PIBB 生物化学与生物物理进展
Progress in Biochemistry and Biophysics
2010, 37(8): 834~846
www.pibb.ac.cn

新一代高通量 RNA 测序数据的处理与分析 *

王 曦[1)] 汪小我[1)] 王立坤[1, 2)] 冯智星[1)] 张学工[1)**]

([1)] 生物信息学教育部重点实验室，清华信息科学与技术国家实验室(筹)生物信息学研究部，清华大学自动化系，北京 100084；
[2)] 吉林大学计算机科学与技术学院，长春 130012)

摘要 随着新一代高通量 DNA 测序技术的快速发展，RNA 测序(RNA-seq)已成为基因表达和转录组分析新的重要手段. RNA-seq 技术产生的海量数据为生物信息学带来了新的机遇和挑战. 有效地对测序数据进行针对性的生物信息学处理和分析，成为 RNA-seq 技术能否在科学探索中发挥重大作用的关键. 以新一代 Illumina/Solexa 测序平台所产生的数据为例，在扼要介绍高通量 RNA-seq 测序流程的基础上，对 RNA-seq 数据处理和分析的方法和现有软件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展望.

关键词 高通量 RNA 测序，转录组，基因表达，数据处理与分析，生物信息学
学科分类号 Q5，Q6，Q7 **DOI:** 10.3724/SP.J.1206.2010.00151

近年来，新一代高通量测序技术得到了突飞猛进的发展，在此基础上，高通量 RNA 测序即 RNA-seq[1–5]也迅速发展. 与基因芯片技术相比，RNA-seq 无需设计探针，能在全基因组范围内以单碱基分辨率检测和量化转录片段，并能应用于基因组图谱尚未完成的物种[6]，具有信噪比高、分辨率高、应用范围广等优势，正成为研究基因表达和转录组的重要实验手段.

RNA-seq 为基因组学的研究带来了高分辨率的海量数据，如何有效处理和分析这些海量数据成为这一新技术能否带来新的科学发现的关键，一些生物信息学方法与软件也应运而生. 本文针对当前 RNA-seq 应用的现实情况，尝试以 Illumina/Solexa 测序平台产生的 mRNA-seq 数据为例，对 RNA 测序数据的产生过程及数据处理和分析的基本流程、关键方法和现有软件进行较全面的介绍，并讨论 RNA-seq 数据分析中存在的挑战.

1 高通量测序技术简介

诞生于 20 世纪 70 年代的 Sanger 法是最早被广泛应用的 DNA 测序技术[7]，也是完成人类基因组计划的基础. 但是，由于它测序通量低，费时费力，科学家们一直在寻求通量更高、速度更快、价格更便宜、自动化程度更高的测序技术. 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[8]. 新一代测序技术又称作深度测序技术，主要特点是测序通量高、测序时间和成本显著下降[9].

把这种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定量检测，统称作 RNA-seq 或 RNA 测序. 目前，在已经推出的几种新一代测序平台中，Illumina/Solexa 测序平台上的 RNA-seq 应用最广，我们以此为例来综述 RNA-seq 数据处理和分析的生物信息学问题和方法.

* 国家自然科学基金资助项目(60702002, 60721003, 30873464, 60905013)和东南大学生物电子学国家重点实验室开放研究基金资助项目.
** 通讯联系人.
Tel: 010-62794919, E-mail: zhangxg@tsinghua.edu.cn
收稿日期：2010-03-25，接受日期：2010-04-30

Illumina/Solexa 测序技术的基本原理是边合成边测序(sequencing by synthesis，SBS)[10−12]，即测序过程是以 DNA 单链为模板，在生成互补链时，利用带荧光标记的 dNTP 发出不同颜色的荧光来确定不同的碱基．新加入 dNTP 的末端被可逆的保护基团封闭，既保证单次反应只能加入一个碱基，又能在该碱基读取完毕后，将保护基团除去，使得下一个反应可继续进行．为了增加荧光强度，使之更易被成像系统所采集，该技术在测序之前还需要对待测片段做桥式扩增(bridge amplification)[13](http://www.illumina.com/)．初期的 Illumina/Solexa 测序技术只能在较短的测序读长上(20～30 碱基)保证较高的正确率．随着技术的改进，目前的读长已经增加到 100 碱基以上．同时，随着双端测序(paired-end，PE)技术的成熟，测序长度更可达到单端测序的 2 倍，测序通量也随之增加．这种测序技术是 Solexa 公司发展起来的，2007 年被 Illumina 公司收购，因此现在通常被称为 Illumina/Solexa 测序技术．近两年来，Illumina/Solexa 测序平台不断升级，相继推出了 GA(Genome Analyzer)、GA IIx、HiSeq 2000 等测序仪．更多关于高通量测序平台的介绍，可以查阅相关文献[9, 14−16].

2 RNA-seq 测序文库制备和测序平台数据输出

本小节针对 Illumina/Solexa 测序平台，对 RNA 测序文库制备标准和平台底层数据产生做一个简单的介绍.

2.1 RNA-seq 测序文库制备

对于 mRNA-seq 实验，从总 RNA 到最终的 cDNA 文库制备完成主要包括以下步骤．首先，用 Poly(T)寡聚核苷酸从总 RNA 中抽取全部带 Poly(A)尾的 RNA，其中的主要部分就是编码基因所转录的 mRNA．将所得 RNA 随机打断成片段，再用随机引物和逆转录酶从 RNA 片段合成 cDNA 片段．然后，对 cDNA 片段进行末端修复并连接测序接头(adapter)，得到将用于测序的 cDNA．在以上过程，将 RNA 随机片段化和采用随机引物进行反转录，都是为了使所得 cDNA 片段较均匀地取自各个转录本．为提高测序效率，一般还需要用电泳切胶法获取长度范围在 200 bp(±25 bp)的 cDNA 片段，再通过 RCR 扩增，得到最终的 cDNA 文库.

在上述文库制备过程中，如果不是只抽取带 Poly(A)尾的 RNA，而是使用全部的 RNA，则 RNA-seq 测得的就是细胞中的全部转录本，如果把带 Poly(A)尾的 RNA 过滤掉，也可以得到非编码的 RNA 转录本，如果从总 RNA 中只提取长度为 21～23 个碱基左右的 RNA，则得到全部的 miRNA (microRNA)转录本, 相应的方法也称作 miRNA-seq.

样品制备最终得到的是双链 cDNA 文库．在后续测序中，测得的每个读段(read)随机地来自双链 cDNA 的某一条链，从读段序列本身无法得知它是与 RNA 方向相同还是倒转互补，在后续的读段定位时需要两个方向都考虑．在新基因识别等应用中，转录本的方向对基因注释尤为重要，需要在文库制备和测序中保留 RNA 的方向信息．最近有文献报道了保留方向信息的 RNA-seq 样品制备方法[17−20].

2.2 测序平台数据输出

将 RNA-seq 测序文库加入流动槽(flow cell)中的各通道(lane)，在桥式 PCR 扩增后，就可以进行测序了．测序过程中，计算机软件同步地对荧光图像数据进行处理，通过分析荧光信号来确定被测碱基，并给出质量评分．按照图像上的位置坐标，计算机程序将同一位置测得的碱基根据测序顺序连成读段(read)．由于荧光图像文件所占有的磁盘空间很大，通常 GA IIx 平台一次实验就能产生上太字节(TB)的图像文件，所以一般情况下不予保留原始的荧光图像数据，而是只保留程序读出的读段数据及对应的质量分值，这就是多数实验室委托测序中心进行 RNA-seq 测序后得到的最原始的数据.

