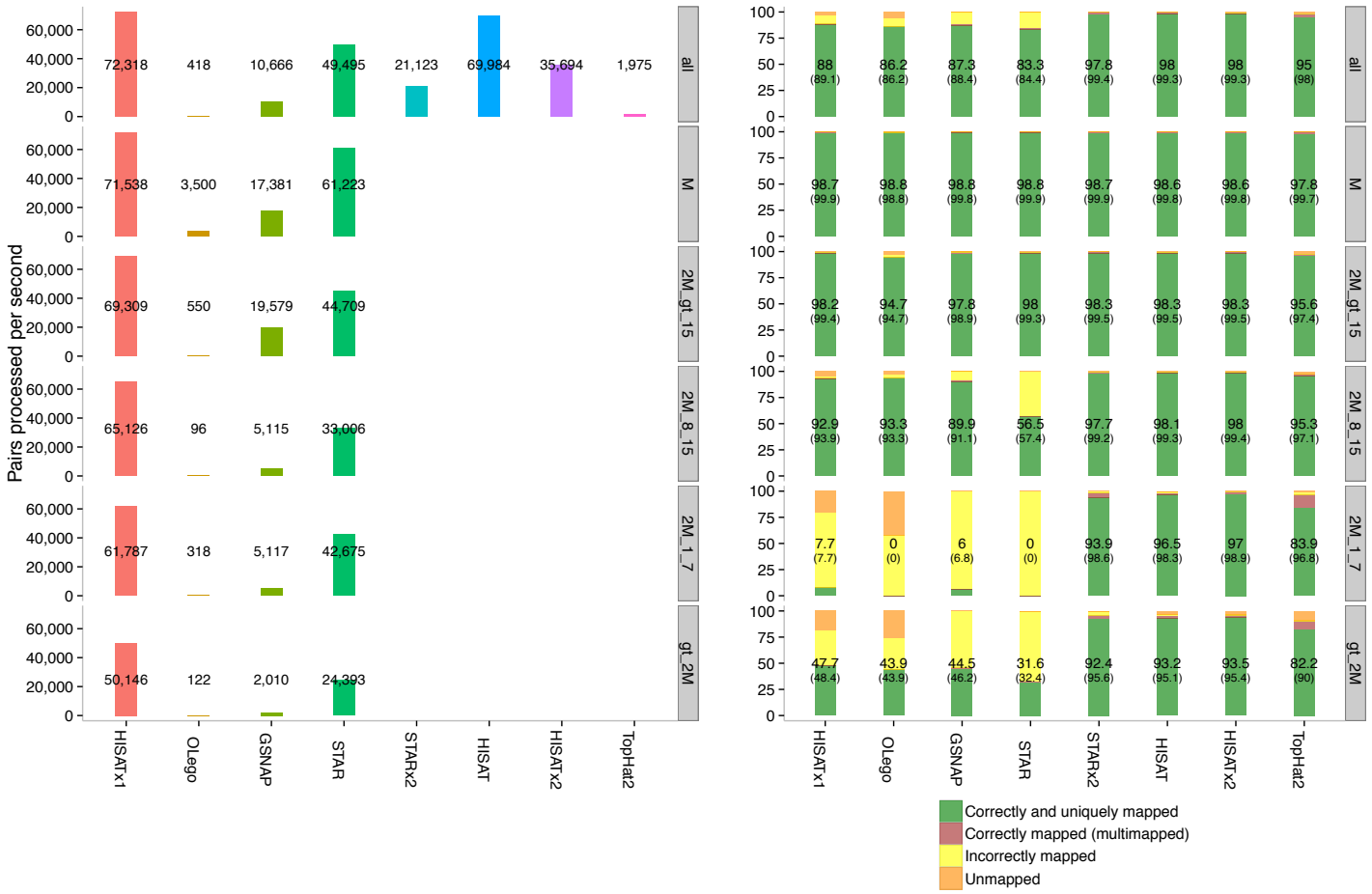
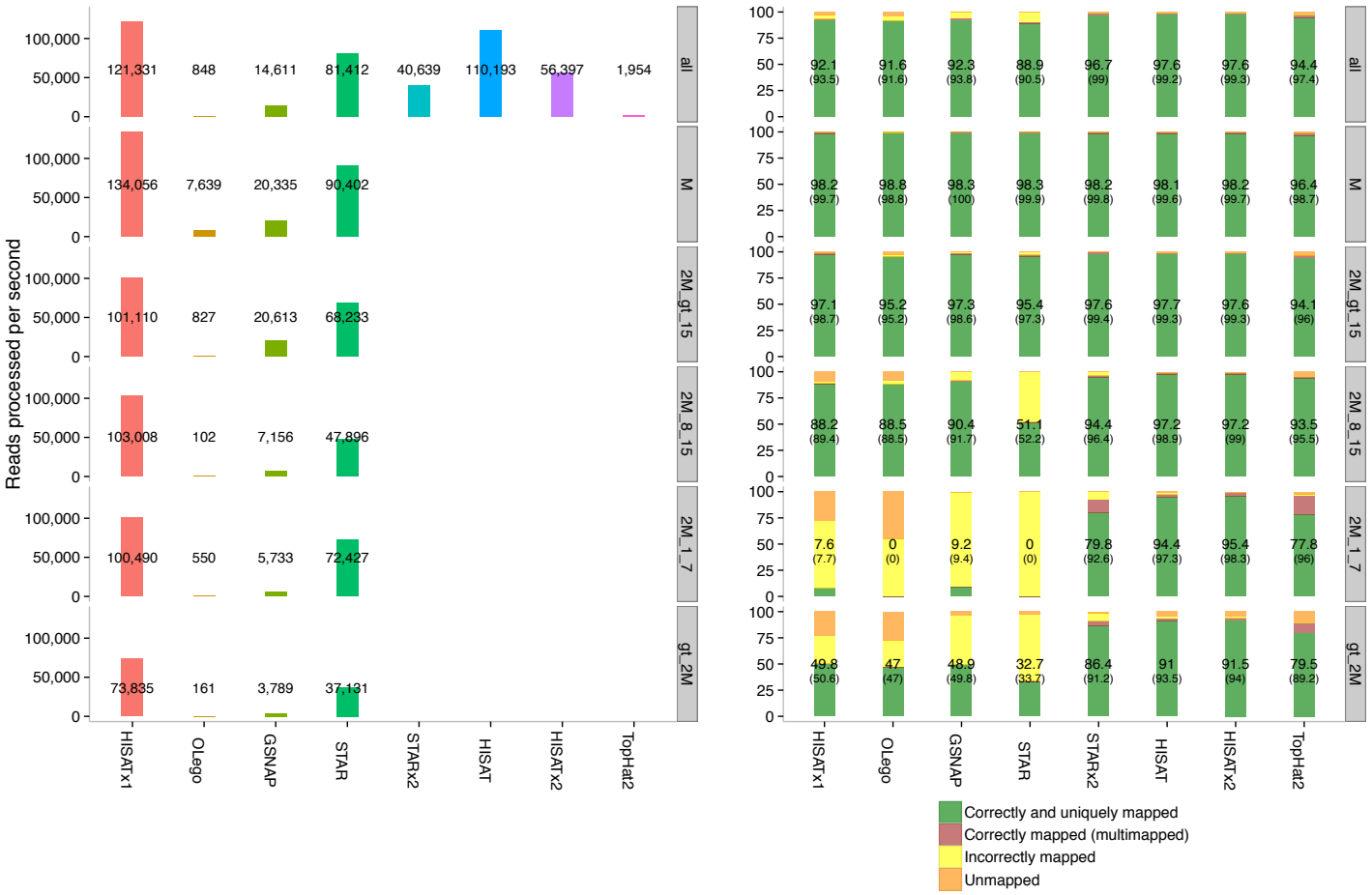


