昆虫 RNA-Seq 数据的分析流程*

刘金定[1,2] 张 赞[1] 黄水清[2] 李 飞[1**]

(1.南京农业大学植物保护学院 南京 210095;2.南京农业大学信息科学技术学院 南京 210095)

摘 要 随着高通量 RNA 测序(RNA-Seq)技术的发展和测序成本迅速下降,RNA-Seq 技术已经成为生物学研究的重要工具,为生物学家全面地了解和研究转录组提供了机遇。高通量测序具有读长短、存在一定比例的测序错误、数据量大等特点,因此 RNA-Seq 数据分析与基因组分析和传统的 EST 数据分析有所不同。本文通过介绍不同的测序平台、原始数据产生和低质量数据过滤的计算流程,对短序列比对、转录组拼接、功能注释、以及差异表达分析进行了研究和分析,最后对 RNA-Seq 在昆虫学研究中的应用进行了综述,并对 RNA-Seq 技术进行了总结和展望。

关键词 高通量 RNA 测序,段序列比对,转录组拼接,基因功能注释,基因表达定量,基因差异表达

Insect RNA-Seq data analysis pipeline

LIU Jin-Ding[1,2] ZHANG Zan[1] HUANG Shui-Qing[2] LI Fei[1**]

(1. Department of Entomology, College of Plant Protection, Nanjing Agricultural University, Nanjing 210095, China;
2. College of Information Science and Technology, Nanjing Agricultural University, Nanjing 210095, China)

Abstract With the rapid development of high-throughput RNA sequencing (RNA-Seq) technology and the rapidly decreasing cost of this method, RNA-Seq is becoming an important tool for biological research, especially investigating gene function at the transcriptome level. RNA-Seq typically reads sequences rapidly with a certain percentage of sequence errors and bias producing a huge amount of data. RNA-Seq data analysis faces lots of challenges. Here, we describe different RNA sequencing platforms, raw data generation processes and data filtering and introduce short sequence alignment, transcriptome assembly, functional annotation and gene expression analysis. Finally, we briefly review the application of RNA-Seq in insects. The prospects of RNA-Seq techniques and their application are also discussed.

Key words high-throughput RNA sequencing, short sequence alignment, transcriptome reconstruction, gene function annotation, gene expression quantification, gene differential expression

利用新一代高通量测序技术,生物学家可在短时间内以较低成本获得大量的基因序列,开创了生物数据获取的新局面。除基因组测序外,目前高通量测序技术还广泛应用于转录组研究中。理论上,转录组测序可获取某一物种特定器官或者组织在某个特定状态下的所有转录本序列以及表达量数据,对从组学水平研究生物学问题具有十分重要的推动作用。本文对高通量 RNA 测序(即 RNA-Seq)的数据产生、转录组拼接、功能注释以及表达分析等各个方面进行介绍,简要综述 RNA-Seq 在昆虫学研究中的应用,对 RNA-Seq 技术进行了总结和展望,为 RNA-Seq 在昆虫学领域中的应用提供参考。

1 高通量 RNA 测序

1.1 高通量测序平台

目前,Roche 公司 454 技术、Illumina 公司 Solexa 技术和 ABI 公司 SOLiD 技术是三大主流测序技术,占据绝大部分市场份额。其中,Illumina Solexa 技术平台性价比高,广泛受到科研人员青睐。2012 年 2 月份统计数据表明,Illumina 占据 80% 左右市场份额。Illumina Solexa 测序技术的基

* 资助项目:国家自然科学基金(31171843);国家高技术研究发展计划("863"计划)(2012AA101505)。

** 通讯作者,E-mail:lifei@njau.edu.cn

收稿日期:2013-07-19,接受日期:2013-08-03

本原理是边合成边测序(sequencing by synthesis, SBS)(Ruparel *et al.*,2005;Seo *et al.*,2005)。为保证准确性,Illumina Solexa 的早期平台只能产生 20~30 bp 的读长。近几年,相继推出了 GA IIX、HiSeq 和 MiSeq 等系列测序仪,测序读长也得到了明显增加。HiSeq2000 于 2010 年推出,其测序通量达到 600 Gb/run,一次运行可以独立测试 16 个样,HiSeq2000 实现了当时行业内最高的测序产量和最快的数据产生速率。2012 年推出的 HiSeq2500 的测序通量进一步提高,可以在 1 d 内完成个人基因组的重测序。MiSeq 于 2011 年推出,属于个人型测序平台,在单个仪器上实现了扩增、测序和数据分析的一体化,是快速、经济高效的遗传分析平台。Roche 454、ABI SOLiD 等其他平台可参考相关文献(Mardis,2008;Fuller *et al.*,2009;Shendure *et al.*,2011)。

1.2　RNA-Seq 测序

以 Illumina Solexa 技术平台为例,简要阐述 RNA-Seq 测序数据的产生过程以及原始测序数据质量过滤的流程。在 Illumina 测序平台上进行 RNA-Seq 测序通常需要经过以下几个步骤:(1)RNA 提取和 RNA 纯化;(2)RNA 反转录为双链 cDNA 序列;(3)序列片段化处理;(4)增加测序接头;(5)PCR 扩增及序列片段选择;(6)测序产生序列文件。具体的测序过程可以参见文献(Wilhelm *et al.*,2010)。在 RNA-Seq 中有两个过程应该注意:(1)RNA 提取和纯化。在提取总 RNA 后,需根据测序目的对总 RNA 进行预处理,在进行 mRNA-Seq 时,需利用 Poly(T)提取总 RNA 中的信使 RNA(mRNA),构建 mRNA 测序文库;如果研究对象为小 RNA,需从总 RNA 中分离出长度小于 200 bp 的小片断 RNA,构成小 RNA 文库。(2)RNA 反转录时的引物设计。在进行 cDNA 反转录时,可以设计两种引物:poly(dT)引物和随机引物。采用 poly(dT)引物可以进一步富集 mRNA,但不利于基因 5′端序列的反转录;采用随机引物进行反转录,虽然不再进一步富集 mRNA,但可以提高 RNA 序列反转录的均一性。