为了便于测序数据的发布和共享，高通量测序数据以 FASTQ 格式来记录所测的碱基读段和质量分数．如图 1 所示，FASTQ 格式以测序读段为单位存储，每条读段占 4 行，其中第 1 行和第 3 行由文件识别标志和读段名(ID)组成(第 1 行以“@”开头而第 3 行以“+”开头；第 3 行中 ID 可以省略，但“+”不能省略)，第 2 行为碱基序列，第 4 行为对应的测序质量分数．关于 FASTQ 格式更多地介绍可参考文献[21]．为方便保存和共享各实验室产生的高通量测序数据，NCBI、EBI、DDBJ 等数据中心建立了大容量的数据库 SRA(Sequence Read Archive, http://www.ncbi.nlm.nih.gov/Traces/sra)来存放共享的测序数据[22−23].

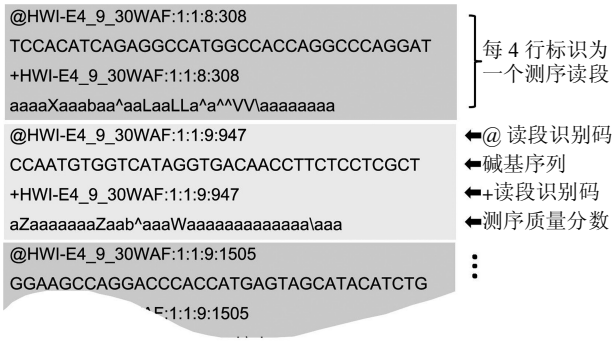

图 1　读段 FASTQ 数据格式示例

Fig. 1　FASTQ format examples

3　RNA-seq 数据的基本处理

RNA-seq 的基本应用是测量一个样本中的基因表达或转录组．有实验表明，新一代高通量测序技术重复数据之间的相关度较高($R^2 \approx 0.96$)[1-2]，因此，如果对同一样本在多个通道上进行了 RNA 测序的技术重复，我们建议可以把几个通道的数据进行合并，这样等效地增加了测序深度．本节讨论单个样本 RNA 测序数据的基本处理流程，如图 2a 所示．

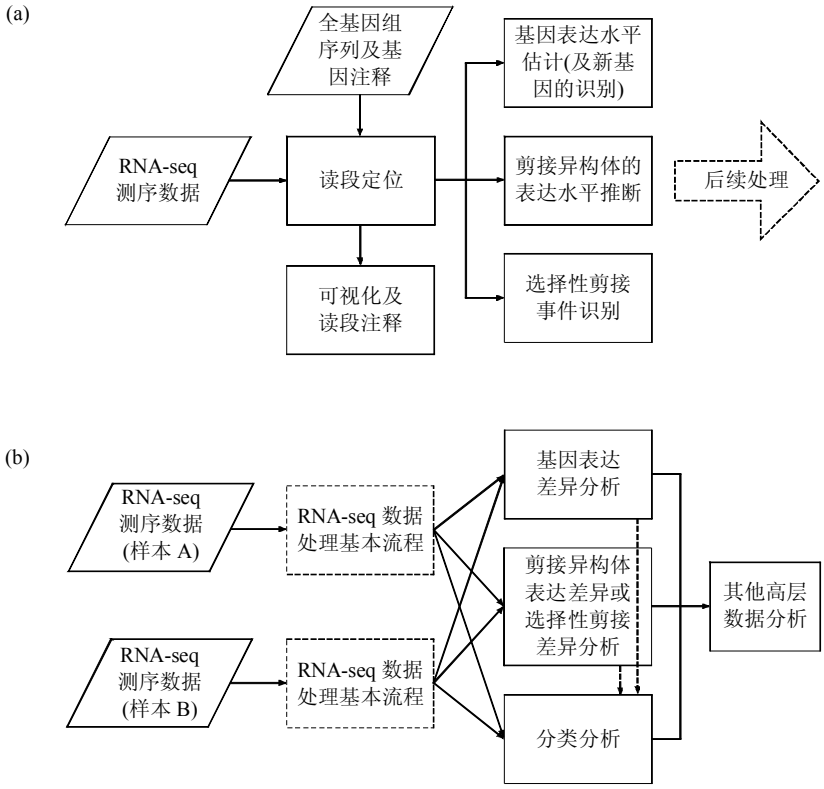

图 2　RNA-seq 数据处理和分析流程图

Fig. 2　The flowchart for RNA-seq data processing and analysis

(a)RNA-seq 数据的基本处理, 其方法介绍见正文第 3 节. (b)两类样本 RNA-seq 数据比较分析的框架, 对应于正文的第 4 节. (b)中虚线框内为(a)所示的流程, 虚线箭头表示可选输入.

3.1　读段定位

获得 RNA-seq 的原始数据后，首先需要将所有测序读段通过序列映射(mapping)定位到参考基因组上，这是所有后续处理和分析的基础．在读段定位之前，有时还需要根据测序数据情况对其做某些基本的预处理．例如，过滤掉测序质量较差的读段、对 miRNA 测序读段数据去除接头序列等．

高通量测序的海量数据对计算机算法的运行时间提出了很高的要求．针对诸如 Illumina/Solexa 等测序平台得到的读段一般较短、且插入删除错误较

少等特点，人们开发了一些短序列定位算法．这些算法主要采用空位种子索引法(spaced-seed indexing)或 Burrows-Wheeler 转换(Burrows-Wheeler Transform，BWT)技术来实现[24]．空位种子索引法首先将读段切分，并选取其中一段或几段作为种子建立搜索索引，再通过查找索引、延展匹配来实现读段定位，通过轮换种子考虑允许出现错配(mismatch)的各种可能的位置组合．BWT 方法通过 B-W 转换[25]将基因组序列按一定规则压缩并建立索引，再通过查找和回溯来定位读段，在查找时可通过碱基替代来实现允许的错配．表 1 列出了目前可免费下载使用的部分短序列定位软件．其中采用空位种子片段索引法的代表是 Maq[26]，而采用 Burrows-Wheeler 转换的代表是 Bowtie[27]．总的来说，采用 BWT 的定位算法在时间效率上要优于空位种子片段索引法[24, 28]．随着读长的增加，允许读段序列中存在插入删除(indel)的定位变得可行而重要．由于以上两类方法对序列中插入删除的处理较为困难，近来人们开发了一些基于改进的 Smith-Waterman 动态规划算法[29]的序列比对工具，如 BFAST[30]、SHRiMP[31]、Mosaik(http://bioinformatics.bc.edu/marthlab/Mosaik)等，但算法速度较慢，大多需采用计算机并行编程技术来解决运行时间的问题．

表 1 适用于 Illumina/Solexa 测序平台的读段定位软件

Table 1 Mappers/aligners for Illumina/Solexa sequencing data

名称	SAM[1)]	质量[2)]	主要采用技术	网址
MAQ[26]	否	是	空位种子	http://maq.sourceforge.net
Bowtie[27]	是	是	BWT	http://bowtie-bio.sourceforge.net/index.shtml
BWA[32]	是	是	BWT	http://bio-bwa.sourceforge.net/bwa.shtml
ZOOM[33]	否	否	空位种子	http://www.bioinformaticssolutions.com/products/zoom
ELAND	否	否	空位种子	http://www.illumina.com/software/genome_analyzer_software.ilmn
SOAP2[34]	否	否	BWT	http://soap.genomics.org.cn
RazerS[35]	否	否	q-grams 过滤	http://www.seqan.de/projects/razers.html
Novoalign	是	是	Needleman-Wunsch 算法 空位种子	http://www.novocraft.com
SHRiMP[31]	否	是	q-grams 过滤 Smith-Waterman 算法	http://compbio.cs.toronto.edu/shrimp
BFAST[30]	是	是	Smith-Waterman 算法 并行编程	https://secure.genome.ucla.edu/index.php/BFAST
Mosaik	是	是	Smith-Waterman 算法 并行编程	http://bioinformatics.bc.edu/marthlab/Mosaik

[1)] SAM: 是否能以 SAM 格式输出; [2)] 质量: 是否提供读段定位质量信息; BWT: Burrows-Wheeler 转换.

