遗传 HEREDITAS (Beijing) 2011 年 11 月, 33(11): 1191－1202
ISSN 0253-9772 www.chinagene.cn

综 述

DOI: 10.3724/SP.J.1005.2011.01191

转录组研究新技术：RNA-Seq 及其应用

祁云霞[1,2], 刘永斌[2], 荣威恒[2]

1. 内蒙古农业大学动物科学学院, 呼和浩特 010018;
2. 内蒙古自治区农牧业科学院, 呼和浩特 010031

摘要: 转录组是特定组织或细胞在某一发育阶段或功能状态下转录出来的所有 RNA 的集合。转录组研究能够从整体水平研究基因功能以及基因结构, 揭示特定生物学过程以及疾病发生过程中的分子机理。RNA-Seq 作为一种新的高效、快捷的转录组研究手段正在改变着人们对转录组的认识。RNA-Seq 利用高通量测序技术对组织或细胞中所有 RNA 反转录而成的 cDNA 文库进行测序, 通过统计相关读段(reads)数计算出不同 RNA 的表达量, 发现新的转录本; 如果有基因组参考序列, 可以把转录本映射回基因组, 确定转录本位置、剪切情况等更为全面的遗传信息, 已广泛应用于生物学研究、医学研究、临床研究和药物研发等。文章主要介绍了 RNA-Seq 原理、技术特点及其应用, 并就 RNA-Seq 技术面临的挑战和未来发展前景进行了讨论, 为今后该技术的研究与应用提供参考。

关键词: RNA-Seq; 转录组; 新一代测序技术

RNA-Seq and its applications: a new technology for transcriptomics

QI Yun-Xia[1, 2], LIU Yong-Bin[2], RONG Wei-Heng[2]

1. *College of Animals Science*, *Inner Mongolia Agriculture University*, *Huhhot* 010018, *China*;
2. *Inner Mongolia Academy of Agriculture-Animal Sciences*, *Huhhot* 010031, *China*

Abstract: The transcriptome is the complete set of transcripts for certain type of cells or tissues in a specific developmental stage or physiological condition. Transcriptome analysis can provide a comprehensive understanding of molecular mechanisms involved in specific biological processes and diseases from the information on gene structure and function. Transcriptome has been challenging due to the efficient and fast procedures of RNA-seq. RNA-seq, refers to the use of high-throughput sequencing technologies to sequence cDNA library transcribed from all RNAs in tissues or cells, can be used to quantify, profile, and discover RNA transcripts by sequence reads. Thus, the transcripts can then be mapped on the reference genome to get comprehensive genetic information, such as transcription localization and alternative splicing status. RNA-Seq has been widely used in biological, medical, clinical and pharmaceutical research. The detailed principles, technical characteristics and applications of RNA-seq are reviewed here, and the challenges and application potentials of RNA-seq in the future are also discussed. This will present the useful information for other researchers.

收稿日期: 2011–01–17; 修回日期: 2011–04–22
基金项目: 内蒙古自然科学基金项目(编号：2010BS0405)和国家现代肉羊产业技术体系(编号：nycytx-39)资助
作者简介: 祁云霞, 在读博士, 研究方向：分子生物学与动物育种。Tel: 15560908924; E-mail: qi_yunxia@163.com
通讯作者: 刘永斌, 博士, 副研究员, 研究方向：分子生物学与牛羊育种。E-mail: ybliu117@126.com
荣威恒, 研究员, 博士生导师, 研究方向：动物遗传育种。E-mail: rongweiheng@126.com
网络出版时间: 2011-7-28 17:21:30
URL: http://www.cnki.net/kcms/detail/11.1913.R.20110728.1721.002.html

Keywords: RNA-Seq; transcriptome; next-generation sequencing (NGS) technology

随着后基因组时代的到来，转录组学、蛋白质组学、代谢组学等各种组学技术相继出现，其中转录组学是率先发展起来以及应用最广泛的技术[1]。遗传学中心法则表明，遗传信息在精密的调控下通过信使RNA(mRNA)从DNA传递到蛋白质。因此，mRNA被认为是DNA与蛋白质之间生物信息传递的一个“桥梁”，而所有表达基因的身份以及其转录水平，综合起来被称作转录组(Transcriptome)[2]。转录组是特定组织或细胞在某一发育阶段或功能状态下转录出来的所有RNA的总和，主要包括mRNA和非编码RNA(non-coding RNA, ncRNA)[2, 3]。

转录组研究是基因功能及结构研究的基础和出发点，了解转录组是解读基因组功能元件和揭示细胞及组织中分子组成所必需的，并且对理解机体发育和疾病具有重要作用。整个转录组分析的主要目标是：对所有的转录产物进行分类；确定基因的转录结构，如其起始位点，5′和3′末端，剪接模式和其他转录后修饰；并量化各转录本在发育过程中和不同条件下(如生理/病理)表达水平的变化[2,3]。

在过去的十几年里，杂交技术的发展，再加上以标签序列为基础的方法的应用，第一次使研究人员对这一领域有了深入的了解，但毋庸置疑，随着新一代测序(Next-generation sequencing, NGS)平台的市场化，RNA-Seq(RNA sequencing)技术的应用已经彻底改变了转录组学的思维方式。RNA-Seq，即RNA测序又称转录组测序，是最近发展起来的利用深度测序技术进行转录组分析的技术[3]，该技术能够在单核苷酸水平对任意物种的整体转录活动进行检测，在分析转录本的结构和表达水平的同时，还能发现未知转录本和稀有转录本，精确地识别可变剪切位点以及cSNP(编码序列单核苷酸多态性)，提供更为全面的转录组信息。相对于传统的芯片杂交平台，RNA-Seq无需预先针对已知序列设计探针，即可对任意物种的整体转录活动进行检测，提供更精确的数字化信号，更高的检测通量以及更广泛的检测范围，是目前深入研究转录组复杂性的强大工具，已广泛应用于生物学研究、医学研究、临床研究和药物研发等。本文在扼要介绍支持RNA-Seq的新一代测序平台的基础上，对RNA-Seq原理、特点以及到目前为止在研究真核生物转录特征方面的进展做一个较为全面的综述，并对其中有待进一步研究的问题进行了展望。

1 RNA 测序技术平台

原则上，所有的高通量测序技术都能进行 RNA 测序。自 2005 年以来，以 Roche 公司的 454 技术、Illumina 公司的 Solexa 技术和 ABI 公司的 SOLiD 技术为标志的新一代测序技术相继诞生，之后 Helicos Biosciences 公司又推出单分子测序(Single molecule sequencing, SMS)技术。新一代测序又称作深度测序或高通量测序，是相对于传统的 Sanger 测序而言，主要特点是测序通量高，测序时间和成本显著下降。各平台测序原理及序列长度的差异决定了各种高通量测序仪具有不同的应用侧重。这就要求我们在熟悉各种高通量测序仪内在技术特点的基础上进行选择。各平台比较见表 1。

1.1 Illumina/Solexa

Illumina公司目前使用最多的测序仪是Genome Analyzer(GA)，其专利核心技术是“DNA簇(DNA cluster)”和“可逆性末端终结(Reversible terminator)”[12]，采用边合成边测序(Sequencing by synthesis, SBS)的原理，测序流程如下[13]：(1)测序文库的构建。将DNA随机打断后在每条DNA链两端加上接头(adapter)；(2)锚定桥接。每一个带接头的DNA片段与测序通道上的接头引物随机结合，添加未标记的dNTP和普通*Taq*酶进行固相桥式 PCR扩增；(3) 产生DNA簇。通过变性和桥式扩增循环在每个测序通道表面获得数百万条密集成簇的待测DNA片段；(4)单碱基延伸测序。将 4 种被标记的dNTP、引物和DNA聚合酶添加到测序通道内以启动测序循环。通过激光的激发，从每个测序通道的测序簇里面产生出对应的荧光，通过判断捕获的荧光颜色记录待测序簇的碱基。

