Title

TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions

Journal

Genome Biology, 14(4)

ISSN

1465-6906

Authors

Kim, Daehwan
Pertea, Geo
Trapnell, Cole
et al.

Publication Date

2013-04-25

DOI

http://dx.doi.org/10.1186/gb-2013-14-4-r36

Genome Biology

METHOD Open Access

TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions

Daehwan Kim[1,2,3*], Geo Pertea[3], Cole Trapnell[5,6], Harold Pimentel[7], Ryan Kelley[8] and Steven L Salzberg[3,4]

Abstract

TopHat is a popular spliced aligner for RNA-sequence (RNA-seq) experiments. In this paper, we describe TopHat2, which incorporates many significant enhancements to TopHat. TopHat2 can align reads of various lengths produced by the latest sequencing technologies, while allowing for variable-length indels with respect to the reference genome. In addition to *de novo* spliced alignment, TopHat2 can align reads across fusion breaks, which can occur after genomic translocations. TopHat2 combines the ability to identify novel splice sites with direct mapping to known transcripts, producing sensitive and accurate alignments, even for highly repetitive genomes or in the presence of pseudogenes. TopHat2 is available at http://ccb.jhu.edu/software/tophat.

Background

RNA-sequencing technologies [1], which sequence the RNA molecules being transcribed in cells, allow exploration of the process of transcription in exquisite detail. One of the primary goals of RNA-sequencing analysis software is to reconstruct the full set of transcripts (isoforms) of genes that were present in the original cells. In addition to the transcript structures, experimenters need to estimate the expression levels for all transcripts. The first step in the analysis process is to map the RNA-sequence (RNA-seq) reads against the reference genome, which provides the location from which the reads originated. In contrast to DNA-sequence alignment, RNA-seq mapping algorithms have two additional challenges. First, because genes in eukaryotic genomes contain introns, and because reads sequenced from mature mRNA transcripts do not include these introns, any RNA-seq alignment program must be able to handle gapped (or spliced) alignment with very large gaps. In mammalian genomes, introns span a very wide range of lengths, typically from 50 to 100,000 bases, which the alignment algorithm must accommodate. Second, the presence of processed pseudogenes, from which some or all introns have been removed, may cause many exon-spanning reads to map incorrectly. This is particularly acute for the human genome, which contains over 14,000 pseudogenes [2].

In the most recent Ensembl GRCh37 gene annotations, the average length of a mature mRNA transcript in the human genome is 2,227 bp long, and the average exon length is 235 bp. The average number of exons per transcript is 9.5. Assuming that sequencing reads are uniformly distributed along a transcript [3], we would expect 33 to 38% of 100 bp reads from an RNA-seq experiment to span two or more exons. Note that this proportion increases significantly as read length increases from 50 to 150 bp (see Additional file 1 for more details).

More important for the alignment problem is that around 20% of junction-spanning reads extend by 10 bp or less into one of the exons they span. These small 'anchors' make it extremely difficult for alignment software to map reads accurately, particularly if the algorithm relies (as most do) on an initial mapping of fixed-length k-mers to the genome. This initial mapping, using exact matches of k-mers, is crucial for narrowing down the search space into small local regions in which a read is likely to align. If a read extends only a few bases into one of two adjacent exons, then it often happens that the read will align equally well, but incorrectly, with the sequence of the intervening intron. For example, as illustrated in Figure 1, suppose that read r spans exons e_1 and e_2, extending only four bases into e_2. Suppose also that that e_2 begins with

* Correspondence: infphilo@umiacs.umd.edu
[1]Center for Bioinformatics and Computational Biology, University of Maryland, College Park, MD, 20742, USA
Full list of author information is available at the end of the article

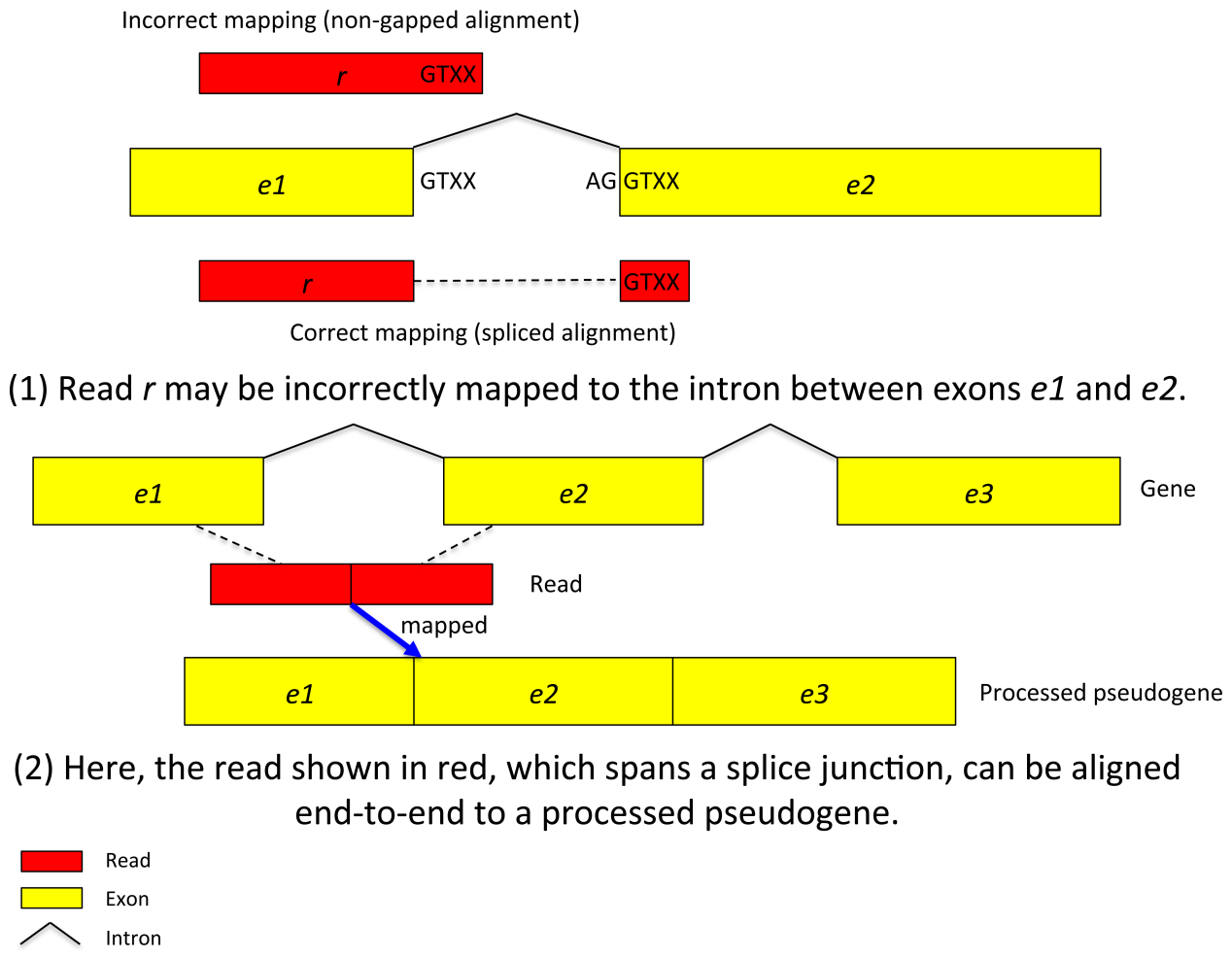

Figure 1 Two possible incorrect alignments of spliced reads. 1) A read extending a few bases into the flanking exon can be aligned to the intron instead of the exon. 2) A read spanning multiple exons from genes with processed pseudogene copies can be aligned to the pseudogene copies instead of the gene from which it originates.

