哈尔滨工业大学计算学部

读书/论文笔记

课程名称：生物信息学

课程类型：选修

项目名称：基因表达与调控 TopHat

班级：2103601

学号：2021112845

姓名：张智雄

设计成绩	报告成绩	指导老师
		刘博

一、 论文的主要研究问题描述

这篇论文主要关注的研究问题是关于 RNA-Seq 数据分析中的剪接位点识别。传统的转录基因序列确定方法使用表达序列标签（ESTs）或全长互补 DNA（cDNA）测序技术，但随着 RNA-Seq 方法的出现，这些传统方法面临着一些挑战。RNA-Seq 利用下一代测序技术对信使 RNA 进行测序，相对于传统方法具有多项优势，如减少偏差、生成更多数据以及直接测量基因表达水平等。然而，RNA-Seq 也存在一些挑战，其中之一就是对短序列的剪接位点进行准确识别。

在 RNA-Seq 实验中，重要的一步是将 NGS“reads”映射到参考转录组或基因组，以识别新的转录本和评估转录本的丰度。然而，由于转录组的不完整性，RNA-Seq 分析通常将数据映射到基因组作为转录组的代理。这种策略使得对转录本进行全面的分析面临挑战，特别是在识别剪接位点时。当前的映射策略针对已知外显子设计了对准过程，但当一个 RNA-Seq read 跨越外显子边界时，映射过程会因 read 的部分未连续映射而失败。

针对这一问题，本文描述了一种名为 TopHat 的软件包，旨在通过大规模映射 RNA-Seq reads 来从头开始识别剪接位点。与现有策略不同，TopHat 不使用评分方案过滤可能的剪接位点，而是对所有位点进行对准，利用高效的 2 位编码和有效利用现代处理器上缓存的数据布局。TopHat 首先将非连接 reads 映射到基因组的外显子，然后再识别可能的新剪接位点。

该研究的重要贡献在于提出了一种新的策略，能够从 RNA-Seq 数据中识别剪接位点，克服了传统方法中的一些限制。通过大规模映射 RNA-Seq reads，TopHat 能够以高效率识别剪接位点，从而为深入理解转录组提供了重要的工具。此外，TopHat 的方法还能够充分利用现代处理器的性能，使其能够在标准台式计算机上运行，进一步提高了其实用性。

二、 论文的主要方法

TopHat 算法是一种用于发现 RNA-Seq 数据中剪接位点的方法，它通过两个主要阶段实现这一目标。

在第一个阶段中，TopHat 使用 Bowtie 将所有的 reads 映射到参考基因组上。未映射到基因组的 reads 被标记为“最初未映射的 reads”或 IUM reads。Bowtie 为每个 read 报告一个或多个对准，其中在 read 的 5'端不超过几个错配（默认为两个），但在 3'端可能有额外的错配，只要 Phred 质量加权的汉明距离小于指定的阈值（默认为 70）。这个策略基于经验观察，即 read 的 5'端包含的测序错误比 3'端少。TopHat 允许 Bowtie 为一个 read 报告多个对准，但会抑制所有对于具有更多对准的 reads，这排除了对于低复杂度序列的对准。

在第二个阶段中，TopHat 利用 Maq 中的组装模块对已映射的 reads 进行组装。它从稀疏一致性中提取连续序列的岛屿，推断它们是潜在的外显子。为了生成岛屿序列，TopHat 调用 Maq assemble 子命令，该命令生成一个包含调用碱基和相应参考碱基的紧凑一致性文件。由于在低覆盖区域中可能包含错误的碱基调用，因此这样的岛屿可能是“伪一致性的”。因为大多数覆盖外显子末端的 reads 也将跨越剪接位点，所以伪一致性中外显子的末端最初将

被少量 reads 覆盖，结果，一个外显子的伪一致性可能在每个末端都缺少一小部分序列。为了捕获来自相邻内含子的供体和受体位点以及这些外显子的序列，TopHat 在每个岛屿的两侧（默认为 45 bp）包含少量参考序列。