RNA-Seq 测序产生的原始数据为图像文件,其数据总量可以达到太字节(TB)大小。利用软件可以从图像中获得核苷酸的信号强度,推测 RNA 序列中的碱基,最终得到 FASTQ 格式的序列文件。在 FASTQ 格式文件中每条 read 序列(测序仪产生的读段序列)对应 4 行,第 1、3 行对应序列标识,第 2、4 行对应碱基序列和测序质量得分,见图 1。碱基序列由 A、C、G、T、N 5 个字母的组合,得分序列对应各种碱基的测序质量。在碱基序列中,N 表示不能确定为哪种具体碱基的字母。得分采用压缩的 ASCII 字符形式表示,ASCII 值越大,则表示测序结果越可靠,FASTQ 格式的说明可参考文献(Mamanova *et al.*,2010)。得到 FASTQ 格式文件后,通常需要进行质量过滤,去除低质量的 reads 序列,以提高后续分析数据的可靠性,降低计算时间。常用的质量过滤标准有:(1)read 序列中各个碱基的平均得分低于某个阈值;(2)N 字母在 read 序列中出现的位置过于靠前。质量过滤标准可根据研究目的而改变,提高或放宽标准的阈值(Wilhelm *et al.*,2010)。对 FASTQ 格式文件处理有专用的软件包,如 fastx_toolkit(http://hannonlab.cshl.edu/fastx_toolkit/)。FASTQ 格式的文件可以被压缩为 SRA(Sequence Read Archive)格式文件,存放在共享数据库中(Kaminuma *et al.*,2010;Shumway *et al.*,2010)。

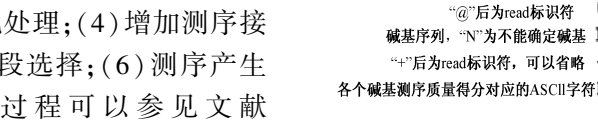

图 1　FASTQ 格式结构

Fig. 1　FASTQ format

2　RNA-Seq 数据分析

利用 RNA-Seq 测序平台获得大量测序数据后,结合研究目的,选择相关分析工具对测序数据进行分析。RNA-Seq 工作在样品准备和测序上花费的时间相对较少,通常 1 周左右就可以完成,而且技术也相对成熟,因此难度不大。RNA-Seq 工作的重点是下游数据分析,时间消耗较长,需要花费几周甚至几个月时间。RNA-Seq 的数据分析主要有:转录组拼接、转录组功能注释,差异表达分析等。其中,转录组拼接是所有后续数据分析工作的前提。常见的 RNA-Seq 数据分析流程如图 2 所示。由于 RNA-Seq 产生的数据量非常巨大,少则几百万条短序列,多则上亿条短序列,对这些序列进行分析需要非常有效的计算方法做支撑,本

文将重点围绕这 4 个方面对相关计算方法和分析工具进行阐述。

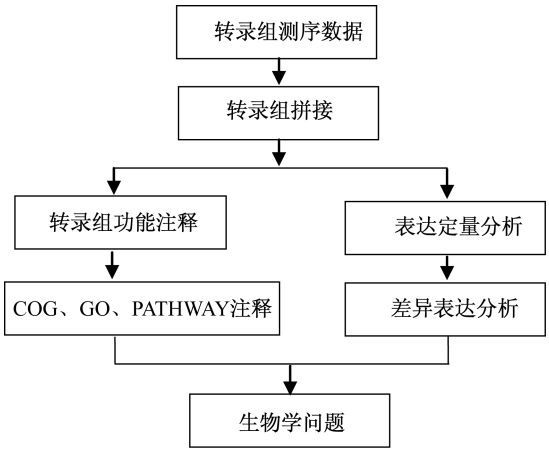

图 2 RNA-Seq 数据分析流程

Fig. 2 The pipeline of RNA-Seq data analysis

2.1 RNA-Seq read 序列比对

短序列比对是把 RNA-Seq 测序获得的 read 序列比对到已知的转录组或基因组上,是基于基因组的转录组拼接、基因表达定量及差异表达分析的重要依据(Mortazavi *et al.*, 2008; Wang *et al.*, 2008; Cloonan *et al.*, 2009; Berger *et al.*, 2010; Griffith *et al.*, 2010)。短序列比对和 EST 比对的流程相似(Kent, 2002; Wu and Watanabe, 2005),但计算复杂度更高,主要有以下 4 个方面原因:(1) RNA-Seq 测序获得的序列较短;(2) RNA-Seq 存在一定的测序错误;(3)存在跨外显子的 read,不能连续比对;(4) RNA-Seq 测序获得的序列数量非常巨大。根据 read 序列是否跨外显子,短序列比对可以分为 un-spliced read alignment 和 spliced read alignment,典型的比对算法工具见表格 1。

表 1 短序列比对工具

Table 1 RNA-Seq read analysis tools

比对类型 Alignment type	软件包 Software packages	网址 Websites
Unspliced	MAQ	http://maq.sourceforge.net
	Stampy	http://www.well.ox.ac.uk/project-stampy
	BWA	http://bio-bwa.sourceforge.net/
	Bowtie	http://bowtie.cbcb.umd.edu
Spliced	SpliceMap	http://www.stanford.edu/group/wonglab/SpliceMap/
	MapSplice	http://www.netlab.uky.edu/p/bioinfo/MapSpliceManual
	Tophat	http://tophat.cbcb.umd.edu/
	GSNAP	http://research-pub.gene.com/gmap/
	QPALMA	http://www2.fml.tuebingen.mpg.de/raetsch/software/qpalma.html

Un-spliced read alignment 是指把来自同一个外显子的 read 序列比对到基因组中,主要有空位种子索引法(spaced-seed indexing)和 Burrows-Wheeler 转换法(Burrows-Wheeler transform, BWT)两大类。空位种子索引法假设每个 read 中至少有一个种子和参考序列完全匹配,通过种子缩小比对序列范围,然后在缩小范围的比对序列中进行扩展,使得 read 序列能够完整比对(Jiang and Wong, 2008; Li *et al.*, 2008a, 2008b; Smith *et al.*, 2008; Homer *et al.*, 2009; Rumble *et al.*, 2009; Rizk and Lavenier, 2010; Lunter and Goodson, 2011),经典的比对工具有 MAQ(Li *et al.*, 2008a), Stampy (Lunter and Goodson, 2011)。Burrows-Wheeler 转换法是把参考序列按照一定规则压缩并建立索引,在通过查找和回溯的方法来定位读段(Langmead *et al.*, 2009; Li and Durbin, 2009; Li *et al.*, 2009b),典型的算法有 BWA (Li and Durbin, 2009), Bowtie(Langmead *et al.*, 2009)。一般而言 Burrows-Wheeler 转换法比空位种子索引法的计算速度更快,但是灵敏度较低(Degner *et al.*,

