BIOINFORMATICS ORIGINAL PAPER Vol. 25 no. 9 2009, pages 1105–1111
doi:10.1093/bioinformatics/btp120

*Sequence analysis*

# TopHat: discovering splice junctions with RNA-Seq

Cole Trapnell[1,*], Lior Pachter[2] and Steven L. Salzberg[1]

[1]Center for Bioinformatics and Computational Biology, University of Maryland, College Park, MD 20742 and [2]Department of Mathematics, University of California, Berkeley, CA 94720, USA

Received on October 23, 2008; revised on February 24, 2009; accepted on February 26, 2009

Advance Access publication March 16, 2009

Associate Editor: Ivo Hofacker

**ABSTRACT**

**Motivation:** A new protocol for sequencing the messenger RNA in a cell, known as RNA-Seq, generates millions of short sequence fragments in a single run. These fragments, or 'reads', can be used to measure levels of gene expression and to identify novel splice variants of genes. However, current software for aligning RNA-Seq data to a genome relies on known splice junctions and cannot identify novel ones. TopHat is an efficient read-mapping algorithm designed to align reads from an RNA-Seq experiment to a reference genome without relying on known splice sites.

**Results:** We mapped the RNA-Seq reads from a recent mammalian RNA-Seq experiment and recovered more than 72% of the splice junctions reported by the annotation-based software from that study, along with nearly 20 000 previously unreported junctions. The TopHat pipeline is much faster than previous systems, mapping nearly 2.2 million reads per CPU hour, which is sufficient to process an entire RNA-Seq experiment in less than a day on a standard desktop computer. We describe several challenges unique to *ab initio* splice site discovery from RNA-Seq reads that will require further algorithm development.

**Availability:** TopHat is free, open-source software available from http://tophat.cbcb.umd.edu

**Contact:** cole@cs.umd.edu

**Supplementary information:** Supplementary data are available at *Bioinformatics* online.

## 1 INTRODUCTION

For many years, the standard method for determining the sequence of transcribed genes has been to capture and sequence messenger RNA using expressed sequence tags (ESTs) (Adams *et al.*, 1993) or full-length complementary DNA (cDNA) sequences using conventional Sanger sequencing technology. Recently a new experimental method, RNA-Seq, has emerged that has a number of advantages over conventional EST sequencing: it uses next-generation sequencing (NGS) technologies that can sample the mRNA with fewer biases, it generates far more data per experiment, and it generates data that can be used as a direct measure of the level of gene expression. Thus RNA-Seq experiments not only capture the transcriptome, they can replace conventional microarray experiments for measuring expression. Compared with microarray technology, RNA-Seq experiments provide much higher resolution measurements of expression at comparable cost (Marioni *et al.*, 2008).

The major drawback of RNA-Seq over conventional EST sequencing is that the sequences themselves are much shorter, typically 25–50 nt versus several hundred nucleotides with older technologies. One of the critical steps in an RNA-Seq experiment is that of mapping the NGS 'reads' to the reference transcriptome. However, because the transcriptomes are incomplete even for well-studied species such as human and mouse, RNA-Seq analyses are forced to map to the reference genome as a proxy for the transcriptome. Mapping to the genome achieves two major objectives of RNA-Seq experiments:

(1) Identification of novel transcripts from the locations of regions covered in the mapping.

(2) Estimation of the abundance of the transcripts from their depth of coverage in the mapping.

Because RNA-Seq reads are short, the first task is challenging. Current mapping strategies (e.g. Cloonan *et al.*, 2008; Marioni *et al.*, 2008; Mortazavi *et al.*, 2008; Sultan *et al.*, 2008) include alignment procedures designed to localize Illumina or SOLiD reads to known exons in the genome. However, whenever an RNA-Seq read spans an exon boundary, part of the read will not map contiguously to the reference, which causes the mapping procedure to fail for that read. The studies cited above solve this problem by concatenating known adjacent exons and then creating synthetic sequence fragments from these spliced transcripts. Reads that do not align to the genome but that map to these synthetic fragments represent evidence for splice junctions between known exons.

We can detect splice sites *ab initio* by identifying reads that span exon junctions, but this strategy presents a number of computational challenges, especially with short read lengths. For rarely transcribed genes, many splice junctions may be spanned by very few reads. Therefore, a splice junction mapping algorithm must be able to identify reads that may have only a few bases on one side of a junction, or else that junction will be missed. Improvements in read length will not completely resolve this problem. However, failing to look for novel junctions at a genome-wide scale wastes much of the potential of RNA-Seq for capturing and describing the transcriptome of a human cell (or other species).