Supplementary Figure 1

Alignment speed and sensitivity of spliced alignment software for 40 million error-free simulated paired-end reads (100 bp long, 20 million pairs).

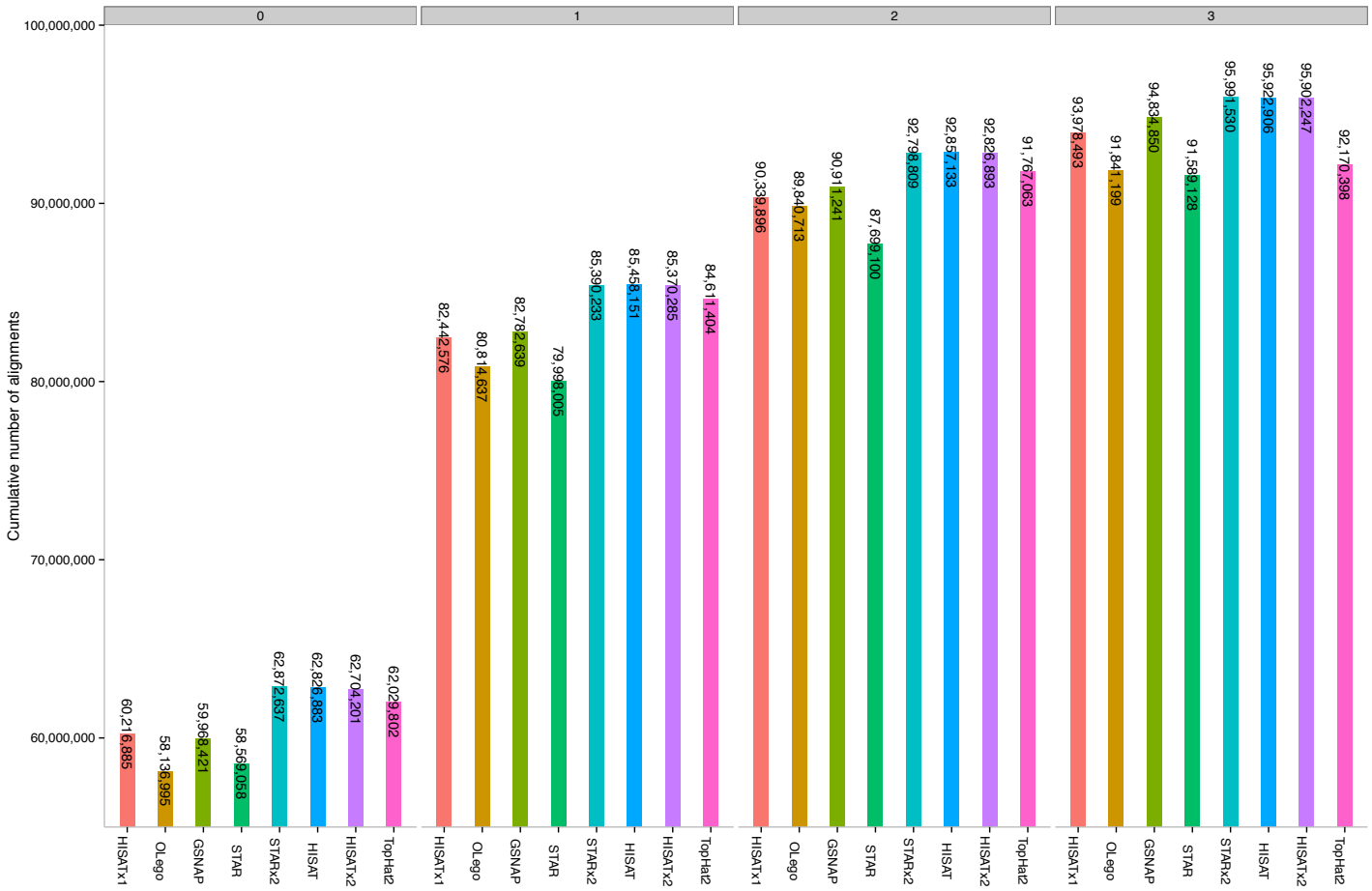
This figure shows the alignment speed and sensitivity for each type of pair (all, M, 2M_gt_15, 2M_8_15, 2M_1_7, gt_2M, where 'all' includes all the pair types). Since a pair consists of left and right reads, the type of a pair is determined by the more difficult read type. The difficulties of read types are given in the following order from easiest to most difficult: M, 2M_gt_15, 2M_8_15, 2M_1_7, and gt_2M. The plot on the left shows the alignment speed of the programs in terms of the number of pairs processed per second. The right plot shows alignment sensitivity. Pairs are categorized as: (1) correctly and uniquely mapped, (2) correctly mapped (multi-mapped), (3) incorrectly mapped, and (4) unmapped. Case (2) covers instances where an aligner mapped a pair to multiple locations and one of the locations was correct. These four categories encompass all of the pairs. The numbers in the right plot represent the percentages of case (1). The numbers inside the parentheses represent the percentages of cases (1) and (2) combined.



Supplementary Figure 2

Alignment speed and sensitivity of spliced alignment software for 20 million simulated single-end reads with a mismatch rate of 0.5% (100 bp long).