在 RNA 测序数据的基因组定位中，一个特殊的问题是跨越两个外显子接合区的读段(junction reads)定位．在真核生物中，成熟的 mRNA 是经过由 mRNA 前体中的外显子经过剪接形成的．如果一个读段跨越了两个外显子，那么就无法将这个读段完整地定位到基因组序列上．而同时，这种跨两个外显子的读段在分析转录本的剪接形式和研究选择性剪接中有重要的作用．为了解决这一问题，人们采取两种典型的策略来进行接合区读段的定位：一是根据已知的基因外显子注释，构建所有可能的外显子接合区序列，与基因组序列一并作为定位的参考基因组；二是不依赖基因注释，而是先利用能完整定位到基因组的读段得到粗略的外显子区域，并结合剪接位点序列构建出可能的剪接位点，然后将不能完整定位的读段分段定位到两个外显子可能的结合区域．Illumina/Solexa 平台提供的 RNA- seq 软件分析包 GApipeline 采用了第一种策略．采用第二种策略的软件有 Tophat[36]和 G-Mo.R-Se[37]等，最新的 Tophat 软件增加了利用已知外显子边界注释信息的选项．

不论是哪种测序平台，测序中都不可避免地存在一定的错误，基因组中又存在单核苷酸多态性等引起的序列变化，所以在读段定位时通常允许一定数量的错配，可以根据不同应用调节允许错配的程度．另一方面，由于基因组中重复序列和高相似度序列的影响，某些读段会出现定位到基因组多个位置的情况．这些因素影响了各个读段到基因组的定位质量，在一些新的读段定位算法中，同时给出每个读段与基因组匹配质量．通常在后续处理前，人们将多定位的读段都过滤掉，也有人尝试用适当的策略把多定位读段“分配”到其中某些位置上[2, 38].

读段定位到基因组后通常采用 SAM(Sequence Alignment/Map)格式或其二进制版本 BAM 格式[39]来存储．二进制版本可大大节省存储空间，但不能直接用普通文本编辑工具显示．关于 SAM 格式的详细介绍，可查阅(http://samtools.sourceforge.net/SAM1.pdf).

3.2 基因表达水平估计

在深度测序技术出现之前，高通量测量不同基因表达水平的主要手段是基因芯片，在此基础上可以对不同组织或者不同发育阶段的基因表达差异和模式进行分析．mRNA-seq 数据最基本的应用也是检测基因的表达水平，与基因芯片数据相比，RNA 测序得到的是数字化的表达信号，具有灵敏度高、分辨率高、无饱和区等优势[40-42].

RNA 测序数据是对提取出的 RNA 转录本中随机进行的短片段测序，如果一个转录本的丰度高，则测序后定位到其对应的基因组区域的读段也就多，可以通过对定位到基因外显子区的读段计数来估计基因表达水平．很显然，读段计数除了与基因真实表达水平成正比，还与基因长度成正比，同时也与测序深度即测序实验中得到的总读段数正相关．为了保持对不同基因和不同实验间估计的基因表达值的可比性，人们提出了 RPM 和 RPKM 的概念[2]．RPM(reads per million reads)即每百万读段中来自于某基因的读段数，考虑了测序深度对读段计数的影响．RPKM(reads per kilo bases per million reads)是每百万读段中来自于某基因每千碱基长度的读段数，公式表示为：

$$RPKM=\frac{\text{基因区读段计数}}{\text{基因长度}\times\text{测序深度}}\times 10^9.$$

RPKM 不仅对测序深度作了归一化，而且对基因长度也作了归一化，使得不同长度的基因在不同测序深度下得到的基因表达水平估计值具有了可比性，是目前最常用的基因表达估计方法．软件 rSeq[43]、DEGseq 软件包[44]和 Cufflinks[45]等都提供了用上述方法进行基因表达水平计算的功能.

根据 RNA-seq 文库制备标准，在不考虑基因结构的理想情况下，读段会均匀地分布在基因上．而实际上，通过对实际数据的可视化分析很容易发现，读段在基因上的分布有着自身的一些模式，呈现出不均匀性(图 3)．这一问题已经引起很多学者的关注[46-48]．造成读段分布出现偏好的原因可能有多个方面：在制备 cDNA文库时，反转录所采用的随机引物对 RNA 序列具有一定的偏好性，使得 cDNA 片段不能够完全均匀地取自各转录本；在 PCR 扩增中，扩增效率与序列的 GC 含量等特征相关，可导致 GC 含量高的 cDNA 片段在文库中拷贝数增加超过其他片段；舍弃多定位的读段也可能导致读段的非均匀分布；等等．如果能对读段分布的不均匀性进行建模并加以校正，可以提高 RNA-seq 推断基因表达量的准确度．但根据对实际数据的观察，对于较长转录本，读段非均匀分布带来的误差很大程度上可相互抵消，用 RPKM 来估计基因的表达水平可以得到比较满意的结果.

3.3 选择性剪接事件识别和剪接异构体表达水平推断

在真核生物中，选择性剪接现象普遍存在．基因转录形成的 mRNA 前体(pre-mRNA)在剪接过程中因去掉不同的内含子区域或保留不同的外显子区域，可形成不同的剪接异构体．根据 RNA-seq 原理，只要测序深度足够深，就能检测到所有转录本的全部序列，包括来自剪接接合区的序列．利用考虑到接合区的读段定位方法，就有可能系统地研究某一组织或某一条件下的基因选择性剪接事件.

前面已经提到，Tophat 等软件定位剪接接合区读段的策略能标定出剪接事件中的两个剪接位点：供体位点和受体位点．通过比较供体位点和受体位点的组合，就能识别选择性剪接事件[4, 49]．图 3 中包含了选择性剪接识别的一个例子．进一步，通过对供体和受体位点的读段计数，结合外显子其他区域的读段数据，还能定量地计算选择性剪接事件之间的比例[50-51].

对于每一个剪接异构体，RNA-seq 数据能在一

定程度上推断其表达水平．比如，可以根据已知外显子组成和各外显子长度对剪接异构体建立数学模型，在测序读段转录本上均匀分布的假设下，利用各外显子上的读段数和接合区读段数求解异构体的表达值．Jiang 等[43]的方法及软件 IsoInfer[52]和 cufflinks[45]都采用了这种思路来实现剪接异构体的表达推断．需要指出的是，某些形式的剪接异构体表达水平在这种方法框架下不可辨识[53].

3.4 新基因的检测

在对 RNA-seq 数据的分析中，人们发现，往往不是所有读段都能定位到已有注释的基因区，说明除了转录噪声或测序错误等的影响外，可能还存在尚未被注释的基因．这里，我们把这种尚未注释的基因称为新基因，包括新的蛋白质编码基因和非编码 RNA 基因．能检测新基因，尤其是低表达基因是 RNA-seq 技术优于基因芯片的特点之一，因为它不需要利用已知基因注释来设计检测探针.

RNA-seq 技术灵敏度高，但样品污染、测序错误等仍可能带来背景噪声．从基因组未注释区域的 RNA 测序读段信号中检测新基因是典型的信号检测问题．如何控制新基因识别的误发现率(FDR)是检测方法的关键．Useq 软件包[54]将 ChIP-seq 数据分析的方法移植到 RNA-seq 数据上，用滑窗的方法来识别测序读段定位富集的区域，给出反映滑窗所在区域读段富集显著程度的 P 值(P-value)及新基因误发现率，通过设定 P 值或误发现率的阈值，可筛选出读段富集的区域，再将相邻区域合并或根据剪接接合区读段将相应区域连接，完成新基因的检测.