One recent method for *ab initio* junction mapping relies on a machine learning strategy to identify junctions. QPALMA (De Bona *et al.*, 2008) trains a support vector machine-like algorithm using known splice junctions from the genome of interest.

*To whom correspondence should be addressed.

While the QPALMA pipeline has organizational similarities to TopHat, there are major differences. First, QPALMA uses a training step that requires a set of known junctions from the reference genome. Second, the QPALMA pipeline's initial mapping phase uses Vmatch (Abouelhoda *et al.*, 2004), a general-purpose suffix array-based alignment program. Vmatch is a flexible, fast aligner, but because it is not designed to map short reads on machines with small main memories, it is substantially slower than other specialized short-read mappers. De Bono *et al*. report that Vmatch maps reads at around 644 400 reads per CPU hour against the 120 Mbp *Arabidopsis thaliana* genome. QPALMA's runtime appears to be dominated by its splice site scoring algorithm; its authors estimate that mapping 71 million RNA-Seq reads to *A.thaliana* would take 400 CPU hours, which is ∼180 000 reads per CPU hour.

In this article, we describe TopHat, a software package that identifies splice sites *ab initio* by large-scale mapping of RNA-Seq reads. TopHat maps reads to splice sites in a mammalian genome at a rate of ∼2.2 million reads per CPU hour. Rather than filtering out possible splice sites with a scoring scheme, TopHat aligns all sites, relying on an efficient 2-bit-per-base encoding and a data layout that effectively uses the cache on modern processors. This strategy works well in practice because TopHat first maps non-junction reads (those contained within exons) using Bowtie (http://bowtie-bio.sourceforge.net), an ultra-fast short-read mapping program (Langmead *et al.*, 2009). Bowtie indexes the reference genome using a technique borrowed from data-compression, the Burrows–Wheeler transform (Burrows and Wheeler, 1994; Ferragina and Manzini, 2001). This memory-efficient data structure allows Bowtie to scan reads against a mammalian genome using around 2 GB of memory (within what is commonly available on a standard desktop computer). Figure 1 illustrates the workflow of TopHat.

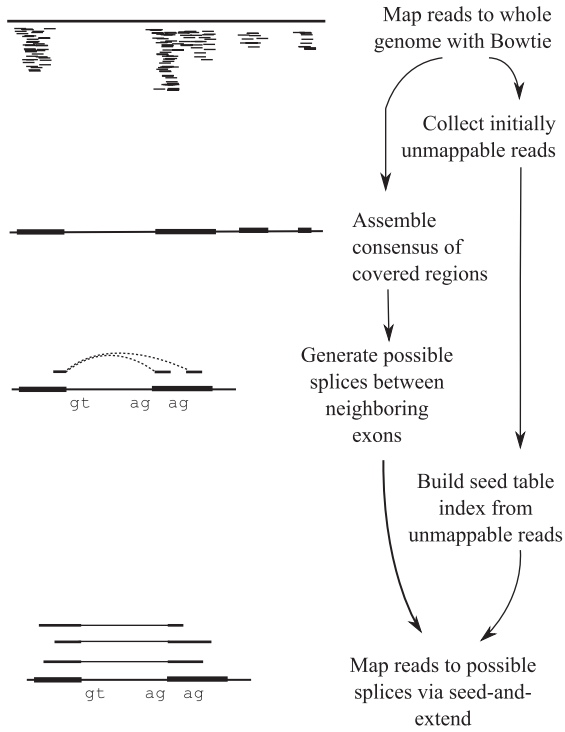


**Fig. 1.** The TopHat pipeline. RNA-Seq reads are mapped against the whole reference genome, and those reads that do not map are set aside. An initial consensus of mapped regions is computed by Maq. Sequences flanking potential donor/acceptor splice sites within neighboring regions are joined to form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

## 2 METHODS

TopHat finds junctions by mapping reads to the reference in two phases. In the first phase, the pipeline maps all reads to the reference genome using Bowtie. All reads that do not map to the genome are set aside as 'initially unmapped reads', or IUM reads. Bowtie reports, for each read, one or more alignment containing no more than a few mismatches (two, by default) in the 5′-most *s* bases of the read. The remaining portion of the read on the 3′ end may have additional mismatches, provided that the Phred-quality-weighted Hamming distance is less than a specified threshold (70 by default). This policy is based on the empirical observation that the 5′ end of a read contains fewer sequencing errors than the 3′ end. (Hillier *et al.*, 2008). TopHat allows Bowtie to report more than one alignment for a read (default = 10), and suppresses all alignments for reads that have more than this number. This policy allows so called 'multireads' from genes with multiple copies to be reported, but excludes alignments to low-complexity sequence, to which failed reads often align. Low complexity reads are not included in the set of IUM reads; they are simply discarded.