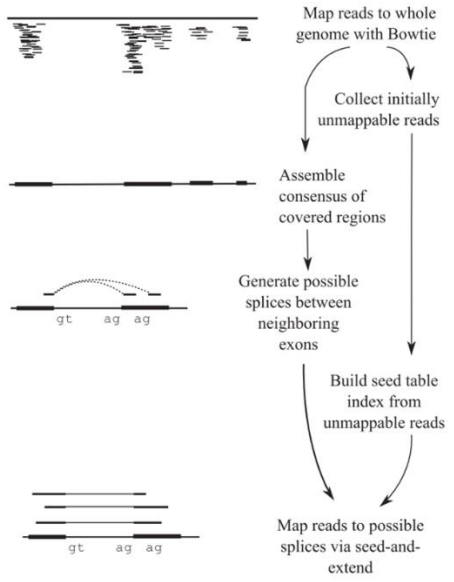

Fig. 1. The TopHat pipeline. RNA-Seq reads are mapped against the whole reference genome, and those reads that do not map are set aside. An initial consensus of mapped regions is computed by Maq. Sequences flanking potential donor/acceptor splice sites within neighboring regions are joined to form potential splice junctions. The IUM reads are indexed and aligned to these splice junction sequences.

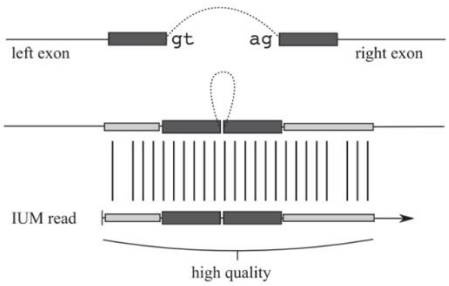

Fig. 3. The seed and extend alignment used to match reads to possible splice sites. For each possible splice site, a seed is formed by combining a small amount of sequence upstream of the donor and downstream of the acceptor. This seed, shown in dark gray, is used to query the index of reads that were not initially mapped by Bowtie. Any read containing the seed is checked for a complete alignment to the exons on either side of the possible splice. In the light gray portion of the alignment, TopHat allows a user-specified number of mismatches. Because reads typically contain low-quality base calls on their 3′ ends, TopHat only examines the first 28 bp on the 5′ end of each read by default.

Table 1. TopHat junction finding under simulated sequencing of transcripts

Depth of sequence coverage	True positives	Total (%)	False positives	Reported (%)
1	1744	17	114	6
5	7666	77	585	7
10	8737	88	428	4
25	9275	93	267	2
50	9351	94	235	2

The simulation sampled a set of transcripts with 9879 true splice junctions.

TopHat 通过调整外显子合并参数来处理低水平转录基因可能存在的外显子间间隙。这个参数定义了允许的最长覆盖缺口的长度，在默认情况下设为 6 bp，这是一个合理的选择，因为在哺乳动物基因组中，小于 70 bp 的内含子很少见。接着，TopHat 列举了岛屿序列中的所有标准供体和受体位点，并考虑了这些位点的成对，这些成对可能形成标准内含子。针对这些位点对，TopHat 使用种子扩展策略来检查 IUM reads 以查找跨越剪接位点的 reads。默认情况下，TopHat 检查长于 70 bp 且短于 20,000 bp 的潜在内含子，但是用户可以调整这些默认值以适应特定情况。为了提高运行速度和避免报告假阳性，TopHat 排除了完全位于单个岛屿内的供体-受体对，除非岛屿被深度测序。这确保了在不损失性能和特异性的情况下检测到连接点。在流水线的岛屿提取阶段，算法为地图中从坐标 i 到 j 的每个跨岛屿计算以下统计量：

$$D_{ij} = \frac{\sum_{m=i}^{j} d_m}{j-i} \cdot \frac{1}{\sum_{m=0}^{n} d_m}$$

其中d_m是 Bowtie 地图中坐标m处的覆盖深度，n是参考基因组的长度。当缩放到范围[0, 1000]时，此值表示岛屿的标准化覆盖深度。观察到，单岛屿连接点往往在具有高D（未显示数据）的岛屿内。因此，TopHat 寻找包含在具有$D \geq 300$的岛屿中的连接点，尽管该参数可以由用户更改。TopHat 根据用户设定的 D 值来决定在岛屿中寻找连接点的策略，高D值将提高运行速度，但可能会错过一些连接点；低D值将增加运行时间，但可能会发现更多