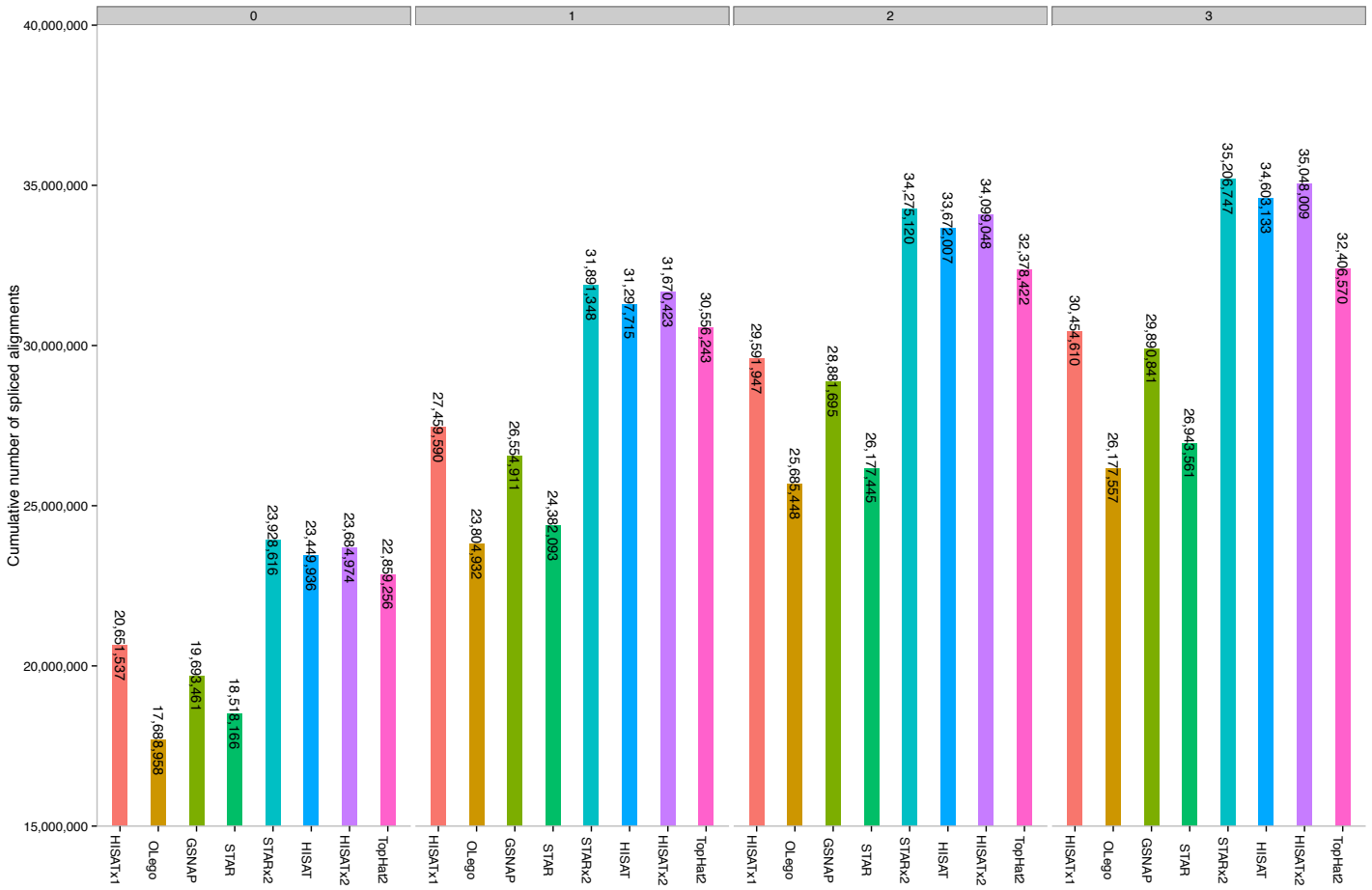
This figure shows the alignment speed and sensitivity for each type of read (all, M, 2M_gt_15, 2M_8_15, 2M_1_7, gt_2M, where 'all' includes all the read types). The plot on the left shows the alignment speed of the programs in terms of the number of reads processed per second. The right plot shows alignment sensitivity. Reads are categorized as: (1) correctly and uniquely mapped, (2) correctly mapped (multi-mapped), (3) incorrectly mapped, and (4) unmapped. Case (2) covers instances where an aligner mapped a read to multiple locations and one of the locations was correct. These four categories encompass all of the reads. The numbers in the right plot represent the percentages of case (1). The numbers inside the parentheses represent the percentages of cases (1) and (2) combined. Note that by looking at both plots, it is easy to see tradeoffs between alignment speed and sensitivity.



Supplementary Figure 3

Alignment results for 109 million reads, each 101 bp long, from a human sample.