GA 作为新一代测序技术平台，具有高准确性、

表 1　主要高通量测序平台比较

平台和机型	Illumina/Solexa GA IIx	Roche/454 GS FLX	ABI/SOLiD SOLiD3	Helicos HeliScope	参考文献
测序原理	可逆染料终结合成测序	焦磷酸合成测序	连接测序	单分子合成测序	[4~7]
平均读长(bp)	100	400	50	35	[4~7]
数据量(Gb/run)	54~60	0.5	100	21～35	[8,9]
每 Mb 费用($)	～2	～60	～2	～1	[10]
仪器价格($)	540,000	500,000	595,000	999,000	[11]
准确率(%)	≥98　99	≥99	≥99.94	≥97～99.8	[9]
主要错误类型	替换	插入，缺失	替换	缺失	[8,10]
运行时间(d)	4	0.35	7	8	[4~7, 11]
优点	性价比高；目前应用最广泛的平台	读长最长；运行速度快	准确率最高	产量高；文库制备简单，不需要 DNA 扩增或连接	/
缺点	读长短	试剂花费高；同源重复序列出错率较高	读长短；运行时间长	失误率高	/

注：由于 NGS 技术发展迅速，费用和运行时间可能会降低和缩短，而序列的长度、数据量和准确率将增加。

高通量、高灵敏度和低运行成本等突出优势，是目前使用最广泛的新一代测序平台。近两年来，Illumina/Solexa 测序平台不断升级，相继推出了 GA IIx、HiSeq 2000 等测序仪。

1.2　Roche/454

454 公司可谓新一代测序技术的奠基人，2003 年底推出了革命性的基于焦磷酸测序法(pyrosequencing)的超高通量基因组测序系统[14]，开创了边合成边测序的先河。之后，454 公司被罗氏诊断公司收购，推出了性能更优的第二代基因组测序系统——Genome Sequencer FLX System。其测序步骤为：(1)测序文库构建。将基因组DNA或待测样品DNA用物理方法打碎成 300~800 bp的片段后，在片段两端加上锚定接头；(2)乳液PCR(emulsion PCR)。每一个带接头的DNA片段与一个磁珠结合，并在小油滴的包裹下进行独立的PCR扩增；(3)PTP载板(Pico Titer Plate)。每一个磁珠进入 454 公司发明的 454 PTP载样板的每一个小孔内；(4)焦磷酸测序。将包含有焦磷酸测序反应激活液的微磁珠加入PTP孔内，通过检测到的光信号确定待测 DNA 的序列[14, 15]。

454 平台的突出优势是读长长，但准确率较低[13]，成本高。尽管如此，对于那些需要较长读长的应用如从头拼接和宏基因组学，它仍是最理想的选择。

1.3　ABI/SOLiD

SOLiD(supported oligo ligation detection)系统在文库构建和PCR扩增方面与GS FLX系统类似，微珠通过接头捕获DNA片段，并进行乳液PCR。接下来的测序则是SOLiD的独特之处[16]：以连接反应取代传统的聚合酶延伸反应。连接反应的底物是 8 碱基单链荧光探针混合物，探针的 5′端标记有荧光，3′端 1～2 位碱基对与 5′端荧光信号的颜色对应，由于 2 个碱基有 16 种组成情况，而只有 4 色荧光，因此每色荧光对应 4 种碱基组成，而碱基序列则通过以下测序循环过程来确定：每次SOLiD测序包括五轮测序反应，每轮测序反应又由多个连接反应组成。第一轮测序的第一次连接反应将参入 1 条探针，测序仪记录下反映该条探针 3′端 1～2 位编码区颜色信息，随后除去 6～8 位碱基及 5′末端荧光基团，这样实际上连接了 5 个碱基，并获得 1～2 位的颜色信息。以此类推，第二次连接反应得到模板上第 6～7 位碱基序列的颜色信息，而第三次连接反应得到第 11～12 位的颜色信息……几个循环之后，引物重置，开始第二轮的测序。由于第二轮测序的引物比第一轮前移一位，所以这轮测序将得到 0～1 位、5～6 位、10～11 位……的颜色信息，五轮测序反应后，就可得到所有位置的颜色信息，并推断出相应的碱基序列。

SOLiD 系统的主要优势在于具有很高的序列读取精确度和数据输出量，相同数据量的测序价格略低于 Solexa 测序的价格。但同样地，由于序列读长较短，测序后数据的装配需要有坚实的生物信息学分析基础。

1.4 Helicos/HeliScope

2008 年, Helicos Biosciences 公司开发了第一台单分子测序仪——HeliScope遗传分析系统, 与上述 3 种高通量测序技术不同的是, 它通过在单一DNA分子组成的阵列上进行合成测序, 跨越了文库制备中基于PCR 扩增的信号放大过程,避免了该过程可能引入的错误, 达到了读取单个荧光分子的能力。其测序流程如下:构建的单链DNA 文库未经扩增, 没有规律地排列在平面基板上。每个测序循环中, DNA 聚合酶和 4 种荧光标记的核苷酸中的一种流入, 按照模板序列延伸DNA 链, 阵列中发生了碱基延伸反应的DNA 链就会发出荧光, 并通过CCD记录下来。经过洗涤, 延伸了的DNA 链上的荧光物质被切除并被移走, 便可以进行下一轮单个碱基的延伸, 荧光标记的切除以及图像的获取[17]。

SMS技术省去了昂贵的DNA扩增步骤, 降低了测序成本, 同时还增加了数据产出量和序列读长。但同时也面临着新的难题, 主要是集中在单分子水平光学信号的检测方面, HeliScope 利用了一项被称为全内反射显微镜(Total internal reflection microscopy, TIRM)的技术来解决这一问题, 只有靠近流通池反应表面很薄的一层空间内的荧光集团才能被消逝波所激发产生荧光[18]。另外该平台原始数据的准确度明显低, 不过应用双末端测序(paired-end sequencing)技术可以显著提高准确率。

2 RNA-Seq 原理

把上述高通量测序技术应用到由mRNA 逆转录生成的cDNA上, 从而获得来自不同基因的mRNA片段在特定样本中的含量, 这就是mRNA测序或mRNA-Seq, 同样原理, 各种类型的转录本都可以用深度测序技术进行高通量检测, 统称作RNA-Seq。该技术[3]首先将细胞中的所有转录产物反转录为cDNA 文库(利用最新的SMS技术可略去这一步, 直接对RNA进行测序[19]), 然后将cDNA 文库中的DNA 随机剪切为小片段(或先将RNA片段化后再反转录), 在cDNA两端加上接头利用新一代高通量测序仪测序, 直到获得足够的序列, 所得序列通过比对(有参考基因组)或从头组装(de novo assembling)(无参考基因组)形成全基因组范围的转录谱(图 1)。

