA，采用传统的直接克隆测序方法就显示出其局限性。近年来，随着大规模平行DNA测序平台的广泛应用，实现了对细胞中各种RNA类型进行高通量序列测定，大大提高了对低丰度snmRNA的检测灵敏度。第二代测序技术主要包括大规模平行测序(massively parallel signature sequencing，MPSS)、焦磷酸测序(pyrose-quencing，454测序)、合成-测序(sequencing-by-synthesis，Solexa)和SOLiD测序等。以上各种“深度测序”技术可以很好地解决传统cDNA克隆测序技术难以检出表达丰度很低的RNA种类的不足。

MPSS技术可以获得上百万计的小RNA的cDNA克隆子序列，主要用于定量评估mRNA表达水平，即测定mRNA所对应的cDNA一端的一段长度在16~20 bp的标签序列，测序序列在样本中出现的频率经过数学标准化处理后可代表该标签序列相应的基因表达水平。Lu等[14]在2005年首先使用该方法测定了模式植物拟南芥(*Arabidopsis thaliana*)种子和开花组织200多万条小分子RNA序列，发现了约7.5万条miRNA序列。

由于MPSS方法测定的序列长度较短，罗氏公司(Roche)推出首个商业用途的第二代测序平台——454测序系统。该系统基于焦磷酸测序方法，可以同时测定40万条长度在250 nt的序列[15]，目前已经广泛地应用在鉴定各类snmRNA领域(表1)，比如，

Ruby等[16]采用该技术测定了秀丽隐杆线虫(*Caenorhabditis elegans*)大约40万条小RNA，鉴定了18个miRNA新基因和5 000多条其他新类型的RNA。Berezikov等[17]利用该技术发现人和黑猩猩脑miRNA的多样性。454平台拥有读长较长的主要优点，因此非常适合用于检测较长类型的snmRNA，如长度在25~30 nt的piRNA，甚至是更长的snoRNA。例如，在2006年，多个实验室采用该测序平台，在人、小鼠[18]、大鼠[19]等哺乳动物的生殖细胞和组织中鉴定了大量的piRNA。454测序系统已经升级到GS FLX Titanium系统，平均读长提升到400 bp，在测序通量(多于100万条序列)、准确性等方面也有较大的提升。

Illumina基因组分析系统,即早前所讲的Solexa测序，是基于大规模平行测序的、采用边合成边测序(sequencing by synthesis,SBS)策略的高通量、高精确性测序平台。尽管有效读长只有26~50 bp，但可获得比454测序更多的数据量(一般为1 Gb以上)，可以同时测定3 000万条长度在50 nt的序列[36]。因此，采用Solexa测序测定长度在25 nt以下的snmRNA类型，比454系统更具优势。目前使用该测序系统来发现新snmRNA的研究多集中在miRNA和siRNA这两类小RNA(表2)。

表1　采用454测序技术鉴定snmRNA的研究

物种	样本来源	小RNA类型	参考文献
秀丽隐杆线虫	不同发育阶段	21U-RNA、内源siRNA	[16]
四种线虫		miRNA、21U RNA	[20]
真涡虫	未分化细胞，成体	miRNA、piRNA	[21]
海绵，海葵		miRNA、piRNA	[22]
果蝇		miRNA	[23]
埃及斑蚊		miRNA	[24]
斑马鱼	卵巢和睾丸	piRNA	[25,26]
斑马鱼	胚胎和成体组织	miRNA	[27]
鸡	鸡胚纤维原细胞	miRNA	[28]
鸡	胚肝	miRNA	[29]
大鼠	睾丸	piRNA	[19]
小鼠	正常与敲除Dicer酶的胚胎干细胞	miRNA	[30]
小鼠	胚胎干细胞	内源shRNA、siRNA和其他小RNA	[31]
小鼠	卵母细胞	假基因来源的siRNA	[32]
小鼠	卵母细胞	内源siRNA	[33]
人、黑猩猩	脑	miRNA	[17]
人、小鼠、大鼠	睾丸	piRNA	[18]
人、鸡、果蝇		转录启始RNA (tiRNA)	[34]
人	血浆	循环miRNA	[35]

另一个测序平台是应用生物系统(Applied Biosystems)公司的SOLiD系统。尽管采用SOLiD测序技术研究小RNA转录本的工作起步得稍微晚些，其有效读长与Illumina基因组分析系统相同，但由于可以提供更多的DNA测序数据量(2~4 Gb)，所以也开始得到广泛应用，比如Goff等[50]应用该技术来研究人类胚胎干细胞和神经元前体中的miRNA。