的连接点。针对每个剪接位点，TopHat 搜索 IUM reads 以查找跨越剪接位点的 reads，采用了种子和扩展策略。该流水线使用查找表对 IUM reads 进行索引，以减少搜索的成本。TopHat 通过每个边缘至少延伸k个碱基来查找跨越剪接位点的所有 reads，其中$k = 5$ bp，默认值。接着，TopHat 对每个可能的剪接位点采取了$2k$-mer“种子”，然后通过扩展这些种子来找到跨越剪接位点的 reads。为了提高灵敏度，用户可以增加 s 以延长高质量区域的长度，但这会增加运行时间。同时，增加k会提高运行速度，但可能会限制 TopHat 在高表达基因中查找连接点。虽然降低s会减少运行时间，但可能会降低灵敏度，而减少k会提高灵敏度，但可能会增加运行时间和报告假阳性的风险。最后，TopHat 对于每个可能的剪接位点都进行了$2k$-mer“种子”的匹配，并通过左岛和右岛的对齐来扩展种子区域的匹配，允许用户指定错配的数量。虽然 TopHat 可能会错过具有种子区域错配的 reads 的剪接对准，但这种速度和灵敏度之间的权衡通常是合理的。

TopHat 算法会报告所有发现的剪接对准，并利用这些对准构建一个非冗余的剪接位点集合。然而，在报告剪接位点之前，会丢弃一些剪接对准，以避免报告假剪接位点。Wang 等人在其大规模 RNA-Seq 研究中发现，人类中存在数百万种替代剪接事件，其中 86%的次要异构体至少以主要异构体的 15%的水平表达。TopHat 的启发式过滤器基于这一观察。对于每个剪接位点，计算了其左右侧相邻区域的平均读取覆盖深度，然后将穿越剪接位点的对准数量除以覆盖更深的一侧的覆盖范围，以估计次要异构体的频率。如果 TopHat 估计剪接位点发生在其两侧外显子的覆盖深度小于 15%的深度，则不会报告该剪接位点。用户可以调整最小次要异构体频率参数，甚至可以完全禁用此过滤器。尽管 TopHat 中的默认值反映了人类 RNA-Seq 研究的结果，但预计在其他哺乳动物中，次要异构体以类似的频率表达，因此默认值在处理其他哺乳动物的 reads 时也是合适的。

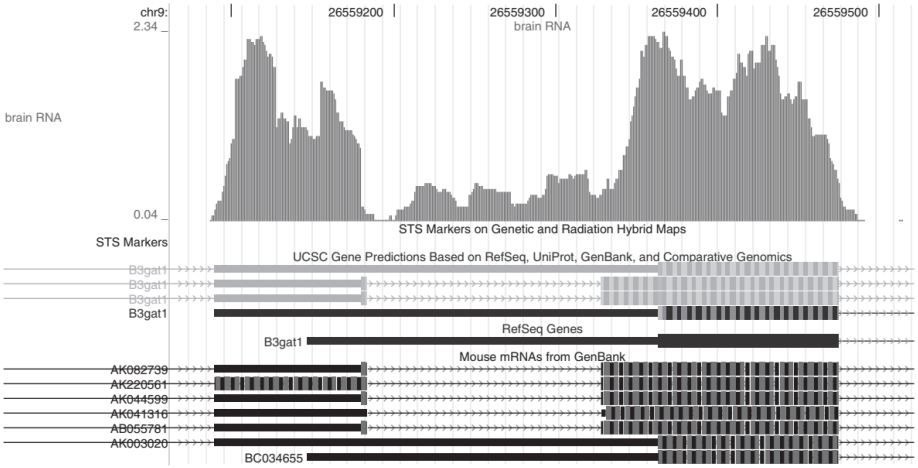

Fig. 2. An intron entirely overlapped by the 5′-UTR of another transcript. Both isoforms are present in the brain tissue RNA sample. The top track is the normalized uniquely mappable read coverage reported by ERANGE for this region (Mortazavi *et al.*, 2008). The lack of a large coverage gap causes TopHat to report a single island containing both exons. TopHat looks for introns within single islands in order to detect this junction.