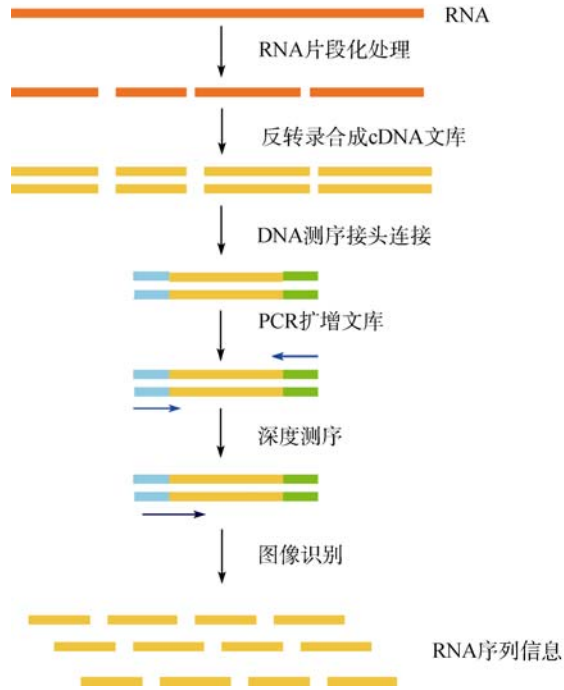

图 1 RNA-Seq 实验流程

此外, 双末端测序是目前各平台广泛采用的一种策略。该技术是将基因组DNA或cDNA打断为一定长度的片断后从两端进行测序, 这样可以从一个片断获得距离已知的两条序列信息, 同时相对于单端测序增加了物理覆盖度(Physical coverage)[20, 21], 因此显著增强了对数据分析的能力。在转录组测序中, 双末端测序使信号可以更好地与转录子联系起来, 例如, 可以更好地区别不同剪切方式[20, 22], 鉴定由染色体重排造成的融合基因[20, 21, 23]等。在全基因组测序中, 双末端测序可以用来确定读段的方向和两个读段之间的距离, 以及基因组重组和结构变异等信息[24]。

3 RNA-Seq 技术优势

目前用于转录组数据获得和分析的方法主要有基于杂交技术的芯片(Gene chip 或 microarray)技术, 基于序列分析的基因表达系列分析(Serial analysis of gene expression, SAGE)和大规模平行信号测序系统(Massively parallel signature sequencing, MPSS), 以及最新提出的 RNA-Seq 技术等。

基因芯片是开发最早也是目前应用较广的高通量转录组检测技术。该技术成本适中, 数据分析软

件较多，整个方法较为成熟，然而基于杂交技术的微阵列技术只限用于已知序列，无法检测新的RNA；而且杂交技术灵敏度有限，难以检测低丰度的目标(需要更多的样品量)和重复序列；也很难检测出融合基因转录、多顺反子转录等异常转录产物[25, 26]。与芯片不同，SAGE不需任何基因序列的信息，能够全局性地检测所有基因的表达水平，除了具有显示基因差异表达谱的作用外，还对那些未知基因特别是那些低拷贝基因的发现起到了巨大的推动作用[27]。MPSS 技术是对SAGE 技术的改进，简化了测序过程，提高了精度，但二者都是基于昂贵的Sanger测序，需要大量的测序工作，技术难度较大，而且涉及酶切、PCR 扩增、克隆等可能会产生碱基偏向性的操作步骤[28]，因而限制了其推广。

相比之下，RNA-Seq技术具有诸多独特优势(表2)。(1)数字化信号：直接测定每个转录本片段序列，单核苷酸分辨率的精确度，可以检测单个碱基差异、基因家族中相似基因以及可变剪接造成的不同转录本的表达[29]，同时不存在传统微阵列杂交的荧光模拟信号带来的交叉反应和背景噪音问题，能覆盖信号超高的动态变化范围。(2)高灵敏度：能够检测到细胞中少至几个拷贝的稀有转录本。(3)任意物种的全基因组分析：无需预先设计特异性探针，因此无需了解物种基因信息，能够直接对任何物种进行转录组分析，这对非模式生物的研究尤为重要，例如Wang等[30]、Xiang等[31] 和 Vera等[32]利用RNA-Seq技术分别对白粉虱、海鲈鱼和蝴蝶转录组进行了研究。同时能够检测未知基因，发现新的转录本，并精确地识别可变剪切位点及cSNP，UTR区域[33, 34]。(4)更广的检测范围：高于 6 个数量级的动态检测范围，能够同时鉴定和定量稀有转录本和正常转录本；而芯片对过低或过高表达的基因缺乏敏感性，因而动态检测范围小[3]。此外，RNA-Seq重复性好[33, 35]，无需技术重复，而且起始样品比芯片技术要少得多[3]，尤其适用于来源极为有限的生物样品分析，如癌症干细胞。

表 2　**RNA-Seq** 与其他转录组学技术比较

技术	芯片	SAGE 和 MPSS	RNA-Seq
原理	杂交	Sanger 测序	高通量测序
信号	荧光模拟信号	数字化信号	数字化信号
分辨率	数个-100 bp	单碱基	单碱基
通量	高	低	高
背景	高	低	低
基因表达定量范围	几十到几百倍	不适用	>8000 倍
全转录组图谱	高	高	相对较低
分析成本			
起始 RNA 用量	多	多	少

4　RNA-Seq 的应用

4.1　转录本结构研究

利用单碱基分辨率的RNA-Seq技术可极大地丰富基因注释的很多方面，包括 5′/3′边界鉴定、UTRs区域鉴定以及新的转录区域鉴定等。Mortazavi等[36]对小鼠的大脑、肝脏和骨骼肌进行了RNA深度测序，分析所得序列，有大于 90%的数据显示落在已知的外显子中，而那些在已知序列之外的信息通过数据分析展示的是从未报道过的RNA剪接形式、3′末端UTRs区、变动的启动子区域及潜在的小RNA前体。2008 年Nagalakshmi等[35]利用RNA-Seq技术分别鉴定出酿酒酵母(*S.cerevisiae*)已知基因中 80%和 85%的 5′边界和 3′边界，同年Wilhelm等[29]使用芯片和RNA-Seq相结合的方法在栗酒裂殖酵母(*S.pombe*)中鉴定出很多 5′和 3′边界。这两项研究导致了许多之前未分析过的 5′和 3′末端UTRs的发现。在酿酒酵母中，发现 3′末端存在广泛的多样性，这些不同的 3′端赋予不同mRNA异构体(isoforms)以不同的属性，如mRNA定位或降解的信号，这反过来又可能与独特的生物学功能相关[29,35]。除了 3′末端多样性外，5′末端UTRs内的上游ORFs (uORFs)名单也大大扩增，从 17 到 340(占酵母基因的 6%)。

2010 年，Zhang等[37]利用配对末端RNA-Seq技术对栽培水稻的 8 个器官进行测序，鉴定出 38 650 个转录单元，通过与之前的芯片结果和已知基因模型比较，检测出 7 232 个之前尚未确定的新转录区。除此之外，还鉴定出 10 595 个新的外显子和 29 751 个新的或延长的 5′和 3′UTRs边界。不久Lu等[38]对栽培水稻的两个亚种(*Oryza sativa indica* 和 *japonica*)进行转录组测序，鉴定出 15 708 个新转录活跃区(nTARs)，并且证明有 6 228 个基因在 5′和/或 3′末端延长至少 50 bp。

RNA-Seq还可对可变剪接(Alternative splicing)进行定量研究。Sultan等[39]利用深度测序对人类细胞系mRNA剪接进行了全局性研究，鉴定出 94 241 个剪接位点，其中有 4 096 个是全新的。该研究还表明，外显子跳跃(Exon skipping)是选择性剪接的一种普遍形式。最新RNA-Seq数据分析显示，至少 48%的水稻基因经历可变剪接[38]，比之前报道的利用RNA-Seq数据分析结果(33%)[37]和EST/cDNA数据分析结果(20%~30%)[40, 41]多；在拟南芥中，至少 42%携带内含子的基因经历可变剪接[42]，多于之前利用EST/cDNA数据分析的 20%[41, 43, 44]到 30%[40],并且这些可变剪接转录本中，大多数是携带成熟前终止密码子的剪接异构体，可能在基因表达调控中发挥重要作用[42]。

