论文

基于 RNA-Seq 数据识别果蝇剪接位点和可变剪接事件

何涛^{①†}, 王端青^{①②†}, 胡亚欧^①, 张颖^①, 邵卫东^②, 汪莉^{①*}, 王玉民^{①*}

① 军事医学科学院生物工程研究所, 北京 100071;

② 苏州大学电子信息学院, 苏州 215006

† 同等贡献

* 联系人, E-mail: liwang@tsinghua.edu.cn; wym66@vip.sina.com

收稿日期: 2011-07-20; 接受日期: 2011-08-22

国家自然科学基金(批准号: 30800644)、国家高技术研究发展计划(批准号: 2007AA022204)和国家“重大新药创制”科技重大专项(批准号: 2008ZXJ09007-001)资助项目

doi: 10.1360/052011-586

摘要 完整基因结构的预测是当前生命科学研究的一个重要基础课题, 其中一个关键环节是剪接位点和各种可变剪接事件的精确识别. 基于转录组测序(RNA-seq)数据, 识别剪接位点和可变剪接事件是近几年随着新一代测序技术发展起来的新技术策略和方法. 本工作基于黑腹果蝇睾丸 RNA-seq 数据, 使用 TopHat 软件成功识别出 39718 个果蝇剪接位点, 其中有 10584 个新剪接位点. 同时, 基于剪接位点的不同组合, 针对各类型可变剪接特征开发出计算识别算法, 成功识别了 8477 个可变剪接事件(其中新识别的可变剪接事件 3922 个), 包括可变供体位点、可变受体位点、内含子保留和外显子缺失 4 种类型. RT-PCR 实验验证了 2 个果蝇基因上新识别的可变剪接事件, 发现了全新的剪接异构体. 进一步表明, RNA-seq 数据可有效应用于识别剪接位点和可变剪接事件, 为深入揭示剪接机制及可变剪接生物学功能提供新思路和新手段.

关键词
剪接位点
可变剪接
黑腹果蝇
RNA-seq

绝大多数真核基因编码序列由外显子和内含子间隔组成. 外显子和内含子之间的边界称作剪接位点, 按它们在内含子两端的位置又可分为 5'剪接位点(位于内含子的 5'端, 也称作供体位点)和 3'剪接位点(位于内含子的 3'端, 也称作受体位点)^[1]. 基因的前体 mRNA 被转录后, 必须通过剪接反应切除内含子, 把外显子连在一起, 形成一个成熟的 mRNA, 由细胞核转运到细胞质中进行翻译. 可变剪接(alternative splicing)是指从一个 mRNA 前体中通过不同的剪接方式(选择不同的剪接位点组合)产生不同的 mRNA 剪

接异构体, 生成具有不同化学性质和生物功能的蛋白亚型的过程^[2]. 可变剪接是高等真核生物中丰富蛋白质多样性的重要机制之一^[3], 非正常的可变剪接会导致各种疾病^[4,5].

剪接位点的精确定位是确定真核生物基因结构的关键, 目前有多种方法可用来在基因组范围内识别剪接位点. 生物实验主要采用外显子连接芯片^[6]和外显子芯片^[7]等高通量技术, 该方法有探针设计和数据分析相对复杂、无法识别未知剪接位点的局限性. 生物信息学已开发出包括从头预测法、基于 EST/

英文引用格式: He T, Wang D Q, Hu Y O, et al. Identification of novel splice sites and alternative splicing events in *Drosophila melanogaster* using RNA-seq data. SCIENTIA SINICA Vitae, 2011, 41: 1016–1023, doi: 10.1360/052011-586