3.5 读段的可视化及注释

对于复杂的组学数据，能尽可能方便地直接观察数据对于数据的分析和解释都非常重要，对新一代测序数据的可视化和交互展示是一个非常重要但容易被人忽视的问题．不深入考查数据的细节，而是满足于对数据的统计分析，是高通量数据应用中经常容易陷入的误区，方便有效的可视化工具能够帮助避免这样的误区．表 2 列出了部分适用于 RNA-seq 数据的全基因组浏览器，其中比较具有代表性的有 UCSC Genome Browser、CisGenome Browser 和 IGV(Integrative Genomics Viewer)等．这些浏览器具有如下特点：a．能在不同尺度下显现单个或多个读段在基因组上的位置，包括来源于剪接接合区的读段；b．能在不同尺度下显示不同区域的读段丰度，以反映不同区域的转录水平或测序效率；c．能显示基因及其剪接异构体的注释信息；d．能显示其他注释信息，例如物种间基因组序列保守性、序列 GC 含量等；e．能直接或间接支持 SAM/BAM 读段定位数据存储格式．UCSC Genome Browser[55]属于基于网络模式的全基因组浏览器，所有数据都需要上载到远程服务器，经过处理后将图形返回客户端显示．图 3 中的例子就是从 UCSC Genome Browser 的显示截取的. CisGenome Browser[56]是典型的本地版基因组浏览器，所有读段数据、注释信息都存于本地文件，因此不需要网络连接，方便内部考查数据用．IGV(http://www.broadinstitute.org/igv)可以说是以上两种模式的融合，既可以从远程服务器端下载各种注释信息，又可以从本地加载注释信息.

表 2 适用于 mRNA-seq 数据的全基因组浏览器

Table 2 The genome browsers/viewers applicable to mRNA-seq data viewing

名称	支持的数据格式	网址
IGV	GFF/GFF3, BED, SAM/BAM, WIG,…	http://www.broadinstitute.org/igv
UCSC Genome Browser[55]	BED, bigBed, BEDGRAPH, GFF, GTF, WIG, bigWig, BAM, …	http://genome.ucsc.edu
CisGenome Browser[56]	BAR, BED, refFlat, FA, …	http://biogibbs.stanford.edu/~jiangh/browser
MochiView	Wig, BED, GFF, FASTA, …	http://johnsonlab.ucsf.edu/sj/mochiview-start
SeqMonk	ELAND, GFF, BED, MAQ, SAM	http://www.bioinformatics.bbsrc.ac.uk/projects/seqmonk
Gambit	BAM, BED, GFF2, GFF3, FASTA, VCF	http://code.google.com/p/gambit-viewer
GenomeGraphs[57]	Expression data, Annotation tracks	http://bioconductor.org/packages/release/bioc/html/GenomeGraphs.html

以上列出的全基因组浏览器均可在 Windows、Linux 和苹果公司的 Mac OS 等计算机平台下运行.

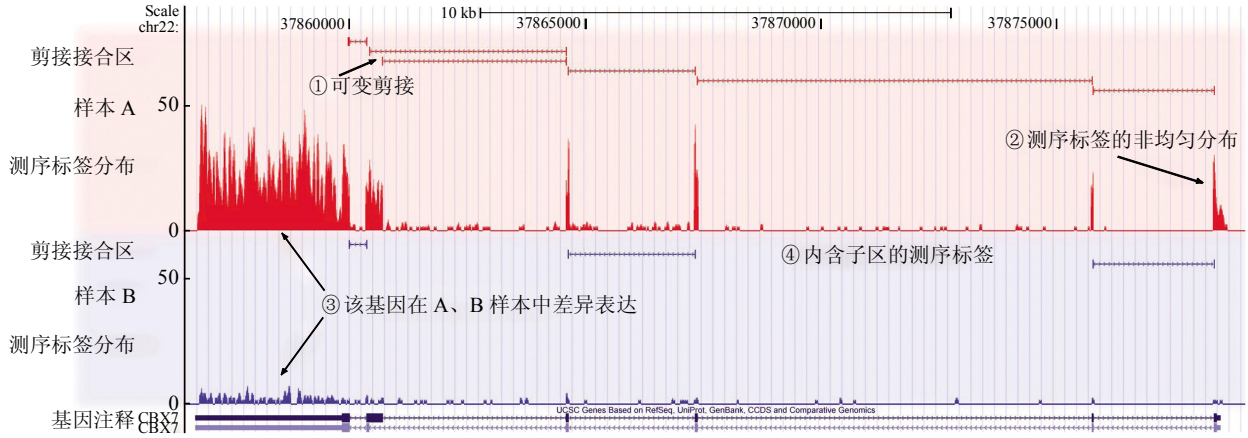

图 3　mRNA-seq 数据可视化示例

Fig. 3　An example for mRNA-seq data visualization

图示区域为人类基因 CBX7. 图中红色表示样本 A 的数据，蓝色表示样本 B. 各轨道(track)依次为: 基因组坐标、样本 A 的剪接接合区、样本 A 的读段分布、样本 B 的剪接接合区、样本 B 的读段分布、UCSC 基因注释. 图中还标识了：①因受体位点不同而形成的选择性剪接; ②基因的 5′端出现读段的非均匀分布; ③在两个样本中, 差异表达基因的读段信号强度不同; ④在基因标注的内含子(intron)区域存在少量不连续的读段.

除对读段的可视化外，用描述统计学方法对实验数据进行分类别统计也十分重要. 例如，统计读段在各个染色体上的分布情况和在注释的外显子、内含子、剪接接合区、基因间区的分布情况等. 目前，已经有一些用于测序数据注释的生物信息学软件，比如 SAMtools[39]、BEDtools[58]等，但由于测序技术发展迅速，用户需求因人而异，用户经常还需要根据需求编写一定的程序或脚本完成或完善注释分析的任务. 对于熟悉图形用户界面的研究人员，还可以利用 UCSC Table Browse[56]和 Galaxy[59-60]来配合完成注释分析. 由于 UCSC Table Browser 集成了大量基因组尺度上的注释信息，而 Galaxy 又为用户提供了书写简单、接口明晰和直观的数据处理流程，这种方法十分方便有效，也为很多学者在展示研究成果时所采用.

以上描述了对于基因组已知的物种进行 RNA-seq 数据处理基本流程，图 2a 给出了其主要步骤. 若研究对象尚未完成基因组测序，则需要采用读段的从头拼装(*de novo* assembly)[6, 61-62]来代替读段定位，后续流程也须做相应的调整. 若 RNA-seq 实验在文库制备时保留了 RNA 的方向信息，则应分别研究来自正链和反链的转录产物，并通过与基因注释比较来检测反义转录本[63]. 最近，RNA-MATE[64]软件在其分析流程中加入如此的处理策略. 此外，通过分析定位到外显子接合区的读段，还可以获取转录本结构，这为研究基因的剪接调控机理提供了重要信息[5]. 而利用 RNA-seq 数据提供的序列信息，通过与 DNA 序列的细致比较可分析转录组的序列差异(如 SNP 等)[65]，从而研究等位基因的表达模式[66-67]及 RNA 编辑[68]等. 最后需要指出，由于 miRNA 在序列和结构上具有一定的特点，miRNA-seq 数据的基本处理流程也与本节所述有所不同，感兴趣的读者可参考软件工具 miRDeep[69]提供的处理策略.