2009)。

Spliced read alignment 是指把跨外显子的 read 序列比对到参考基因组中,这种比对通常用于真核生物的 RNA-Seq 测序分析中,在基因组上定位跨外显子的 reads 可提高表达定量分析的可靠性,而且有助于转录本的基因结构识别(Black,2003; Guttman *et al.*,2010)。在基因组上定位跨外显子 reads 的方法通常分为两大类:"外显子优先法"和"种子扩充法"。外显子优先法通常分为两步执行:首先利用 Un-spliced read alignment 算法把 reads 比对到基因组上;然后把不能比对到基因组上的 reads 分割成更短的片段进行单独比对。该方法第一步确定基因组上转录区域,大大缩小的后续短片段比对的范围。然后,利用内含子剪接信号特征(供体、受体位置的碱基构成)快速定位跨越外显子的 reads,确定外显子边界。典型的外显子优先算法有 MapSplice(Wang *et al.*,2010a)、SpliceMap(Au *et al.*,2010)和 TopHat(Trapnell *et al.*,2009)等。种子扩充法(De Bona *et al.*,2008;Wu and Nacu,2010)把 reads 分割成更短的种子序列,定位到基因组上,缩小了参考序列比对的范围,然后利用更为灵敏的方法去检查、扩充、合并比对区域,最终确定比对位置。典型的种子扩充法算法有 GSNAP(Wu and Nacu,2010)和 QPALMA(De Bona *et al.*,2008)。由于假基因的存在,跨外显子的 reads 往往会因为外显子优先的原则被定位到假基因上,因此外显子优先算法获得的拼接位点数量往往少于种子扩充法,但是外显子优先算法的计算速度更快。

由于 reads 序列数量巨大,因此短 reads 序列比对输出结果占用空间很大,存储、处理以及检索输出结果都不方便。目前短 reads 序列比对结果通常以 SAM(Sequence Alignment/Map)格式或者压缩的二进制 BAM 格式来存储(Li *et al.*,2009a)。对 SAM 或者 BAM 格式的输出结果处理可采用 samtools(http://samtools.sourceforge.net)工具辅助操作。

2.2　转录组拼接

转录组拼接是从 RNA-Seq 测序获得的短 reads 序列拼接出转录本。目前转录组拼接方法可分为基因组导向法和从头组装法。基因组导向法依赖于 reads 在参考基因组上的位置信息去拼接转录本,而从头组装法则依赖于 reads 之间的重叠区进行序列拼接(Zerbino and Birney,2008; Guttman *et al.*,2010; Martin *et al.*,2010; Robertson *et al.*,2010; Trapnell *et al.*,2010; Garber *et al.*,2011; Grabherr *et al.*,2011)。

基因组导向法把 RNA-Seq 测序的 reads 序列通过前述方法比对到基因组上,获得基因组上的转录区域,然后利用不能连续比对到单个基因组区域的 reads 在转录区域间建立联系,构建一个由转录区域组成的有向图,最后获得转录本被归约为求解图上的路径。算法构建图的数量对应着基因座的数量,图上的路径对应基因座上的一个转录本。典型的两个算法有 Cufflinks(Trapnell *et al.*,2010)和 Scripture(Guttman *et al.*,2010)。这两种算法均利用拼接比对的 reads 位置信息直接组装转录本,但两者有区别:Cufflinks 求解转录本时,用最少数量的路径覆盖整个图,而 Scripture 则求解出图上所有的路径。因此,Scriputre 拼接的转录本数量比 Cufflinks 多,假阳性偏高,但灵敏度很高,理论上单拷贝转录本也可以被发现。另一个基因组导向法算法是 G. Mor. Se(Denoeud *et al.*,2008),该算法利用 read 比对外显子的区域,然后结合 reads 的位置信息把外显子连接起来构建转录本。G. Mor. Se 算法比较适合于早期 RNA-Seq 测序序列较短的实验,此算法拼接出来的转录本通常不完整,目前很少使用。

从头组装法利用 reads 序列间的重叠区直接拼接获得转录本。由于获得 reads 之间的重叠关系会随着 reads 数量增加而快速增长,导致计算量成指数级上升。目前处理这类问题一般是利用 de Bruijn 图求解(Pevzner,1989; Zerbino and Birney,2008; Surget-Groba and Montoya-Burgos,2010)。算法基本思想为:首先把上百万的 reads 序列归约为固定数量 k-mers,然后利用 k-mers 之间的 k-1 个字母重叠性构建出若干 de Bruijn 图,最后在 de Bruijn 图上求解贯穿图的路径。每个 de Bruijn 图对应一个基因,在 de Bruijn 图上求解的路径对应转录本。在 de Bruijn 图上求解转录本时,不会盲目列举路径,通常会根据路径上 read 和两端序列(PE)reads 覆盖情况确定转录本。典型的从头组装算法有 transAbyss(Robertson *et al.*,2010), Trinity(Grabherr *et al.*,2011), Velvet(Zerbino and Birney,2008), CAP3(Huang and Madan,1999),

Newbler(Margulies *et al.*,2005)等。

基因组导向法仅针对基因组序列已知的物种进行转录组拼接,其拼接转录组的质量很大程度上取决于基因组的质量。如果参考基因组不完整,必然导致拼接的转录组不完整。从头组装法则适合于任意物种的转录组拼接,但当测序错误和物种个体之间的 SNP 等都会对从头组装结果产生不利影响,降低转录组拼接的可靠性。两种拼接方法各有优缺点,在分析过程中应根据实验目的做出具体选择,比如对一个具有不完整基因组的物种进行转录组拼接,采用两种方法相结合是一个不错的选择。常见的转录组组装工具见表 2。

表 2 常见转录组组装工具

Table 2 The softwares used for transcriptome assembly

软件包 Software packages	网址 Websites
transAbyss	http://www.bcgsc.ca/platform/bioinfo/software/trans-abyss
Trinity	http://trinityrnaseq.sourceforge.net/
Velvet	http://www.ebi.ac.uk/~zerbino/velvet/
CAP3	http://seq.cs.iastate.edu/
Newbler	http://454.com/products/analysis-software/index.asp
Scripture	http://www.broadinstitute.org/software/scripture/
Cufflinks	http://cufflinks.cbcb.umd.edu/
G. Mor. Se	http://genomebiology.com/content/9/12/R175#B45