cDNA 序列比对法和基于 RNA-seq 数据识别等多种有效的方法. 从头预测算法主要采用支持向量机^[8]、概率模型^[9]、马尔可夫模型^[10]和神经网络^[11]等智能信息处理技术预测剪接位点. 这类方法能够不依赖于转录丰度, 直接在全基因组范围内快速识别潜在的剪接位点, 但其预测精度依赖于数学模型和参数体系的选择, 假阳性率普遍较高. EST 技术长期以来被作为新转录体检测和基因结构识别的“核心方法”和“黄金标准”, 基于序列比对从序列片段中识别出剪接位点和可变剪接事件^[12], 但该技术价格较贵, 具有克隆偏性, 并且不同转录区域的 EST/cDNA 覆盖度差异较大. RNA-seq 技术是全新的转录组研究方法, 基本上克服了上述技术的弊端和缺陷, 无需预先设计探针, 可对任意物种的整体转录活动进行检测, 发现新基因、新剪接位点和可变剪接事件, 对转录体结构的分析有了明显的提高^[13]. 已有研究者将 RNA-Seq 技术成功应用于剪接位点和可变剪接事件识别. Ali Mortazavi 等人^[14]通过 RNA-seq 技术检测到 145000 个小鼠 mRNA 剪接形式, 总共有 3500 个基因发生了可变剪接; Sultan 等人^[15]通过 RNA-seq 技术发现 94241 个人类 mRNA 剪接形式, 其中 4096 个为新的剪接形式; Pan 等人^[16]基于 RNA-seq 数据, 在约 20% 的人类多外显子基因中识别出新的剪接形式, 并推测高达 95% 的人类多外显子基因均发生可变剪接. Ramani 等人^[17]用线虫 RNA-seq 数据识别出 71% 的已标注剪接位点, 并发现上千新的剪接位点和几百个受发育时序调控的可变剪接事件.

黑腹果蝇是生物学研究中最重要模式生物之一. 可变剪接事件在果蝇中普遍存在, Stolc 等人^[18]发现至少 40% 的果蝇基因发生可变剪接. 研究发现, 可变剪接在果蝇的性别决定、神经分布和眼睛发育等多方面发挥重要作用^[19]. Gan 等人^[20]基于 RNA-seq 技术研究雄性和雌性果蝇的性腺转录组, 发现了一批性别差异的剪接异构体, 研究结果表明剪接事件在未分化的性腺细胞中比分化后的显著活跃. Graveley 等人^[21]基于 RNA-seq 技术, 揭示了果蝇多个发育阶段的转录组, 发现了 22965 个新剪接位点, 在果蝇 12295 个多外显子基因中, 高达 60.7% 的基因均发生可变剪接. 由此可知, 果蝇转录组实际上比原有的认识还要复杂的多, 需要更多的研究工作揭示果蝇转录组的复杂性. 本研究基于果蝇睾丸 RNA-seq 数据, 识别了 29134 个已知剪接位点和 10584 个新剪接位点.

基于剪接位点的不同组合, 针对各类型可变剪接特征开发出计算识别算法, 成功识别了 8477 个可变剪接事件, 其中 46.3% 为新识别的可变剪接事件. RT-PCR 实验成功验证了 2 个果蝇基因上新识别的可变剪接事件, 发现了全新的剪接异构体.

1 材料与方法

1.1 数据来源和已知果蝇剪接位点数据集的构建

果蝇基因组序列(dm3 版)下载自 UCSC 网站; 果蝇睾丸组织的 RNA-Seq 数据^[22](SRR054704)下载自 NCBI SRA 数据库, 是由 Solexa 双末端(pair-end)测序获得的 75 bp 的转录组序列, 共计 32585670 条; 基因注释数据下载自 FlyBase 数据库和 UCSC 网站, 合并了 FlyBase 和 UCSC 标注的内含子, 去冗余后得到 52989 个内含子, 提取供体和受体位点, 构建了果蝇已知剪接位点数据集.

1.2 剪接位点的识别

目前已有多款基于 RNA-seq 数据识别剪接位点的软件^[23], 其中 TopHat^[24]是一款被广泛应用的软件, 不依赖已知标注信息, 将第一步无法比对到基因组上的序列进一步分割成小片段, 同第一步比对序列拼接聚类建立的剪接接合区进行二次比对, 结合两次比对识别剪接位点. 该软件在内部调用第三方比对软件 bowtie^[25]. 本研究工作中, 果蝇剪接位点识别采用 TopHat(v-0.1.4)和 bowtie(v-0.12.7)软件.