4　多类样本 mRNA-seq 数据间的比较分析

很多 RNA-seq 实验的目的是为了比较两种或多种样本中基因表达或整个转录组的差异，如比较癌症组织和正常组织的转录组差异等. 这些差异既包括通常意义下的差异表达基因，也主要包括选择性剪接模式的差异、剪接异构体表达的差异、非编码转录本的差异等. 这些差异一般可以用一些统计假设检验方法检测，但这种检验有时会受到测序深度、基因长度等因素的影响[70-71]，需要对结果进行仔细分析，消除可能的混杂因素，必要时可以用读段的绝对表达值倍数变化(fold-change)来作为补充. 图 2b 给出了两类样本数据分析的框架.

4.1　差异表达基因的识别

虽然新一代测序相对第一代测序的单位成本大大降低，但是，利用 RNA 测序进行基因表达研究的成本仍很高，因此，很多实验室没有条件进行样本重复．如果两类样本均没有生物重复，例如只对两个细胞系各进行一次 mRNA 样本测序，则可以用随机采样模型通过假设检验来分析差异表达．对于某个基因，如果一个读段来自于这个基因，我们称事件 A 发生．对于一次 RNA-seq 实验，事件 A 发生的概率可以用这个基因上的读段数 n 除以所有基因上的读段总数 N 来估计，即 RPM．事件 A 发生的概率反应了这个基因的表达水平．如果要判断某个基因在两个样本中的表达水平是否一致，就可以通过检验事件 A 在两种条件下发生的概率是否一致来实现，采用似然比检验[1]、Fisher 精确检验[72]以及基于 MA 图的统计检验方法[44]等．同样，也可用 RPKM 作为统计量来进行假设检验分析，由于是比较同一个基因在两个样本间的差异，基因长度的影响被抵消，用 RPKM 和用 RPM 得到的结果相似．对无生物重复的 RNA-seq 数据进行差异表达基因分析，已经有几个公开发表的软件，包括 DEGseq[44]、Useq[54]、Cufflinks[45]中的 Cuffdiff 模块等．图 4 展示了我们开发的 DEGseq 软件提供的多种差异表达基因识别方法的应用例子．

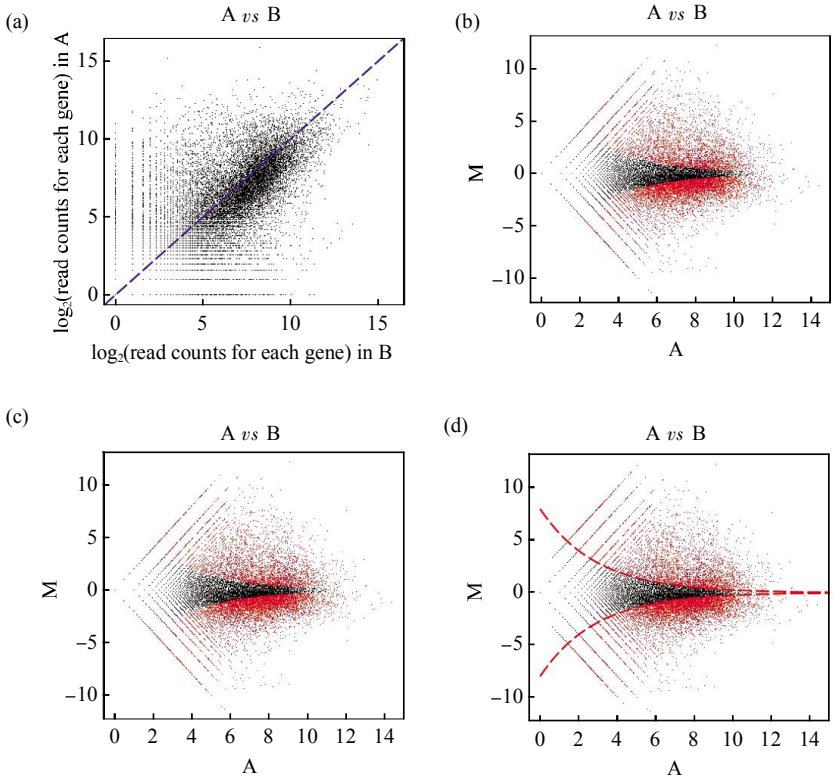

图 4　用 DEGseq 软件包识别差异表达基因的结果

Fig. 4　The results given by DEGseq for differentially expressed gene identification

(a)各基因在样本 A 和样本 B 中表达水平的散点图. (b), (c), (d)图中红点表示分别用 FET、LRT 和 MARS 方法得到的差异表达基因. FET: Fisher's Exact Test, Fisher 精确检验. LRT: Likelihood Ratio Test, 似然比检验. MARS: MA-plot-based method with Random Sampling model, 基于 MA 图的随机采样模型.

如果每一类样本都包含了若干生物重复，如病人和正常人对照研究，则可以沿用基因芯片数据分析中的很多方法．比如，可以用 t 检验结合倍数变化的方法来分析差异表达．如果两类样本具有配对的信息，也可以通过整合每对样本分析结果来实现．其步骤为，先在每对样本中识别出差异表达的基因，再寻找这若干组差异表达基因之间的相同者，或用投票的方法来为基因的差异程度打分．针对某些 RNA-seq 数据生物样本量小，R 软件包 DEGseq[44]和 edgeR[73]等还专门提供了基于改进模型的统计方法．此外，一类将分类器与特征选择包裹在一起的方法也同样适用于此类问题(见 **4.3**)．

4.2 差异表达剪接异构体的识别

差异表达剪接异构体的识别方法与差异表达基因的识别相似．如果把剪接异构体看成是独立的基因，那么前面讨论的用于识别差异表达基因的方法对剪接异构体完全适用．但是，注意到来自于同一个基因的剪接异构体并不独立，某些假设检验的基本条件并不满足，得到的结果就不一定正确．此外，由于现在剪接异构体表达推断的方法还不够成熟，加之在基因结构不可辨识的剪接异构体上作表达推断会出现病态结果，差异表达剪接异构体的识别问题还处于探索的阶段．目前，在剪接异构体表达水平可辨识且读段覆盖度较高的基因上，BASIS[74]方法通过贝叶斯模型来推断差异表达的剪接异构体．

换一个角度，剪接异构体由选择性剪接造成，如果剪接异构体的表达有差异，那么导致这些异构体的选择性剪接事件及异构体特异的外显子的表达也会有差异．因此，对差异表达剪接异构体的识别可以转变为分析选择性剪接事件和外显子表达的差异[75]．外显子表达差异的分析可以完全利用基因表达差异的分析方法．而剪接接合区也可以看成是一个较短的“外显子”(长度一般与测序长度相当)．不过，由于外显子长度较基因的长度短，对应的读段数量较少，差异识别的敏感度会有所下降．Solas 方法[75]就是根据类似的原理，采用统计学假设检验的方法来识别差异表达的剪接异构体．

4.3 对样本的分类分析

通过统计方法识别出来的差异表达基因及剪接异构体能否有效地区别两类样本，可以通过分类分析进一步证实．如果把每个基因(或剪接异构体)的表达值作为特征，则差异表达基因(或剪接异构体)的选取也就是特征筛选的过程．把前面用统计方法等检测出来的差异表达基因(或剪接异构体)用于分类分析，常被称为过滤法．另一类基于分类器的包裹法，例如 R-SVM[76]、SVM-RFE[77]等，可以根据每个特征在分类器中所占的权重来筛选特征，因此也可以用于差异表达基因(或剪接异构体)的识别．分类的性能可以用交叉验证(cross-validation，CV)方法来评估．需要特别注意的是，交叉验证应该包括对特征选择步骤的交叉验证，防止发生信息泄露而导致评估结果过于乐观．具体做法是：将样本按一定的策略分成两份，一份(通常是样本数多的一份)用于特征选取和分类器训练，而用余下的样本进行分类器性能的估计；重复以上步骤多次，就得到交叉验证错误率．必要时还可以用随机置换检验(permutation test)来推断所得错误率的统计显著性[76]．当样本数较小时，可以采用留一法交叉验证(leave-one-out cross-validation，LOOCV)．