TopHat then assembles the mapped reads using the assembly module in Maq (Li *et al.*, 2008). TopHat extracts the sequences for the resulting islands of contiguous sequence from the sparse consensus, inferring them to be putative exons. To generate the island sequences, Tophat invokes the Maq `assemble` subcommand (with the `-s` flag) which produces a compact consensus file containing called bases and the corresponding reference bases. Because the consensus may include incorrect base calls due to sequencing errors in low-coverage regions, such islands may be a 'pseudoconsensus': for any low-coverage or low-quality positions, TopHat uses the reference genome to call the base. Because most reads covering the ends of exons will also span splice junctions, the ends of exons in the pseudoconsensus will initially be covered by few reads, and as a result, an exon's pseudoconsensus will likely be missing a small amount of sequence on each end. In order to capture this sequence along with donor and acceptor sites from flanking introns, TopHat includes a small amount of flanking sequence from the reference on both sides of each island (default = 45 bp).

Because genes transcribed at low levels will be sequenced at low coverage, the exons in these genes may have gaps. TopHat has a parameter that controls when two distinct but nearby exons should be merged into a single exon. This parameter defines the length of the longest allowable coverage gap in a single island. Because introns shorter than 70 bp are rare in mammalian genomes such as mouse (Pozzoli *et al.*, 2007), any value less than 70 bp for this parameter is reasonable. To be conservative, the TopHat default is 6 bp.

To map reads to splice junctions, TopHat first enumerates all canonical donor and acceptor sites within the island sequences (as well as their reverse complements). Next, it considers all pairings of these sites that could form canonical (GT–AG) introns between neighboring (but not necessarily adjacent) islands. Each possible intron is checked against the IUM reads for reads that span the splice junction, as described below. By default, TopHat only examines potential introns longer than 70 bp and shorter than 20 000 bp, but these default minimum and maximum intron lengths can be adjusted by the user. These values describe the vast majority of known eukaryotic introns. For example, more than 93% of mouse introns in the UCSC known gene set fall within this range. However, users willing to make a small sacrifice in sensitivity will see substantially lower running time by reducing the maximum intron length. To improve running times and avoid reporting false positives, the program excludes donor–acceptor pairs that fall entirely within a single island, unless the island is very deeply sequenced. An example of a 'single island' junction is illustrated in Figure 2. The gene shown has two alternate transcripts, one of which has an intron that coincides with the

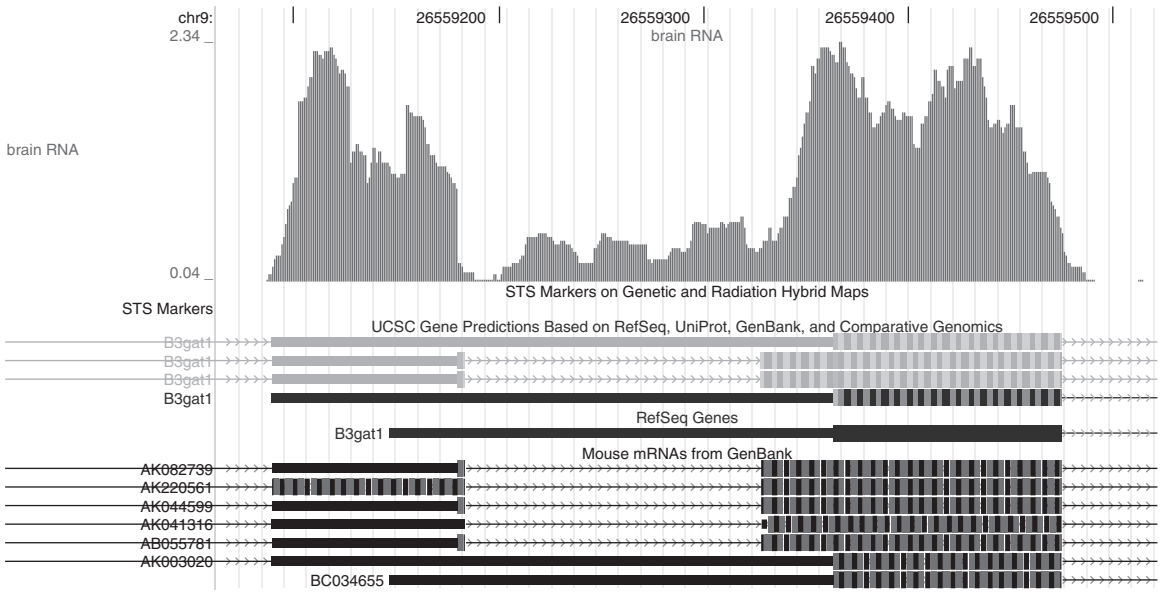


**Fig. 2.** An intron entirely overlapped by the 5′-UTR of another transcript. Both isoforms are present in the brain tissue RNA sample. The top track is the normalized uniquely mappable read coverage reported by ERANGE for this region (Mortazavi *et al.*, 2008). The lack of a large coverage gap causes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