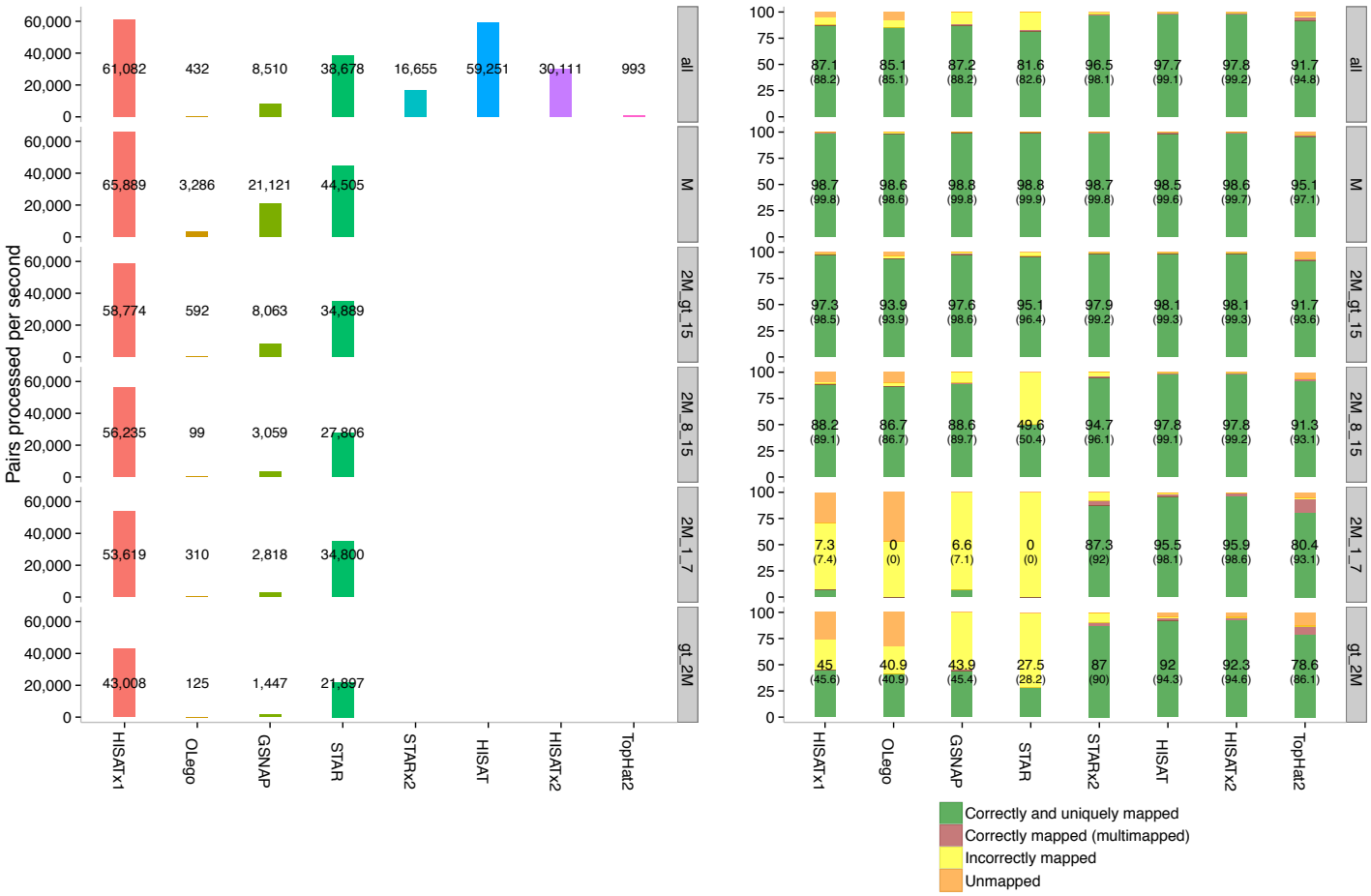
Shown are the cumulative numbers of alignments up to a given edit distance. Edit distance is defined here simply as the number of differences ('edits') between the read and the reference sequence. The leftmost panel shows reads that matched exactly (with an edit distance of 0). The next panel (labelled "1") shows the number of reads that aligned with either 0 or 1 mismatches; similarly for the panels labelled 2 and 3. Note that GSNAP and STAR report soft-clipped alignments where bases on the ends of reads are left unaligned. To compute edit distances for these alignments, we re-aligned the soft-clipped bases to their corresponding locations in the reference genome and calculated the number of mismatches.



Supplementary Figure 4

Alignment results of spliced alignment software for 109 million real reads (101 bp long).

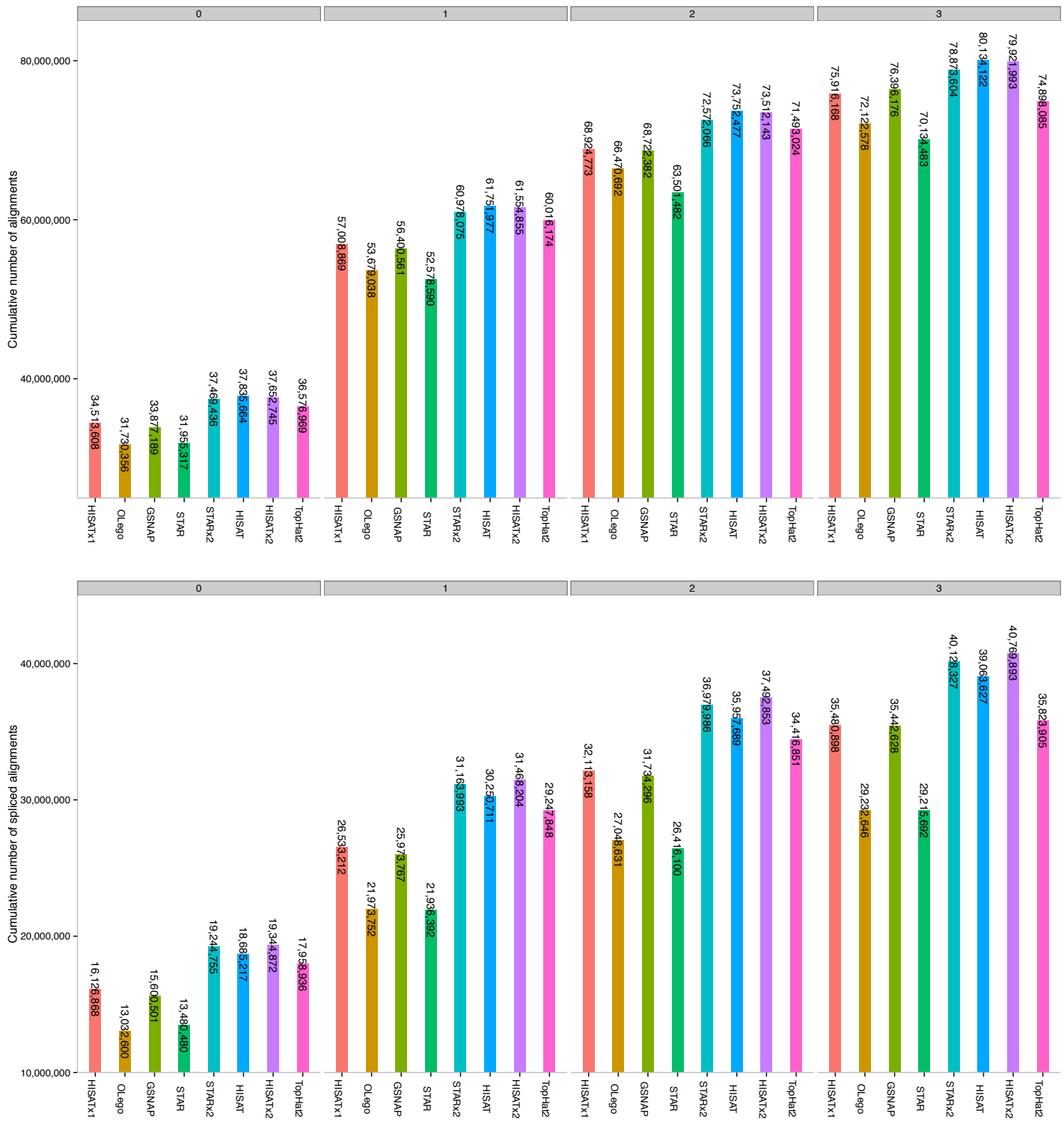
This figure shows the cumulative number of spliced alignments up to a given edit distance (0, 1, 2, and 3) whose splice sites are known in gene annotations.