4.4 其他高层分析方法

检测差异表达的基因或差异表达异构体是人们认识所研究的生物问题机理的第一步，接下来需要从功能上研究这些差异转录现象的分子机理．这与在基因芯片应用中所面临的是同样的生物学问题，对芯片数据分析结果的后续处理方法，都可以借鉴到测序数据上来．如何进一步地从机理来解释结果，还需结合已知生物学知识进行后续分析．人们对基因芯片得到的基因表达数据进行分析的很多方法都可以用到 RNA-seq 数据上来，比如利用机器学习方法进行分类和特征选择，对差异表达的基因进行 GO(gene ontology)[78]类别富集分析、信号通路富集分析等，一些常用的分析工具包括 GoMiner[79]、DAVID[80]和 VisANT[81]等．

需要说明的是，在各种以差异表达基因为基础的分析中，由于基因表达水平都是通过读段计数来估计的，表达水平较高或转录本较长的基因拥有更多的读段，更容易被多数统计方法识别为差异表达基因[70]．这种偏好可能对后续分析带来影响．以 GO 类别富集分析为例，这种偏好将导致长基因占主导的功能类别更有可能被识别为富集的功能．这将对生物机理的研究带来误导．最近，Young 等[71]发展了一种 GOseq 方法，针对这一偏好对 GO 类别富集分析做了改进．

5 RNA-seq 数据处理中的生物信息学挑战

高通量测序技术的发展十分迅速，这要求相应的数据处理与分析方法快速跟进．正是这些方法，架起了高通量实验数据与科学问题之间的桥梁．这种桥梁作用正日趋重要，也为生物信息学带来了挑战[7, 71]．这里，我们重点讨论两方面的挑战：a．如何实现剪接接合区读段的准确定位？b．在数据处理各阶段中，如何对 RNA-seq 数据的系统误差和固有偏好建模或补偿，以消除它们可能带来的错误推断及结论？

5.1 剪接接合区读段的定位

测序技术的一个发展趋势是测序长度不断增加．随着读长的增加，RNA-seq 中来自剪接接合区的读段会越来越多．我们粗略估算，按照人类基因组 refSeq 基因注释，一般情况下，如果测序读长

为 50 个碱基，则约有 10%的读段来自剪接接合区．而当测序长度达到 100 个碱基时，这个比例将达到 25%左右．对这些剪接接合区读段的分析，将使我们能够更准确地检测剪接事件和推断剪接异构体的表达水平，大大推进人们对选择性剪接的研究．

在 RNA-seq 出现的早期，人们没有意识到剪接接合区读段的重要性．因为当时的读长只有 20～30 个碱基，来自剪接接合区的读段所占比例甚小．当时读段定位的通常做法是，先将读段与全基因组序列做映射定位，再考虑不能定位的读段是否来自于剪接接合区[2]．这种做法虽然在一定程度上保证了读段定位的比率，但由于基因组中重复序列和相似序列的存在，部分接合区读段有可能在容许错配的情况下被定位到基因组上其他位置，从而失去了定位到正确的剪接接合区的机会．

在读段定位时，如果要同时考虑基因组序列和剪接接合区序列，就要利用已知的剪接事件注释，这是目前软件通用的方法．然而，包括人类在内的各物种的基因注释信息都还有待完善，也没有较完整的剪接组(splicome)数据库，能够不依赖注释信息和对剪接机理的现有认识，高效、准确地定位所有已知和未知的接合区读段，仍然是对读段映射定位算法的一个挑战．

5.2 系统噪声和偏好的分析

虽然深度测序技术的准确性较以前的技术有了很大提高，但仍然存在错误和噪声．比如从图 3 中可以看到，内含子区内有一些不连续的读段，很可能由系统噪声造成，如样品污染、测序错误和不恰当的读段定位策略等．从图 3 还能看出，外显子区域内的读段信号分布也很不均匀．有文献报道，序列组成尤其是 GC 含量[46]、RNA 二级结构[2]等也有可能是导致读段不均匀分布的原因．这些噪声和分布偏好将影响新基因的识别和对剪接异构体形式和表达水平的推断．

合理地建模 RNA-seq 数据中的系统噪声和偏好是解决上述问题最有效的办法．基本的思路可以是：首先根据实验原理寻找可能产生系统噪声或偏差的因素，并尽可能将这些因素转化成可量化的特征，如序列特征、二级结构等；然后，将用实验数据对这些特征做统计分析，构造和训练模型，用模型来对数据进行校正．需要注意的是，某些偏好是由当前的测序技术和分析方法共同造成的，难以完全消除[71]．在这种情况下，后续处理和解释时需要充分意识到这种偏好可能对生物学结论带来的影响，必要时通过补充其他实验来验证和修正通过高通量测序得到的生物学结论．

6 总结与展望

本文以 Illumina/Solexa 测序平台为例，尝试对新一代测序技术的 RNA-seq 数据处理和分析方法做了较为全面的梳理，并对各个环节上可用的软件进行了汇总．高通量测序是正在飞速发展的技术，相应的生物信息学方法也在快速发展，这里讨论的是 RNA-seq 中一些代表性的方法和问题，希望能对正在或即将采用 RNA-seq 实验进行科学研究的学者和进行 RNA 测序数据处理的同行提供参考．

RNA 测序和基因芯片有很多共同的应用领域，尽管相对还不是很成熟，RNA-seq 技术在很多方面已经表现出了优势，有人甚至预言基因芯片时代即将结束[36]．但也有报道认为，RNA-seq 数据在基因表达水平的估计上和基因芯片相比没有明显的优势[60]，加上测序的成本目前还远高于芯片实验的成本，所以更多人认为测序和基因芯片将长期共存，以各自不同的特点在现代组学研究中发挥作用．

新一代高通量测序技术的应用面非常广[82]，RNA-seq 只是其中一个方面，除此之外，基因组的从头测序和重测序[83-84]、染色质免疫沉淀测序(ChIP-seq)[85-86]、甲基化测序(Methyl-seq)[87-88]等技术都同样有着广泛的应用．尤其是，用 ChIP-seq 研究蛋白质与 DNA 的相互作用，能够得到高分辨率的转录因子结合数据和组蛋白修饰等表观遗传学数据．发展有效的生物信息学方法，将 ChIP-seq 数据与 RNA-seq 得到的转录组数据进行综合分析，将大大推进人们对复杂的基因转录调控系统的认识．

致谢 感谢本实验室刘霖曦、谢芃、孟璐等同学对本工作有意义的讨论，感谢斯坦福大学 Wing H. Wong 教授、Hui Jiang 博士和 Jun Li 同学等的讨论和帮助．

参 考 文 献

[1] 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**(9): 1509−1517

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

[3] Nagalakshmi U, Wang Z, Waern K, *et al*. The transcriptional landscape of the yeast genome defined by RNA sequencing. Science, 2008, **320**(5881): 1344-1349

[4] Sultan M, Schulz M H, Richard H, *et al*. A global view of gene activity and alternative splicing by deep sequencing of the human transcriptome. Science, 2008, **321**(5891): 956-960

[5] Wang E T, Sandberg R, Luo S, *et al*. Alternative isoform regulation in human tissue transcriptomes. Nature, 2008, **456**(7221): 470-476