本研究对 TopHat 软件参数做如下设置: (1) 果蝇内含子大部分在 10 kb 以内, 因此内含子最大搜索距离设置为 15 kb; (2) TopHat 比对种子区最多允许出现一个错配, 且剪接位点上下游 8 bp 区域内不允许有错配; (3) 比对到基因组上的序列位置唯一.

1.3 可变剪接事件的识别

可变剪接事件共有 5 种基本类型, 分别是可变供体位点(alternative donor site)、可变受体位点(alternative acceptor site)、内含子保留型(intron retention)、外显子缺失型(exon skipping)和外显子互斥型(mutually exclusive exon). 另外也有分为 7 种形式的, 包括前面 5 种类型加上可变的起始或末端外显子, 而后两种形式更有可能是可变启动子和可变 polyA 位点造成的, 可进行专门分析. 如图 1 所示, 结合已知剪接

图 1 可变剪接类型及计算识别策略

图中黑色方块代表组成性外显子, 红色方块代表可变外显子, 蓝色和黄色方块分别代表供体位点和受体位点, 黑色实线代表内含子. 第一列图示了 4 种可变剪接类型, 第二列穷举 TopHat 识别的供体和受体位点组合, 第三列是基于位点组合识别的潜在可变剪接事件

位点数据集和基于 RNA-Seq 数据新识别的剪接位点, 通过比较供体位点和受体位点的组合, 我们建立了识别其中 4 种可变剪接事件的计算策略, 具体如下:

(1) 可变供体位点各剪接形式都有相同的受体位点和不同的供体位点, 不同的供体位点改变了其上游外显子的长度. 如图 1(A)所示, 4 个剪接位点包含不同的供体位点 D1, D2, D3 和相同的受体位点 A1, 可组合为 3 种剪接形式 D1-A1, D2-A1 和 D3-A1, 即识别出 2 个可变剪接事件, 可变外显子区分别为 D1-D2 和 D1-D3.

(2) 可变受体位点各剪接形式都有不同的受体位点和相同的供体位点, 不同的受体位点改变了其下游外显子的长度. 如图 1(B)所示, 4 个剪接位点包含相同的供体位点 D1 和不同的受体位点 A1, A2 和 A3, 可组合为 3 种剪接形式 D1-A1, D1-A2 和 D1-A3, 即识别出 2 个可变剪接事件, 可变外显子区分别为 A1-A3 和 A2-A3.

(3) 内含子保留型可变剪接是完整的内含子有时会被当作外显子中的一段保留的剪接事件, 其受体和供体位点必须落在同一个外显子上. 如图 1(C)所示, D1 和 A1 分别为剪接事件的供体和受体位点, 若它们又落在同一个外显子上, 即识别出一个可变剪接事件, 可变外显子区为 D1-A1.

(4) 外显子缺失型可变剪接是指某一完整外显子可能被某种剪接形式跨过从而作为内含子去除的剪接事件. 如图 1(D)所示, 6 个剪接位点包含不同的供体位点 D1, D2, D3 和不同的受体位点 A1, A2 和 A3, 穷举位点 D1 和 A3 之间所有可能的外显子, 得到 A1-D2, A1-D3 和 A2-D3 3 个假定的外显子, 其中若 A1-D3 有已知外显子支持, 则识别出 3 个可变剪接事件, 可变外显子为 A1-D2, A2-D3 和 A1-D3; 若 A1-D3 无已知外显子支持, 由于其间含有 5'和 3'剪接位点, 本研究中认为此类假定的外显子不存在, 故识别出 2 个可变剪接事件, 可变外显子为 A1-D2 和 A2-D3.