由测序得到的数据通常需要生物信息学分析，尤其是应用第二代测序技术所获得的海量数据。一般来说，分析内容包括将测序数据与基因组序列进行序列比对分析，按照基因组注释信息，如编码序列、ncRNA、基因组重复序列等，将测序数据进行筛选与分类，然后采用合适的计算机程序进行结构预测、基因组织形式和分布特征分析。以下主要针对snmRNA的计算机预测分析方法进行介绍如下。

2　计算机程序预测

与实验RNA组学相对应，采用计算机方法寻找包括专利调控元件在内的结构单元和序列元件，识别snmRNA的各种生物信息学技术称为“计算RNA组学”(computational RNomics)[4]。它是获得新类型snmRNA另外一种有效的途径。大部分的ncRNA基因的计算机算法都基于二级结构特征、热力学性

表2 应用Solexa测序平台鉴定snmRNA的研究

物种	样本来源	小RNA类型	参考文献
海绵、海葵		miRNA、piRNA	[22]
海鞘	未受精卵、分裂期、幼年期胚胎与成体	来源于miRNA前体的小RNA (moRs)	[37]
秀丽隐杆线虫	单细胞期胚胎	miRNA、26G-RNA	[38]
果蝇	S2细胞、卵巢	内源siRNA	[39]
果蝇	头、S2细胞等体细胞	转座子和mRNA来源的内源siRNA	[40]
果蝇	Ago3突变体	体细胞来源piRNA	[41]
爪蟾	卵母细胞、肝、皮肤	miRNA、piRNA	[42]
鸡	禽流感病毒感染肺和气管	miRNA	[43]
鸡	5、7、9胚龄胚胎	miRNA	[44]
鸡	3~5 d胚内外胚层	miRNA	[45]
小鼠	卵母细胞	假基因来源的siRNA	[32]
人	胚胎干细胞	miRNA	[46]
人	胚胎干细胞	miRNA	[47]
人	急性淋巴白血病骨髓	miRNA	[48]
哺乳动物	血清、血浆	miRNA	[49]

质、保守功能序列或者结构元件，或者根据系统发育(phylogeny)保守性，通过比较基因组学分析来预测。

2.1 从头预测

大部分ncRNA从头预测(*De novo*)是基于二级结构特征开发的，如Rivas 和Eddy[51]采用比较序列分析算法开发QRNA来寻找新的结构性RNA基因，成功地预测大肠杆菌和酵母等比较基因组序列中的候选ncRNA基因，成为第一个实现在基因组水平预测ncRNA的计算机程序。Pedersen等[52]基于比较基因组学方法开发了可以检测具有高度保守的功能性RNA程序Evofold，在人类基因组中发现了数以万计的候选RNA结构，其中包括169个miRNA候选分子。由于snmRNA及其他ncRNA都缺乏蛋白质编码基因所具有的启动子、终止密码子、开放读码框、剪接信号、多聚腺苷酸等具有可供识别的信号，因此采用计算机从头预测具有结构特征的RNA基因面临很多困难，其最大的问题就是假阳性率高，特别是大的单基因组序列的ncRNA从头预测算法缺乏完整的理论和特定的罚分模型，所以在应用中受到一定的限制[53]。

2.2 同源搜索策略

在基因组水平上，采用计算机程序能可靠地预测ncRNA的方法，目前仅限于搜索已知RNA的同源基因,即同源搜索策略(homology-based search)。同源搜索算法基于比对(alignment-based)的方法,这对预测序列上具有物种间保守性的snmRNA特别有效。例如Weber[54]利用同源搜索方法来预测直系同源或者是旁系同源的miRNA基因，分别在人类和小鼠基因组中鉴定了35个和45个miRNA候选基因。事实上，许多同源的miRNA基因在生物进化过程中会发生突变，但其前体的二级结构仍然保守。预测该类型的snmRNA往往需要同时考虑序列与结构的保守性，如Wang等[55]基于序列和结构对比原理开发更为灵敏的程序MiRalign,在甘比亚疟蚊(*Anopheles gambiae*)基因组中检测到59个新的miRNA基因。

2.3 预测特定类型snmRNA的程序

到目前为止，仍缺乏一个能在基因组水平上准确且通用地鉴定各类RNA基因的预测算法。绝大多数snmRNA预测程序主要基于snmRNA的序列保守元件、二级结构或者折叠自由能等特征来预测特定的snmRNA类型。例如，snoRNA的预测主要根据其一级序列的保守元件、典型二级结构和与靶基因反义互补的序列等特征；而miRNA的前体结构比其他任一随机获得的RNA序列结构具有更低的自由能，可以根据茎环结构特征，结合最低自由能(MFE)原理将miRNA前体与其他结构RNA加以区分[56]。表3中列举了预测tRNA、snoRNA和miRNA的部分主要工具。其中，tRNAscan-SE是较为成功的一个。它用来识别基因组中的tRNA及相关序列，在人类基因组检测tRNA时其成功率达到99.5%左右，并且实现零假阳性率。我们于2006年开发了能在哺乳动物基因组中预测包括“孤儿”snoRNA在内的snoRNA预测软件包snoSeeker[57]，该程序在预测人