[6] Birzele F, Schaub J, Rust W, *et al*. Into the unknown: expression profiling without genome sequence information in CHO by next generation sequencing. Nucleic Acids Res, 2010, doi:10.1093/nar/gkq 116

[7] Sanger F, Nicklen S, Coulson A R. DNA sequencing with chain-terminating inhibitors. Proc Natl Acad Sci USA, 1977, **74** (12): 5463-5467

[8] Margulies M, Egholm M, Altman W E, *et al*. Genome sequencing in microfabricated high-density picolitre reactors. Nature, 2005, **437**(7057): 376-380

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

[10] Ruparel H, Bi L, Li Z, *et al*. Design and synthesis of a 3′-O-allyl photocleavable fluorescent nucleotide as a reversible terminator for DNA sequencing by synthesis. Proc Natl Acad Sci USA, 2005, **102**(17): 5932-5937

[11] Seo T S, Bai X, Kim D H, *et al*. Four-color DNA sequencing by synthesis on a chip using photocleavable fluorescent nucleotides. Proc Natl Acad Sci USA, 2005, **102**(17): 5926-5931

[12] Ju J, Kim D H, Bi L, *et al*. Four-color DNA sequencing by synthesis using cleavable fluorescent nucleotide reversible terminators. Proc Natl Acad Sci USA, 2006, **103**(52): 19635-19640

[13] Fedurco M, Romieu A, Williams S, *et al*. BTA, a novel reagent for DNA attachment on glass and efficient generation of solid-phase amplified DNA colonies. Nucleic Acids Res, 2006, **34**(3): e22

[14] Shendure J A, Porreca G J, Church G M. Overview of DNA sequencing strategies//Ausubel F M, Brent R, Kingston R E, *et al*. Current Protocols in Molecular Biology. USA: John Wiley and Sons, Inc., 2008: Unit 7.1

[15] Mardis E R. Next-generation DNA sequencing methods. Annu Rev Genomics Hum Genet, 2008, **9**: 387-402

[16] Fuller C W, Middendorf L R, Benner S A, *et al*. The challenges of sequencing by synthesis. Nat Biotechnol, 2009, **27**(11): 1013-1023

[17] Croucher N J, Fookes M C, Perkins T T, *et al*. A simple method for directional transcriptome sequencing using Illumina technology. Nucleic Acids Res, 2009, **37**(22): e148

[18] Parkhomchuk D, Borodina T, Amstislavskiy V, *et al*. Transcriptome analysis by strand-specific sequencing of complementary DNA. Nucleic Acids Res, 2009, **37**(18): e123

[19] Perkins T T, Kingsley R A, Fookes M C, *et al*. A strand-specific RNA-Seq analysis of the transcriptome of the typhoid bacillus Salmonella typhi. PLoS Genet, 2009, **5**(7): e1000569

[20] Mamanova L, Andrews R M, James K D, *et al*. FRT-seq: amplification-free, strand-specific transcriptome sequencing. Nat Methods, 2010, **7**(2): 130-132

[21] Cock P J, Fields C J, Goto N, *et al*. The Sanger FASTQ file format for sequences with quality scores, and the Solexa/Illumina FASTQ variants. Nucleic Acids Res, 2010, 38(6): 1767-1771

[22] Shumway M, Cochrane G, Sugawara H. Archiving next generation sequencing data. Nucleic Acids Res, 2009, **38** (Database issue): D870-871

[23] Kaminuma E, Mashima J, Kodama Y, *et al*. DDBJ launches a new archive database with analytical tools for next-generation sequence data. Nucleic Acids Res, 2009, **38**(Database issue): D33-38

[24] Trapnell C, Salzberg S L. How to map billions of short reads onto genomes. Nat Biotechnol, 2009, **27**(5): 455-457

[25] Burrows M, Wheeler D J. A Block Sorting Lossless Data Compression Algorithm [M/OL]. Technical Report 124. Digital Systems Research Center, 1994. http://Citeseerx.ist.psu.edu/viewdoc/summary?(doi=10.1.1.37.6774)

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

[27] 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

[28] Flicek P, Birney E. Sense from sequence reads: methods for alignment and assembly. Nat Methods, 2009, **6**(11 Suppl): S6-S12

[29] Smith T F, Waterman M S. Identification of common molecular subsequences. J Mol Biol, 1981, **147**(1): 195-197

[30] Homer N, Merriman B, Nelson S F. BFAST: an alignment tool for large scale genome resequencing. PLoS One, 2009, **4**(11): e7767

[31] Rumble S M, Lacroute P, Dalca A V, *et al*. SHRiMP: accurate mapping of short color-space reads. PLoS Comput Biol, 2009, **5**(5): e1000386

[32] Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics, 2009, **25**(14): 1754-1760

[33] Lin H, Zhang Z, Zhang M Q, *et al*. ZOOM! Zillions of oligos mapped. Bioinformatics, 2008, **24**(21): 2431-2437

[34] Li R, Yu C, Li Y, *et al*. SOAP2: an improved ultrafast tool for short read alignment. Bioinformatics, 2009, **25**(15): 1966-1967

[35] Weese D, Emde A K, Rausch T, *et al*. RazerS--fast read mapping with sensitivity control. Genome Res, 2009, **19**(9): 1646-1654

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

[37] Denoeud F, Aury J M, Da Silva C, *et al*. Annotating genomes with massive-scale RNA sequencing. Genome Biol, 2008, **9**(12): R175

[38] Li B, Ruotti V, Stewart R M, *et al*. RNA-Seq gene expression estimation with read mapping uncertainty. Bioinformatics, 2010, **26**(4): 493-500

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

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

[41] Shendure J. The beginning of the end for microarrays?. Nat

Methods, 2008, **5**(7): 585−587

[42] ′t Hoen P A, Ariyurek Y, Thygesen H H, *et al*. Deep sequencing-based expression analysis shows major advances in robustness, resolution and inter-lab portability over five microarray platforms. Nucleic Acids Res, 2008, **36**(21): e141

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

[44] 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

[45] 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 Biotechnol, 2010, **28**(5): 511−515

[46] Dohm J C, Lottaz C, Borodina T, *et al*. Substantial biases in ultra-short read data sets from high-throughput DNA sequencing. Nucleic Acids Res, 2008, **36**(16): e105

[47] Li J, Jiang H, Wong W H. Modeling non-uniformity in short-read rates in RNA-Seq data. Genome Biol, 2010, **11**(5): R50

[48] Hansen K D, Brenner S E, Dudoit S. Biases in Illumina transcriptome sequencing caused by random hexamer priming. Nucleic Acids Res, 2010, doi:10.1093/nar/gkq 224

[49] Pan Q, Shai O, Lee L J, *et al*. Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing. Nat Genet, 2008, **40**(12): 1413−1415

[50] Wang L, Xi Y, Yu J, *et al*. A statistical method for the detection of alternative splicing using RNA-seq. PLoS One, 2010, **5**(1): e8529

[51] Blekhman R, Marioni J C, Zumbo P, *et al*. Sex-specific and lineage-specific alternative splicing in primates. Genome Res, 2010, **20**(2): 180−189

[52] Feng J, Li W, Jiang T. Inference of isoforms from short sequence reads//Proc. 14th Annual International Conference on Research in Computational Molecular Biology (RECOMB), Lisbon, Portugal, 2010: 138−157

[53] Hiller D, Jiang H, Xu W, *et al*. Identifiability of isoform deconvolution from junction arrays and RNA-Seq. Bioinformatics, 2009, **25**(23): 3056−3059

[54] Nix D A, Courdy S J, Boucher K M. Empirical methods for controlling false positives and estimating confidence in ChIP-Seq peaks. BMC Bioinformatics, 2008, **9**: 523