三、 论文的主要实验结果

将 TopHat 与 ERANGE 在一项最近的 RNA-Seq 研究中使用的 47,781,892 个 25 bp 长的 reads 上进行了比较，该研究使用了小鼠（Mus musculus）脑组织。为了跨越剪接位点对准 reads，ERANGE 在参考基因组中附加了一组跨越序列，其中包含所有注释的剪接位点。对于每个剪接位点，从该位点周围的外显子中提取长度为 L−4 的序列，然后将它们串联起来创建一个跨越序列。这在 M.musculus 中总共构成了 205,151 个接头。Mortazavi 等人将 reads 修剪为 25 bp，因此选择了$s = 25$和$k = 5$，这导致 TopHat 报告了由 read 的 5'端的 25 bp 跨越的接头，每端至少有 5 bp 在接头的两侧。还要求 reads 在剪接位点两侧的外显子序列上完全匹配。此外，仅使用岛屿“伪一致性”序列的参考碱基呼叫。这可能会阻止 TopHat 识别一些具有外显子序列中 SNP 的剪接位点。然而，在岛屿中不正确的碱基呼叫，特别是在岛屿端点附近，会导致更多的剪接位点被错过，而使用组装的岛屿中的参考碱基可以大大减少这个问题。

对于每个基因，ERANGE 报告每百万已映射 reads 的外显子每千碱基的读数（RPKM），这是转录活性的度量。作者将 15.0 和 25.0 视为中等和高水平的转录，ERANGE 在具有正 RPKM 的基因中报告了 108,674 个剪接位点，在具有$RPKM \geq 15.0$的基因中报告了 37,675 个剪接位点。TopHat 报告了 15.0 RPKM 以上的基因中的 ERANGE 剪接位点的 81.9%，以及所有 ERANGE 剪接位点的 72.2%。图 4 显示了 TopHat 在不同基因的 RPKM 下检测剪接位点的灵敏度。TopHat 的能力甚至在具有非常低 RPKM 的基因中检测剪接位点的例子如图 6 所示。在 ERANGE 报告但 TopHat 未报告的 30,121 个剪接位点中，15,689 个（52%）位于表达量低于 5 RPKM 的基因中，可能是由于覆盖不足而被忽略。另外，3209 个（10%）未报告的剪接位点的$RPKM \geq 5.0$，但两端相隔超过 20,000 bp。基于次要异构体比例的过滤排除了 4560 个（15%）剪接位点。TopHat 检测到 ERANGE 排除的数千个已知剪接位点，可能是在其多读“救援”阶段期间排除的，该阶段根据相对表达水平随机将每个剪接多读分配给匹配的基因。

为了评估 TopHat 在识别真实剪接位点而不报告假阳性的能力，模拟了对另类剪接基因进行 Illumina 短读测序的结果。EMBL-EBI 另类剪接转录本数据库（ASTD）包含来自小鼠染色体 7 的 1295 个转录本。这些是由 Maq 的短读模拟器生成的。该模拟器计算读数质量分数的经验分布，并使用这些分数在其生成的读数中生成测序错误。使用了 Mortazavi 等人研究中的 reads 来训练模拟器，因此模拟读数上的测序错误分布应与真实 reads 类似。从 ASTD 转录本中生成了模拟序列，其中包含 9879 个剪接位点，覆盖率分别为 1、5、10、25 和 50 倍。TopHat 在每个覆盖水平下的剪接预测如表 1 所示。TopHat 在小鼠染色体 7 上捕获了 9879 个 ASTD 剪接位点的最多 94%。当转录本的覆盖率低于 5 倍时，灵敏度会下降。即使在深度测序的转录本中，TopHat 也几乎不报告假阳性。

UCSC 基因模型相对保守，因此使用 BLAT 搜索了 GenBank 小鼠 EST 数据库，以查找先前未报告的剪接位点。还搜索了数据库以获取已知剪接位点和随机生成的剪接位点作为阳

性和阴性对照，分别。阳性对照组来自 Mortazavi 等人作为 ERANGE 研究的一部分构建的 205,151 个剪接位点序列。第二组包含 TopHat 报告的先前未报告的剪接位点序列。阴性对照组由第二组的剪接位点序列的左半部分和右半部分的随机配对组成。每个组中的所有序列都是 42 bp 长，每组包含 1000 个随机选择的序列。图 5 显示了每个序列与 GenBank 小鼠 EST 数据库中最佳 BLAST 命中的 E 值分布。预计，几乎所有已知的剪接位点都通过对 EST 的高质量命中得到了确认。预期的是，在“随机配对”阴性对照中，序列缺乏与高质量的 EST 命中。搜索的 1000 个 TopHat 剪接位点中有超过 11％实际上与小鼠 EST 具有高质量的命中。总共，19,722 个 UCSC 基因模型中不存在的剪接位点中有 2543 个具有小鼠 EST 的命中，E 值小于1×10^{-6}。