GTXX, and the intervening intron also begins with GTXX, where X stands for any of A, G, C, and T. Then *r* might align perfectly to e_1 and the first four bases of the intron, and the alignment algorithm will fail to find the spliced alignment of *r*.

In order to handle this problem, the software TopHat2 uses a two-step procedure. First, similar to TopHat1 [4], it detects potential splice sites for introns (detailed further in Methods). It uses these candidate splice sites in a subsequent step to correctly align multiexon-spanning reads. Some RNA-seq aligners, including GSNAP [5], RUM [6], and STAR [7], map reads independently of the alignments of other reads, which may explain their lower sensitivity for these spliced reads (see Results and discussion). MapSplice [8] uses a two-step approach similar to TopHat2.

RNA-seq read alignment is further complicated by the presence of processed pseudogenes in the reference genome. Pseudogenes often have highly similar sequences to functional, intron-containing genes. In most cases, the pseudogene versions are not transcribed [9], although this suggestion has recently been disputed [10]. The crucial problem for alignment is that reads spanning multiple exons can be mapped perfectly or near-perfectly to the pseudogene version of a functional gene. For example, suppose a read *r* spans two exons of a given gene. If the aligner tries to align the read globally (end to end), then it will find an alignment to the pseudogene copy (Figure 1). If the spliced alignment phase, which usually occurs later, does not attempt to realign *r*, then the pseudogene copy will 'absorb' all reads spanning the splice sites for that gene. TopHat2 can feed *r* into the spliced alignment phase even when *r* has been aligned end to end, allowing it to circumvent this problem (see Methods and Results and discussion sections).

We also note that, in our analysis of RNA-seq reads from multiple human samples [11,12], genes with

processed pseudogenes seem to be expressed at higher levels compared with other genes (see Results and discussion). Although this observation has not been explored thoroughly, a plausible explanation is that genes with higher levels of expression may, over the course of evolution, have had an increased chance of being picked up by transposons and re-integrated into the genome, creating pseudogene copies.

Concerning the human genome, for which there are relatively comprehensive annotations of protein-coding genes, the annotations can be used to map reads more accurately by aligning the reads preferentially to real genes rather than pseudogenes. GSNAP [5] and STAR [7] also make use of annotation, although they use it in a more limited fashion in order to detect splice sites. TopHat2 can use the full-length transcripts defined by annotations during its initial mapping phase, which produces significant gains in sensitivity and accuracy (see Results and discussion).

Transcripts from a target genome may differ substantially from the reference genome, possibly containing insertions, deletions, and other structural variations [13,14]. For such regions, previous spliced alignment programs (including the original TopHat) sometimes fail to find a proper alignment. In TopHat2, we implemented new procedures to ensure that reads are aligned with true insertions and deletions (indels). Indels due to sequencing errors will be discovered by Bowtie2 [15], the underlying mapping engine of TopHat2, which can detect short indels very efficiently. In addition, very large deletions, inversions on the same chromosome, and translocations involving different chromosomes are detected by the TopHat-Fusion algorithms [16], which are now incorporated into TopHat2 and available by a simple command-line switch.

TopHat2 also includes new algorithms to handle more diverse types of sequencing data. This includes the ability to handle reads generated by ABI SOLiD technology (Life Technologies, Carlsbad, CA, USA) using its 'color space' representation. To accomplish this, TopHat2 uses a reference genome translated entirely into color space in order to take advantage of the error-correction capability of that format. TopHat2 also handles datasets in which the reads have variable lengths, allowing the experimenter to merge datasets from multiple sequencing runs with different lengths.

Results and discussion

TopHat2 can use either Bowtie [17] or Bowtie2 [15] as its core read-alignment engine. TopHat2 has its own indel-finding algorithm, which enhances indel-finding ability of Bowtie2 in the context of spliced alignments. In order to evaluate TopHat2 and compare it with other methods, we ran multiple computational experiments using both real and simulated RNA-seq data.

For the simulations, we created multiple sets of 40,000,000 paired-end reads, 100 bp in length, from the entire human genome (release GRCh37). Instead of trying to precisely mimic real RNA-seq experiments, which may not be possible in any practical sense, we generated data with relatively simple settings and expression levels, calculated using a model from the Flux Simulator system [18], as follows. For the first test set, we generated reads from the known transcripts on the entire human genome without introducing any mismatches or indels. We then generated additional datasets, in which we included 1) insertions and deletions into the known transcripts at random locations, and 2) insertions and deletions in the reads themselves to mimic sequencing errors (see Additional file 1). Each of these experimental errors was introduced to test different capabilities of TopHat2 and other RNA-seq aligners. Following the simulations, we evaluated the programs using a recent, real RNA-seq dataset.

Alignments of simulated reads (error-free)

We generated 40,000,000 paired-end reads and performed two sets of experiments, using: 1) 20,000,000 'left' reads from the paired-end dataset (Table 1) and 2) 20,000,000 pairs of reads (Table 2). Reads spanning multiple exons are called junction reads, and our single-end data contained 6,862,278 such reads (34.3%). The most challenging alignments are those for which a junction read extends by 10 bp or less into one of the exons, which we call short-anchored reads; 1,448,022 of the single-end reads (7.2%) fell into this category. We report the accuracy junction reads and short-anchored reads separately (Table 1, Table 2).

We also tested 20,000,000 read pairs (40,000,000 reads), of which 9,491,394 (47.5%) had at least one read spanning multiple exons; 2,702,624 of these pairs (13.5%) had at least one short-anchored read that extended by10 bp or less into one of its exons. Table 2 shows the results of mapping these reads with TopHat2 and other programs.

As shown in Table 1, TopHat2 correctly aligned more than 98% of the reads, which was higher than with any of the other methods, whose accuracy ranged from 88 to 97%. The difference was more pronounced for junction reads, with TopHat2 being able to align more than 94% of the reads, whereas the other methods ranged in accuracy from 65 to 92%. GSNAP, RUM, and STAR had particular difficulty aligning short-anchored reads, aligning only 26%, 8.6%, and 3.5%, respectively, while MapSplice performed considerably better, aligning 75.6% of these reads. By contrast, using Bowtie1 as its main aligner, TopHat2 aligned 93.7% of the short-anchored reads (Table 1). Both TopHat2 and MapSplice use a two-step algorithm, first detecting potential splice sites,

and then using these sites to map reads. This two-step method may explain their superior performance at mapping reads with short anchors.

The results for the paired reads (Table 2) were similar to those for the unpaired reads. TopHat2 aligned the highest percentage of reads (96.7%), followed by MapSplice (92%), with the other methods ranging from 79 to 88%. The difference widened again for junction reads, with TopHat2 aligning 93%, followed by MapSplice (86%), GSNAP (76%), STAR (69%), and RUM (56%). Most striking of all was the performance on short-anchored reads, which most of the methods had great difficulty aligning correctly; TopHat2 aligned 90% of these, MapSplice aligned 72%, and the other methods aligned only 3 to 22%.

We assessed alignment rates for reads, junction reads, and junction reads with small anchors for a variety of read lengths (50, 100, 150, and 200 bp) (see Additional file 1, Figure S1). TopHat2 consistently outperformed all the other aligners for each read length. Comparing the alignment performance for junction reads with one to three mismatches, TopHat2 and MapSplice showed the highest recall rates (see Additional file 1, Table S2).

Alignments of simulated reads with short indels

We tested the spliced alignment programs using reads with small indels (1 to 3 bp), using two sets of simulated reads: (1) true indels, in which the transcripts were modified by inserting or deleting one to three bases at random locations; and (2) indels caused by sequencing errors, in which indels are randomly inserted into the reads. As before, all transcripts were simulated from known genes from the entire human genome. We intentionally used a relatively high rate of indels to test the mapping capabilities of the programs in the presence of these types of mutations.