[55] Kent W J, Sugnet C W, Furey T S, *et al*. The human genome browser at UCSC. Genome Res, 2002, **12**(6): 996−1006

[56] Ji H, Jiang H, Ma W, *et al*. An integrated software system for analyzing ChIP-chip and ChIP-seq data. Nat Biotechnol, 2008, **26**(11): 1293−1300

[57] Durinck S, Bullard J, Spellman P T, *et al*. GenomeGraphs: integrated genomic data visualization with R. BMC Bioinformatics, 2009, **10**: 2

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

[59] Blankenberg D, Taylor J, Schenck I, *et al*. A framework for collaborative analysis of ENCODE data: making large-scale analyses biologist-friendly. Genome Res, 2007, **17**(6): 960−964

[60] Taylor J, Schenck I, Blankenberg D, *et al*. Using galaxy to perform large-scale interactive data analyses//Baxevanis A D, Stein L D, Stormo G D, *et al*. Current Protocols in Bioinformatics. USA: John Wiley and Sens, Inc., 2007, Unit 10.5

[61] Zerbino D R, Birney E. Velvet: algorithms for de novo short read assembly using de Bruijn graphs. Genome Res, 2008, **18**(5): 821−829

[62] Birol I, Jackman S D, Nielsen C B, *et al*. De novo transcriptome assembly with ABySS. Bioinformatics, 2009, **25**(21): 2872−2877

[63] Katayama S, Tomaru Y, Kasukawa T, *et al*. Antisense transcription in the mammalian transcriptome. Science, 2005, **309**(5740): 1564−1566

[64] Cloonan N, Xu Q, Faulkner G J, *et al*. RNA-MATE: a recursive mapping strategy for high-throughput RNA-sequencing data. Bioinformatics, 2009, **25**(19): 2615−2616

[65] Levin J Z, Berger M F, Adiconis X, *et al*. Targeted next-generation sequencing of a cancer transcriptome enhances detection of sequence variants and novel fusion transcripts. Genome Biol, 2009, **10**(10): R115

[66] Pickrell J K, Marioni J C, Pai A A, *et al*. Understanding mechanisms underlying human gene expression variation with RNA sequencing. Nature, 2010, **464**(7289): 768−772

[67] Montgomery S B, Sammeth M, Gutierrez-Arcelus M, *et al*. Transcriptome genetics using second generation sequencing in a Caucasian population. Nature, 2010, **464**(7289): 773−777

[68] Li J B, Levanon E Y, Yoon J K, *et al*. Genome-wide identification of human RNA editing sites by parallel DNA capturing and sequencing. Science, 2009, **324**(5931): 1210−1213

[69] Friedlander M R, Chen W, Adamidi C, *et al*. Discovering microRNAs from deep sequencing data using miRDeep. Nat Biotechnol, 2008, **26**(4): 407−415

[70] Oshlack A, Wakefield M J. Transcript length bias in RNA-seq data confounds systems biology. Biol Direct, 2009, **4**: 14

[71] Young M D, Wakefield M J, Smyth G K, *et al*. Gene ontology analysis for RNA-seq: accounting for selection bias. Genome Biol, 2010, **11**(2): R14

[72] Bloom J S, Khan Z, Kruglyak L, *et al*. Measuring differential gene expression by short read sequencing: quantitative comparison to 2-channel gene expression microarrays. BMC Genomics, 2009, **10**: 221

[73] 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

[74] Zheng S, Chen L, A hierarchical Bayesian model for comparing transcriptomes at the individual transcript isoform level. Nucleic Acids Res, 2009, **37**(10): e75

[75] Richard H, Schulz M H, Sultan M, *et al*. Prediction of alternative isoforms from exon expression levels in RNA-Seq experiments. Nucleic Acids Res, 2010, doi:10.1093/nar/gkq 041

[76] Zhang X, Lu X, Shi Q, *et al*. Recursive SVM feature selection and sample classification for mass-spectrometry and microarray data.

BMC Bioinformatics, 2006, **7**: 197

[77] Guyon I, Weston J, Barnhill S, *et al*. Gene selection for cancer classification using support vector machines. Machine Learning, 2002, **46**(1): 389-422

[78] Ashburner M, Ball C A, Blake J A, *et al*. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat Genet, 2000, **25**(1): 25-29

[79] Zeeberg B R, Feng W, Wang G, *et al*. GoMiner: a resource for biological interpretation of genomic and proteomic data. Genome Biol, 2003, **4**(4): R28

[80] Dennis G, Jr., Sherman B T, Hosack D A, *et al*. DAVID: Database for annotation, visualization, and integrated discovery. Genome Biol, 2003, **4**(5): P3

[81] Hu Z, Hung J H, Wang Y, *et al*. VisANT 3.5: multi-scale network visualization, analysis and inference based on the gene ontology. Nucleic Acids Res, 2009, **37**(Web Server issue): W115-121

[82] Wold B, Myers R M. Sequence census methods for functional genomics. Nat Methods, 2008, **5**(1): 19-21

[83] Li R, Fan W, Tian G, *et al*. The sequence and de novo assembly of the giant panda genome. Nature, 2010, **463**(7279): 311-317

[84] Wang J, Wang W, Li R, *et al*. The diploid genome sequence of an Asian individual. Nature, 2008, **456**(7218): 60-65

[85] Mardis E R. ChIP-seq: welcome to the new frontier. Nat Methods, 2007, **4**(8): 613-614

[86] Park P J. ChIP-seq: advantages and challenges of a maturing technology. Nat Rev Genet, 2009, **10**(10): 669-680

[87] Brunner A L, Johnson D S, Kim S W, *et al*. Distinct DNA methylation patterns characterize differentiated human embryonic stem cells and developing human fetal liver. Genome Res, 2009, **19**(6): 1044-1056

[88] Deng J, Shoemaker R, Xie B, *et al*. Targeted bisulfite sequencing reveals changes in DNA methylation associated with nuclear reprogramming. Nat Biotechnol, 2009, **27**(4): 353-360

A Review on The Processing and Analysis of Next-generation RNA-seq Data*

WANG Xi[1)], WANG Xiao-Wo[1)], WANG Li-Kun[1,2)], FENG Zhi-Xing[1)], ZHANG Xue-Gong[1)**]

([1)] *Ministry of Education Key Laboratory of Bioinformatics and Bioinformatics Div, TNLIST / Department of Automation, Tsinghua University, Beijing* 100084, China;

[2)] *College of Computer Science and Technology, Jilin University, Changchun* 130012, China)

Abstract With the rapid development of the next-generation sequencing (NGS) technology, high-throughput RNA sequencing or RNA-seq is becoming a key experimental approach in the study of gene expression and transcriptome. The overwhelming amount of RNA-seq data brings new opportunities and challenges for bioinformatics. The efficient and effective processing and analysis of RNA-seq data is becoming the bottleneck for turning the possibilities provided by the new technology into real scientific discovery. A general description of the typical RNA-seq protocol was given. A complete review of major methods and available software in the processing and analysis of RNA-seq data were presented, using the Illumina/Solexa platform as an example. Questions that are still open and awaiting further research are also discussed.

Key words high-throughput RNA sequencing, transcriptome, gene expression, data processing and analysis, bioinformatics

DOI: 10.3724/SP.J.1206.2010.00151

*This work was supported by grants from The National Natural Science Foundation of China (60702002, 60721003, 30873464, 60905013) and The Open Research Fund of State Key Laboratory of Bioelectronics, Southeast University.

**Corresponding author.

Tel: 86-10-62794919, E-mail: zhangxg@tsinghua.edu.cn

Received: March 25, 2010 Accepted: April 30, 2010