1.4 新剪接位点和可变剪接事件的实验验证

野生型雄性果蝇由中国农业大学赵章武教授提供. 利用常规 Trizol 法提取总 RNA, cDNA 第一链的合成用 M-MLV 反转录酶(Takara)完成. 使用 LaTaq 扩增酶(Takara)以合成的 cDNA 为模板进行 PCR 扩增. 反应条件如下: 94℃预变性 5 min; 94℃变性 30 s, 54℃退火 30 s, 72℃延伸 2.5 min, 进行 31 个循环; 72℃再延伸 10 min. 以普通琼脂糖凝胶 DNA 回收试剂盒(TIANGEN, 北京)回收纯化 PCR 产物, 将该产物与 pMD18-T 载体(Takara)进行连接, 16℃反应 5 h, 转化至大肠杆菌 DH5α, 蓝白斑筛选并进行 PCR 鉴定. 引物合成由北京奥科生物技术有限责任公司完成. 测序由北京天一辉远生物科技有限公司进行. 验证实验所用引物如下: CG5450_F: 5'-CAAGAGCCGGA-TAGACAGAGC-3', CG5450_R: 5'-TTGGCAGCAGAGTGGCTAAGT-3'; CG9008_F: 5'-AGTTTCAATCG-GTCGCTCGTT-3', CG9008-R: 5'-TACAACCTTCTTG-ATTGGGTGC-3'; CG31601_F: 5'-TGTCCAGTTCAC-AATCATCAG-3', CG31601-R: 5'-TTCTCTATTTGC-TGCTCTACG-3'.

2 结果

2.1 基于果蝇 RNA-Seq 数据的全基因组范围剪接位点识别

使用 TopHat 将果蝇睾丸 RNA-seq 数据比对到其

基因组(dm3), 在限制种子区最多允许出现一个错配, 并且剪接位点上下游 8 bp 区域内不允许有错配, 28038277(86.04%)条转录组测序 read 在第一轮比对中比对到基因组唯一位置, 2078497(6.38%)条 read 二次比对时匹配到剪接位点结合区, 识别出 39718 个剪接位点.

如表 1 所示, 同构建的果蝇已知剪接位点数据集进行比较, 发现识别的剪接位点中有 29134 个为已知剪接位点, 覆盖了已知果蝇剪接位点的 55.0%, 分布在果蝇 8201 个基因上. 在重标注的已知剪接位点中, read 支持数大于 5 的占多数(69.16%); 未在已知剪接位点数据集中的剪接位点有 10584 个, 这些位点是从果蝇睾丸转录组中识别出的新剪接位点. 在新剪接位点中, read 支持数大于等于 5 的有 2186 个(20.66%), 新剪接位点的 reads 支持数相比已知剪接位点普遍偏低, 大多只有少数 reads 支持. Ramani 等人^[17]在基于 RNA-Seq 识别线虫可变剪接的工作中, 对仅有一条 read 支持的可变剪接事件进一步分析, 发现超过 90% 的事件在增加测序覆盖度后均得到确认, 该文认为仅有一条 read 支持的剪接位点和可变剪接事件仍是可靠的.

2.2 基于果蝇 RNA-Seq 数据的全基因组范围可变剪接事件的识别

结合已知剪接位点数据集和基于 RNA-Seq 数据新识别的剪接位点, 通过比较供体位点和受体位点

表 1 TopHat 比对结果和剪接位点识别结果统计

对比和剪接位点识别结果	Read 和剪接位点数目	比例
果蝇睾丸 RNA-Seq 总 read 数	32585670	
比对到唯一位置的 read	28038277	86.04%
比对到剪接位点的 read	2078497	6.38%
其他 read(测序质量低\比对到多个位置或未比 对上的 read)	2468896	7.58%
识别出的剪接位点	39718	
重标注已知剪接位点	29134	
剪接位点(支持 read<5)	8986	30.84%
剪接位点(支持 read>5)	4832	16.59%
剪接位点(支持 read>10)	2819	9.68%
剪接位点(支持 read>15)	1853	6.36%
剪接位点(支持 read>20)	10644	36.53%
新识别的剪接位点	10584	
剪接位点(支持 read<5)	8398	79.34%
剪接位点(支持 read>5)	950	8.98%
剪接位点(支持 read>10)	377	3.56%
剪接位点(支持 read>15)	186	1.76%
剪接位点(支持 read>20)	673	6.36%

的组合, 我们针对各类型可变剪接特征开发出识别算法, 成功识别出 8477 个可变剪接事件(已知可变剪接事件 4555 个和新可变剪接事件 3922 个, 详见表 2), 包括可变供体位点、可变受体位点、内含子保留和外显子缺失 4 种类型.