4.2 转录本结构变异研究

在发现序列差异(如融合基因鉴定、编码序列多态性研究)方面，RNA-Seq也展示了其很大的潜力。Zhang等[37]在对水稻转录组进行测序时发现了 234 件转录融合事件，可能是由反式剪接所产生。其中，173 件发生在染色体之间，即两个RNA前体来自不同的染色体；其余 61 件发生在染色体内部。Shah等[45]对雌激素受体-α-正转移性乳腺小叶肿瘤的基因组进行了重测序，在DNA水平上发现了 32 个非同义突变，结合基因组和转录组数据，他们找到了 2 个未报道过的RNA编辑事件(引导重新编码SRP9 和COG3 的氨基酸序列)。上述单核苷酸的突变成为了原发性早期、中期乳腺癌的特征之一，亦是癌症病变过程的重要因素。Sugarbaker等[46]利用mRNA深度测序对恶性胸膜瘤和对照样品进行比较，发现了肿瘤中存在的 15 个不同的点突变。由于大多数与疾病相关的单核苷酸变异都发生在蛋白编码区，Chepelev等[47]利用RNA-Seq对人Jurkat T细胞和 $CD4^+$ T细胞外显子进行测序，分别检测到 12 176 和 10 621 个SNVs(单核苷酸变异体)，其中 4 703 和 2 952 个是全新的。

4.3 基因表达水平研究

自 20 世纪 90 年代中期, DNA芯片已被用于大规模的基因表达水平研究。然而基于杂交技术的微阵列技术只限用于已知序列，无法检测新的mRNA；而且杂交技术灵敏度有限，难以检测低丰度的目标，也无法捕捉到目的基因表达水平的微小变化——而这恰恰是研究在刺激下或环境变化时的生物反应所必需的。由于RNA-Seq技术是定量的，它可以比芯片更准确地确定RNA的表达水平。原则上，RNA-Seq有可能确定细胞群中的每一个分子的绝对数量，并对实验之间的结果进行直接比较。Marioni等[48]对RNA-Seq和芯片技术在检测差异表达基因方面进行了比较，研究人员利用Illumina测序平台对肝脏和肾脏RNA样品进行测序，并与使用相同RNA样品的芯片(Affymetrix公司)结果比较。发现，在相同的错误发现率(False discovery rate, FDR)的情况下，RNA-Seq比芯片多检测出 30%的差异表达基因。研究结果还表明，Illumina的测序数据具有高度的重复性，技术的变化相对较小。最近Xiang等[31]用RNA-Seq和DGE技术分析了海鲈鱼受细菌攻击前后的转录组谱，发现在受到哈维氏弧菌攻击后海鲈鱼的转录组谱是变化的，有 1 224 个转录本表现出具有显著意义的上调或下调表达，这一结果表明具有先天性的调节免疫适应性的组分和转录组的改变，在鱼类和其他脊椎动物模式中都是全面的保守存在的。

RNA-Seq一个特别强大的优势是它可以捕捉不同组织或状态下的转录组动态变化而无需对数据集进行复杂的标准化[29,33,36]。RNA-Seq已被用来准确地监测酵母营养生长[35]、酵母减数分裂[29]、小鼠胚胎干细胞分化[33]期间和白粉虱发育过程[30]中的基因表达，来跟踪发育过程中基因表达变化，并提供不同组织间基因差异表达的“数字化测量”。

4.4 非编码区域功能研究

转录组学研究的一个重要方面就是发现和分析ncRNA。目前高通量实验揭示，至少 93%以上的人类基因组可转录为RNA[49]，除了不到 2%的序列用于编码蛋白[50]，其余 91%的基因组可转录为非蛋白编码的RNA分子，即ncRNA。ncRNA按其功能可分为看家ncRNA和调节ncRNA。前者通常稳定表达，发挥着一系列对细胞存活至关重要的功能，主要包括转移RNA(tRNA)、核糖体RNA(rRNA)、小核RNA(snRNA)及小核仁RNA(snoRNA)等；后者主要包括长链ncRNA(lncRNA)和以microRNA为代表的小ncRNA(small ncRNA)，在表观遗传、转录及转录

后等多个层面调控基因表达[51]。过去几年里对ncRNA的研究大部分集中于小ncRNA。microRNA是长约 19~23 个核苷酸(nt)的内源性非编码RNA，是动植物中基因表达至关重要的转录后调控因子[52]，与之相关的一类非编码RNA——siRNA，长度为 21~24 nt，是植物中小RNA分子的主要类型[53]。尽管microRNAs和siRNAs大小相似，都参与基因表达的转录后调控，但其生物合成和具体功能是不同的[54]。

新一代测序技术不涉及克隆等步骤，所产生的读长与成熟的microRNA和siRNA长度兼容，这使得RNA-Seq在ncRNA测序研究中比MPSS具有几个重要优势：降低了程序的复杂性和成本，并极大地增加了通量和覆盖深度。到目前为止，利用 454 技术对小ncRNA的分析研究已有很多报道，涉及从低等生物到高等生物的诸多物种，如藻类[55]、病毒[56]、植物[57]以及灵长类动物[58]，这些研究都发现了许多新的小RNA。重要的是，在利用 454 技术研究小RNAs的同时促成了一类新的小RNA的发现，称为Piwi-interacting RNAs(PiwiRNAs)。这种PiwiRNAs在哺乳动物的睾丸中表达，可能是哺乳动物和其他物种生殖细胞发育所需要的[59]。Illumina 和 SOLiD平台单次运行所产生数据量比 454 平台多，可能提供更深的小RNA测序覆盖度。Morin等[60]利用Illumina测序技术对胚体分化前后的人类胚胎干细胞的小RNA文库进行了测序。这项研究从每个文库中获得了超过 600 万的短序列，并鉴定出 334 个已知的和 104 个新的microRNA基因。Zhang等[37]最近也利用Illumina测序技术在水稻中鉴定出 181 个之前未报道的microRNA，更新了当前水稻microRNA的估计数。

lncRNA由于序列保守性较低曾经一度被认为是转录噪音，直到近几年随着研究的不断深入才证明，lncRNA具有明确的生物学功能[61, 62]，与癌症、冠心病及神经退行性疾病等多种疾病密切相关[63]。目前对lncRNA的认识尚处于初级阶段，一般认为大于 200 nt且缺乏蛋白编码能力的转录本为lncRNA，迄今发现的lncRNA长度从几百nt到十几万nt不等，大概可以分为 5 种类型：正义的，反义的，内含子型的和基因间的[51]。lncRNA可以通过不同模式发挥多种分子功能，如调节转录模式，调节蛋白质活力，具有结构和组织功能，改变RNA加工方式和作为一些小RNA的前体[64]。Peng等[65]首次利用新一代测序技术，在全转录组范围分析了呼吸道病毒(SARS病毒和H1N1 型流感病毒)感染后，宿主细胞lncRNA的变化情况。研究人员采用not-so-random引物法(可对PolyA和非PolyA、编码和非编码转录本进行分析)，对病毒感染后的小鼠肺部细胞构建cDNA文库进行深度测序，结果发现，呼吸道病毒感染引发大量lncRNA的表达发生变化。提示这些RNA分子可能参与了宿主细胞的先天性免疫。

4.5　低丰度全新转录本发现

以往利用转座子标签和芯片技术的研究表明，在酵母、果蝇和人类的基因组中有许多新的转录区存在[3]。但是由于交叉杂交，芯片结果的准确性是不确定的，RNA-Seq不受背景噪音问题的困扰，已证实至少 75%，也许大于 90%的酿酒酵母和栗酒裂殖酵母的基因组表达[29, 35]。此外，RNA-Seq结果表明，在每一个检测的基因组中都存在大量的新转录区域，包括酿酒酵母[35]、栗酒裂殖酵母[29]、拟南芥[42, 66]、水稻[37, 38]、小鼠[33]、人[34]、和人体白色念珠菌[67]等基因组。在酿酒酵母和栗酒裂殖酵母中分别发现了487个和453个新转录本，其中酿酒酵母的新转录本中有一半用芯片技术没有鉴定出来[29, 35]。在水稻中检测出的 7 232 个新转录区，其大部分转录水平都低于已知的cDNA基因[37]。