Supplementary Figure 5

Alignment speed and sensitivity of spliced-alignment software for 40 million simulated paired-end reads with a mismatch rate of 0.5% (100 bp long, 20 million pairs).

This figure shows the alignment speed and sensitivity for each type of pair (all, M, 2M_gt_15, 2M_8_15, 2M_1_7, gt_2M, where 'all' includes all the pair types). Since a pair consists of left and right reads, the type of a pair is determined by the more difficult read type. The difficulties of read types are given in the following order from easiest to most difficult: M, 2M_gt_15, 2M_8_15, 2M_1_7, and gt_2M. The plot on the left shows the alignment speed of the programs in terms of the number of pairs processed per second. The right plot shows alignment sensitivity. Pairs are categorized as: (1) correctly and uniquely mapped, (2) correctly mapped (multi-mapped), (3) incorrectly mapped, and (4) unmapped. Case (2) covers instances where an aligner mapped a pair to multiple locations and one of the locations was correct. These four categories encompass all of the pairs. The numbers in the right plot represent the percentages of case (1). The numbers inside the parentheses represent the percentages of cases (1) and (2) combined.

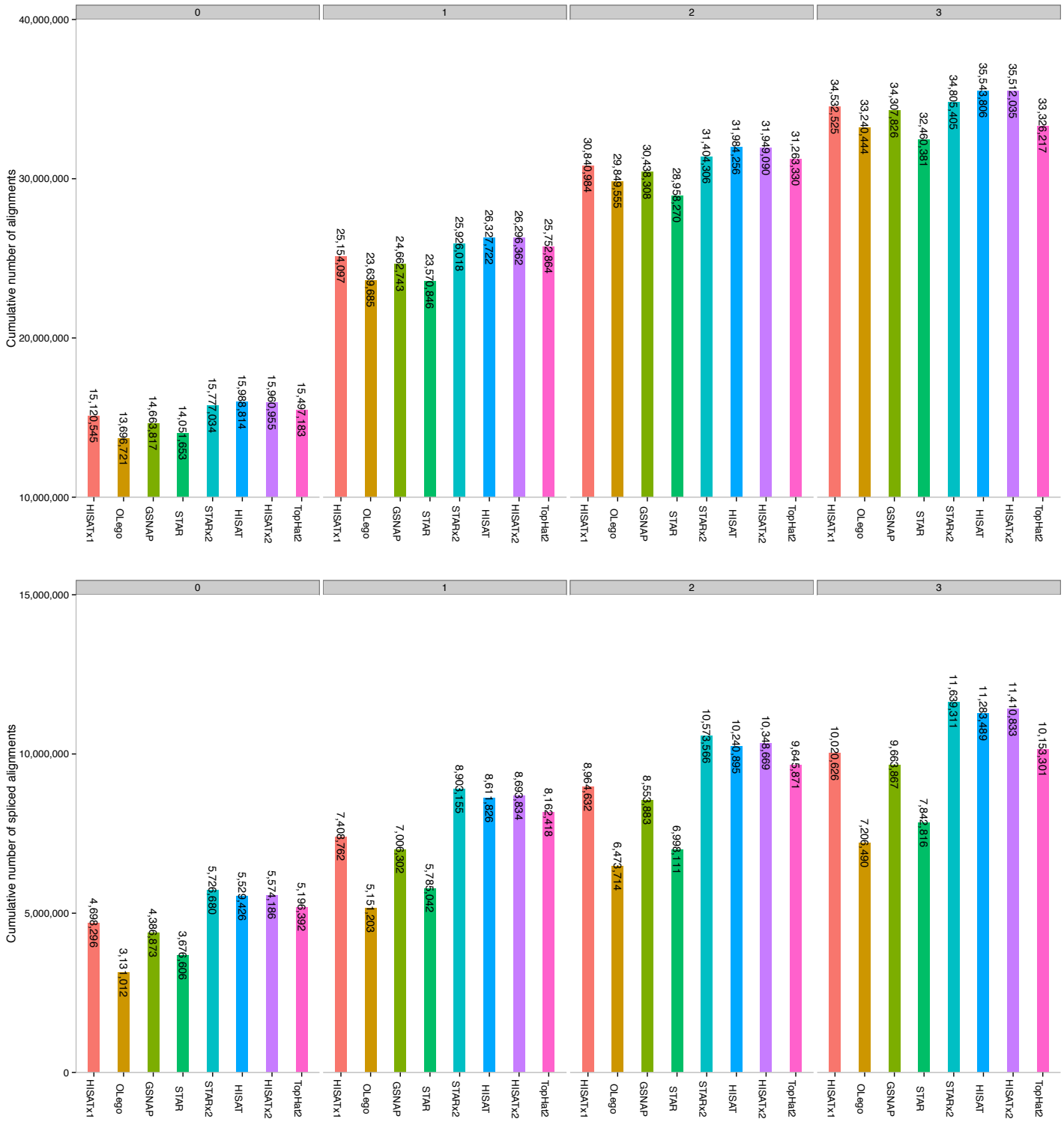


Supplementary Figure 6

Alignment results of spliced alignment software for ~218 million real paired-end reads (~109 million pairs).

This figure shows two plots: (1) the cumulative number of alignments up to a given edit distance (0, 1, 2, and 3) and (2) the cumulative number of spliced alignments whose splice sites are known in gene annotations. Note these alignments are pair alignments with the

combined edit distance from the left and the right alignments. Spliced alignments are those whose read alignment is a spliced alignment.