基于果蝇睾丸 RNA-Seq 数据识别的 8477 个可变剪接事件, 发生在共计 3705 个基因上(约占已知基因的 25.8%, 3705/14381), 特别是新识别的 3922 个可变剪接事件进一步丰富了果蝇的蛋白质产物. 例如果蝇 *Cdlic2* 基因在睾丸中特异性表达^[26], 它参与精子个体化的两个独立过程并起着重要作用^[27]. 依据 UCSC 数据库注释(图 2(D)), 该基因有 2 种已知的剪接异构体, 本研究又在该基因上识别出 1 个新的可变受体位点, 使该基因第二个外显子缩短了 12 nt, 导致 *Cdlic2* 基因的 5'-UTR 缩短. RT-PCR 实验验证出 2 个已知的剪接异构体, 未验证出此可变受体位点, 可能是由于如此小的差异(12 nt)在电泳条带上较难被区分, 从而导致未挑选到含有该可变受体位点的单克隆进行测序. 但是该可变受体位点得到已知果蝇 EST 序列(AI946956)的支持. 果蝇 *CG9008* 基因参与高尔基体和脂肪颗粒的组织^[28]. 依据 UCSC 数据库注释(如图 2(C)所示), 该基因有 2 种已知的剪接异构体, 本研究又在该基因上识别了新的内含子保留型可变剪接事件, 使处于 CDS 区的外显子增加了 207 nt, 从而导致编码的蛋白质增加了 69 个氨基酸, RT-PCR 及 PCR 测序成功验证了该内含子保留型可变剪接事件, 发现了一种全新的剪接异构体 *CG9008_N1*. 依据 UCSC 数据库注释(如图 2(B)所示), 果蝇 *CG31601* 基因不发生可变剪接, 我们在该基因上识别了新的内含子保留型可变剪接事件, 使处于 CDS 区的外显子减少了 54 nt, 从而导致编码的蛋白质减少了 18 个氨基酸, RT-PCR 及 PCR 测序成功验证了该内含子保留型可变剪接事件, 发现了一种全新的剪接异构体

CG31601_N1, 同时已知果蝇 EST 序列(BT125056)也支持该可变剪接事件.

3 讨论

RNA-Seq 技术能够精确地判断转录本的边界区域, 发现小的外显子和未注释的基因^[13,14], 并且能够揭示基因序列的变化如可变剪接^[14-17]等. 可以说, RNA-Seq 技术是转录组学研究的革命性工具^[13]. 本研究基于果蝇睾丸 RNA-Seq 数据, 成功识别出 39718 个果蝇剪接位点, 其中有 10584 个新剪接位点. 基于剪接位点的不同组合, 计算识别了可变供体位点、可变受体位点、内含子保留和外显子缺失 4 种类型的 8477 个可变剪接事件, 其中 3922 个为新可变剪接事件. RT-PCR 实验验证了 2 个果蝇基因上新识别的可变剪接事件, 发现了全新的剪接异构体. 本工作为进一步研究黑腹果蝇剪接机制及可变剪接生物学功能奠定了基础.

在本工作识别出的大量新剪接位点中, 如表 1 所示, 可以发现大部分新剪接位点的 read 支持数普遍较低, 可能是由于这些剪接位点所在的转录体本身的转录丰度低, 或者剪接位点自身的剪接效率较低, 传统实验技术较难检测, 有赖于新一代测序技术的深度挖掘才被识别出来; 本工作识别到的已知剪接位点仅占全部已知剪接位点的 55%, 一方面可能是软件比对和识别参数设置所限; 另一方面可能是果蝇基因并未全部在其雄性睾丸中转录, 或者本工作基于的原始测序数据深度不足, 仍有基因或剪接结合区未能被覆盖. 使用序列联配方法识别可变剪接作为一种重要研究手段, 是对实验研究的必要补充, 显著降低实验周期和消耗. 目前基于 RNA-Seq 技术的可变剪接研究只是一个开端, 还有许多值得探索和改进的地方.