UTR of the other transcript. The figure shows the normalized coverage of the intron and its flanking exons by uniquely mappable reads as reported by Mortazavi *et al*. Both transcripts are clearly present in the RNA-Seq sample, and TopHat reports the entire region as a single island. In order to detect such junctions without sacrificing performance and specificity, TopHat looks for introns within islands that are deeply sequenced. During the island extraction phase of the pipeline, the algorithm computes the following statistic for each island spanning coordinates $i$ to $j$ in the map:

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

where $d_m$ is the depth of coverage at coordinate $m$ in the Bowtie map, and $n$ is the length of the reference genome. When scaled to range [0, 1000], this value represents the normalized depth of coverage for an island. We observed that single-island junctions tend to fall within islands with high $D$ (data not shown). TopHat thus looks for junctions contained in islands with $D \geq 300$, though this parameter can be changed by the user. A high $D$ -value will prevent TopHat from looking for junctions within single islands, which will improve running time. A low $D$ -value will force TopHat to look within many islands, slowing the pipeline, but potentially finding more junctions.

For each splice junction, Tophat searches the IUM reads in order to find reads that span junctions using a seed-and-extend strategy. The pipeline indexes the IUM reads using a simple lookup table to amortize the cost of searching for a spliced alignment over many reads. As illustrated in Figure 3, TopHat finds any reads that span splice junctions by at least $k$ bases on each side (where $k = 5$ bp by default), so the table is keyed by $2k$-mers, where each $2k$-mer is associated with reads that contain that $2k$-mer. For each read, the table contains $(s - 2k + 1)$ entries corresponding to possible positions where a splice may fall within a read, where $s$ is the length of the high-quality region on the 5′ end (default = 28 bp). Users with longer reads may wish to increase $s$ to improve sensitivity. Lowering $s$ will improve running time, but may reduce sensitivity. Increasing $k$ will improve running time, but may limit TopHat to finding junctions only in highly expressed (and thus deeply covered) genes. Reducing it will dramatically increase running time, and while sensitivity will improve, the program may report more false positives. Next TopHat takes each possible splice junction and makes a $2k$-mer 'seed' for it by concatenating the $k$ bases downstream of the acceptor to the $k$ bases upstream of the donor. The IUM read index is then queried with this $2k$-mer to find all reads which contain the seed. This exact $2k$-mer match is extended to find all reads that span the splice junction. To extend the exact match for the seed region, TopHat aligns the portions of the read to the left and right of the seed with the left island and right island, respectively, allowing a user-specified number of mismatches. TopHat will miss spliced alignments to reads with mismatches in the seed region of the splice junction, but we expect this tradeoff between speed and sensitivity will be favorable for most users.

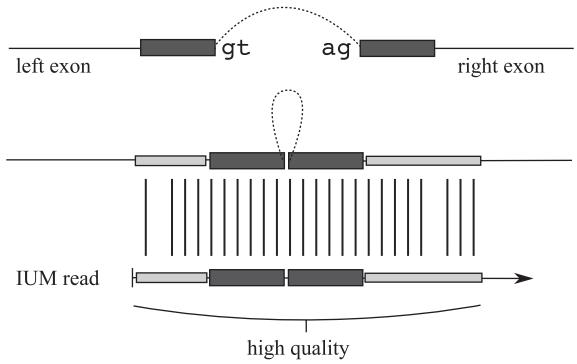


**Fig. 3.** The seed and extend alignment used to match reads to possible splice sites. For each possible splice site, a seed is formed by combining a small amount of sequence upstream of the donor and downstream of the acceptor. This seed, shown in dark gray, is used to query the index of reads that 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.