5　RNA-Seq 面临的挑战及发展前景

随着测序技术的不断进步，我们能够对转录组开展更为深入的测序工作，能够发现更多、更可靠的转录子，目前的大规模并行测序技术已经彻底改变了我们对转录组的研究方法，测序结果的质量也在不断提高，得到的信息量也在爆炸式增长。然而和其他所有新生技术一样，RNA-Seq技术也面临着一系列新问题：其一是庞大的数据量所带来的信息学难题，比如如何最好地诠释和比对鉴定多个类似的同源基因，如何确定最佳测序量，获得高质量的转录图谱等[68]；其二是如何针对更复杂的转录组来识别和追踪所有基因中罕见RNA亚型的表达变化。有可能提前实现这一目标的将是使用配对末端测序和单分子测序等更新的测序技术，以及使用更长的读段来增加测序深度和覆盖度[3]；其三，目前的高

通量测序技术大都需要较多的样品起始量，这使得来源极为有限的生物样品分析受到限制，因此如何对单细胞或少量细胞进行转录组测序是一个亟待解决的问题。最近这方面的研究也取得了一定进展，如Tang等[69]建立了一种mRNA-Seq方案，它以PCR为基础扩增单个细胞的mRNA转录组，成功分析了取自小鼠四细胞胚胎时期的单个卵裂球的转录组。然而该方法只能捕捉带有poly(A)尾巴的mRNA，也不能检测绝大多数较长的mRNA(大于 3 kb)的 5′末端，同时也不能保留原转录子的方向信息。除此之外还有一些新的针对低数量细胞进行转录组研究的技术正在不断被开发[20]。最后，标准的RNA-Seq技术不能提供序列转录的方向信息，而这对于转录组注释尤为重要，采用single-strand sequencing[70]和strand-specific sequencing[71]技术能很好的解决这一问题，或将成为RNA-Seq技术发展的一个重要方向。

虽然 RNA-Seq 技术还面临着种种困难，但作为一个刚刚起步的新技术 RNA-Seq 已经显示出其他转录组学技术无可比拟的优势：既能提供单碱基分辨率的转录组注释又能提供全基因组范围的“数字化”的基因表达谱，而且其成本通常比芯片和大规模的 Sanger EST 测序要低，有人甚至提出了 RNA-Seq 最终取代基因芯片的猜测。然而就目前来看，作为两个高通量的转录组学研究技术，在应用的某些方面既存在重叠和竞争也存在优势互补，一种技术能弥补另一种技术遗漏的部分，通常对一个生物学问题的回答需要不同实验技术的协同配合，例如序列捕获(Sequence Capture)技术就是结合了芯片和深度测序，利用芯片探针捕获待测片段，再用深度测序技术分析核酸序列。但基因芯片的缺点，就在于它是一个“封闭系统”，它只能检测人们已知序列的特征(或有限的变异)；而 RNA-Seq 的强项，就在于它是一个“开放系统”，它的发现能力和寻找新的信息的能力从本质上高于芯片技术，相信随着相关学科的进一步发展和测序成本的进一步降低，RNA-Seq 必将在转录组学研究领域占主导地位。

参考文献(References):

[1] Lockhart DJ, Winzeler EA. Genomics, gene expression and DNA arrays. *Nature*, 2000, 405(6788): 827–836.DOI

[2] Costa V, Angelini C, De Feis I, Ciccodicola A. Uncovering the complexity of transcriptomes with RNA-Seq. *J Biomed Biotechnol*, 2010, 2010: 853916. DOI

[3] Wang Z, Gerstein M, Snyder M. RNA-Seq: a revolutionary tool for transcriptomics. *Nat Rev Genet*, 2009, 10(1): 57–63. DOI

[4] 454 Home Page. http://www.454.com/indecx.asp. DOI

[5] Illumina Home Page. http://www.illumina.com/. DOI

[6] Applied Biosystems Home Page. http://www.appliedbiosystems.com.cn/. DOI

[7] Helicos Home Page. http://www.helicosbio.com/. DOI

[8] Magi A, Benelli M, Gozzini A, Girolami F, Torricelli F, Brandi ML. Bioinformatics for next generation sequencing data. *Genes*, 2010, 1(2): 294–307. DOI

[9] Nowrousian M. Next-generation sequencing techniques for eukaryotic microorganisms: sequencing-based solutions to biological problems. *Eukaryot Cell*, 2010, 9(9): 1300–1310. DOI

[10] Shendure J, Ji H. Next-generation DNA sequencing. *Nat Biotechnol*, 2008, 26(10): 1135–1145. DOI

[11] Metzker ML. Sequencing technologies-the next generation. *Nat Rev Genet*, 2010, 11(1): 31–46. DOI

[12] Bentley DR, Balasubramanian S, Swerdlow HP, Smith GP, Milton J, Brown CG, Hall KP, Evers DJ, Barnes CL, Bignell HR, Boutell JM, Bryant J, Carter RJ, Keira Cheetham R, Cox AJ, Ellis DJ, Flatbush MR, Gormley NA, Humphray SJ, Irving LJ, Karbelashvili MS, Kirk SM, Li H, Liu XH, Maisinger KS, Murray LJ, Obradovic B, Ost T, Parkinson ML, Pratt MR, Rasolonjatovo IMJ, Reed MT, Rigatti R, Rodighiero C, Ross MT, Sabot A, Sankar SV, Scally A, Schroth GP, Smith ME, Smith VP, Spiridou A, Torrance PE, Tzonev SS, Vermaas EH, Walter K, Wu XL, Zhang L, Alam MD, Anastasi C, Aniebo IC, Bailey DMD, Bancarz IR, Banerjee S, Barbour SG, Baybayan PA, Benoit VA, Benson KF, Bevis C, Black PJ, Boodhun A, Brennan JS, Bridgham JA, Brown RC, Brown AA, Buermann DH, Bundu AA, Burrows JC, Carter NP, Castillo N, Chiara E Catenazzi M, Chang S, Neil Cooley R, Crake NR, Dada OO, Diakoumakos KD, Dominguez-Fernandez B, Earnshaw DJ, Egbujor UC, Elmore DW, Etchin SS, Ewan MR, Fedurco M, Fraser LJ, Fuentes Fajardo KV, Scott Furey W, George D, Gietzen KJ, Goddard CP, Golda GS, Granieri PA, Green DE, Gustafson DL, Hansen NF, Harnish K, Haudenschild CD, Heyer NI, Hims MM, Ho JT, Horgan AM, Hoschler K, Hurwitz S, Ivanov DV, Johnson MQ, James T, Huw Jones TA, Kang GD, Kerelska TH, Kersey AD, Khrebtukova I, Kindwall AP, Kingsbury Z, Kokko-Gonzales PI, Kumar A, Laurent MA, Lawley CT, Lee SE, Lee X, Liao AK, Loch JA, Lok M, Luo S, Mam-

men RM, Martin JW, McCauley PG, McNitt P, Mehta P, Moon KW, Mullens JW, Newington T, Ning ZM, Ling Ng B, Novo SM, O'Neill MJ, Osborne MA, Osnowski A, Ostadan O, Paraschos LL, Pickering L, Pike AC, Pike AC, Chris Pinkard D, Pliskin DP, Podhasky J, Quijano VJ, Raczy C, Rae VH, Rawlings SR, Chiva Rodriguez A, Roe PM, Rogers J, Rogert Bacigalupo MC, Romanov N, Romieu A, Roth RK, Rourke NJ, Ruediger ST, Rusman E, Sanches-Kuiper RM, Schenker MR, Seoane JM, Shaw RJ, Shiver MK, Short SW, Sizto NL, Sluis JP, Smith MA, Ernest Sohna Sohna J, Spence EJ, Stevens K, Sutton N, Szajkowski L, Tregidgo CL, Turcatti G, Vandevondele S, Verhovsky Y, Virk SM, Wakelin S, Walcott GC, Wang JW, Worsley GJ, Yan JY, Yau L, Zuerlein M, Rogers J, Mullikin JC, Hurles ME, McCooke NJ, West JS, Oaks FL, Lundberg PL, Klenerman D, Durbin R, Smith AJ. Accurate whole human genome sequencing using reversible terminator chemistry. *Nature*, 2008, 456(7218): 53–59. DOI