For single-end reads, RUM, GSNAP, and TopHat2 performed similarly, with 69 to 82% accuracy (recall) rates for true indels and 62 to 83% for reads with indel-sequencing errors (Table 3, Table 4). STAR and MapSplice showed relatively lower recall rates for both datasets. Note that when used with the original Bowtie program (a non-

Table 1 Performance of TopHat2 and other spliced aligners on a set of 20 million 100-bp, single-end reads, simulated based on transcripts from the entire human genome.

Program	No. of mapped reads	Correctly mapped reads, %	Incorrectly mapped reads, %	Unmapped reads, %	Correct junction reads, %[a]	Correct short-anchored reads, %[b]
TopHat2 + Bowtie1	19,826,638	98.31	0.82	0.87	95.28	93.69
TopHat2 + Bowtie2	19,826,673	98.03	1.10	0.87	94.28	89.67
TopHat1.14	19,616,874	94.64	3.45	1.91	84.44	44.08
GSNAP	19,997,255	94.21	5.77	0.02	83.15	26.01
RUM	19,555,823	88.11	9.67	2.22	65.35	8.59
MapSplice	19,872,372	97.28	2.08	0.64	92.09	75.57
STAR	19,087,508	92.14	3.30	4.56	77.17	3.54

[a]There were 6,862,278 reads spanning one or more splice junctions; the alignment accuracy of junction reads refers to this set.

[b]There were 1,448,022 reads extending 10 bp or less into one exon; the alignment accuracy of the short-anchored reads is based on these alignments.

Table 2 Performance of TopHat2 and other spliced aligners on a set of 20 million pairs of 100-bp reads, simulated based on transcripts from the entire human genome.

Program	No. of mapped pairs	Correctly mapped pairs, %	Incorrectly mapped pairs, %	Unmapped pairs, %	Correct junction pairs, %[a]	Correct short-anchored pairs, %[b]
TopHat2 + Bowtie1	19,683,426	96.70	1.72	1.58	93.31	90.09
TopHat2 + Bowtie2	19,686,006	96.19	2.24	1.57	92.03	85.88
TopHat1.14	19,219,055	89.57	6.53	3.90	78.36	40.39
GSNAP	19,999,867	88.84	11.16	0.00	76.55	22.87
RUM	19,869,579	79.07	20.28	0.65	56.28	8.42
MapSplice	19,342,087	92.03	4.68	3.29	86.53	72.48
STAR	19,951,620	85.21	14.55	0.24	68.94	3.16

[a]There were 9,491,394 pairs of reads classified as junction pairs.

[b]There were 2,702,624 pairs containing short-anchored reads.

Table 3 Performance of TopHat2 and other spliced aligners on single-end reads containing insertions and deletions (indels) of 1 to 3 bp.

Program	Reads with true indels (1,428,499)[a]		Reads with sequencing-error indels (1,525,657)[a]	
	Accuracy, %	Accuracy on 351,465 reads with boundary indels, %[b]	Accuracy, %	Accuracy on 357,334 reads with boundary indels, %[b]
TopHat2 + Bowtie1	70.9	16.8	12.1	2.8
TopHat2 + Bowtie2	63.7	25.2	62.6	21.2
GSNAP	82.7	71.9	83.1	71.8
RUM	69.4	43.0	70.3	45.4
MapSplice	27.3	3.7	27.5	3.8
STAR	46.6	16.9	47.7	17.1

[a]The number of reads containing each type of error is indicated in the column header. [b]Boundary indels occur within 25 bp of an exon boundary. Percentages refer only to the reads of each type, not to the entire dataset.

Table 4 Performance of TopHat2 and other spliced aligners on paired reads in which at least one of the reads contained insertions and deletions (indels) of 1 to 3 bp.

Program	Pairs with true indels (2,754,313)[a]		Pairs with sequencing-error indels (2,934,043)[a]	
	Accuracy, %	Accuracy on 685,937 pairs with boundary indels, %[b]	Accuracy, %	Accuracy on 695,771 pairs with boundary indels, %[b]
TopHat2 + Bowtie1	69.8	16.3	14.0	3.1
TopHat2 + Bowtie2	62.3	24.0	60.8	19.8
GSNAP	77.0	63.8	77.8	64.8
RUM	60.3	34.3	61.3	36.0
MapSplice	25.5	3.4	25.0	3.2
STAR	53.4	19.2	54.9	21.4

[a]The number of pairs containing each type of error is indicated in the column header.

[b]Boundary indels occur within 25 bp of an exon boundary. Percentages refer only to the pairs of each type, not to the entire dataset.

gapped aligner), TopHat2 was able to map 'true' indel reads using its own indel-finding algorithms.

For paired-end reads with indels, GSNAP had the highest rate of correct alignments (77%), followed by TopHat2 (60 to 69%), RUM (60 to 61%), and STAR (53 to 54%). MapSplice showed the lowest accuracy for both single-end and paired-end reads.

We defined boundary indels as those within 25 bp of a splice site. We separately computed the accuracy on all reads with boundary indels (Table 3, Table 4).

Alignment of a large set of real RNA-seq reads

Any test of alignment algorithms should use real data to provide a measure of the likely performance in practice. For these experiments, we used a recently released set of RNA-seq reads gathered across a time-course experiment reported by Chen *et al.* ([11]; GEO accession number: GSM818582). These data comprise 130,705,578 million paired-end reads in 65,352,789 pairs. All reads are 101 bp in length.

Because we did not know the true alignments for this RNA-seq dataset, we used the following objective criteria to evaluate each program: 1) the cumulative number of alignments with edit distances of 0, 1, 2, and 3 for each read; and 2) the cumulative number of spliced alignments that agree with the annotation for the corresponding human genes, taken from the Ensembl GRCh37 release of the human genome.

For each program, we aligned the paired-end reads with and without the known gene annotations, where possible. RUM is designed to run with these annotations, whereas MapSplice maps strictly without them. We then evaluated the mapping results in terms of the number of read or paired-read mappings.

TopHat2 consists of three mapping steps: 1) transcriptome mapping, which is used only when annotation is provided; 2) genome mapping; and 3) spliced mapping (see Methods for details). TopHat2 uses a remapping edit distance threshold *t*, specified by the user, as follows. If a read aligns to the transcriptome in step 1) with an edit

distance of less than t, TopHat2 will not remap the read in subsequent steps. Otherwise, TopHat2 will try to realign the read in steps 2) and 3), and then, depending on the resulting edit distance, it will use the read to detect novel splice sites. A setting of $t = 0$ means that TopHat2 will realign every read in all three steps. When we used $t = 0$ (Figure 2: 'TopHat2 realignment 0') on the real data, we consistently obtained better mapping results in terms of edit distance and the number of alignments corresponding to known splice sites (Figures 2; Figure 3; see Additional file 1, Figures S3-S4) for read and pair alignments, respectively (see Additional file 1, Tables S5-S6 and Tables S10-S11).

Figure 2 shows the alignment performance for each program both with and without using annotations, where all the programs were configured to report alignments with edit distances of up to 3 (and more in some programs). We compared the *de novo* alignments of reads for edit distances of 0, 1, 2, and 3. As expected, all programs found more alignments as the maximum permissible edit distance increased. For an edit distance of 0, which allows only perfect matches, TopHat2 mapped noticeably fewer reads without its new realignment function than it did with the function. This occurs because TopHat2 first aligns reads end to end with Bowtie2 before trying spliced alignments. Thus if a read is aligned end to end with, for example, one to three mismatches, then without the realignment function, TopHat2 accepts that alignment and may miss a spliced alignment with fewer mismatches.