2.3 蛋白功能注释

蛋白功能注释是指通过同源比对的方法将转录组序列进行功能注释。同源比对进行功能注释的理论依据是“序列决定结构、结构决定功能”,主要涉及 3 个方面:(1)比对程序;(2)参考序列数据库;(3)相关参数。同源比对算法很多,其中 blast(Altschul *et al.*,1990)是转录组功能注释中最常用的程序。可用的参考序列数据库有 NCBI 的非冗余蛋白数据库(non-redundant protein database,nr database)、Swiss-Prot 等。两个数据库各有优缺点:nr 数据库中的序列数量多,但是注释准确性不如 Swiss-Prot 可靠;Swiss-Prot 数据库中序列少,但是功能注释可靠性高。用公共数据库对转录组进行功能注释时,一般用 1E-5 作为 E-VALUE 参数的阈值,然后取最佳比对序列的功能作为转录组序列的功能注释。由于公共数据库中序列总数非常大,因此选择最优命中的结果作为序列功能注释结果一般可以接受。如果自定义参考数据库,那么对注释结果要审慎。主要原因是因为自定义参考数据库中序列总数少,真正同源的蛋白序列可能不在参考数据库中,即使最佳命中的参考序列对应的 E-VALUE 值低于设定阈值,注释结果仍需要进一步确认。因此,在自定义参考数据库时,对注释结果不仅要考虑 E-VALUE 值,还要评估其他参数,如比对长度、得分等,甚至还要观察注释数据库的大小。

除了对每条转录本进行功能注释之外,对转录组的注释还有 COG 注释、GO 注释以及 PATHWAY 注释。COG 注释可以根据基因功能和进化关系对转录组中的序列进行分类,进而可以宏观地认识和比较物种的转录组构成。COG 注释可以通过在线服务(http://www.ncbi.nlm.nih.gov/COG/)实现,也可以下载 COG 数据库,利用 blast 在本地实现(Tatusov *et al.*,1997;Koonin *et al.*,1998)。GO 注释指的是从分子功能、生物学过程以及细胞组成 3 个方面对基因进行注释。GO 注释词汇来自于基因本体联合会所建立的数据库,保证了对基因注释结果的一致性。目前广泛使用的 GO 注释软件有 BLAST2GO(Conesa *et al.*,2005)、GOstat(Beissbarth and Speed,2004)和 Annot8r(Schmid and Blaxter,2008)等。PATHWAY 注释主要指将转录组包含的生物代谢

或调控信息构建成网络,从而进行基因间的相互作用分析。目前进行 pathway 分析的数据库主要有 KEGG(Ogata *et al.*,1998;Kanehisa,2002),PID(Schaefer *et al.*,2009),BioCyc(Karp *et al.*,2005)等。以 KEGG 为例,KEGG 保存了可靠的参考信号通路。根据转录组的序列相似性把转录组比对到参考信号通路上,从而构建转录组所包含的 pathway。这种方法获得的代谢路径只能提供参考,对预测的信号通路还需生物学验证。详细的 COG 注释、GO 注释以及 PATHWAY 注释可以参考文献(Tatusov *et al.*,1997;Koonin *et al.*,1998;Ogata *et al.*,1998;Kanehisa,2002;Karp *et al.*,2005;Beissbarth and Speed,2004;Conesa *et al.*,2005;Schmid and Blaxter,2008;Schaefer *et al.*,2009)和相关工具网站。

2.4　表达定量和表达差异分析

2.4.1　表达定量　基因芯片技术具有通量高、成本低的特点,曾广泛地应用于生物学研究。但基因芯片技术有两个明显的缺点(Wang *et al.*,2009;Marguerat and Bahler,2010):1)依赖已知的序列构建芯片,不能检测已知序列之外的基因表达量;(2)依赖荧光信号,导致低表达的基因难以检测。RNA-Seq 的基因表达分析技术是基于对 read 的计数,对低表达的基因也能够检测,具有灵敏度高、分辨率高、无饱和区等优点(Hoen *et al.*,2008;Shendure,2008;Wang *et al.*,2009)。

RNA-Seq 用于表达定量研究应考虑两个系统差异:1)read 数量随基因长度不同而不同;2)RNA-Seq 测序深度不同,测序获得的 reads 总数也不同。通常采用 RPKM 指标来消除这两种系统差异。RPKM(Reads Per KilobasesPer Million reads)指的是每 1 百万个 reads 中,比对到每 1 kb 碱基外显子上的 reads 数。当 reads 来自于 PE 测序数据时,类似的归一化方法有 FPKM(Fragments Per Kilobase Per million reads)(Mortazavi *et al.*,2008;Trapnell *et al.*,2010)。经归一化处理后,转录本表达量将不会受到测序深度和基因长度的影响,从而使得不同长度、不同测序深度下的基因表达量具有可比性。

当一个基因由于选择性剪接具有多个转录本或者存在基因重复时,部分 reads 可能被比对到多个转录本或者多个基因上,这种比对的不确定性会影响表达量结果分析的准确性。可供选择的方法是用只比对到一个参考位置的 reads 数量计算表达量,典型的算法有 Alexa-seq(Griffith *et al.*,2010)。这种表达定量算法虽然在一定程度上解决了 reads 归属不确定性的问题,但是对那些没有特有外显子(或者外显子片段)的转录本而言,这种方法则无法确定该转录本的表达量。比如,一个基因有 3 个外显子,由于选择性剪接具有 3 个转录本:转录本 1 由外显子 1、2 拼接得到,转录本 2 由外显子 1、3 拼接得到,而转录本 3 由外显子 2、3 拼接得到。在此情况下,来自该基因的任意 read 都不会唯一比对到某个转录本上,Alexa-seq 算法则无法评估这 3 个转录本的表达量。另一类表达量计算方法采用似然函数估计法处理 reads 不确定性的问题,典型的算法有 Cufflinks(Trapnell *et al.*,2010),MISO(Katz *et al.*,2010)和 RSEM(Wang *et al.*,2010c)。但是,由于这类算法基于统计分析方法,当基因表达量过低时会影响基因表达量的估计准确性。