[13] Mardis ER. The impact of next-generation sequencing technology on genetics. *Trends Genet*, 2008, 24(3): 133–141. DOI

[14] Margulies M, Egholm M, Altman WE, Attiya S, Bader JS, Bemben LA, Berka J, Braverman MS, Chen YJ, Chen ZT, Dewell SB, Du L, Fierro JM, Gomes XV, Godwin BC, He W, Helgesen S, Ho CH, Irzyk GP, Jando SC, Alenquer ML, Jarvie TP, Jirage KB, Kim JB, Knight JR, Lanza JR, Leamon JH, Lefkowitz SM, Lei M, Li J, Lohman KL, Lu H, Makhijani VB, McDade KE, McKenna MP, Myers EW, Nickerson E, Nobile JR, Plant R, Puc BP, Ronan MT, Roth GT, Sarkis GJ, Simons JF, Simpson JW, Srinivasan M, Tartaro KR, Tomasz A, Vogt KA, Volkmer GA, Wang SH, Wang Y, Weiner MP, Yu PG, Begley RF, Rothberg JM. Genome sequencing in microfabricated high-density picolitre reactors. *Nature*, 2005, 437(7057): 376–380. DOI

[15] Ronaghi M, Uhlén M, Nyrén P. A sequencing method based on real-time pyrophosphate. *Science*, 1998, 281(5375): 363–365. DOI

[16] Smih DR, Quinlan AR, Peckhham HE, Makowsky K, Tao W, Woolf B, Shen L, Donahue WF, Tusneem N, Stromberg MP, Stewart DA, Zhang L, Ranade SS, Warner JB, Lee CC, Coleman BE, Zhang Z, McLaughlin SF, Malek JA, Sorenson JM, Blanchard AP, Chapman J, Hillman D, Chen F, Rokhsar DS, McKernan KJ, Jeffries TW, Marth GT, Richardson PM. Rapid whole-genome mutational profiling using next-generation sequencing technologies. *Genome Res*, 2008, 18(10): 1638–1642. DOI

[17] Harris TD, Buzby PR, Babcock H, Beer E, Bowers J, Braslavsky I, Causey M, Colonell J, Dimeo J, Efcavitch JW, Giladi E, Gill J, Healy J, Jarosz M, Lapen D, Moulton K, Quake SR, Steinmann K, Thayer E, Tyurina A, Ward R, Weiss H, Xie Z. Single-molecule DNA sequencing of a viral genome. *Science*, 2008, 320(5872): 106–109. DOI

[18] Harris TD, Buzby PR, Jarosz M, Gill J, Weiss H, Lapidus SN. Optical train and method for TIRF single molecule detection and analysis. US patent application, 20070070349, 2007. DOI

[19] Haas BJ, Zody MC. Advancing RNA-Seq analysis. *Nat Biotechnol*, 2010, 28(5): 421–423. DOI

[20] Ozsolak F, Milos PM. RNA sequencing: advances, challenges and opportunities. *Nat Rev Genet*, 2011, 12(2): 87–98. DOI

[21] Maher CA, Palanisamy N, Brenner JC, Cao XH, Kalyana-Sundaram S, Luo SJ, Khrebtukova I, Barrette TR, Grasso C, Yu JD, Lonigro RJ, Schroth G, Kumar-Sinha C, Chinnaiyan AM. Chimeric transcript discovery by paired-end transcriptome sequencing. *Proc Natl Acad Sci USA*, 2009, 106(30): 12353–12358. DOI

[22] Au KF, Jiang H, Lin L, Xing Y, Wong WH. Detection of splice junctions from paired-end RNA-seq data by SpliceMap. *Nucleic Acids Res*, 2010, 38(14): 4570–4578. DOI

[23] Edgren H, Murumagi A, Kangaspeska S, Nicorici D, Hongisto V, Kleivi K, Rye IH, Nyberg S, Wolf M, Borresen-Dale AL, Kallioniemi O. Identification of fusion genes in breast cancer by paired-end RNA-sequencing. *Genome Biol*, 2011, 12(1): R6. DOI

[24] Bashir A, Volik S, Collins C, Bafna V, Raphael BJ. Evaluation of paired-end sequencing strategies for detection of genome rearrangements in cancer. *PLoS Comput Biol*, 2008, 4(4): e1000051. DOI

[25] Okoniewski MJ, Miller CJ. Hybridization interactions between probesets in short oligo microarrays lead to spurious correlations. *BMC Bioinformatics*, 2006, 7(1): 276. DOI

[26] Royce TE, Rozowsky JS, Gerstein MB. Toward a universal microarray: prediction of gene expression through nearest-neighbor probe sequence identification. *Nucleic Acids Res*, 2007, 35(15): e99. DOI

[27] Velculescu VE, Zhang L, Vogelstein B, Kinzler KW. Serial analysis of gene expression. *Science*, 1995, 270(5235): 484–487. DOI

[28] Brenner S, Johnson M, Bridgham J, Golda G, Lloyd DH, Johnson D, Luo SJ, McCurdy S, Foy M, Ewan M, Roth R, George D, Eletr S, Albrecht G, Vermaas E, Williams SR, Moon K, Burcham T, Pallas M, DuBridge RB, Kirchner J,

Fearon K, Mao J, Corcoran K. Gene expression analysis by massively parallel signature sequencing (MPSS) on microbead arrays. *Nat Biotechnol*, 2000, 18(6): 630–634. DOI

[29] Wilhelm BT, Marguerat S, Watt S, Schubert F, Wood V, Goodhead I, Penkett CJ, JaneRogers J, Bähler J. Dynamic repertoire of a eukaryotic transcriptome surveyed at single-nucleotide resolution. *Nature*, 2008, 453(7199): 1239–1243. DOI

[30] Wang XW, Luan JB, Li JM, Bao YY, Zhang CX, Liu SS. De novo characterization of a whitefly transcriptome and analysis of its gene expression during development. *BMC Genomics*, 2010, 11(1): 400. DOI

[31] Xiang LX, He D, Dong WR, Zhang YW, Shao JZ. Deep sequencing-based transcriptome profiling analysis of bacteria-challenged Lateolabrax japonicus reveals insight into the immune-relevant genes in marine fish. *BMC Genomics*, 2010, 11(1): 472. DOI

[32] Vera JC, Wheat CW, Fescemyer HW, Frilander MJ, Crawford DL, Hanski I, Marden JH. Rapid transcriptome characterization for a nonmodel organism using 454 pyrosequencing. *Mol Ecol*, 2008, 17(7): 1636–1647. DOI

[33] Cloonan N, Forrest Alistair RR, Kolle G, Gardiner BBA, Faulkner GJ, Brown MK, Taylor DF, Steptoe AL, Wani S, Bethel G, Robertson AJ, Perkins AC, Bruce SJ, Lee CC, Ranade SS, Peckham HE, Manning JM, McKernan KJ, Grimmond SE. Stem cell transcriptome profiling via massive-scale mRNA sequencing. *Nat Methods*, 2008, 5(7): 613–619. DOI

[34] Morin RD, Bainbridge M, Fejes A, Hirst M, Krzywinski M, Pugh TJ, McDonald H, Varhol R, Jones SJM, Marra MA. Profiling the HeLa S3 transcriptome using randomly primed cDNA and massively parallel short-read sequencing. *Biotechniques*, 2008, 45(1): 81–94. DOI

[35] Nagalakshmi U, Wang Z, Waern K, Shou C, Raha D, Gerstein M, Snyder M. The transcriptional landscape of the yeast genome defined by RNA sequencing. *Science*, 2008, 320(5881): 1344–1349. DOI