类box C/D snoRNA时灵敏度可达到90%以上。另外一类算法是利用RNA的功能域(motif)或基于其在基因组上的轮廓特征(profile-based)来预测所有的结构型RNA，比如ERPIN[58]可以预测包括miRNA[59]等多种RNA基因。

最近，许多研究都倾向于使用机器学习算法(machine-learningalgorithms)包括支持向量机(support vector machines，SVM)、隐马尔科夫模型(hidden Markov models，HMM)、神经网络算法来预测snmRNA。特别是SVM被广泛应用于获取miRNA的

表3　预测特定类型snmRNA的部分程序

程序/软件	snmRNA类型	使用物种基因组	操作系统	参考文献
tRNAscan-SE	tRNA、tRNA来源的重复元件、tRNA假基因	真核和原核生物	UNIX(本地安装)、网络服务	[60]
snoScan	box C/D snoRNA	酵母、现支持哺乳动物和真细菌	网络服务	[61]
snoGPS	box H/ACA snoRNA	酵母、现支持人和真细菌	UNIX(本地安装)、网络服务	[62]
snoSeeker	snoRNA和“孤儿”snoRNA、snoRT基因	人和其他脊椎动物	Windows，Linux(本地安装)、网络服务	[57]
SnoReport	snoRNA和“孤儿”snoRNA	秀丽隐杆线虫、人、果蝇、利什曼原虫	Linux(本地安装)	[63]
MiRscan	miRNA	秀丽隐杆线虫、脊椎动物	网络服务	[64,65]
MiRAlign	miRNA	现支持动、植物	网络服务	[55]
miRFinder	miRNA	脊椎动物、果蝇	Windows，Linux	[66]
ProMiR/ProMiR II	miRNA	人和其他脊椎动物	网络服务	[67,68]
RNAmicro	miRNA	哺乳动物、尾索动物、线虫	Linux(本地安装)、网络服务	[69]
microPred	miRNA	人类	Linux(本地安装)，网络服务	[70]

各种典型特征，如RNAmicro[69]和miPred[71]等程序。新的预测算法的开发应用大大地推动了snmRNA研究的发展，特别是对于发掘具有物种特异性或表达特异性的等采用常规的实验方法难以检测的snmRNA基因。

2.4　大规模小RNA测序数据分析方法

尽管第二代测序技术可以在一定程度上解决表达丰度低，或者是只在某些特殊组织或分化、发育状态下表达的snmRNA难以克隆的问题，但“深度测序”对采用生物信息学方法处理海量数据方面也提出了更高的要求，如怎么有效地分析和管理这些海量数据，如何在这些数据中将不同类型的snmRNA加以区分等。目前NCBI的GEO数据库中已经积累了包括454、Solexa、SOLid等测序平台来源的小RNA海量测序数据。如何有效处理和发掘小RNA“深度测序”数据是当前迫切解决的问题。

目前已开发出来的用于分析小RNA“深度测序”数据的程序主要集中在以下两类功能：

首先，将小RNA测序数据快速地定位(mapping)到基因组序列。大部分测序数据能否与基因组匹配，是判断该测序实验是否成功，也是迈入分析测序数据大门最重要的第一步。由于传统的序列比对软件，诸如BLAST[72]和BLAT[73]不能有效地处理新一代测序技术所产生的如此庞大的数据，众多针对小RNA测序对比的新算法也应运而生。表4列出几款具有代表性的分析程序。其中Bowtie应用了新的索引策略(burrows-wheeler索引)，大大提高了内存效率和运算速度，并且支持多个操作系统，从而得到了广泛引用。

其次，可视化处理小RNA数据。处理新一代测序技术产生的海量数据过程中产生的另一个重要问题就是数据的可读性差。目前，研究人员已开发出各类可视化程序来解决该问题。EagleView[78]是第一款为下一代测序技术而设计的可视化工具。随后开发的多种软件功能比较类似，主要用于基因组的组装和序列多态性的检测，但在运算处理、内存效

表4 处理下一代测序数据基因组比对程序

程序	基本索引策略	特点	计算机操作平台	参考文献
MAQ	哈希表算法	内存占用最少，但速度最慢	32位或64位Linux	[74]
SOAP	种子、哈希查找表算法	速度比MAQ快，但内存占用偏多	64位Linux/Unix，Mac	[75]
Bowtie	Burrows-Wheeler转化(BWT)索引和全文微空间(FM)索引	速度最快，占用内存少	Windows，Mac OS X，Linux和Solaris	[76]
SOAP2	Burrows-Wheeler 转化(BWT)压缩索引	速度比SOAP大大提高，占用内存更少	64位Linux	[77]