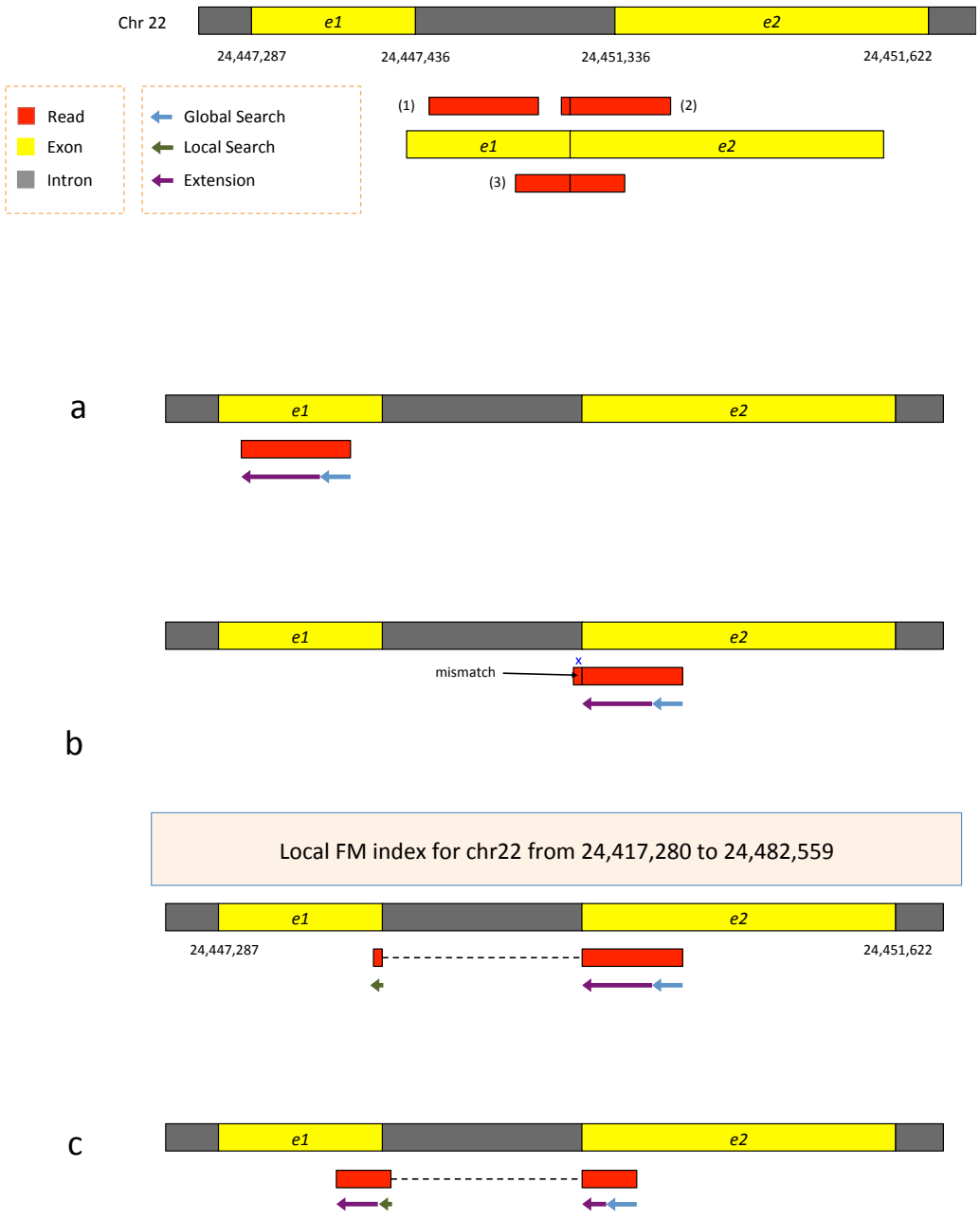


Supplementary Figure 7

Alignment results of spliced alignment software for ~126 million real paired-end reads (~63 million pairs).

This figure shows two plots: (1) the cumulative number of alignments up to a given edit distance (0, 1, 2, and 3) and (2) the cumulative number of spliced alignments whose splice sites are known in gene annotations. Note these alignments are pair alignments with the

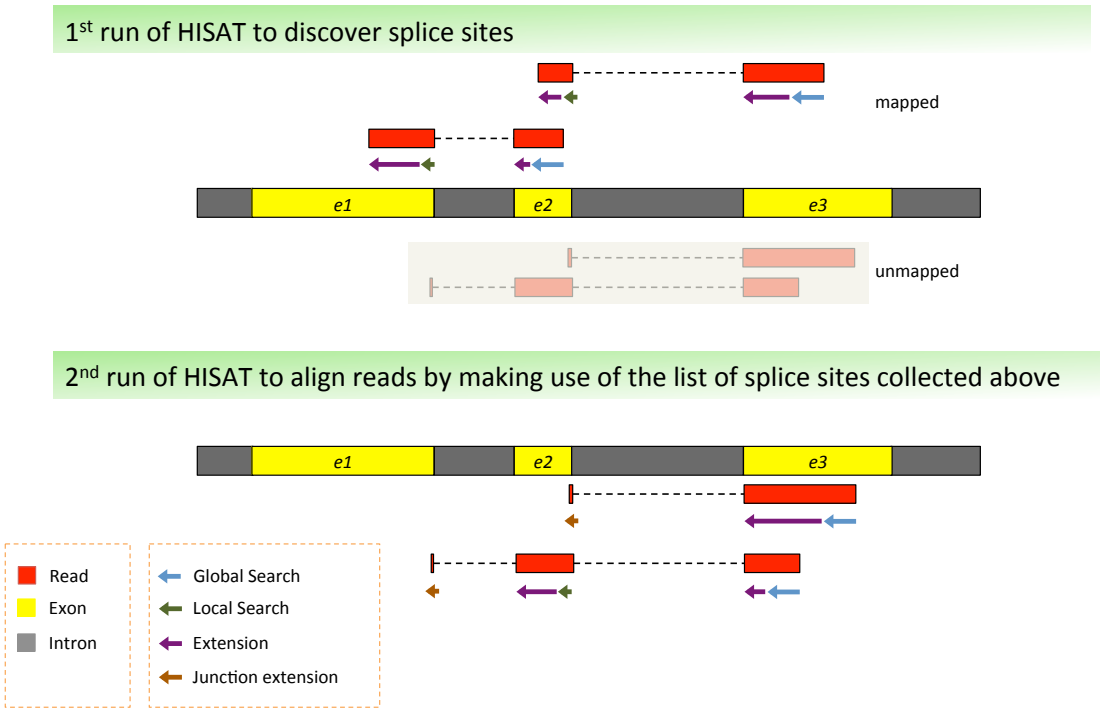
combined edit distance from the left and the right alignments. Spliced alignments are those whose read alignment is a spliced alignment.



Supplementary Figure 8

Three working examples demonstrating how HISAT applies its hierarchical indexing for fast and sensitive alignment.