[36] Mortazavi A, Williams BA, McCue K, Schaeffer L, Wold B. Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat Methods*, 2008, 5(7): 621–628. DOI

[37] Zhang GJ, Guo GW, Hu XD, Zhang Y, Li QY, Li RQ, Zhuang RH, Lu ZK, He ZQ, Fang XD, Chen L, Tian W, Tao Y, Kristiansen K, Zhang XQ, Li SG, Yang HM, Wang J, Wang J. Deep RNA sequencing at single base-pair resolution reveals high complexity of the rice transcriptome. *Genome Res*, 2010, 20(5): 646–654. DOI

[38] Lu TT, Lu GJ, Fan DL, Zhu CR, Li W, Zhao Q, Feng Q, Zhao Y, Guo YL, Li WJ, Huang XH, Han B. Function annotation of the rice transcriptome at single-nucleotide resolution by RNA-seq. *Genome Res*, 2010, 20(9): 1238–1249. DOI

[39] Sultan M, Schulz MH, Richard H, Magen A, Klingenhoff A, Scherf M, Seifert M, Borodina T, Soldatov A, Parkhomchuk D, Schmidt D, O'Keeffe S, Haas S, Vingron M, Lehrach H, Yaspo ML. A global view of gene activity and alternative splicing by deep sequencing of the human transcriptome. *Science*, 2008, 321(5891): 956–960. DOI

[40] Campbell MA, Haas BJ, Hamilton JP, Mount SM, Buell CR. Comprehensive analysis of alternative splicing in rice and comparative analyses with *Arabidopsis*. *BMC Genomics*, 2006, 7: 327. DOI

[41] Wang BB, Brendel V. Genomewide comparative analysis of alternative splicing in plants. *Proc Natl Acad Sci USA*, 2006, 103(18): 7175–7180. DOI

[42] Filichkin SA, Priest HD, Givan SA, Shen R, Bryant DW, Fox SE, Wong WK, Mockler TC. Genome-wide mapping of alternative splicing in *Arabidopsis* thaliana. *Genome Res*, 2010, 20(1): 45–58. DOI

[43] Chen FC, Wang SS, Chaw SM, Huang YT, Chuang TJ. Plant Gene and Alternatively Spliced Variant Annotator. A plant genome annotation pipeline for rice gene and alternatively spliced variant identification with cross-species expressed sequence tag conservation from seven plant species. *Plant Physiol*, 2007, 143(3): 1086–1095. DOI

[44] Barbazuk WB, Fu Y, McGinnis KM. Genome-wide analyses of alternative splicing in plants: opportunities and challenges. *Genome Res*, 2008, 18(9): 1381–1392. DOI

[45] Shah SP, Morin RD, Khattra J, Prentice L, Pugh T, Burleigh A, Delaney A, Gelmon K, Guliany R, Senz J, Steidl C, Holt RA, Jones S, Sun M, Leung G, Moore R, Severson T, Taylor GA, Teschendorff AE, Tse K, Turashvili G, Varhol R, Warren RL, Watson P, Zhao YJ, Caldas C, Huntsman D, Hirst M, Marra MA, Aparicio S. Mutational evolution in a lobular breast tumour profiled at single nucleotide resolution. *Nature*, 2009, 461(7265): 809–813.DOI

[46] Sugarbaker DJ, Richards WG, Gordon GJ, Dong LS, De Rienzo A, Maulik G, Glickman JN, Chirieac LR, Hartman ML, Taillon BE, Du L, Bouffard P, Kingsmore SF, Miller NA, Farmer AD, Jensen RV, Gullans SR, Bueno R. Transcriptome sequencing of malignant pleural mesothelioma tumors. *Proc Natl Acad Sci USA*, 2008, 105(9): 3521–3526. DOI

[47] Chepelev I, Wei G, Tang QS, Zhao KJ. Detection of single nucleotide variations in expressed exons of the human genome using RNA-Seq. *Nucleic Acids Res*, 2009, 37(16):

e106. DOI

[48] Marioni JC, Mason CE, Mane SM, Stephens M, Gilad Y. RNA-seq: an assessment of technical reproducibility and comparison with gene expression arrays. *Genome Res*, 2008, 18(9): 1509–1517. DOI

[49] Birney E, Stamatoyannopoulos JA, Dutta A, Guigó R, Gingeras TR, Margulies EH, Weng Z, Snyder M, Dermitzakis ET, Thurman RE, Kuehn MS, Taylor CM, Neph S, Koch CM, Asthana S, Malhotra A, Adzhubei I, Greenbaum JA, Andrews RM, Flicek P, Boyle PJ, Cao H, Carter NP, Clelland GK, Davis S, Day N, Dhami P, Dillon SC, Dorschner MO, Fiegler H, Giresi PG, Goldy J, Hawrylycz M, Haydock A, Humbert R, James KD, Johnson BE, Johnson EM, Frum TT, Rosenzweig ER, Karnani N, Lee K, Lefebvre GC, Navas PA, Neri F, Parker SC, Sabo PJ, Sandstrom R, Shafer A, Vetrie D, Weaver M, Wilcox S, Yu M, Collins FS, Dekker J, Lieb JD, Tullius TD, Crawford GE, Sunyaev S, Noble WS, Dunham I, Denoeud F, Reymond A, Kapranov P, Rozowsky J, Zheng D, Castelo R, Frankish A, Harrow J, Ghosh S, Sandelin A, Hofacker IL, Baertsch R, Keefe D, Dike S, Cheng J, Hirsch HA, Sekinger EA, Lagarde J, Abril JF, Shahab A, Flamm C, Fried C, Hackermüller J, Hertel J, Lindemeyer M, Missal K, Tanzer A, Washietl S, Korbel J, Emanuelsson O, Pedersen JS, Holroyd N, Taylor R, Swarbreck D, Matthews N, Dickson MC, Thomas DJ, Weirauch MT, Gilbert J, Drenkow J, Bell I, Zhao X, Srinivasan KG, Sung WK, Ooi HS, Chiu KP, Foissac S, Alioto T, Brent M, Pachter L, Tress ML, Valencia A, Choo SW, Choo CY, Ucla C, Manzano C, Wyss C, Cheung E, Clark TG, Brown JB, Ganesh M, Patel S, Tammana H, Chrast J, Henrichsen CN, Kai C, Kawai J, Nagalakshmi U, Wu J, Lian Z, Lian J, Newburger P, Zhang X, Bickel P, Mattick JS, Carninci P, Hayashizaki Y, Weissman S, Hubbard T, Myers RM, Rogers J, Stadler PF, Lowe TM, Wei CL, Ruan Y, Struhl K, Gerstein M, Antonarakis SE, Fu Y, Green ED, Karaöz U, Siepel A, Taylor J, Liefer LA, Wetterstrand KA, Good PJ, Feingold EA, Guyer MS, Cooper GM, Asimenos G, Dewey CN, Hou M, Nikolaev S, Montoya-Burgos JI, Löytynoja A, Whelan S, Pardi F, Massingham T, Huang H, Zhang NR, Holmes I, Mullikin JC, Ureta-Vidal A, Paten B, Seringhaus M, Church D, Rosenbloom K, Kent WJ, Stone EA; NISC Comparative Sequencing Program; Baylor College of Medicine Human Genome Sequencing Center; Washington University Genome Sequencing Center; Broad Institute; Children's Hospital Oakland Research Institute, Batzoglou S, Goldman N, Hardison RC, Haussler D, Miller W, Sidow A, Trinklein ND, Zhang ZD, Barrera L, Stuart R, King DC, Ameur A, Enroth S, Bieda MC, Kim J, Bhinge AA, Jiang N, Liu J, Yao F, Vega VB, Lee CW, Ng P, Shahab A, Yang A, Moqtaderi Z, Zhu Z, Xu X, Squazzo S, Oberley MJ, Inman D, Singer MA, Richmond TA, Munn KJ, Rada-Iglesias A, Wallerman O, Komorowski J, Fowler JC, Couttet P, Bruce AW, Dovey OM, Ellis PD, Langford CF, Nix DA, Euskirchen G, Hartman S, Urban AE, Kraus P, Van Calcar S, Heintzman N, Kim TH, Wang K, Qu C, Hon G, Luna R, Glass CK, Rosenfeld MG, Aldred SF, Cooper SJ, Halees A, Lin JM, Shulha HP, Zhang X, Xu M, Haidar JN, Yu Y, Ruan Y, Iyer VR, Green RD, Wadelius C, Farnham PJ, Ren B, Harte RA, Hinrichs AS, Trumbower H, Clawson H, Hillman-Jackson J, Zweig AS, Smith K, Thakkapallayil A, Barber G, Kuhn RM, Karolchik D, Armengol L, Bird CP, de Bakker PI, Kern AD, Lopez-Bigas N, Martin JD, Stranger BE, Woodroffe A, Davydov E, Dimas A, Eyras E, Hallgrímsdóttir IB, Huppert J, Zody MC, Abecasis GR, Estivill X, Bouffard GG, Guan X, Hansen NF, Idol JR, Maduro VV, Maskeri B, McDowell JC, Park M, Thomas PJ, Young AC, Blakesley RW, Muzny DM, Sodergren E, Wheeler DA, Worley KC, Jiang H, Weinstock GM, Gibbs RA, Graves T, Fulton R, Mardis ER, Wilson RK, Clamp M, Cuff J, Gnerre S, Jaffe DB, Chang JL, Lindblad-Toh K, Lander ES, Koriabine M, Nefedov M, Osoegawa K, Yoshinaga Y, Zhu B, de Jong PJ. Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature*, 2007, 447(7146): 799–816. DOI