By contrast, TopHat2 with $t = 0$ mapped the largest number of reads for all edit distances, followed in most cases by GSNAP. Note that for alignments with an edit distance of up to 3, TopHat2 without realignment discovered almost as many alignments as GSNAP.

When alignment methods were run with the assistance of gene annotations (Figure 2, right panel), the results were somewhat better than the *de novo* alignments. TopHat2 with or without realignment produced the highest number of mappings, followed by GSNAP, RUM, and STAR. The realignment procedure gave a much smaller advantage to TopHat2 in these experiments.

One way to estimate the accuracy of mappings is to compare alignments to known splice sites. We compared all aligners on only those reads that required splitting, counting how many known (Figure 3, left) and known plus novel (Figure 3, right) splice sites they identified. For *de novo* alignment, TopHat2 with realignment had the highest sensitivity, followed by MapSplice. Consistent with our tests on simulated reads, GSNAP and STAR showed relatively lower alignment rates. When using annotation, TopHat2 without realignment showed the highest mapping rate, slightly outperforming TopHat2 with realignment. GSNAP and STAR, which performed less well, map reads against substrings containing splice sites rather than whole transcripts. Direct mapping against whole transcripts, as TopHat2 does, worked well, especially when mapping reads spanning small exons, where a single read might span more than two exons.

Based on these results, we suggest two alternative strategies for alignment with TopHat2. First, if gene annotations are available, as they are for the human genome and some model organisms, then these annotations should be used with TopHat2, even without realignment. Alternatively, if annotations are unavailable or incomplete, then we recommend using TopHat2 with its realignment algorithm to produce the most complete set of alignments.

The run time and the peak memory usage of the programs used in this study varied greatly. We compared performance on all programs using the Chen *et al*. data [11] of 130 million reads (see Additional file 1, Table S8). Overall, STAR is much faster (32 minutes) than the other programs, which required from 8 to 55 hours. However, STAR requires a large amount of real memory, at least 28 GB, whereas most of the other programs required less than 8 GB.

Effects of pseudogenes on RNA-seq mapping

The Ensembl gene annotations (release 66) contain 32,439 genes, including non-coding RNA genes, and over 14,000 pseudogenes. Of the real genes, we found that 872 (2.7%) genes had pseudogene copies; that is, at least one transcript (or isoform) can be aligned to a pseudogene with at least 80% identity across the full length of the transcript. Using data from the Chen *et al*. study [11] and from the Illumina Body Map project [12], we found that genes with pseudogene copies seem to have higher expression levels than those without pseudogene copies. Table 5 shows the proportion of reads mapping to genes with pseudogenes, using both the raw count and a normalized count divided by the length of the transcript. Although only 2.7% of genes have pseudogene copies, these genes account for 22.5% (non-normalized) or 26.9% (normalized) of the RNA-seq reads in the Chen *et al*. data. In the RNA-seq experiments from the Illumina Body Map (the white blood sample only), we saw 19.1% (normalized) of reads mapping to genes with pseudogenes (see Additional file 1; Table S12). From both RNA-seq experiments, we noted that genes with multiple pseudogene copies were more abundantly expressed than those with a single pseudogene copy. We ran a similar analysis looking only at the 20,417 protein-coding genes in Ensembl, with similar results: 22% of read pairs (26 times the number expected) mapped to genes with processed pseudogenes (see Additional file 1; Table S13).

Figure 4 shows various mapping results from TopHat2 with and without realignments at various edit distances.

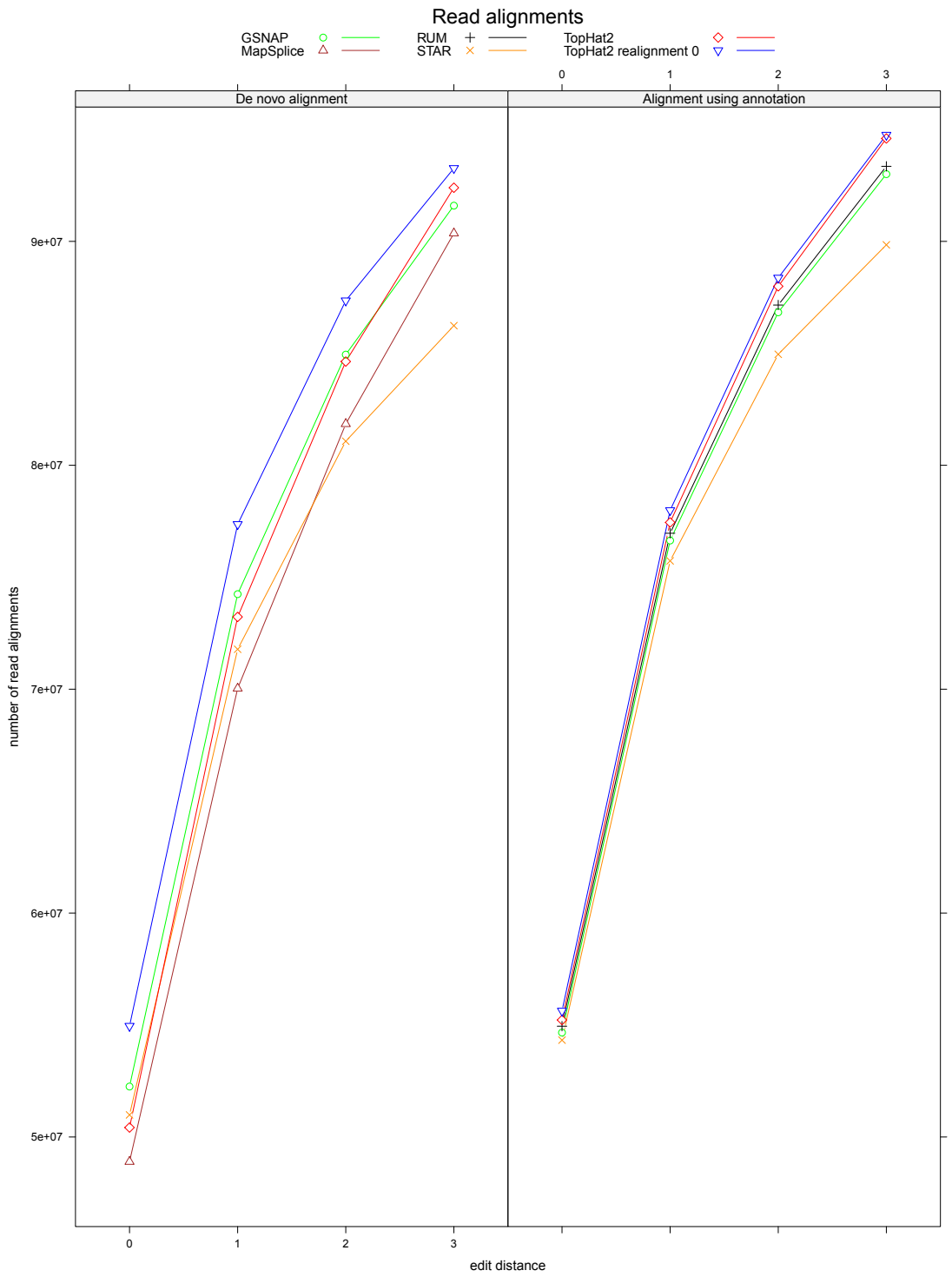

Figure 2 The number of read alignments from TopHat2, GSNAP, RUM, MapSplice, and STAR. Tthe RNA-seq reads are from Chen *et al.* [11]. TopHat2 was run with and without realignment (realignment edit distance of 0). TopHat2, GSNAP, and STAR were run in both *de novo* and gene-mapping modes, while MapSplice was run only in *de novo* mode and RUM was run only in gene-mapping mode. The number of alignments at each edit distance is cumulative; for instance, the number of alignments at an edit distance of 2 includes all the alignments with edit distance of 0, 1, or 2.