率和跨平台使用等方面有较大的改进，如MaqView[79]、Tablet[80]，并且开始出现基于网页浏览器形式的可视化在线分析程序，如LookSeq[81]。最近，我们构建专门用于整理、注释高通量小RNA测序数据、具有可视化界面的综合性分析数据库deepBase[82]。该平台不仅展示了7个模式生物基因组中复杂的小RNA转录组图谱，还可以用于发掘新的非编码RNA甚至是新类的非编码RNA。

遗憾的是，目前专门开发用于从高通量小RNA测序数据中鉴定特定类型snmRNA的算法或程序还比较少，Rajewsky实验室开发的miRDeep[83]是其中一款用来预测miRNA的程序。miRdeep采用概率模型评估符合miRNA加工特征的RNA测序转录本，比如考虑Dicer酶的加工，来源于茎环结构成熟miRNA区域的测序丰度比其他区域更多等，在鉴定已知线虫miRNA时获得较高的灵敏度(89%)，但鉴定人类已知miRNA的灵敏度只有72%。最近，Hackenberg等[84]开发的基于网页服务器的分析工具miRanalyzer，可分析通过Solexa或454测序平台获得的不同文库来源的miRNA数据。

除了miRNA、snoRNA和tRNA等计算机预测算法外，目前对于内源siRNA和piRNA等新类型snmRNA的计算机预测则主要依赖于序列碱基组成特点、片段长度、基因组分布和是否由双链结构加工而来等特征进行分析，缺乏成熟的算法。动物内源siRNA长度集中在21 nt左右，基因组正义与反义链均有分布。小鼠等哺乳动物中约30%的内源siRNA来源于反转座子、转座元件，也有一些是由与特定转录本形成互补配对的假基因加工而来[32]。果蝇的内源siRNA则绝大部分起源于转座子或mRNA[40]，有些来源于异染色质序列、基因间隔区，少部分来源于具有反向重复的发夹结构RNA(hairpin RNAs，hpRNAs)转录本[85]。相对siRNA来说，piRNA较长，一般为24~30 nt，并且其序列的5’末端多以U开头。与内源siRNA相似，部分piRNA(如小鼠粗线期piRNA)[18, 86]和大部分果蝇和斑马鱼piRNA来源于基因组重复元件，尤其是各类转座元件或反转座子，在基因组中呈现不连续性分布，并且大部分piRNA是由基因组中单一序列的基因簇产生[87]。早期在果蝇生殖细胞中发现的大部分重复序列相关siRNA(repeat-associated siRNAs，rasiRNAs)[88, 89]、线虫的21U-RNAs[16, 90]，其实就是现在的piRNA。有趣的是，部分哺乳动物卵巢来源的内源siRNA和piRNA在基因组上的位置相互重叠，这表明部分的内源siRNA有可能从piRNA加工而来，当然也不排除是piRNA的降解产物。因此，在判别内源siRNA和piRNA起源方面仍需要严谨地分析。

出人意料的是，最近研究发现，原本用以发掘miRNA、siRNA等小RNA的数据中还蕴藏着发现其他新类型的RNA的新契机。如Babiarz等[31]从Solexa测序数据中发现具有特殊二级结构的miRNA前体。但随后的研究表明，其实该“前体”就是snoRNA ACA45，并且还发现了这些由snoRNA加工而来的小分子具有类似miRNA的翻译抑制功能[91]。此后，陆续有利用大规模小RNA测序数据发掘新的snoRNA基因的报道[92-94]。除了snoRNA，其他新类型的snmRNA也陆续被发现：比如Taft等[34]发现了来源于转录起始位点的长度集中在18 nt的转录启始RNA(transcription initiation RNAs,tiRNAs)。因此，利用大规模测序数据来发掘其他类型RNA包括长的ncRNA是今后研究的一个重要方向。

不管是应用测序技术，还是通过生物信息学预测方法所鉴定的候选snmRNA，一般还需要根据研究对象的不同，结合基因芯片、northern杂交、实时定量PCR或RT-PCR、核酸酶保护实验等进行验证。比如，果蝇内源siRNA和piRNA的3’末端具有类似的2’-*O*-甲基化修饰，而动物miRNA则没有

这种修饰，可以通过$NaIO_4$处理和β-elimination反应，判断电泳后的不同迁移速率加以区分[95, 96]。因此，采用实验与计算机分析相结合的方法是目前snmRNA鉴定最有效的研究策略，这对于分析具有复杂来源的snmRNA尤其重要。