[50] Clamp M, Fry B, Kamal M, Xie XH, Cuff J, Lin MF, Kellis M, Lindblad-Toh K, Lander ES. Distinguishing protein-coding and noncoding genes in the human genome. *Proc Natl Acad Sci USA*, 2007, 104(49): 19428–19433. DOI

[51] Ponting CP, Oliver PL, Reik W. Evolution and functions of long noncoding RNAs. *Cell*, 2009, 136(4): 629–641. DOI

[52] Filipowicz W, Bhattacharyya SN, Sonenberg N. Mechanisms of post-transcriptional regulation by microRNAs: are the answers in sight? *Nat Rev Genet*, 2008, 9(2): 102–114. DOI

[53] Lu C, Tej SS, Luo SJ, Haudenschild CD, Meyers BC, Green PJ. Elucidation of the small RNA component of the transcriptome. *Science*, 2005, 309(5740): 1567–1569. DOI

[54] Xie ZX, Johansen LK, Gustafson AM, Kasschau KD, Lellis AD, Zilberman D, Jacobsen SE, Carrington JC. Genetic and functional diversication of small RNA pathways in plants. *PloS Biol*, 2004, 2(5): 642–652. DOI

[55] Zhao T, Li GL, Mi SJ, Li S, Hannon GJ, Wang XJ, Qi YJ. A complex system of small RNAs in the unicellular green

alga *Chlamydomonas reinhardtii*. *Genes Dev*, 2007, 21(10): 1190–1203. DOI

[56] Burnside J, Bernberg E, Anderson A, Lu C, Meyers BC, Green PJ, Jain N, Isaacs G, Morgan RW. Marek's disease virus encodes microRNAs that map to meq and the latency-associated transcript. *J Virol*, 2006, 80(17): 8778–8786. DOI

[57] Yao YY, Guo GG, Ni ZF, Sunkar R, Du JK, Zhu JK, Sun QX. Cloning and characterization of microRNAs from wheat (*Triticum aestivum* L.). *Genome Biol*, 2007, 8(6): R96. DOI

[58] Berezikov E, Thuemmler F, van Laake LW, Kondova I, Bontrop R, Cuppen E, Plasterk RH. Diversity of microRNAs in human and chimpanzee brain. *Nat Genet*, 2006, 38(12): 1375–1377. DOI

[59] Lau NC, Seto AG, Kim J, Kuramochi-Miyagawa S, Nakano T, Bartel DP, Kingston RE. Characterization of the piRNA complex from rat testes. *Science*, 2006, 313(5785): 363–367. DOI

[60] Morin RD, O'Connor MD, Griffith M, Kuchenbauer F, Delaney A, Prabhu AL, Zhao YJ, McDonald H, Zeng T, Hirst M, Eaves CJ, Marra MA. Application of massively parallel sequencing to microRNA profiling and discovery in human embryonic stem cells. *Genome Res*, 2008, 18(4): 610–621. DOI

[61] Guttman M, Amit I, Garber M, French C, Lin MF, Feldser D, Huarte M, Zuk O, Carey BW, Cassady JP, Cabili MN, Jaenisch R, Mikkelsen TS, Jacks T, Hacohen N, Bernstein BE, Kellis M, Regev A, Rinn JL, Lander ES. Chromatin signature reveals over a thousand highly conserved large non-coding RNAs in mammals. *Nature*, 2009, 458(7235): 223–227. DOI

[62] Khalil AM, Guttman M, Huarte M, Garber M, Raj A, Rivea Morales D, Thomas K, Presser A, Bernstein BE, van Oudenaarden A, Regev A, Lander ES, Rinn JL. 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. DOI

[63] 黄文涛, 郭向前, 戴甲培, 陈润生. MicroRNA, lncRNA与神经退行性疾病. 生物化学与生物物理进展, 2010, 37(8): 826–833. DOI

[64] Wilusz JE, Sunwoo H, Spector DL. Long noncoding RNAs: functional surprises from the RNA world. *Genes Dev*, 2009, 23(13): 1494–1504. DOI

[65] Peng XX, Gralinski L, Armour CD, Ferris MT, Thomas MJ, Proll S, Bradel-Tretheway BG, Korth MJ, Castle JC, Biery MC, Bouzek HK, Haynor DR, Frieman MB, Heise M, Raymond CK, Baric RS, Katze MG. Unique signatures of long noncoding RNA expression in response to virus infection and altered innate immune signaling. *MBio*, 2010, 1(5): e00206–10. DOI

[66] Lister R, O'Malley RC, Tonti-Filippini J, Gregory BD, Berry CC, Millar AH, Ecker JR. Highly integrated single-base resolution maps of the epigenome in *Arabidopsis*. *Cell*, 2008, 133(3): 523–536. DOI

[67] Bruno VM, Wang Z, Marjani SL, Euskirchen GM, Martin J, Sherlock G, Snyder M. Comprehensive annotation of the transcriptome of the human fungal pathogen *Candida albicans* using RNA-seq. *Genome Res*, 2010, 20(10): 1451–1458. DOI

[68] Vliet VA. Next generation sequencing of microbial transcriptomes: challenges and opportunities. *FEMS Microbiol Lett*, 2010, 302(1): 1–7. DOI

[69] Tang FC, Barbacioru C, Wang YZ, Nordman E, Lee C, Xu NL, Wang XH, Bodeau J, Tuch BB, Siddiqui A, Lao KQ, Surani MA. mRNA-Seq whole-transcriptome analysis of a single cell. *Nat Methods*, 2009, 6(5): 377–382. DOI

[70] Croucher NJ, Fookes MC, Perkins TT, Turner DJ, Marguerat SB, Keane T, Quail MA, He M, Assefa S, Bähler J, Kingsley RA, Parkhill J, Bentley SD, Dougan G, Thomson NR. A simple method for directional transcriptome sequencing using Illumina technology. *Nucleic Acids Res*, 2009, 37(22): e148. DOI

[71] Vivancos AP, Güell M, Dohm JC, Serrano L, Himmelbauer H. Strand-specific deep sequencing of the transcriptome. *Genome Res*, 2010, 20(7): 989–999. DOI