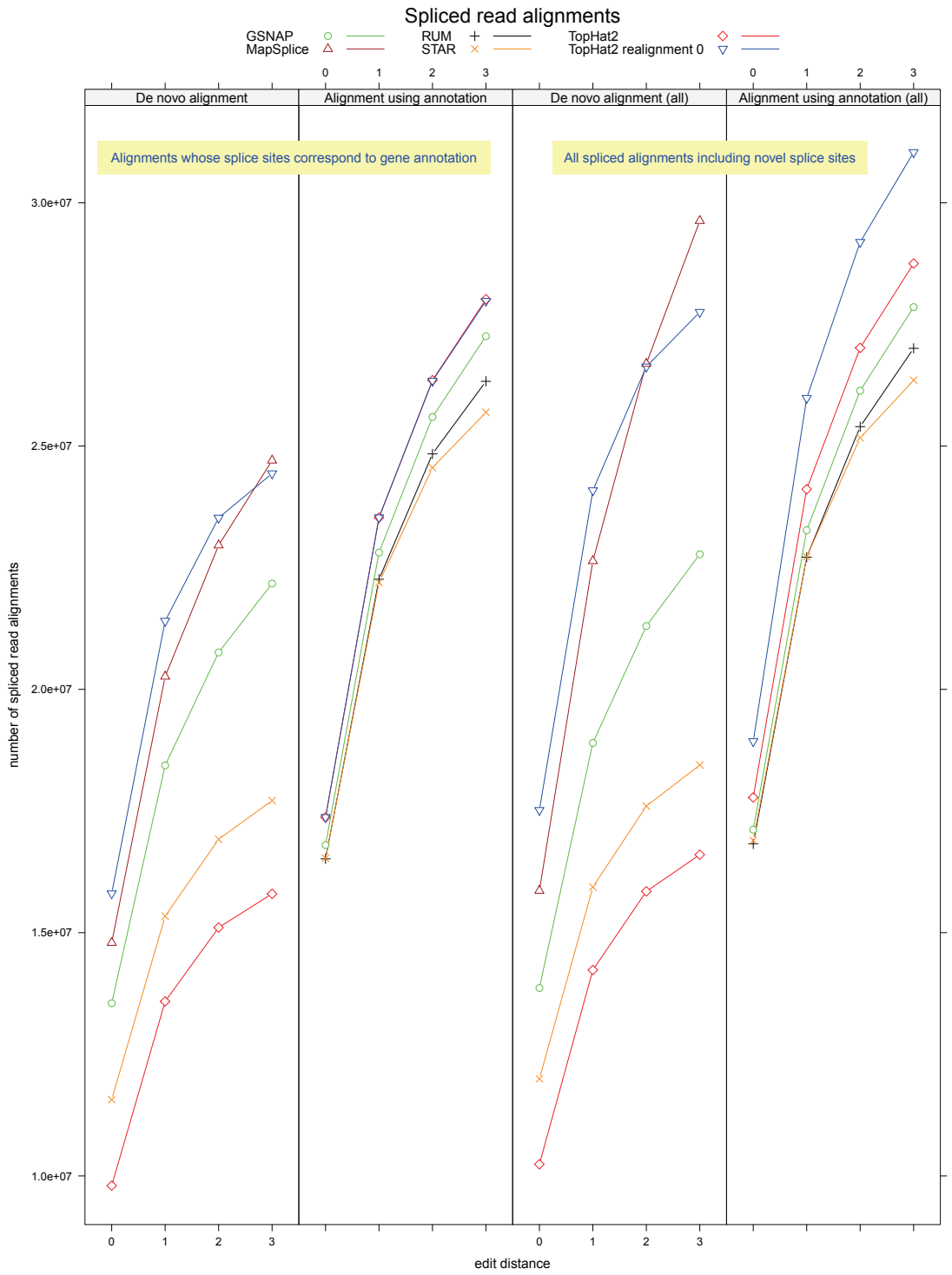

Figure 3 The number of spliced-read alignments from TopHat2, GSNAP, RUM, MapSplice, and STAR. The RNA-seq reads are from Chen *et al.* [11]. TopHat2, GSNAP, and STAR were run in both *de novo* and gene-mapping modes while MapSplice was run only in *de novo* mode and RUM was run only in gene-mapping mode. For each mapping mode, the two panels on the left show the number of spliced alignments whose splice sites were found in the gene annotations, and the two panels on the right show the number of all spliced alignments including novel splice sites.

Table 5 Expression levels of genes with pseudogene copies from Chen *et al*. [11].[a]

Number of pseudogene copies[b]	Gene with pseudogene copies (%)[c]	Pair count, %[d]	Ratio[e]	Normalized count, %[f]	Normalized ratio[f]
1	553 (1.7%)	6.85	× 4.02	9.37	× 5.49
2	113 (0.4)	5.15	× 14.79	5.20	× 14.93
3	49 (0.2)	1.27	× 8.38	1.96	× 12.99
4	27 (0.1)	2.27	× 27.32	2.28	× 27.35
≥ 5	130 (0.4)	6.91	× 17.24	8.08	× 20.16
Total (≥ 1)	872/32,439 (2.7)	22.45	× 8.35	26.88	× 10.00

[a]Using Bowtie2, we aligned RNA-seq paired-end reads to 32,439 annotated genes.

[b]Number of pseudogene copies of a gene. The first row shows genes that have just one pseudogene, followed by rows for genes with two, three, four, and at least five pseudogene copies.

[c]Number of genes with the specified number of pseudogene copies; for example, 553 genes (1.7% of all genes) have one pseudogene copy.

[d]Percentage of read pairs that were mapped to genes with pseudogene copies.

[e]Ratio of columns 3 and 2.

[f]These two columns were similarly defined using a normalized count, where the number of reads mapping to each gene was normalized to account for gene length.

As we allowed TopHat2 to realign more reads, it found the spliced alignments that were otherwise hidden by pseudogene alignments. This in turn substantially increased its mapping rates for known splice sites.

The completeness of human gene annotations

Using the *de novo* mapping mode in TopHat2, GSNAP, MapSplice, and STAR, we looked at how many spliced alignments were found in the Ensembl annotations. The proportions of spliced mappings to known splice sites were 88 to 90%, 97%, 96%, and 83 to 93% in TopHat2, GSNAP, STAR, and MapSplice, respectively (Figure 5). Although this analysis considered only the RNA-seq data from Chen *et al.* [11], the TopHat2 result suggests that many additional spliced alignments, up to 12%, might remain to be discovered. Most of the novel splicing events in these alignments were supported by 10 or more reads that extended for 50 or more bases on each side.

Conclusions

Discovery of new genes and transcripts is a major objective in many RNA-seq experiments. Deep RNA-seq experiments continue to uncover previously unseen elements of the transcriptome, even in well-studied organisms. Mapping reads to the genome is a core step in such screens, and the accuracy of mapping software can determine the accuracy of downstream steps such as gene and transcript discovery or expression quantification.

We have described TopHat2, which provides major accuracy improvements over previous versions and over other RNA-seq mapping tools. Because TopHat2 is built around Bowtie2, it can now align reads across small indels with high accuracy, a feature crucial for studies assessing the effects of genetic mutations on gene and transcript expression. TopHat2 is engineered to work well with a wide range of RNA-seq experimental designs, and it is optimized for the widely available long paired-end reads. These reads pose new challenges because they can span multiple splice sites rather than just one or two; we estimate that nearly half of reads 150 bp long will span two or more human exons. The algorithmic improvements in TopHat2 address this challenge, maintaining both accuracy and speed. Other refinements to the algorithm increase accuracy for reads that span a junction with only a small (≤ 10 bp) overhang, reducing errors in downstream transcript assembly using tools such as Cufflinks. TopHat2 also makes powerful use of available gene annotations, which allow it to avoid erroneously mapping reads to pseudogenes, and generally improve its overall alignment accuracy. Annotation also allows TopHat2 to better align reads that cover microexons, non-canonical splice sites, and other 'unusual' features of eukaryotic transcriptomes.