在对先前未报告且缺乏小鼠 EST 高质量命中的剪接位点进行分类检查后，发现它们可分为三类：连接两个已知外显子的剪接位点、连接已知外显子与新外显子的剪接位点，以及连接两个新外显子的剪接位点。这些剪接位点中，有 10,499 个连接了新的外显子，6077 个连接了一个新的外显子与一个已知的外显子，以及 603 个连接了一对已知的外显子。其中，一个示例是发生在 ADP-核糖基化因子 Arfgef1 中的剪接位点，该基因在囊泡运输中起重要作用。TopHat 报告了 Arfgef1 中几个以前未知的剪接位点，并表明 Arfgef1 是一种选择性剪接。此外，将 TopHat 与基于 RNA-Seq reads 的 de novo 组装策略进行了比较，发现 Velvet+GMAP 方法的灵敏度约为 20％，检测到了 ERANGE 报告的所有剪接位点。尽管该方法在转录值大于 25.0 的基因中检测到约 40％的剪接位点，但随着转录值的进一步增加，其检测率下降。推测许多高度转录的基因具有几个不同的异构体，这些基因中的剪接位点可能会导致 Velvet 在被多个异构体共享的转录本剪接位点处断开 contigs。整个 TopHat 运行耗时 21 小时 50 分钟，在 3.0 GHz Intel Xeon 5160 处理器上使用了不到 4 GB 的 RAM，在每个 CPU 小时内的吞吐量接近 220 万个 reads。

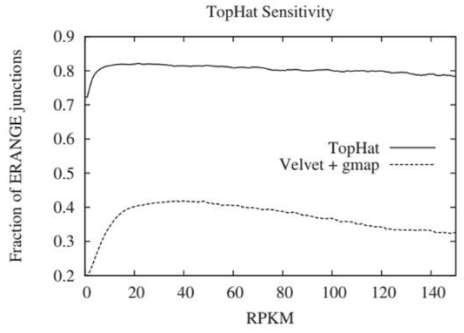

Fig. 4. TopHat sensitivity as RPKM varies. For genes transcribed above 15.0 RPKM, TopHat detects more than 80% reported by ERANGE in the *M. musculus* brain tissue study. TopHat detects more than 72% of all junctions observed by ERANGE, including those in genes expressed at only a single transcript per cell. A *de novo* assembly of the RNA-Seq reads, followed by spliced alignment of the assembled transcripts produces markedly poorer sensitivity, detecting around 40% of junctions in genes transcribed above 25.0 RPKM, but comparatively few junctions in more highly transcribed genes.

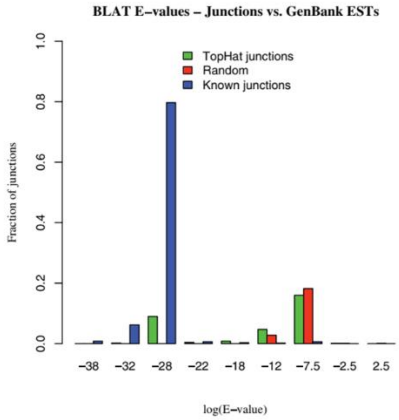

Fig. 5. The BLAT E-value distribution of known, previously unreported, and randomly generated splice junction sequences when searched against GenBank mouse ESTs. As expected, known junctions have high-quality BLAT hits to the EST database. Randomly-generated junction sequences do not. High-quality BLAT hits for more than 11% of the junctions identified by TopHat suggest that the UCSC gene models for mouse are incomplete. These junctions are almost certainly genuine, and because the mouse EST database is not complete, 11% is only a lower bound on the specificity of TopHat.