2.4.2　差异表达分析　基因表达差分析是指找出不同时间点、不同组织或者不同处理条件下具有差异表达的基因。利用基因芯片技术进行表达差异分析时,基因芯片上荧光信号通常被转换成服从正太分布的随机变量加以分析。由于 RNA-Seq 技术基于 read 计数,受测序错误、测序偏好性以及基因长度的影响,因此基于基因芯片技术的统计分析方法不适合于 RNA-Seq 分析。

早期,利用 RNA-Seq 数据进行基因表达差异分析时,通常采用泊松分布模型去统计分析测序获得的 reads。泊松分布模型在多个 RNA-Seq 实验上得到检验(Marioni *et al.*,2008;Bullard *et al.*,2010;Wang *et al.*,2010b),尤其适合于 Illumina GA 测序产生的数据分析,典型的软件有 DEGseq(Wang *et al.*,2010b)。虽然泊松分布模型在技术重复时表现了很好的鲁棒性,但在分析生物学重复数据时,分析结果并不理想,检测的差异表达基因假阳性偏高,其主要原因是建库取样差异性所致(Langmead *et al.*,2010;Oshlack *et al.*,2010)。理论上,通过更多的生物学重复可以把取样错误率估算,但是需要较高的 RNA-Seq 实验成本。为了克服生物学重复不足导致差异表达分析准确性下降的问题,另一些算法采用负二项分布模型代替泊松分布模型,同时采用多参数估计的方法进

行表达差异分析,在很大程度上改善了生物学重复不足带来的误差(Anders and Huber, 2010; Robinson *et al.*, 2010; Trapnell *et al.*, 2010; Wang *et al.*, 2010b),典型的分析软件有 EdgeR(Robinson *et al.*, 2010)、DESeq(Wang *et al.*, 2010b)、Cuffdiff(Anders and Huber, 2010)等,常见的基因差异表达分析软件工具见表 3。另外,BitSeq 采用贝叶斯推断方法解决 reads 不确定性的问题,可应用于技术重复和生物学重复的 RNA-seq 数据分析中(Glaus *et al.*, 2012)。

表 3 常见表达定量和差异表达分析软件

Table 3 Common used software for gene expression analysis

表达分析类型 Expression analysis type	软件包 Software packages	网址 Websites
定量分析 Expression quantification analysis	Alexa-seq	http://www.alexaplatform.org/alexa_seq/
	Cufflinks	http://cufflinks.cbcb.umd.edu/manual.html
	MISO	http://genes.mit.edu/burgelab/miso/
	RSEM	http://deweylab.biostat.wisc.edu/rsem/
差异表达分析 Differential expression analysis	EdgeR	http://www.bioconductor.org/packages/release/bioc/html/edgeR.html
	DESeq	http://bioconductor.org/packages/release/bioc/html/DESeq.html
	Cuffdiff	http://cufflinks.cbcb.umd.edu/
	BitSeq	http://code.google.com/p/bitseq/downloads/list

2.5 RNA-Seq 数据分析工具的选择和使用

在每一个 RNA-Seq 数据分析步骤中,都有多个分析软件可供选择。这些软件都有其自身的优点和缺点,甚至和 RNA-Seq 数据产生的技术流程有紧密的关联性。因此,在进行 RNA-Seq 数据分析时,对软件的选择要慎重,要充分了解软件特点和适用对象。如果软件分析工具选择不当,必然会导致后续分析结果不可靠。此外,各个分析软件往往有多个参数,这些参数值的可以分为必选参数和优化参数。必选参数指的是用户必须提供的基本参数,如输入输出文件名以及文件格式等参数。而优化参数用于指导软件运行,体现使用者对软件各种性能指标的要求。一般而言,软件分析工具的优化参数均有默认参数,为软件开发者在综合考虑各种性能指标后的设置。在实际使用过程中,这些优化参数的设置依赖于经验。在不具有足够经验的情况下,可直接选择默认参数值。熟悉软件特点的情况下,可通过对小规模数据进行多次模拟,评估执行结果再选择最优参数值。

3 RNA-Seq 在昆虫学研究中的重要应用

RNA-Seq 正越来越多地应用于昆虫学研究中,推动着昆虫学的发展,其应用主要包括以下 3 个方面。

(1)提供表达证据,完善昆虫基因组注释。RNA-Seq 可以捕捉细胞内任意表达水平的基因,获得的序列数据是基因表达的最直接证据。不同时期、不同组织的 RNA-Seq 序列数据比对到基因组上,可以完善基因组的注释,比如修正基因的结构、发现新的基因、识别新的选择性剪接等。例如,利用 RNA-Seq 数据比对到黑腹果蝇基因组上,根据比对结果修正了 30% 的黑腹果蝇基因结构,新发现了 319 个转录本,识别了大量可变剪接事件(Daines *et al.*, 2011)。在锥虫基因组测序完成 5 年后,再次利用 RNA-seq 数据对锥虫基因组注释结果进行修订完善(Kolev *et al.*, 2010)。早在 2004 年,家蚕就完成了基因组测序工作,2012 年利用 RNA-Seq 数据纠正了 1 140 个已知的基因结构,发现了几千个新的可变剪接。RNA-Seq 还可以直接用于新测序基因组的基因注释中,如小菜蛾、帝王蝶、二化螟等昆虫基因组的注释(Zhan

et al.,2011; You *et al.*,2013)。结果表明,RNA-Seq 数据作为基因表达证据用于基因组注释,可以提高注释结果的可靠性。

(2)提供序列资源,促进昆虫学研究。目前只有 43 种昆虫的基因组已测序发表(如蝇、蚊子、家蚕等)。对于大部分昆虫而言,基因组测序仍尚未能完成。转录组虽然不如基因组包含的遗传信息丰富、完整,但 RNA-Seq 获取转录组的成本低、效率高,是目前获得基因序列的最有效方式。通过 RNA-Seq 获得大量群居蝗虫和独居蝗虫的转录组数据,发现了多个与发育相关的基因以及一些重要的代谢通路,为蝗虫防治研究奠定了重要的数据基础(Chen *et al.*,2010)。二化螟中肠转录组数据公布后改善了其基因数量严重不足的状况,为二化螟解毒代谢研究提供了数据资源(Ma *et al.*,2012)。在昆虫进化研究方面,利用来自几个物种的直系同源基因序列,通过进化分析确定物种之间的进化关系可以大大提高进化分析结果的可靠性。例如,Jimenez-Guri 等(2013)利用转录组中的直系同源基因进行了进化分析,成功地为几个非果蝇的双翅目昆虫确定了进化关系。