We have shown that TopHat2 performs well over a wide range of read lengths, making it a good fit for most RNA-seq experimental designs. This scalability suggests that as read lengths grow, TopHat2 will continue to report accurate, sensitive alignment results and allow for robust downstream analysis. We believe that TopHat2 reports more accurate alignments than competing tools, using fewer computational resources. RNA-seq experiments are becoming increasingly common and are now routinely used by many biologists. We expect that TopHat2 will provide these scientists with accurate results for use with expression analysis, gene discovery, and many other applications.

Methods

Given RNA-seq reads as input, TopHat2 begins by mapping reads against the known transcriptome, if an annotation file is provided. This transcriptome mapping improves the overall sensitivity and accuracy of the mapping. It also gives the whole pipeline a significant

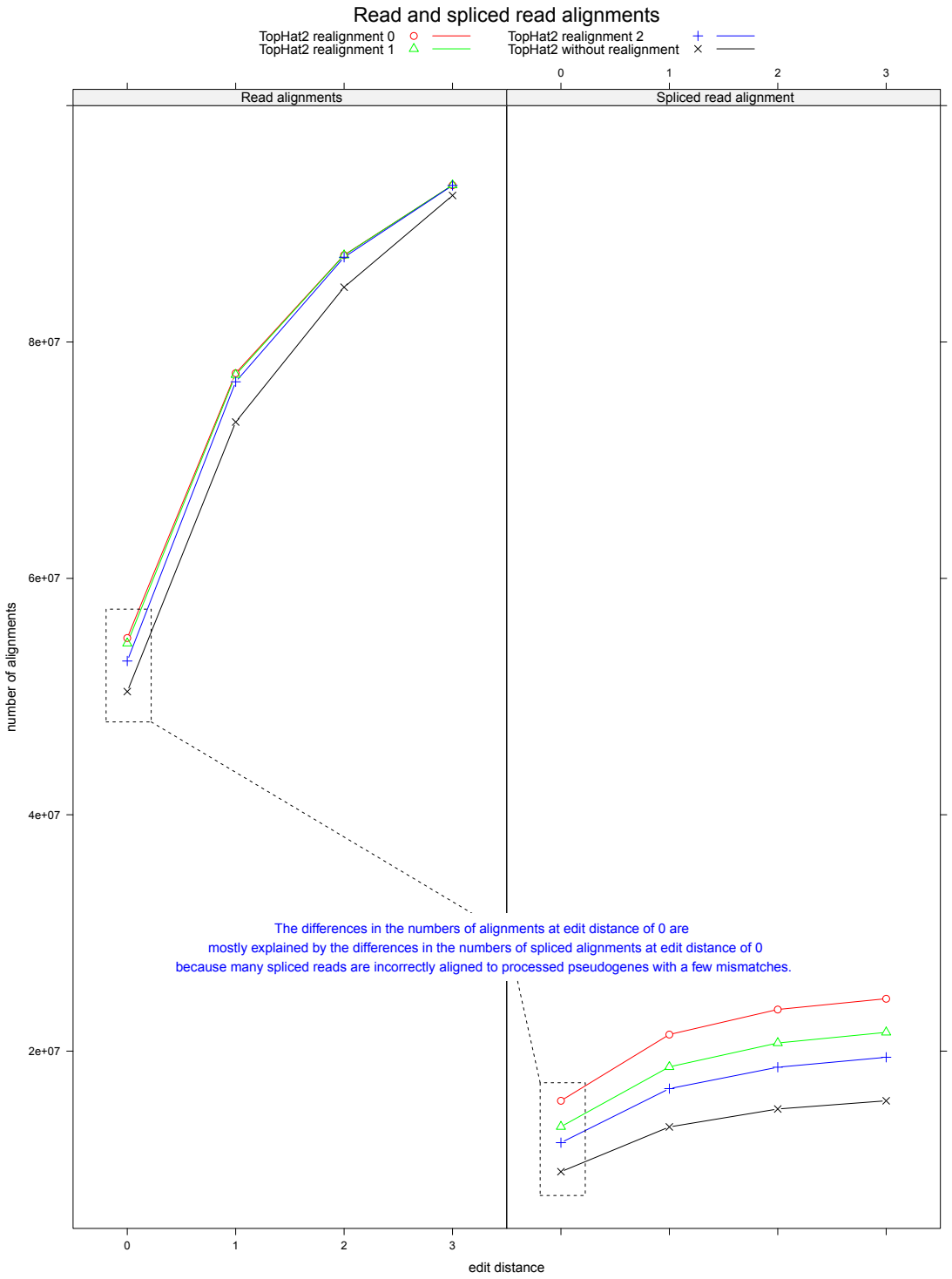

Figure 4 The number of read and spliced-read alignments from TopHat2, using different realignment edit distances and no realignment. Edit distances of 0, 1, and 2 were used. As TopHat2 allows more realignment from no realignment to 2 to 1 to 0, the number of read alignments and spliced-read alignments increases, so that the differences in the numbers of read alignments from TopHat run with different realignment edit distance are mostly explained by the increase in the number of spliced-read alignments.