3　结语和展望

新型测序技术的广泛应用极大地推动了snmRNA研究。有理由相信，将来会有更多新类型的snmRNA被发现，这必定会重塑整个“RNA世界”。尽管如此，目前snmRNA的系统鉴定方面仍存在诸多问题没有解决，如在应用“深度测序”技术鉴定小RNA时，究竟要达到多“深”的测序程度，才能最大限度地发现各种低丰度的小RNA，piRNA转录本前体不具有诸如miRNA前体的结构特征，尽管研究人员已对piRNA的计算机分析进行了有益地尝试，但目前来讲仍有待于进一步改善算法，提高准确性。再如起源于转录启动区域、长度各异的PAS RNA(promoter-associated short RNA)、TSS RNA(transcription start site RNA)和tiRNA是否功能不同，不属于同一类型；还是由于实验操作的问题，如克隆方法等不同而导致它们长度差异。由于基因组中存在大量的具有形成茎环结构的序列片段，对于通过测序得到的小RNA片段，是否能折叠成类似于miRNA前体特征的小RNA，都是以miRNA的方式起作用，还是存在另外的机制。到目前为止，科学家所发现的小RNA长度最短的约18 nt，那是否就是有功能的小RNA的最短长度极限；还是另外有更短的功能性RNA。显然，在snmRNA的系统识别领域，我们仅仅揭开了“RNA世界”的一角面纱，很多机制还有待于进一步阐明。

ÖÂÐ»¸ÐÐ»±¾ÊµÑéÊÒÁÎ½¨ÓÑ¡¢Ö£ÁèÁæ¶Ô±¾ÎÄÌá³ö½¨ÉèÐÔÒâ¼û£

[参　考　文　献]

[1] Eddy SR. Non-coding RNA genes and the modern RNA world. Nat Rev Genet, 2001, 2(12): 919-29

[2] Kiss-Laszlo Z, Henry Y, Bachellerie JP, et al. Site-specific ribose methylation of preribosomal RNA: a novel function for small nucleolar RNAs. Cell, 1996, 85(7): 1077-88

[3] Huttenhofer A, Kiefmann M, Meier-Ewert S, et al. RNomics: an experimental approach that identifies 201 candidates for novel, small, non-messenger RNAs in mouse. EMBO J, 2001, 20(11): 2943-53

[4] Huttenhofer A, Brosius J, Bachellerie JP. RNomics: identification and function of small, non-messenger RNAs. Curr Opin Chem Biol, 2002, 6(6): 835-43

[5] Shendure J, Ji H. Next-generation DNA sequencing. Nat Biotechnol, 2008, 26(10): 1135-45

[6] Chen CL, Liang D, Zhou H, et al. The high diversity of snoRNAs in plants: identification and comparative study of 120 snoRNA genes from *Oryza sativa*. Nucleic Acids Res, 2003, 31(10): 2601-13

[7] Wang JF, Zhou H, Chen YQ, et al. Identification of 20 microRNAs from *Oryza sativa*. Nucleic Acids Res, 2004, 32 (5): 1688-95

[8] Fu H, Tie Y, Xu C, et al. Identification of human fetal liver miRNAs by a novel method. FEBS Lett, 2005, 579(17): 3849-54

[9] Kiss AM, Jady BE, Bertrand E, et al. Human box H/ACA pseudouridylation guide RNA machinery. Mol Cell Biol, 2004, 24(13): 5797-807

[10] Miyoshi K, Tsukumo H, Nagami T, et al. Slicer function of *Drosophila* Argonautes and its involvement in RISC formation. Genes Dev, 2005, 19(23): 2837-48

[11] Kawamura Y, Saito K, Kin T, et al. *Drosophila* endogenous small RNAs bind to Argonaute 2 in somatic cells. Nature, 2008, 453(7196): 793-7

[12] Gu AD, Zhou H, Yu CH, et al. A novel experimental approach for systematic identification of box H/ACA snoRNAs from eukaryotes. Nucleic Acids Res, 2005, 33(22): e194

[13] Shao P, Yang JH, Zhou H, et al. Genome-wide analysis of chicken snoRNAs provides unique implications for the evolution of vertebrate snoRNAs. BMC Genomics, 2009, 10: 86

[14] Lu C, Tej SS, Luo S, et al. Elucidation of the small RNA component of the transcriptome. Science, 2005, 309(5740): 1567-9

[15] Margulies M, Egholm M, Altman WE, et al. Genome sequencing in microfabricated high-density picolitre reactors. Nature, 2005, 437(7057): 376-80

[16] Ruby JG, Jan C, Player C, et al. Large-scale sequencing reveals 21U-RNAs and additional microRNAs and endogenous siRNAs in *C. elegans*. Cell, 2006, 127(6): 1193-207

[17] Berezikov E, Thuemmler F, van Laake LW, et al. Diversity of microRNAs in human and chimpanzee brain. Nat Genet, 2006, 38(12): 1375-7

[18] Girard A, Sachidanandam R, Hannon GJ, et al. A germline-specific class of small RNAs binds mammalian Piwi proteins. Nature, 2006, 442(7099): 199-202