新一代测序技术产生的 read 比 cDNA/EST 序列

表 2 识别的已知可变剪接和新可变剪接事件

可变剪接类型	重标注已知的可变剪接事件			新识别的可变剪接事件		
	总计	支持 read>5	支持 read>10	总计	支持 read>5	支持 read>10
可变供体位点	1274	368(28.9%)	260(20.4%)	1158	273(23.6%)	147(12.7%)
可变受体位点	1664	495(29.7%)	370(22.2%)	1176	291(24.7)	165(14%)
内含子保留	684	464(67.8%)	353(51.6%)	1344	316(23.5%)	191(14.2%)
外显子缺失	933 ^{a)}	295(63.9%)	222(48%)	244	52(21.3%)	24(9.8%)
总计	4555			3922		

a) 外显子缺失型可变剪接中, 剪接事件由已知数据支持无 RNA-seq 数据覆盖时无法统计 read 支持数, 占 44.6% (416/933)

图 2 实验验证 3 个果蝇基因上的新可变剪接事件

(A) RT-PCR 扩增的各基因不同剪接异构体的电泳图; (B) CG31601 基因的 RT-PCR 结果和原有数据库(FlyBase 和 GenBank)标注信息; (C) CG9008 基因的 RT-PCR 结果和原有数据库(FlyBase)标注信息; (D) Cdic2 基因的 RT-PCR 结果和原有数据库(FlyBase 和 GenBank)标注信息

的长度短很多, 通常用于转录组表达分析的 read 长度在 30 nt 左右. 序列过短将限制其进一步应用于剪接位点和可变剪接事件的识别, 一方面因为读段长度只有 20~30 个碱基时, 来自剪接接合区的读段所占比例很小; 另一方面, 目前基于 RNA-Seq 数据识别剪接位点的软件会将无法连续比对到基因组的 read 片段化, 如 TopHat 默认将 read 切成 25 nt 长的多个

片段进行二次比对, 故其推荐应用于剪接位点识别的 RNA-Seq 读段长度最小为 75 nt. 现在各测序平台都支持双末端测序, 产生的 reads 长度可达 2×100 bp, 随着这一技术的不断发展, 势必产生更长的序列, 跨剪接接合区 read 数量更多, 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, 大大推进人们对可变剪接的研究.

参考文献

- 1 朱玉贤, 李毅. 现代分子生物学. 第 2 版. 北京: 高等教育出版社, 2005. 243-246
- 2 Grabowski P J, Black D L. Alternative RNA splicing in the nervous system. Prog Neurobiol, 2001, 65: 289-308
- 3 Douglas L, Black. Mechanisms of alternative pre-messenger RNA splicing. Annu Rev Biochem, 2003, 72: 291-336
- 4 Wang G S, Cooper T A. Splicing in disease: Disruption of the splicing code and the decoding machinery. Nat Rev Genet, 2007, 8: 749-761
- 5 Cooper T A, Wan L L, Dreyfuss G. RNA and disease. Cell, 2009, 136: 777-793