The algorithm reports all of the spliced alignments it finds, and then builds a set of non-redundant splice junctions using these alignments. However, some spliced alignments are discarded prior to reporting junctions in order to avoid reporting false junctions. In their large-scale RNA-Seq study, Wang *et al.* (2008) reported millions of alternative splicing events in humans and observed that 86% of the minor isoforms were expressed at at least 15% of the level of the major isoform. TopHat's heuristic filter for spliced alignments is based on this observation. For each junction, the average depth of read coverage is computed for the left and right flanking regions of the junction separately. The number of alignments crossing the junction is divided by the coverage of the more deeply covered side to obtain an estimate of the minor isoform frequency. If TopHat estimates that the splice junction occurs at <15% of the depth of coverage of the exons flanking it, the junction is not reported. The minimum minor isoform frequency parameter is adjustable by the user, and may be entirely disabled. While the default value in TopHat reflects a result from a human RNA-Seq study, we expect that minor isoforms are expressed at similar frequencies in other mammals, and that the value will be suitable when the software is used to process reads from other mammals.

## 3 RESULTS

We compared TopHat with ERANGE on a set of 47 781 892 reads, each 25 bp long, from a recent RNA-Seq study using *Mus musculus* brain tissue (Mortazavi *et al.*, 2008). To align reads across splice junctions, ERANGE appends to the reference genome a set of spanning sequences that contain all annotated splice sites. For each splice site, a sequence of length $L-4$ (for reads of length $L$) is extracted from the exons flanking that site, and these are concatenated to create a spanning sequence. This constituted a total of 205 151 junctions for *M.musculus*. Mortazavi *et al*. trimmed reads to 25 bp, so we chose $s=25$ and $k=5$, which caused TopHat to report junctions spanned by the 25 bp on the 5′ end of a read, with at least 5 bp on each side of the junction. We also required reads to match the exon sequence on each side of the junction exactly. In addition, we used only reference base calls for the island 'pseudoconsensus' sequences. This may have prevented TopHat from identifying some junctions with SNPs in the flanking exon sequence. However, incorrect base calls in islands, especially near island endpoints, would cause many more junctions to be missed, a problem that was greatly reduced by the use of the reference bases within our assembled islands.

For each gene, ERANGE reports the number of mapped reads per kilobase of exon per million mapped reads (RPKM), a measure of transcription activity. The authors characterize 15.0 and 25.0 as moderate and high levels of transcription, respectively. ERANGE reported 108 674 splice junctions in genes with positive RPKM, and 37 675 junctions in genes with RPKM $\geq$ 15.0. TopHat reported 81.9% of the ERANGE junctions in genes above 15.0 RPKM, and 72.2% of all ERANGE junctions. Figure 4 shows how TopHat's sensitivity in detecting junctions varies with the RPKM of the genes. An example of TopHat's ability to detect junctions even in genes with very low RPKM is illustrated in Figure 6. Of the 30 121 junctions reported by ERANGE and not reported by TopHat, 15 689 (52%) fell within genes expressed below 5 RPKM and were likely missed due to lack of coverage. A further 3209 (10%) of the missed junctions had RPKM $\geq$5.0 but had endpoints more than 20 000 bp apart. Filtering based on minor isoform fraction excluded 4560 (15%) junctions. TopHat detected several thousand known splice junctions that ERANGE excluded, presumably during its multiread 'rescue' phase, where it randomly assigns each spliced multiread to matched

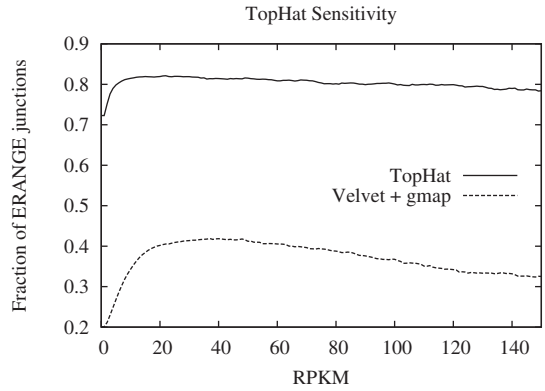


**Fig. 4.** TopHat sensitivity as RPKM varies. For genes transcribed above 15.0 RPKM, TopHat detects more than 80% reported by ERANGE in the *M. musculus* brain tissue study. TopHat detects more than 72% of all junctions observed by ERANGE, including those in genes expressed at only a single transcript 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.

**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.

genes according to their relative expression levels. Of the 104 711 junctions reported by TopHat, 84 988 are listed among the UCSC gene models for *M. musculus*, or 81.1%. The remaining 19 722 may represent novel junctions.