[19] Lau NC, Seto AG, Kim J, et al. Characterization of the piRNA complex from rat testes. Science, 2006, 313(5785): 363-7

[20] de Wit E, Linsen SE, Cuppen E, et al. Repertoire and evolution of miRNA genes in four divergent nematode species. Genome Res, 2009, 19(11): 2064-74

[21] Friedlander MR, Adamidi C, Han T, et al. High-resolution profiling and discovery of planarian small RNAs. Proc Natl Acad Sci USA, 2009, 106(28): 11546-51

[22] Grimson A, Srivastava M, Fahey B, et al. Early origins and evolution of microRNAs and Piwi-interacting RNAs in animals. Nature, 2008, 455(7217): 1193-7

[23] Lu J, Shen Y, Wu Q, et al. The birth and death of microRNA

genes in *Drosophila*. Nat Genet, 2008, 40(3): 351-5
[24] Li S, Mead EA, Liang S, et al. Direct sequencing and expression analysis of a large number of miRNAs in *Aedes aegypti* and a multi-species survey of novel mosquito miRNAs. BMC Genomics, 2009, 10: 581
[25] Houwing S, Kamminga LM, Berezikov E, et al. A role for Piwi and piRNAs in germ cell maintenance and transposon silencing in zebrafish. Cell, 2007, 129(1): 69-82
[26] Houwing S, Berezikov E, Ketting RF. Zili is required for germ cell differentiation and meiosis in zebrafish. EMBO J, 2008, 27(20): 2702-11
[27] Soares AR, Pereira PM, Santos B, et al. Parallel DNA pyrosequencing unveils new zebrafish microRNAs. BMC Genomics, 2009, 10: 195
[28] Burnside J, Ouyang M, Anderson A, et al. Deep sequencing of chicken microRNAs. BMC Genomics, 2008, 9: 185
[29] Hicks JA, Trakooljul N, Liu HC. Discovery of chicken microRNAs associated with lipogenesis and cell proliferation. Physiol Genomics, 2010, 41: 185-93
[30] Calabrese JM, Seila AC, Yeo GW, et al. RNA sequence analysis defines Dicer's role in mouse embryonic stem cells. Proc Natl Acad Sci USA, 2007, 104(46): 18097-102
[31] Babiarz JE, Ruby JG, Wang Y, et al. Mouse ES cells express endogenous shRNAs, siRNAs, and other microprocessor-independent, Dicer-dependent small RNAs. Genes Dev, 2008, 22(20): 2773-85
[32] Tam OH, Aravin AA, Stein P, et al. Pseudogene-derived small interfering RNAs regulate gene expression in mouse oocytes. Nature, 2008, 453(7194): 534-8
[33] Watanabe T, Totoki Y, Toyoda A, et al. Endogenous siRNAs from naturally formed dsRNAs regulate transcripts in mouse oocytes. Nature, 2008, 453(7194): 539-43
[34] Taft RJ, Glazov EA, Cloonan N, et al. Tiny RNAs associated with transcription start sites in animals. Nat Genet, 2009, 41(5): 572-8
[35] Mitchell PS, Parkin RK, Kroh EM, et al. Circulating microRNAs as stable blood-based markers for cancer detection. Proc Natl Acad Sci USA, 2008, 105(30): 10513-8
[36] Bentley DR. Whole-genome re-sequencing. Curr Opin Genet Dev, 2006, 16(6): 545-52
[37] Shi W, Hendrix D, Levine M, et al. A distinct class of small RNAs arises from pre-miRNA-proximal regions in a simple chordate. Nat Struct Mol Biol, 2009, 16(2): 183-9
[38] Stoeckius M, Maaskola J, Colombo T, et al. Large-scale sorting of *C. elegans* embryos reveals the dynamics of small RNA expression. Nat Methods, 2009, 6(10): 745-51
[39] Czech B, Malone CD, Zhou R, et al. An endogenous small interfering RNA pathway in *Drosophila*. Nature, 2008, 453(7196): 798-802
[40] Ghildiyal M, Seitz H, Horwich MD, et al. Endogenous siRNAs derived from transposons and mRNAs in *Drosophila* somatic cells. Science, 2008, 320(5879): 1077-81
[41] Li C, Vagin VV, Lee S, et al. Collapse of germline piRNAs in the absence of Argonaute3 reveals somatic piRNAs in flies. Cell, 2009, 137(3): 509-21
[42] Armisen J, Gilchrist MJ, Wilczynska A, et al. Abundant and dynamically expressed miRNAs, piRNAs, and other small RNAs in the vertebrate *Xenopus* tropicalis. Genome Res, 2009, 19(10): 1766-75
[43] Wang Y, Brahmakshatriya V, Zhu H, et al. Identification of differentially expressed miRNAs in chicken lung and trachea with avian influenza virus infection by a deep sequencing approach. BMC Genomics, 2009, 10: 512
[44] Glazov EA, Cottee PA, Barris WC, et al. A microRNA catalog of the developing chicken embryo identified by a deep sequencing approach. Genome Res, 2008, 18(6): 957-64
[45] Rathjen T, Pais H, Sweetman D, et al. High throughput sequencing of microRNAs in chicken somites. FEBS Lett, 2009, 583(9): 1422-6
[46] Marson A, Levine SS, Cole MF, et al. Connecting microRNA genes to the core transcriptional regulatory circuitry of embryonic stem cells. Cell, 2008, 134(3): 521-33
[47] Morin RD, O'Connor MD, Griffith M, et al. Application of massively parallel sequencing to microRNA profiling and discovery in human embryonic stem cells. Genome Res, 2008, 18(4): 610-21
[48] Zhang H, Yang JH, Zheng YS, et al. Genome-wide analysis of small RNA and novel microRNA discovery in human acute lymphoblastic leukemia based on extensive sequencing approach. PLoS ONE, 2009, 4(9): e6849
[49] Chen X, Ba Y, Ma L, et al. Characterization of microRNAs in serum: a novel class of biomarkers for diagnosis of cancer and other diseases. Cell Res, 2008, 18(10): 997-1006
[50] Goff LA, Davila J, Swerdel MR, et al. Ago2 immunoprecipitation identifies predicted microRNAs in human embryonic stem cells and neural precursors. PLoS ONE, 2009, 4(9): e7192
[51] Rivas E, Eddy SR. Noncoding RNA gene detection using comparative sequence analysis. BMC Bioinformatics, 2001, 2: 8
[52] Pedersen JS, 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
[53] Griffiths-Jones S. Annotating noncoding RNA genes. Annu Rev Genomics Hum Genet, 2007, 8: 279-98
[54] Weber MJ. New human and mouse microRNA genes found by homology search. FEBS J, 2005, 272(1): 59-73
[55] Wang X, Zhang J, Li F, et al. MicroRNA identification based on sequence and structure alignment. Bioinformatics, 2005, 21(18): 3610-4
[56] Bonnet E, Wuyts J, Rouze P, et al. Evidence that microRNA precursors, unlike other non-coding RNAs, have lower folding free energies than random sequences. Bioinformatics, 2004, 20(17): 2911-7
[57] Yang JH, Zhang XC, Huang ZP, et al. snoSeeker: an advanced computational package for screening of guide and orphan snoRNA genes in the human genome. Nucleic Acids Res, 2006, 34(18): 5112-23
[58] Lambert A, Fontaine JF, Legendre M, et al. The ERPIN server: an interface to profile-based RNA motif identification. Nucleic Acids Res, 2004, 32(Web Server issue): W160-5
[59] Legendre M, Lambert A, Gautheret D. Profile-based detection of microRNA precursors in animal genomes. Bioinformatics, 2005, 21(7): 841-5
[60] Lowe TM, Eddy SR. tRNAscan-SE: a program for improved