- 6 Johnson J M, Castle J, Garrett-Engle P, et al. Genome-wide survey of human alternative pre-mRNA splicing with exon junction microarrays. *Science*, 2003, 302: 2141–2144
- 7 Clark T A, Schweitzer A C, Chen T X, et al. Discovery of tissue-specific exons using comprehensive human exon microarrays. *Genome Biol*, 2007, 8: R64
- 8 Sonnenburg S, Schweikert G, Philips P, et al. Accurate splice site prediction using support vector machines. *BMC Bioinformatics*, 2007, 10: S7
- 9 Pertea M, Lin X Y, Salzberg S L. GeneSplicer: A new computational method for splice site detection. *Nucleic Acids Res*, 2001, 29: 1185–1190
- 10 Zhang Q W, Peng Q K, Li K K, et al. Splice sites detection by combining markov and hidden markov model. *Biomed Eng Inform*, 2009, 9: 1–5
- 11 Reese M G. Application of a time-delay neural network to promoter annotation in the *Drosophila melanogaster*. *Comput Chem*, 2001, 26: 51–56
- 12 Iida K, Fukami-Kobayashi K, Toyoda A, et al. Analysis of multiple occurrences of alternative splicing events in *Arabidopsis thaliana* using novel sequenced full-length cDNAs. *DNA Res*, 2009, 16: 155–164
- 13 Wang Z, Gerstein M, Snyder M. RNA-Seq: A revolutionary tool for transcriptomics. *Nat Rev Genet*, 2009, 10: 57–63
- 14 Mortazavi A, Williams B A, McCue K, et al. Mapping and quantifying mammalian transcriptomes by RNA-seq. *Nat Methods*, 2008, 5: 621–628
- 15 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: 956–960
- 16 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: 1413–1415
- 17 Ramani A K, Calarco J A, Pan Q, et al. Genome-wide analysis of alternative splicing in *Caenorhabditis elegans*. *Genome Res*, 2011, 21: 342–348
- 18 Stolc V, Gauhar Z, Mason C, et al. A gene expression map for the euchromatic genome of *Drosophila melanogaster*. *Science*, 2004, 306: 655–660
- 19 Black D L. Mechanisms of alternative pre-messenger RNA splicing. *Annu Rev Biochem*, 2003, 72: 291–336
- 20 Gan Q, Chepelev I, Wei G, et al. Dynamic regulation of alternative splicing and chromatin structure in *Drosophila gonads* revealed by RNA-seq. *Cell Res*, 2010, 20: 763–783
- 21 Graveley B R, Brooks A N, Carlson J W, et al. The developmental transcriptome of *Drosophila melanogaster*. *Nature*, 2011, 471: 473–479
- 22 Yun D, Li Z, Shuang Y, et al. A young *Drosophila* duplicate gene plays essential roles in spermatogenesis by regulating several Y-linked male fertility genes. *PLoS Genet*, 2010, 6: e1001255
- 23 Pepke S, Wold B, Mortazavi A. Computation for ChIP-seq and RNA-seq studies. *Nat Methods*, 2009, 6: S22–S32
- 24 Trapnell C, Pachter L, Salzberg S L. TopHat: Discovering splice junctions with RNA-seq. *Bioinformatics*, 2009, 25: 1105–1111
- 25 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: R25
- 26 Dorus S, Freeman Z N, Parker E R, et al. Recent origins of sperm genes in *Drosophila*. *Mol Biol Evol*, 2008, 25: 2157–2166
- 27 Ghosh-Roy A, Desai B S, Ray K. Dynein light chain 1 regulates dynamin-mediated F-actin assembly during sperm individualization in *Drosophila*. *Mol Biol Cell*, 2005, 16: 3107–3116
- 28 Kondylis V, Tang Y, Fuchs F, et al. Identification of ER proteins involved in the functional organisation of the early secretory pathway in *Drosophila* cells by a targeted RNAi screen. *PLoS One*, 2011, 6: e17173

Identification of Novel Splice Sites and Alternative Splicing Events in *Drosophila melanogaster* Using RNA-seq Data

HE Tao¹, WANG DuanQing^{1,2}, HU YaOu¹, ZHANG Ying¹, SHAO WeiDong², WANG Li¹
& WANG YuMin¹

¹ Institute of Biotechnology, Academy of Military Medical Sciences, Beijing 100071, China;

² School of Electronics & Information, Soochow University, Suzhou 215006, China

Gene structure prediction is the first and most fundamental step to genome analysis and annotation. Splice site and alternative splicing (AS) prediction is particularly challenging for eukaryotes. With the Next Generation sequencing technologies, RNA-seq has been used in identification of splice site and alternative splicing. In this work, 39718 fruit fly splice sites were identified based on *Drosophila melanogaster* testis RNA-seq data by using Tophat software, of which 10584 were new discoveries. By different donor/acceptor splice site combinations, a computational identification method has been developed and applied to predict 8477 alternative splicing events (containing four distinct classes of AS events: alternative donor site, alternative acceptor site, intron retention and exon skipping). RT-PCR successfully validated novel alternative splicing events and new isoforms in two genes. Our result indicates that RNA-seq was not only an effective and accurate method for splice site and AS event detection, but also a new technique for deciphering molecular mechanism of RNA splicing further.

splice site, alternative splicing, *Drosophila melanogaster*, RNA-seq

doi: 10.1360/052011-586