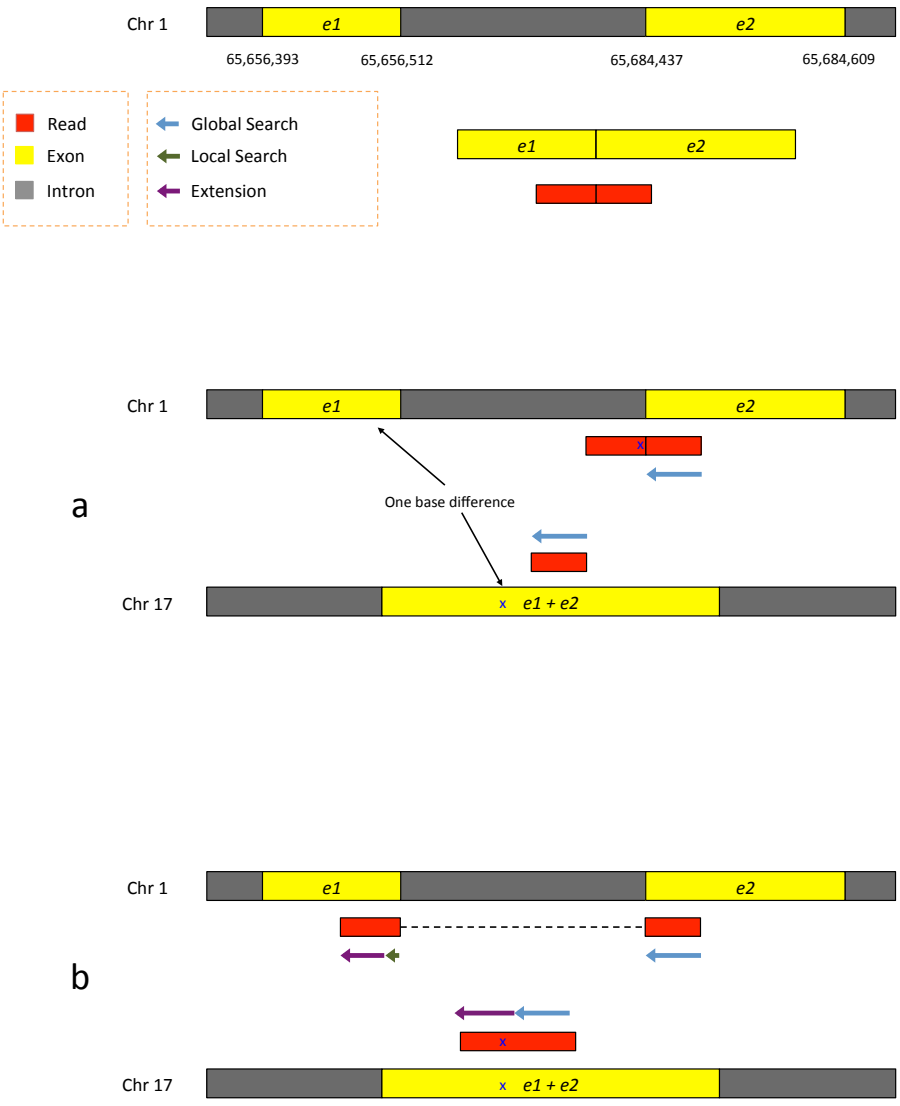
The examples include alignment of one exonic read and two junction reads (one an intermediate-anchored read and the other a long-anchored read). Reads are error-free and 100-bp long.



Supplementary Figure 9

Two-step approach version of HISAT to allow alignment of junction reads with small anchors.

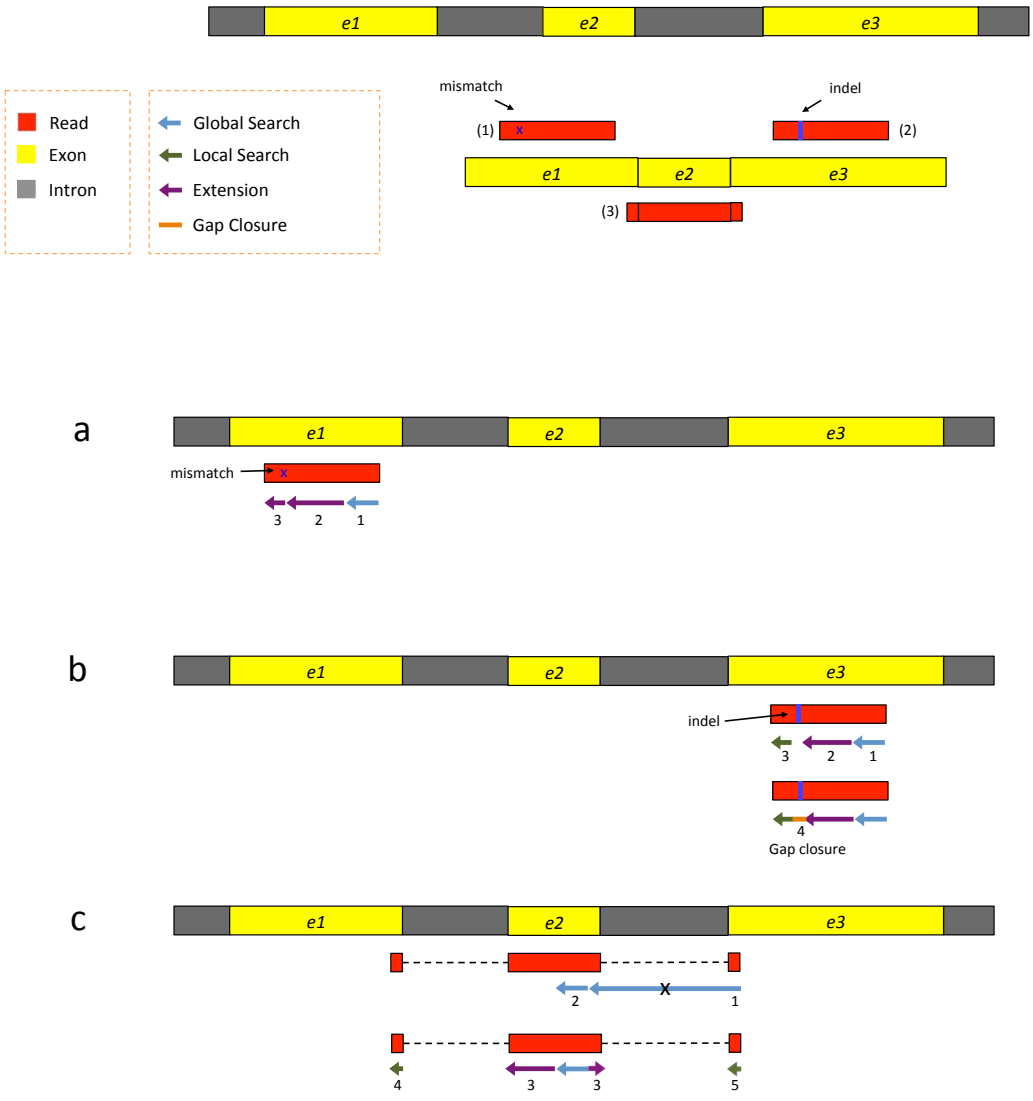
This figure shows how to align reads with short anchors (1-7 bp) by making use of splice sites found by reads with long anchors.