detection of transfer RNA genes in genomic sequence. Nucleic Acids Res, 1997, 25(5): 955-64

[61] Lowe TM, Eddy SR. A computational screen for methylation guide snoRNAs in yeast. Science, 1999, 283(5405): 1168-71

[62] Schattner P, Decatur WA, Davis CA, et al. Genome-wide searching for pseudouridylation guide snoRNAs: analysis of the *Saccharomyces cerevisiae* genome. Nucleic Acids Res, 2004, 32(14): 4281-96

[63] Hertel J, Hofacker IL, Stadler PF. SnoReport: computational identification of snoRNAs with unknown targets. Bioinformatics, 2008, 24(2): 158-64

[64] Lim LP, Lau NC, Weinstein EG, et al. The microRNAs of *Caenorhabditis elegans*. Genes Dev, 2003, 17(8): 991-1008

[65] Lim LP, Glasner ME, Yekta S, et al. Vertebrate microRNA genes. Science, 2003, 299(5612): 1540

[66] Huang TH, Fan B, Rothschild MF, et al. MiRFinder: an improved approach and software implementation for genome-wide fast microRNA precursor scans. BMC Bioinformatics, 2007, 8: 341

[67] Nam JW, Shin KR, Han J, et al. Human microRNA prediction through a probabilistic co-learning model of sequence and structure. Nucleic Acids Res, 2005, 33(11): 3570-81

[68] Nam JW, Kim J, Kim SK, et al. ProMiR II: a web server for the probabilistic prediction of clustered, nonclustered, conserved and nonconserved microRNAs. Nucleic Acids Res, 2006, 34(Web Server issue): W455-8