(3)差异表达分析,预测基因功能。通过分析不同处理、不同条件下、不同组织间的基因表达差异性,可以将那些显著差异表达的基因与某些生物学功能关联起来,从而为深入研究昆虫体内相关的分子生物学机制奠定基础。例如,分别对褐飞虱和稻纵卷叶螟的不同发育阶段建库并进行 RNA-Seq 测序,进行差异表达分析,获得了大量与发育阶段相关且具有显著表达差异的基因(Xue *et al.*,2010;Li *et al.*,2012),这些成果为害虫防治研究提供了重要的参考。2011 年,Bonizzoni 等对埃及伊蚊吸血之后不同时间段的基因表达量进行了分析,发现与吸血后的生化反应相关的调控基因,这在疾病传播控制研究中具有十分重要的应用价值。

4　总结和展望

RNA-Seq 为转录组学研究提供了一个很好的契机,不仅提供了转录组序列的全貌,而且提供了基因表达信息,正逐步成为昆虫分子生物学研究的主要工具。美国科学家发起了 ik5 计划,拟开展 5 000 种昆虫的基因组测序项目,丰富昆虫基因资源。但由于基因组测序成本高,昆虫种类繁多,RNA-Seq 仍然是获取基因资源的主要方式。随着 RNA-Seq 技术进一步发展,其测序读长和测序准确性将会得到改善,可以更深入具体地研究昆虫的各种复杂生命现象,对害虫防治和资源昆虫利用起到重要的推动作用。此外,由于 RNA-Seq 数据分析工具和产生的数据特征紧密相关,因此在 RNA-Seq 技术不断成熟并被广泛应用时,需要更多针对昆虫基因数据进行优化的分析软件,将给生物信息学研究者提供广阔的空间。

参考文献(References)

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ, 1990. Basic local alignment search tool. *J. Mol. Biol.*, 215(3):403 – 410.

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

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

Beissbarth T, Speed TP, 2004. GOstat: find statistically overrepresented Gene Ontologies within a group of genes. *Bioinformatics*, 20(9):1464 – 1465.

Berger MF, Levin JZ, Vijayendran K, Sivachenko A, Adiconis X, Maguire J, Johnson LA, Robinson J, Verhaak RG, Sougnez C, Onofrio RC, Ziaugra L, Cibulskis K, Laine E, Barretina J, Winckler W, Fisher DE, Getz G, Meyerson M, Jaffe DB, Gabriel SB, Lander ES, Dummer R, Gnirke A, Nusbaum C, Garraway LA, 2010. Integrative analysis of the melanoma transcriptome. *Genome Res.*, 20(4):413 – 427.

Black DL, 2003. Mechanisms of alternative pre-messenger RNA splicing. *Annu. Rev. Biochem.*, 72:291 – 336.

Bonizzoni M, Dunn WA, Campbell CL, Olson KE, Dimon MT, Marinotti O, James AA, 2011. RNA-seq analyses of blood-induced changes in gene expression in the mosquito vector species, *Aedes aegypti*. *BMC Genomics*, 12:82.

Bullard JH, Purdom E, Hansen KD, Dudoit S, 2010. Evaluation of statistical methods for normalization and differential expression in mRNA-Seq experiments. *BMC Bioinformatics*, 11:94.

Chen S, Yang P, Jiang F, Wei Y, Ma Z, Kang L, 2010. De novo analysis of transcriptome dynamics in the migratory locust during the development of phase traits. *PLoS ONE*, 5(12):e15633.

Cloonan N, Xu Q, Faulkner GJ, Taylor DF, Tang DT, Kolle

G, Grimmond SM, 2009. RNA-MATE: a recursive mapping strategy for high-throughput RNA-sequencing data. *Bioinformatics*, 25(19): 2615 – 2616.

Conesa A, Gotz S, Garcia-Gomez JM, Terol J, Talon M, Robles M, 2005. Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. *Bioinformatics*, 21(18): 3674 – 3676.

Daines B, Wang H, Wang L, Li Y, Han Y, Emmert D, Gelbart W, Wang X, Li W, Gibbs R, Chen R, 2011. The *Drosophila melanogaster* transcriptome by paired-end RNA sequencing. *Genome Res.*, 21(2): 315 – 324.

De Bona F, Ossowski S, Schneeberger K, Ratsch, G, 2008. Optimal spliced alignments of short sequence reads. *Bioinformatics*, 24(16): 174 – 180.

Degner JF, Marioni JC, Pai AA, Pickrell JK, Nkadori E, Gilad Y, Pritchard JK, 2009. Effect of read-mapping biases on detecting allele-specific expression from RNA-sequencing data. *Bioinformatics*, 25(24): 3207 – 3212.

Denoeud F, Aury JM, Da Silva C, Noel B, Rogier O, Delledonne M, Morgante M, Valle G, Wincker P, Scarpelli C, Jaillon O, Artiguenave F, 2008. Annotating genomes with massive-scale RNA sequencing. *Genome Biol.*, 9(12): R175.

Fuller CW, Middendorf LR, Benner SA, Church GM, Harris T, Huang X, Jovanovich SB, Nelson JR, Schloss JA, Schwartz DC, Vezenov DV, 2009. The challenges of sequencing by synthesis. *Nat. Biotechnol.*, 27(11): 1013 – 1023.

Garber M, Grabherr MG, Guttman M, Trapnell C, 2011. Computational methods for transcriptome annotation and quantification using RNA-seq. *Nat. Methods*, 8(6): 469 – 477.

Glaus P, Honkela A, Rattray M, 2012. Identifying differentially expressed transcripts from RNA-seq data with biological variation. *Bioinformatics*, 28(13): 1721 – 1728.

Grabherr MG, Haas BJ, Yassour M, Levin JZ, Thompson DA, Amit I, Adiconis X, Fan L, Raychowdhury R, Zeng Q, Chen Z, Mauceli E, Hacohen N, Gnirke A, Rhind N, di Palma F, Birren BW, Nusbaum C, Lindblad-Toh K, Friedman N, Regev A, 2011. Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nat. Biotechnol.*, 29(7): 644 – 652.