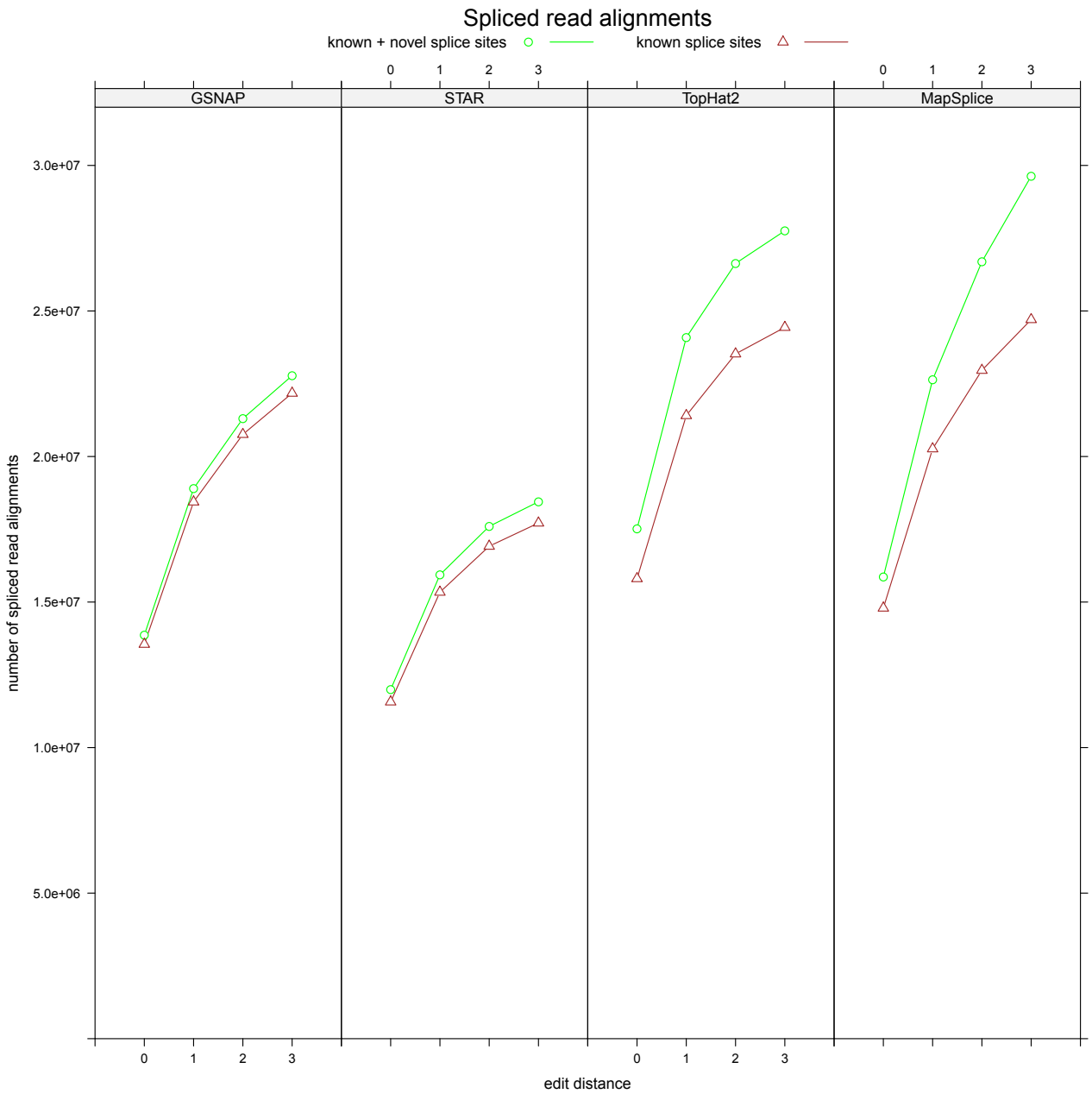

Figure 5 The number of spliced-read alignments from TopHat2, GSNAP, STAR, and MapSplice without using gene annotation. The number of read alignments whose splice sites were found in the gene annotations are shown in brown, and the number of all spliced-read alignments including novel splice sites are shown in green.

speed increase, owing to the much smaller size of the transcriptome compared with that of the genome (see Figure 6).

After the transcriptome-mapping step, some reads remain unmapped because they are derived from unknown transcripts not present in the annotation, or because they contain many miscalled bases. In addition, there may be poorly aligned reads that have been mapped to the wrong location. TopHat2 aligns these unmapped or potentially misaligned reads against the genome (Figure 6, step 2). Any reads contained entirely within exons will be mapped, whereas other spanning introns may not be.

Using unmapped reads from step 2, TopHat2 tries to find novel splice sites that are based on known junction signals (GT-AG, GC-AG, and AT-AC). TopHat2 also provides an option to allow users to remap some of the mapped reads, depending on the edit distance values of these reads; that is, those reads whose edit distance is

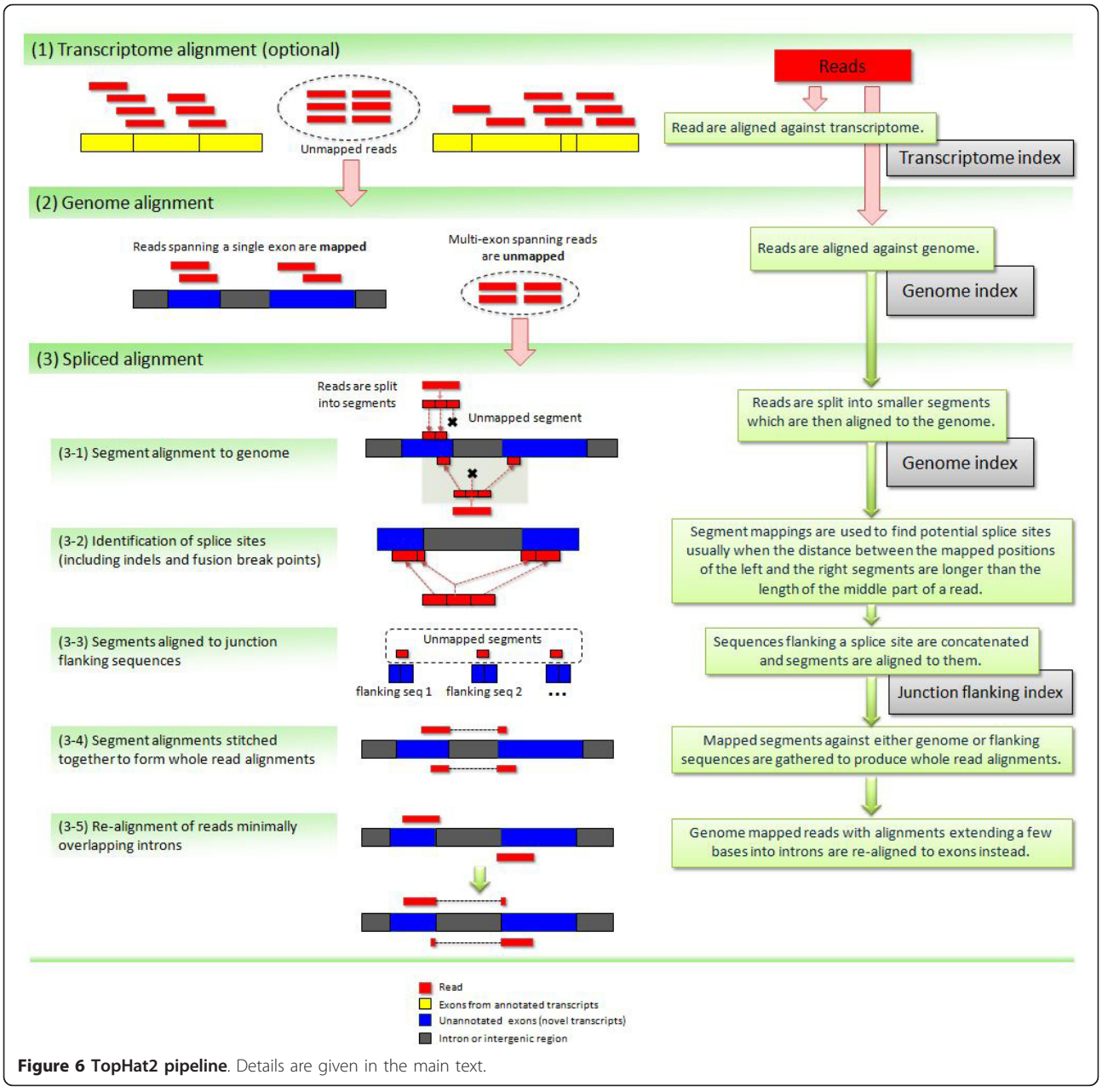

Figure 6 TopHat2 pipeline. Details are given in the main text.

greater than or equal to a user-provided threshold will be treated as unmapped reads. To accomplish this, the unmapped reads (and previously mapped reads with low alignment scores) are split into smaller non-overlapping segments (25 bp each by default) which are then aligned against the genome (Figure 6, step 3). Tophat2 examines any cases in which the left and right segments of the same read are mapped within a user-defined maximum intron size (usually between 50 and 100,000 bp). When this pattern is detected, TopHat2 re-aligns the entire read sequence to that genomic region in order to identify the most likely locations of the splice sites (Figure 6). Using a similar approach, indels and fusion breakpoints are also detected in this step.

The genomic sequences flanking these splice sites are concatenated, and the resulting spliced sequences are collected as a set of potential transcript fragments. Any reads not mapped in the previous stages (or mapped very poorly) are then re-aligned with Bowtie2 [15] against this novel transcriptome.

After these steps, some of the reads may have been aligned incorrectly by extending an exonic alignment a few bases into the adjacent intron (Figure 1; Figure 6, steps 3 to 5). TopHat2 checks if such alignments extend

into the introns identified in the split-alignment phase; if so, it can realign these reads to the adjacent exons instead.