To assess TopHat's ability to identify true junctions without reporting false positives, we simulated the results of Illumina short-read sequencing of alternatively spliced genes at several depths. The EMBL-EBI Alternative Splicing Transcript Database (ASTD) (Le Texier *et al.*, 2006) contains 1295 transcripts from mouse chromosome 7. These were generated by the short-read simulator from Maq. The simulator computes an empirical distribution of read quality scores and uses these to generate sequencing errors in the reads it produces. We trained the simulator using the reads from the Mortazavi *et al*. study, so the sequencing error profile on simulated reads should be similar to the real reads. We generated simulated sequence from the ASTD transcripts, which contained 9879 splice junctions, at 1-, 5-, 10-, 25- and 50-fold coverage. TopHat's junction predictions at each coverage level are summarized in Table 1. TopHat captures up to 94% of the 9879 ASTD splice junctions on mouse chromosome 7. Sensitivity suffers when transcripts are sequenced

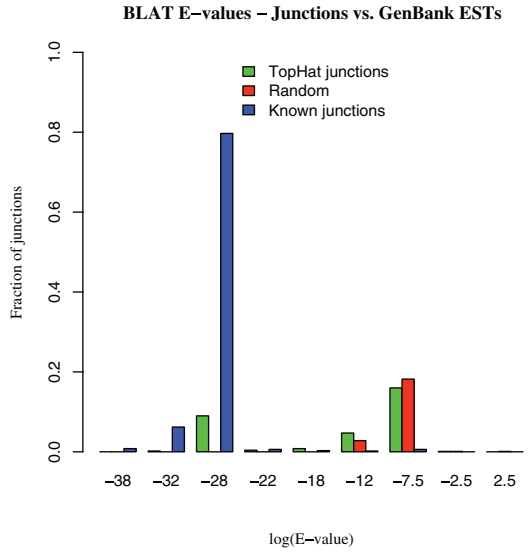


**Fig. 5.** The BLAT E-value distribution of known, previously unreported, and randomly generated splice junction sequences when searched against GenBank mouse ESTs. As expected, known junctions have high-quality BLAT hits to the EST database. Randomly-generated junction sequences do not. High-quality 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.

at less than 5-fold coverage. TopHat reports few false positives even in deeply sequenced transcripts.

The UCSC gene models are relatively conservative, so we searched the GenBank mouse EST database using BLAT (Kent, 2002) for the previously unreported junctions. We also searched the database for known junctions and randomly generated junctions as positive and negative controls, respectively. The positive control group was drawn from the 205 151 junction sequences constructed by Mortazavi *et al*. as part of the ERANGE study. The second set consisted of previously unreported junction sequences reported by TopHat. The negative control consisted of random pairings of the left and right halves of junction sequences from the second group. All sequences in each of the three groups were 42 bp long, and each group contained 1000 sequences chosen randomly. Figure 5 shows the distribution of *E*-values for each sequence's best BLAST hit against the GenBank mouse EST database. As expected, nearly all of the known junctions are confirmed by high-quality hits to ESTs. Also expected is the lack of high-quality hits for sequences in the 'random-pairing' negative control. More than 11% of the 1000 TopHat junctions we searched for actually have high-quality hits to mouse ESTs. In total, 2543 of the 19 722 junctions not in UCSC gene models had hits to mouse ESTs with $E$-value $< 1 \times 10^{-6}$.

We examined the previously unreported junctions that lacked high-quality hits to mouse EST by dividing them into three categories: junctions between two known exons, junctions between a known exon and a novel one and junctions between two novel exons. Of the 17 719 junctions without EST hits, 10 499 joined novel exons, 6077 joined a novel exon with a known one and 603 joined a pair of known exons. One example of a junction from the second category is occurred in the ADP-ribosylation factor *Arfgef1*, which is important in vesicular trafficking (Morinaga *et al.*, 1996). The junction in Figure 7 skips two of the gene's 38 exons. TopHat reported several junctions in *Arfgef1* that were previously unknown and indicates that *Arfgef1* is alternatively spliced.

We also compared TopHat to a simple strategy based on *de novo* assembly of RNA-Seq reads. The advantage of such a strategy is that, like TopHat, no known junctions or gene models are needed. We ran the Velvet short-read assembler (Zerbino and Birney, 2008) (version 0.7.11, -k=21) on our RNA-Seq reads to produce 149 628 transcript contigs with N50 = 131. We then aligned these contigs back to the mouse reference genome using the spliced alignment program GMAP (Wu and Watanabe, 2005), one of the leading methods for alignment of ESTs and full-length cDNAs to genomic DNA. The sensitivity of the Velvet+GMAP method is shown in Figure 4. The method detects around 20% of all junctions reported by ERANGE. While the method detects around 40% of junction in genes transcribed above an RPKM value of 25.0, its detection rate decreases as RPKM further increases. We speculate that many of these highly transcribed genes have several alternate isoforms, and that junctions in these genes may cause Velvet to break contigs at the transcript junctions shared by multiple isoforms.