Griffith M, Griffith OL, Mwenifumbo J, Goya R, Morrissy AS, Morin RD, Corbett R, Tang MJ, Hou YC, Pugh TJ, Robertson G, Chittaranjan S, Ally A, Asano JK, Chan SY, Li HI, McDonald H, Teague K, Zhao Y, Zeng T, Delaney A, Hirst M, Morin GB, Jones SJ, Tai IT, Marra MA, 2010. Alternative expression analysis by RNA sequencing. *Nat. Methods*, 7(10): 843 – 847.

Guttman M, Garber M, Levin JZ, Donaghey J, Robinson J, Adiconis X, Fan L, Koziol MJ, Gnirke A, Nusbaum C, Rinn JL, Lander ES, Regev A, 2010. Ab initio reconstruction of cell type-specific transcriptomes in mouse reveals the conserved multi-exonic structure of lincRNAs. *Nat. Biotechnol.*, 28(5): 503 – 510.

Homer N, Merriman B, Nelson SF, 2009. BFAST: an alignment tool for large scale genome resequencing. *PLoS ONE*, 4(11): e7767.

Hoen PA, Ariyurek Y, Thygesen HH, Vreugdenhil E, Vossen RH, de Menezes RX, Boer JM, van Ommen GJ, den Dunnen JT, 2008. Deep sequencing-based expression analysis shows major advances in robustness, resolution and inter-lab portability over five microarray platforms. *Nucleic Acids Res.*, 36(21): e141.

Huang X, Madan A, 1999. CAP3: A DNA sequence assembly program. *Genome Res.*, 9(9): 868 – 877.

Jiang H, Wong WH, 2008. SeqMap: mapping massive amount of oligonucleotides to the genome. *Bioinformatics*, 24(20): 2395 – 2396.

Jimenez-Guri E, Huerta-Cepas J, Cozzuto L, Wotton KR, Kang H, Himmelbauer H, Roma G, Gabaldon T, Jaeger J, 2013. Comparative transcriptomics of early dipteran development. *BMC Genomics*, 14: 123.

Kaminuma E, Mashima J, Kodama Y, Gojobori T, Ogasawara O, Okubo K, Takagi T, Nakamura Y, 2010. DDBJ launches a new archive database with analytical tools for next-generation sequence data. *Nucleic Acids Res.*, 38 (Database issue): D33 – 38.

Kanehisa M, 2002. The KEGG database. *Novartis Found Symp*, 247: 91 – 101; discussion 101 – 103, 119 – 128, 244 – 152.

Karp PD, Ouzounis CA, Moore-Kochlacs C, Goldovsky L, Kaipa P, Ahren D, Tsoka S, Darzentas N, Kunin V, Lopez-Bigas N, 2005. Expansion of the BioCyc collection of pathway/genome databases to 160 genomes. *Nucleic Acids Res.*, 33(19): 6083 – 6089.

Katz Y, Wang ET, Airoldi EM, Burge CB, 2010. Analysis and design of RNA sequencing experiments for identifying isoform regulation. *Nat. Methods*, 7(12): 1009 – 1015.

Kent WJ, 2002. BLAT-the BLAST-like alignment tool. *Genome Res.*, 12(4): 656 – 664.

Kolev NG, Franklin JB, Carmi S, Shi H, Michaeli S, Tschudi C, 2010. The transcriptome of the human pathogen Trypanosoma brucei at single-nucleotide resolution. *PLoS*

Pathog., 6(9):e1001090.

Koonin EV, Tatusov RL, Galperin MY, 1998. Beyond complete genomes:from sequence to structure and function. *Curr. Opin. Struct Biol.*, 8(3):355 – 363.

Langmead B, Hansen KD, Leek JT, 2010. Cloud-scale RNA-sequencing differential expression analysis with Myrna. *Genome Biol.*, 11(8):R83.

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

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

Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R, 2009a. The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, 25(16):2078 – 2079.

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

Li R, Li Y, Kristiansen K, Wang J, 2008b. SOAP:short oligonucleotide alignment program. *Bioinformatics*, 24(5):713 – 714.

Li R, Yu C, Li Y, Lam TW, Yiu SM, Kristiansen K, Wang J, 2009b. SOAP2:an improved ultrafast tool for short read alignment. *Bioinformatics*, 25(15):1966 – 1967.

Li SW, Yang H, Liu YF, Liao QR, Du J, Jin DC, 2012. Transcriptome and gene expression analysis of the rice leaf folder, *Cnaphalocrosis medinalis*. *PLoS ONE*, 7(11):e47401.

Lunter G, Goodson M, 2011. Stampy:a statistical algorithm for sensitive and fast mapping of Illumina sequence reads. *Genome Res.*, 21(6):936 – 939.

Ma W, Zhang Z, Peng C, Wang X, Li F, Lin Y, 2012. Exploring the midgut transcriptome and brush border membrane vesicle proteome of the rice stem borer, *Chilo suppressalis* (Walker). *PLoS ONE*, 7(5):e38151.

Mamanova L, Andrews RM, James KD, Sheridan EM, Ellis PD, Langford CF, Ost TW, Collins JE, Turner DJ, 2010. FRT-seq:amplification-free, strand-specific transcriptome sequencing. *Nat. Methods*, 7(2):130 – 132.

Mardis ER. 2008. Next-generation DNA sequencing methods. *Annu. Rev. Genomics Hum. Genet.*, 9:387 – 402.

Marguerat S, Bahler J, 2010. RNA-seq:from technology to biology. *Cell Mol. Life Sci.*, 67(4):569 – 579.

Margulies M, Egholm M, Altman WE, Attiya S, Bader JS, Bemben LA, Berka J, Braverman MS, Chen YJ, Chen Z, 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 P, Begley RF, Rothberg JM, 2005. Genome sequencing in microfabricated high-density picolitre reactors. *Nature*, 437(7057):376 – 380.

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

Martin J, Bruno VM, Fang Z, Meng X, Blow M, Zhang T, Sherlock G, Snyder M, Wang Z, 2010. Rnnotator:an automated de novo transcriptome assembly pipeline from stranded RNA-Seq reads. *BMC Genomics*, 11:663.

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

Ogata H, Goto S, Fujibuchi W, Kanehisa M, 1998. Computation with the KEGG pathway database. *Biosystems*, 47(1/2):119 – 128.

Oshlack A, Robinson MD, Young MD, 2010. From RNA-seq reads to differential expression results. *Genome Biol.*, 11(12):220.