In the final stage, TopHat2 divides the reads into those with unique alignments and those with multiple alignments. For the multi-mapped reads, TopHat2 gathers statistical information (for example, the number of supporting reads) about the relevant splice junctions, insertions, and deletions, which it uses to recalculate the alignment score for each read. Based on these new alignment scores, TopHat2 reports the most likely alignment locations for such multi-mapped reads.

For paired-end reads, TopHat2 processes the two reads separately through the same mapping stages described above. In the final stage, the independently aligned reads are analyzed together to produce paired alignments, taking into consideration additional factors including fragment length and orientation.

For the experiments described in this study, the program version numbers were: TopHat2 2.0.8, TopHat1 1.1.4, GSNAP 2013-01-23, RUM 1.12_01, MapSplice 1.15.2, and STAR 2.3.0e. For the specific parameters for each program, see Additional file 1, Table S9, and for the source code of TopHat 2.0.8, see Additional file 2.

Additional material

Additional File 1: Supplementary material.

Additional File 2: TopHat2 source code.

Abbreviations

bp: Base pair; indel: Insertion or deletion; RNA-seq: RNA sequence;

Authors' contributions

DK, SLS, GP, and CT performed the analysis and discussed the results of TopHat2. DK, GP, and CT were mainly responsible for implementing TopHat2. HP implemented the transcriptome-mapping algorithms, with help from GP and DK. RK implemented the indel-alignment algorithms, with help from DK. DK, SLS, GP, and CT wrote the manuscript. All authors read and approved the final manuscript.

Acknowledgements

We thank Lior Patcher and Adam Roberts for their invaluable contributions to our discussions on the TopHat2 pipeline. This work is supported in part by the National Human Genome Research Institute (NIH) under grants R01-HG006102 and R01-HG006677.

Author details

[1]Center for Bioinformatics and Computational Biology, University of Maryland, College Park, MD, 20742, USA. [2]Department of Computer Science, University of Maryland, College Park, MD 20742, USA. [3]Center for Computational Biology, McKusick-Nathans Institute of Genetic Medicine, Johns Hopkins University School of Medicine, 733 N. Broadway, Baltimore, MD, 21205, USA. [4]Department of Biostatistics, Bloomberg School of Public Health, Johns Hopkins University, 615 N. Wolfe Street, Baltimore, MD, 21205, USA. [5]Broad Institute of MIT and Harvard, 7 Cambridge Center, Cambridge, MA, 02142, USA. [6]Department of Stem Cell and Regenerative Biology, Harvard University, 7 Divinity Ave., Cambridge, MA, 02142, USA. [7]Department of Electrical Engineering and Computer Science, University of California, 101 Sproul Hall, Berkeley, CA, 94720, USA. [8]Illumina Inc., 5200 Illumina Way, San Diego, CA, 92122, USA.

Received: 15 November 2012 Revised: 5 April 2013
Accepted: 25 April 2013 Published: 25 April 2013

References

1. Mortazavi A, Williams BA, McCue K, Schaeffer L, Wold B: **Mapping and quantifying mammalian transcriptomes by RNA-Seq.** *Nat Methods* 2008, **5**:621-628.
2. Pei B, Sisu C, Frankish A, Howald C, Habegger L, Mu XJ, Harte R, Balasubramanian S, Tanzer A, Diekhans M, Reymond A, Hubbard TJ, Harrow J, Gerstein MB: **The GENCODE pseudogene resource.** *Genome Biol* 2012, **13**:R51.
3. Roberts A, Trapnell C, Donaghey J, Rinn JL, Pachter L: **Improving RNA-Seq expression estimates by correcting for fragment bias.** *Genome Biol* 2011, **12**:R22.
4. Trapnell C, Pachter L, Salzberg SL: **TopHat: discovering splice junctions with RNA-Seq.** *Bioinformatics* 2009, **25**:1105-1111.
5. Wu TD, Nacu S: **Fast and SNP-tolerant detection of complex variants and splicing in short reads.** *Bioinformatics* 2010, **26**:873-881.
6. Grant GR, Farkas MH, Pizarro AD, Lahens NF, Schug J, Brunk BP, Stoeckert CJ, Hogenesch JB, Pierce EA: **Comparative analysis of RNA-Seq alignment algorithms and the RNA-Seq unified mapper (RUM).** *Bioinformatics* 2011, **27**:2518-2528.
7. Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, Batut P, Chaisson M, Gingeras TR: **STAR: ultrafast universal RNA-seq aligner.** *Bioinformatics* 2013, **29**:15-21.
8. 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: **MapSplice: accurate mapping of RNA-seq reads for splice junction discovery.** *Nucleic Acids Res* 2010, **38**:e178.
9. Zhang Z, Harrison PM, Liu Y, Gerstein M: **Millions of years of evolution preserved: a comprehensive catalog of the processed pseudogenes in the human genome.** *Genome Res* 2003, **13**:2541-2558.
10. Kalyana-Sundaram S, Kumar-Sinha C, Shankar S, Robinson DR, Wu YM, Cao X, Asangani IA, Kothari V, Prensner JR, Lonigro RJ, Iyer MK, Barrette T, Shanmugam A, Dhanasekaran SM, Palanisamy N, Chinnaiyan AM: **Expressed pseudogenes in the transcriptional landscape of human cancers.** *Cell* 2012, **149**:1622-1634.
11. Chen R, Mias GI, Li-Pook-Than J, Jiang L, Lam HY, Miriami E, Karczewski KJ, Hariharan M, Dewey FE, Cheng Y, Clark MJ, Im H, Habegger L, Balasubramanian S, O'Huallachain M, Dudley JT, Hillenmeyer S, Haraksingh R, Sharon D, Euskirchen G, Lacroute P, Bettinger K, Boyle AP, Kasowski M, Grubert F, Seki S, Garcia M, Whirl-Carrillo M, Gallardo M, Blasco MA, *et al*: **Personal omics profiling reveals dynamic molecular and medical phenotypes.** *Cell* 2012, **148**:1293-1307.
12. **The Illumina Body Map 2.0 data..** [http://www.ebi.ac.uk/arrayexpress/browse.html?keywords=E-MTAB-513&expandefo=on].
13. Xing J, Zhang Y, Han K, Salem AH, Sen SK, Huff CD, Zhou Q, Kirkness EF, Levy S, Batzer MA, Jorde LB: **Mobile elements create structural variation: analysis of a complete human genome.** *Genome Res* 2009, **19**:1516-1526.
14. Levy S, Sutton G, Ng PC, Feuk L, Halpern AL, Walenz BP, Axelrod N, Huang J, Kirkness EF, Denisov G, Lin Y, MacDonald JR, Pang AW, Shago M, Stockwell TB, Tsiamouri A, Bafna V, Bansal V, Kravitz SA, Busam DA, Beeson KY, McIntosh TC, Remington KA, Abril JF, Gill J, Borman J, Rogers YH, Frazier ME, Scherer SW, Strausberg RL, *et al*: **The diploid genome sequence of an individual human.** *PLoS Biol* 2007, **5**:e254.
15. Langmead B, Salzberg SL: **Fast gapped-read alignment with Bowtie 2.** *Nat Methods* 2012, **9**:357-359.
16. Kim D, Salzberg SL: **TopHat-Fusion: an algorithm for discovery of novel fusion transcripts.** *Genome Biol* 2011, **12**:R72.
17. Langmead B, Trapnell C, Pop M, Salzberg SL: **Ultrafast and memory-efficient alignment of short DNA sequences to the human genome.** *Genome Biol* 2009, **10**:R25.
18. Griebel T, Zacher B, Ribeca P, Raineri E, Lacroix V, Guigo R, Sammeth M: **Modelling and simulating generic RNA-Seq experiments with the flux simulator.** *Nucleic acids Res* 2012.

doi:10.1186/gb-2013-14-4-r36
Cite this article as: Kim *et al.*: TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biology* 2013 **14**:R36.