The entire TopHat run took 21 h, 50 min on a 3.0 GHz Intel Xeon 5160 processor, using <4 GB of RAM, a throughput of nearly 2.2 million reads per CPU hour.

## 4 DISCUSSION

In our comparison, TopHat reported more than 72% of all exon splice junctions captured by the ERANGE annotation-based analysis pipeline, including junctions from genes transcribed at around one transcript per cell. TopHat captured around 80% of splice junctions in more actively transcribed genes. More significant is its ability to detect novel splice junctions. While it is difficult to assess how many of TopHat's 19 722 newly discovered junctions are genuine, TopHat's alignment parameters for this run were quite strict: only exact matches were reported for splice junctions, and reads were required to have relatively long anchors on each side of the splice site. Close inspection of junctions strengthened the case that many are true splices. The TopHat pipeline processed an entire RNA-Seq run in less than a day on a single processor of a standard workstation. ERANGE is appropriate for high-quality measurement of gene expression in mammalian RNA-Seq projects, provided that a reliable annotation of exon–exon junctions is available. QPALMA can accurately align short reads across junctions without an annotation, but makes such substantial sacrifices in speed that it may not be practical for large mammalian projects. TopHat thus represents a significant advance over previous RNA-Seq splice detection methods, both in its performance and its ability to find junctions *de novo*.

The TopHat pipeline and its default parameter values are designed for detecting junctions even in genes transcribed at very low levels. However, the system may fail to detect junctions for a variety of reasons. The most common reason for missing a junction is that the transcript has very low sequencing coverage, in which case there might be no read that straddles the junction with sufficient sequence on each side. Junctions spanning very long introns or introns with non-canonical donor and acceptor sites (such as GC–AG introns) will also be missed. As discussed in Section 2, TopHat can

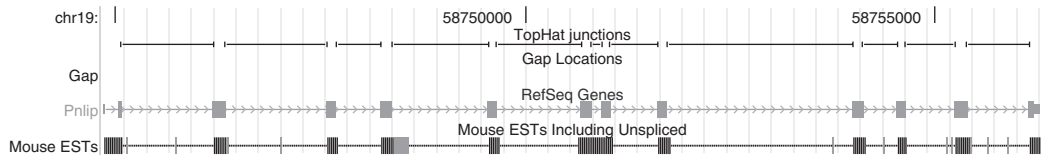


**Fig. 6.** TopHat detects junctions in genes transcribed at very low levels. The gene *Pnlip* was transcribed at only 7.88 RPKM in the brain tissue according to ERANGE, and yet TopHat reports the complete known gene model.

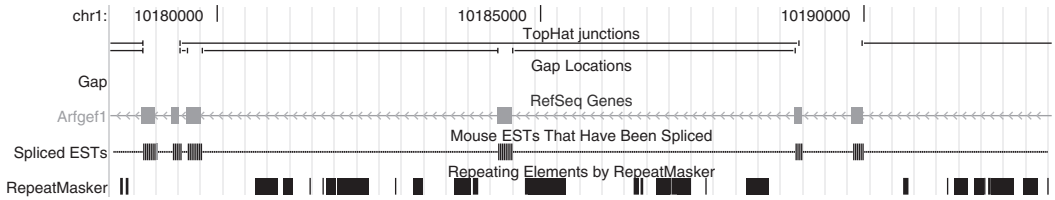


**Fig. 7.** A previously unreported splice junction detected by TopHat is shown as the topmost horizontal line. This junction skips two exons in the ADP-ribosylation gene *Arfgef1*. As explained in Section 2, islands of read coverage in the Bowtie mapping are extended by 45 bp on either side.

also miss single-island junctions in islands with a low normalized depth of coverage. Single-island junctions can occur when the UTR of one isoform entirely overlaps an intron from another isoform, as illustrated in Figure 2. They may also occur when a transcript is incompletely processed. While several thousand known junctions were captured by TopHat but not reported by ERANGE, this merely reflects differences in the goal of the two programs. ERANGE is primarily meant to quantitate gene expression, while TopHat aims to identify junctions. For reads with multiple spliced alignments, ERANGE assigns each read to a single position, in order to increase the accuracy of its expression estimates. Were TopHat to do this, its sensitivity would suffer slightly.

In the near future, new RNA-Seq protocols that produce paired-end reads will make TopHat's task easier. Splice detection rates will improve, and false positives should become much less common, as mate-pair information can drastically reduce the number of possible splices that must be considered. The current version of TopHat looks for splice junctions between all islands within a certain distance of each other on each strand of the reference. A version of TopHat that made use of mate pairs might consider only pairings of islands where one read from a mate pair maps to each island. The alignment constraints between splices and reads can also be relaxed: longer introns and those with non-canonical donor and acceptors sites will be readily detectable.