四、 论文方法的优缺点分析

TopHat 在对 ERANGE 基于注释的分析流程捕获的所有外显子剪接位点中报告了超过72%，包括来自约每个细胞一个转录本的基因的剪接位点。在更活跃转录的基因中，TopHat 捕获了约 80%的剪接位点。更重要的是，TopHat 能够检测到新的剪接位点。尽管很难评估 TopHat 的 19,722 个新发现的剪接位点中有多少是真实的，但本次运行的 TopHat 的对准参数相当严格：仅报告剪接位点的精确匹配，并且要求 reads 在剪接位点的每一侧具有相对较长的锚定序列。对剪接位点的仔细检查加强了许多剪接位点是真实的判断。TopHat 流水线在标准工作站的单个处理器上不到一天就可以处理完整个 RNA-Seq 运行。对于哺乳动物 RNA-Seq 项目中基因表达的高质量测量，只要可靠的外显子-外显子剪接位点注释可用，ERANGE 是合适的。QPALMA 能够准确对齐没有注释的短 reads 跨越剪接位点，但在速度方面做出了巨大的牺牲，因此对于大型哺乳动物项目可能并不实用。因此，在性能上和在发现剪接位点的能力上，TopHat 在 RNA-Seq 剪接检测方法方面代表了重大进步。

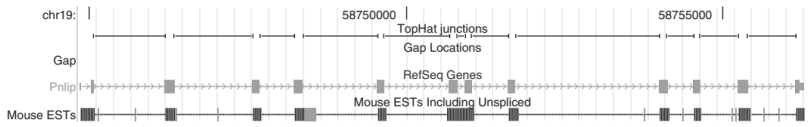

Fig. 6. TopHat detects junctions in genes transcribed at very low levels. The gene *Pnlip* was transcribed at only 7.88 RPKM in the brain tissue according to ERANGE, and yet TopHat reports the complete known gene model.

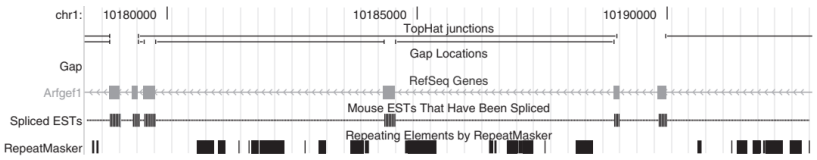

Fig. 7. A previously unreported splice junction detected by TopHat is shown as the topmost horizontal line. This junction skips two exons in the ADP-ribosylation gene *Arfgef1*. As explained in Section 2, islands of read coverage in the Bowtie mapping are extended by 45 bp on either side.

TopHat 流水线及其默认参数值设计用于在转录水平非常低的基因中检测剪接位点。然而，系统可能因各种原因而无法检测到剪接位点。其中，最常见的原因是转录本的测序覆盖率非常低，这可能导致没有足够的序列跨越剪接位点的 reads 以满足每一侧的序列要求。此外，TopHat 也可能错过跨越非常长的内含子或具有非经典供体和受体位点的内含子的剪接位点。另外，在具有低归一化覆盖深度的 island 中，TopHat 也可能错过单个 island 剪接位点。这种情况通常发生在一个异构体的 UTR 完全重叠另一个异构体的内含子时，或者当一个转录本被不完全处理时。虽然 TopHat 捕获了几千个已知的剪接位点，但未被 ERANGE 报告，但这只是反映了两个程序目标不同的差异。ERANGE 主要用于量化基因表达，而 TopHat 旨在识别剪接位点。对于具有多个剪接对准的 reads，ERANGE 将每个 read 分配到一个位置，以增加其表达估计的准确性。如果 TopHat 这样做，其灵敏度会稍微降低。

未来不久，新的 RNA-Seq 协议将产生成对末端 reads，这将使 TopHat 的任务变得更加容易。这将提高剪接检测率，并且假阳性应该会减少，因为 mate-pair 信息可以大大减少必须考虑的可能的剪接数。目前的 TopHat 版本会查找每个 strands 上距离内的所有 islands 之

间的剪接位点。而使用 mate pair 的 TopHat 版本可能只考虑其中一个 mate pair 的每个 island 的配对。此外，剪接和 reads 之间的对齐约束也可以放宽：较长的内含子和具有非经典供体和受体位点的内含子将很容易被检测到。

近期，TopHat 将专注于为外显子提供碱基分辨率的注释，以及对这些外显子的表达进行近似定量。这项任务并不容易，因为必须区分编码区域、UTR 和非编码 RNA。然而，RNA-Seq 在检测转录区域方面的分辨率和经济性大大减少了计算基因预测方法必须考虑的序列量。相信这样的方法将在不久的将来取得巨大成功。目前的流水线无法识别微小外显子（比单个 read 更短），因为它们不会被初始的 Bowtie 映射捕获。使用 IUM reads 的额外映射阶段应该能够捕获其中许多微小外显子。