Pevzner PA, 1989, 1-Tuple DNA sequencing:computer analysis. *J. Biomol. Struct. Dyn.*, 7(1):63 – 73.

Rizk G, Lavenier D, 2010. GASSST:global alignment short sequence search tool. *Bioinformatics*, 26(20):2534 – 2540.

Robertson G, Schein J, Chiu R, Corbett R, Field M, Jackman SD, Mungall K, Lee S, Okada HM, Qian JQ, Griffith M, Raymond A, Thiessen N, Cezard T, Butterfield YS, Newsome R, Chan SK, She R, Varhol R, Kamoh B, Prabhu AL, Tam A, Zhao Y, Moore RA, Hirst M, Marra MA, Jones SJ, Hoodless PA, Birol I, 2010. De novo assembly and analysis of RNA-seq data. *Nat. Methods*, 7(11):909 – 912.

Robinson MD, McCarthy DJ, Smyth GK, 2010. edgeR:a bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, 26(1):139 – 140.

Rumble SM, Lacroute P, Dalca AV, Fiume M, Sidow A, Brudno M, 2009. SHRiMP: accurate mapping of short color-space reads. *PLoS Comput. Biol.*, 5(5):e1000386.

Ruparel H, Bi L, Li Z, Bai X, Kim DH, Turro NJ, Ju J, 2005. Design and synthesis of a 3′-O-allyl photocleavable fluorescent nucleotide as a reversible terminator for DNA sequencing by synthesis. *PNAS*, 102(17):5932 – 5937.

Schaefer CF, Anthony K, Krupa S, Buchoff J, Day M, Hannay T, Buetow KH, 2009. PID:the Pathway Interaction Database. *Nucleic Acids Res.*, 37(Database issue):D674 – 679.

Schmid R, Blaxter ML, 2008. annot8r:GO, EC and KEGG annotation of EST datasets. *BMC Bioinformatics*, 9:180.

Seo TS, Bai X, Kim DH, Meng Q, Shi S, Ruparel H, Li Z, Turro NJ, Ju J, 2005. Four-color DNA sequencing by synthesis on a chip using photocleavable fluorescent nucleotides. *PNAS*, 102(17):5926 – 5931.

Shendure J, 2008. The beginning of the end for microarrays? *Nat. Methods*, 5(7):585 – 587.

Shendure JA, Porreca GJ, Church GM, Gardner AF, Hendrickson CL, Kieleczawa J, Slatko BE, 2011. Overview of DNA sequencing strategies. *Curr. Protoc. Mol. Biol.*, 7(7):1.

Shumway M, Cochrane G, Sugawara H, 2010. Archiving next generation sequencing data. *Nucleic Acids Res.*, 38(Database issue):D870 – 871.

Smith AD, Xuan Z, Zhang MQ, 2008. Using quality scores and longer reads improves accuracy of Solexa read mapping. *BMC Bioinformatics*, 9:128.

Surget-Groba Y, Montoya-Burgos JI, 2010. Optimization of de novo transcriptome assembly from next-generation sequencing data. *Genome Res.*, 20(10):1432 – 1440.

Tatusov RL, Koonin EV, Lipman DJ, 1997. A genomic perspective on protein families. *Science*, 278(5338):631 – 637.

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

Trapnell C, Williams BA, Pertea G, Mortazavi A, Kwan G, van Baren MJ, Salzberg SL, Wold BJ, Pachter L, 2010. Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat. Biotechnol.*, 28(5):511 – 515.

Wang ET, Sandberg R, Luo S, Khrebtukova I, Zhang L, Mayr C, Kingsmore SF, Schroth GP, Burge CB, 2008. Alternative isoform regulation in human tissue transcriptomes. *Nature*, 456(7221):470 – 476.

Wang K, Singh D, Zeng Z, Coleman SJ, Huang Y, Savich GL, He X, Mieczkowski P, Grimm SA, Perou CM, MacLeod JN, Chiang DY, Prins JF, Liu J, 2010a. MapSplice:accurate mapping of RNA-seq reads for splice junction discovery. *Nucleic Acids Res.*, 38(18):e178.

Wang L, Feng Z, Wang X, Wang X, Zhang X, 2010b. DEGseq: an R package for identifying differentially expressed genes from RNA-seq data. *Bioinformatics*, 26(1):136 – 138.

Wang X, Wu Z, Zhang X, 2010c. Isoform abundance inference provides a more accurate estimation of gene expression levels in RNA-seq. *J. Bioinform. Comput. Biol.*, 8(Suppl 1):177 – 192.

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

Wilhelm BT, Marguerat S, Goodhead I, Bahler J, 2010. Defining transcribed regions using RNA-seq. *Nat. Protoc.*, 5(2):255 – 266.

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

Wu TD, Watanabe CK, 2005. GMAP: a genomic mapping and alignment program for mRNA and EST sequences. *Bioinformatics*, 21(9):1859 – 1875.

Xue J, Bao YY, Li BL, Cheng YB, Peng ZY, Liu H, Xu HJ, Zhu ZR, Lou YG, Cheng JA, Zhang CX, 2010. Transcriptome analysis of the brown planthopper Nilaparvata lugens. *PLoS ONE*, 5(12):e14233.

You M, Yue Z, He W, Yang X, Yang G, Xie M, Zhan D, Baxter SW, Vasseur L, Gurr GM, Douglas CJ, Bai J, Wang P, Cui K, Huang S, Li X, Zhou Q, Wu Z, Chen Q, Liu C, Wang B, Li X, Xu X, Lu C, Hu M, Davey JW, Smith SM, Chen M, Xia X, Tang W, Ke F, Zheng D, Hu Y, Song F, You Y, Ma X, Peng L, Zheng Y, Liang Y, Chen Y, Yu L, Zhang Y, Liu Y, Li G, Fang L, Li J, Zhou X, Luo Y, Gou C, Wang J, Wang J, Yang H, Wang J, 2013. A heterozygous moth genome provides insights into herbivory and detoxification. *Nat. Genet.*, 45(2):220 – 225.

Zerbino DR, Birney E, 2008. Velvet:algorithms for de novo short read assembly using de Bruijn graphs. *Genome Res.*, 18(5):821 – 829.

Zhan S, Merlin C, Boore JL, Reppert SM, 2011. The monarch butterfly genome yields insights into long-distance migration. *Cell*, 147(5):1171 – 1185.