In the nearer term, TopHat will aim to provide base-pair resolution exon annotations along with approximate quantitation of expression for those exons. This task is not without difficulty, since coding regions must still be distinguished from UTRs and non-coding RNAs. However, the resolution and economy of RNA-Seq in detecting transcribed regions dramatically reduces the amount of sequence that must be considered by a computational gene prediction approach. We are confident that such methods will see great success in the near future. The current pipeline has no means of identifying microexons (shorter than a single read) because they will not be captured by the initial Bowtie mapping. An additional mapping phase using IUM reads should be able to capture many of these microexons.

## 5 SOFTWARE

TopHat is implemented in C++ and Python and runs on Linux and Mac OS X. It makes substantial use of previously described tools, including Bowtie (Langmead *et al.*, 2009), Maq (Li *et al.*, 2008) and the SeqAn library (Döring *et al.*, 2008).

## ACKNOWLEDGEMENTS

We thank Adam Phillippy, Geo Pertea, Ben Langmead, Kasper Hansen, Angela Brooks and Ali Mortazavi for helpful technical discussions. We thank Diane Trout, Ali Mortazavi, Brian Williams, Kenneth McCue, Lorian Schaeffer and Barbara Wold for making their data available for our case study.

*Funding:* National Institues of Health (R01-LM06845, R01-GM083873 to S.L.S.); National Science Foundation (CCF 0347992 to L.P.).

*Conflict of Interest*: none declared.

## REFERENCES

Abouelhoda,M. *et al.* (2004) Replacing suffix trees with enhanced suffix arrays. *J. Discrete Alg.*, **2**, 53–86.

Adams,M.D. *et al.* (1993) Rapid cDNA sequencing (expressed sequence tags) from a directionally cloned human infant brain cDNA library. *Nat. Genet.*, **4**, 373–380.

Burrows,M. and Wheeler,D. (1994) A block sorting lossless data compression algorithm. *Technical Report 124*, DEC, Digital Systems Research Center, Palo Alto, California.

Cloonan,N. *et al.* (2008) Stem cell transcriptome profiling via massive-scale mRNA sequencing. *Nat. Meth.*, **5**, 613–619.

De Bona,F. *et al.* (2008) Optimal spliced alignments of short sequence reads. *Bioinformatics*, **24**, i174–i180.

Döring,A. *et al.* (2008) Seqan an efficient, generic c++ library for sequence analysis. *BMC Bioinformatics*, **9**, 11.

Ferragina,P. and Manzini,G. (2001) An experimental study of an opportunistic index. In *Proceedings of the Twelfth Annual ACM-SIAM Symposium on Discrete Algorithms*. Washington, D.C. USA, pp. 269–278.

Hillier,L.W. *et al.* (2008) Whole-genome sequencing and variant discovery in C. elegans. *Nat. Meth.*, **5**, 183–188.

Kent,W.J. (2002) Blat—the blast-like alignment tool. *Genome Res.*, **12**, 656–664.

Langmead,B. *et al.* (2009) Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol.*, **10**, R25.

Le Texier,V. *et al.* (2006) Alttrans: transcript pattern variants annotated for both alternative splicing and alternative polyadenylation. *BMC Bioinformatics*, **7**, 169.

Li,H. *et al.* (2008) Mapping short dna sequencing reads and calling variants using mapping quality scores. *Genome Res.*, **18**, 1851–1858.

Marioni,J. *et al.* (2008) RNA-Seq: an assessment of technical reproducibility and comparison with gene expression arrays. *Genome Res.*, **18**, 1509–1517.

Morinaga,N. *et al.* (1996) Isolation of a brefeldin A-inhibited guanine nucleotide-exchange protein for ADP ribosylation factor (ARF) 1 and ARF3 that contains a Sec7-like domain. *Proc. Natl Acad. Sci. USA*, **93**, 12856–12860.

Mortazavi,A. *et al.* (2008) Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat. Meth.*, **5**, 621–628.

Pozzoli,U. *et. al.* (2007) Intron size in mammals: complexity comes to terms with economy. *Trends Genet.*, **23**, 20–24.

Sultan,M. *et al.* (2008) A global view of gene activity and alternative splicing by deep sequencing of the human transcriptome. *Science*, **321**, 956.

Wang,E.T. *et al.* (2008) Alternative isoform regulation in human tissue transcriptomes. *Nature*, **456**, 470–476.

Wu,T.D. and Watanabe,C.K. (2005) GMAP: a genomic mapping and alignment program for mRNA and EST sequences. *Bioinformatics*, **21**, 1859–1875.

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