[69] Hertel J, Stadler PF. Hairpins in a haystack: recognizing microRNA precursors in comparative genomics data. Bioinformatics, 2006, 22(14): e197-202

[70] Batuwita R, Palade V. microPred: effective classification of pre-miRNAs for human miRNA gene prediction. Bioinformatics, 2009, 25(8): 989-95

[71] Ng KL, Mishra SK. De novo SVM classification of precursor microRNAs from genomic pseudo hairpins using global and intrinsic folding measures. Bioinformatics, 2007, 23(11): 1321-30

[72] Altschul SF, Gish W, Miller W, et al. Basic local alignment search tool. J Mol Biol, 1990, 215(3): 403-10

[73] Kent WJ. BLAT-the BLAST-like alignment tool. Genome Res, 2002, 12(4): 656-64

[74] Li H, Ruan J, Durbin R. Mapping short DNA sequencing reads and calling variants using mapping quality scores. Genome Res, 2008, 18(11): 1851-8

[75] Li R, Li Y, Kristiansen K, et al. SOAP: short oligonucleotide alignment program. Bioinformatics, 2008, 24(5): 713-4

[76] 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

[77] Li R, Yu C, Li Y, et al. SOAP2: an improved ultrafast tool for short read alignment. Bioinformatics, 2009, 25(15): 1966-7

[78] Huang W, Marth G. EagleView: a genome assembly viewer for next-generation sequencing technologies. Genome Res, 2008, 18(9): 1538-43

[79] Bao H, Guo H, Wang J, et al. Mapview: visualization of short reads alignment on a desktop computer. Bioinformatics, 2009, 25(12): 1554-5

[80] Milne I, Bayer M, Cardle L, et al. Tablet-next generation sequence assembly visualization. Bioinformatics, 2010, 26(3): 401-2

[81] Manske HM, Kwiatkowski DP. LookSeq: a browser-based viewer for deep sequencing data. Genome Res, 2009, 19(11): 2125-32

[82] Yang JH, Shao P, Zhou H, et al. deepBase: a database for deeply annotating and mining deep sequencing data. Nucleic Acids Res, 2010, 38(Database issue): D123-30

[83] Friedlander MR, Chen W, Adamidi C, et al. Discovering microRNAs from deep sequencing data using miRDeep. Nat Biotechnol, 2008, 26(4): 407-15

[84] Hackenberg M, Sturm M, Langenberger D, et al. miRanalyzer: a microRNA detection and analysis tool for next-generation sequencing experiments. Nucleic Acids Res, 2009, 37(Web Server issue): W68-76

[85] Okamura K, Chung WJ, Ruby JG, et al. The *Drosophila* hairpin RNA pathway generates endogenous short interfering RNAs. Nature, 2008, 453(7196): 803-6

[86] Aravin A, Gaidatzis D, Pfeffer S, et al. A novel class of small RNAs bind to MILI protein in mouse testes. Nature, 2006, 442(7099): 203-7

[87] Aravin AA, Hannon GJ. Small RNA silencing pathways in germ and stem cells. Cold Spring Harb Symp Quant Biol, 2008, 73: 283-90

[88] Aravin AA, Naumova NM, Tulin AV, et al. Double-stranded RNA-mediated silencing of genomic tandem repeats and transposable elements in the *D. melanogaster* germline. Curr Biol, 2001, 11(13): 1017-27

[89] Aravin AA, Lagos-Quintana M, Yalcin A, et al. The small RNA profile during *Drosophila melanogaster* development. Dev Cell, 2003, 5(2): 337-50

[90] Batista PJ, Ruby JG, Claycomb JM, et al. PRG-1 and 21U-RNAs interact to form the piRNA complex required for fertility in *C. elegans*. Mol Cell, 2008, 31(1): 67-78

[91] Ender C, Krek A, Friedlander MR, et al. A human snoRNA with microRNA-like functions. Mol Cell, 2008, 32(4): 519-28

[92] Chen HM, Wu SH. Mining small RNA sequencing data: a new approach to identify small nucleolar RNAs in *Arabidopsis*. Nucleic Acids Res, 2009, 37(9): e69

[93] Taft RJ, Glazov EA, Lassmann T, et al. Small RNAs derived from snoRNAs. RNA, 2009, 15(7): 1233-40

[94] Saraiya AA, Wang CC. snoRNA, a novel precursor of microRNA in *Giardia lamblia*. PLoS Pathog, 2008, 4(11): e1000224

[95] Vagin VV, Sigova A, Li C, et al. A distinct small RNA pathway silences selfish genetic elements in the germline. Science, 2006, 313(5785): 320-4

[96] Okamura K, Balla S, Martin R, et al. Two distinct mechanisms generate endogenous siRNAs from bidirectional transcription in *Drosophila melanogaster*: Nat Struct Mol Biol, 2008, 15(9): 998