Supplementary Figure 10

Alignment of junction reads in the presence of processed pseudogenes.

This figure shows how to correctly align reads that would otherwise be mapped incorrectly to processed pseudogenes.



Supplementary Figure 11

Three more examples demonstrating how HISAT applies its hierarchical indexing for reads involving mismatches, indels and three exons.

The examples include alignment of one exonic read with one mismatch, one exonic read with an indel, and three exon spanning reads with two small anchors on both sides. Reads are 100-bp long.

Supplementary Table 1

Program	No. of splice sites reported	No. of true splice sites reported	Sensitivity (%)	Precision (%)
HISATx1	95,732	91,050	97.7	95.1
HISATx2	94,217	91,210	97.9	96.8
HISAT	94,121	91,159	97.8	96.9
STAR	96,326	90,535	97.1	94.0
STARx2	95,404	90,590	97.2	95.0
GSNAP	97,385	91,171	97.8	93.6
OLego	91,107	87,701	94.1	96.3
TopHat2	109,276	84,866	91.1	77.7

Sensitivity and precision of splice sites reported by spliced alignment software for 40M simulated error-free paired-end reads (20 million pairs) from the entire human genome.

The number of known splice sites included in the simulated paired-end reads (20 million pairs) is 93,199. Sensitivity is the percentage of true splice sites found out of the total that were present. Precision (or positive predictive value) is the percentage of reported splice sites that are correct.

Supplementary Table 2

Program	Correctly and uniquely aligned reads	Correctly mapped reads (multi-mapped)	Incorrectly mapped reads	Unmapped reads
HISATx1	92.1%	1.5%	3.7%	2.8%
HISATx2	97.6%	1.7%	0.2%	0.5%
HISAT	97.6%	1.7%	0.3%	0.5%
STAR	88.9%	1.6%	9.2%	0.3%
STARx2	96.7%	2.3%	0.9%	0.2%
GSNAP	92.3%	1.5%	6.0%	0.1%
OLego	91.6%	0.0%	4.4%	4.0%
TopHat2	94.4%	3.1%	0.1%	2.4%

Alignment results of spliced alignment software for 20 million simulated 100-bp reads with a mismatch rate of 0.5%.
This table shows the alignment results for all the reads (all, M, 2M_gt_15, 2M_8_15, 2M_1_7, gt_2M). Reads are categorized as: (1) correctly and uniquely mapped, (2) correctly mapped (multi-mapped), (3) incorrectly mapped, and (4) unmapped. Case (2) covers instances where an aligner mapped a read to multiple locations and one of the locations was correct. These four categories encompass all of the reads.

Supplementary Table 3

	Alignment precision for non-GT/AG splice sites (exact matching)	Alignment precision for non-GT/AG splice sites (±5-bp window)
HISATx1	49,244 (39%)	66,112 (52%)
HISATx2	72,120 (57%)	95,702 (75%)
HISAT	73,738 (58%)	96,083 (75%)
STAR	49,276 (39%)	69,520 (55%)
STARx2	68,337 (54%)	91,398 (72%)
GSNAP	64,688 (51%)	69,572 (55%)
TopHat2	68,863 (54%)	70,934 (56%)

Alignment precision involving non-GT/AG splice sites reported by spliced alignment software for 20 million simulated single-end reads from the human genome, with a mismatch rate of 0.5%.

There were 127,287 reads spanning non-GT/AG splice sites in the simulated 20 million single-end reads. Alignment precision measures the percentage of reads that are aligned correctly. Reads that mapped to multiple locations were considered correct if one of the mapped locations is correct. Column 2 shows the precision of each program if alignments were required to map precisely to the non-consensus splice site. Column 3 shows precision with a relaxed criterion, counting alignments as correct if they match within 5 bp of the non-consensus splice site. Note that OLego is not included in the table because it does not predict non-GT/AG splice sites.

Supplementary Table 4

Program	No. of splice sites reported	No. of true splice sites reported	Sensitivity (%)	Precision (%)
HISATx1	98,345	91,059	97.3	92.6
HISATx2	96,246	91,253	97.6	94.8
HISAT	96,186	91,209	97.5	94.8
STAR	102,488	90,624	96.9	88.4
STARx2	98,496	90,683	96.9	92.1
GSNAP	100,741	91,508	97.8	90.8
OLego	93,851	88,171	94.3	93.9
TopHat2	114,036	84,995	90.9	74.5

Sensitivity and precision of splice sites reported by spliced alignment software for 40M simulated paired-end reads (20 million pairs) from the entire human genome, with a mismatch rate of 0.5%.

The number of known splice sites included in the simulated paired-end reads (20 million pairs) is 93,543. Sensitivity is the percentage of true splice sites found out of the total that were present. Precision (or positive predictive value) is the percentage of reported splice sites that are correct.

Supplementary Table 5

Program	Run time (minutes)
HISATx1	46.2
HISATx2	96.7
HISAT	55.8
STAR	52.8
STARx2	147.4
GSNAP	1365.3
OLego	1978.7
TopHat2	2416.5

Run-time of the alignment software for ~218 million real paired-end reads (~109 million pairs)

Supplementary Table 6

Program	Run time (minutes)
HISATx1	31
HISATx2	64.5
HISAT	34.6
STAR	35.8
STARx2	93.1
GSNAP	2588.5
OLego	1187.9
TopHat2	1606.0

Run-time of the alignment software for ~126 million real paired-end reads (~63 million pairs)

Supplementary Table 7

Program	Version	Parameters
HISATx1	0.1.2-beta	hisat-align -p 3 --no-temp-splicesite <index> -1 <read_1> -2 <read_2>
HISAT (default)		hisat-align -p 3 <index> -1 <read_1> -2 <read_2>
HISATx2		First pass) hisat-align -p 3 --novel-splicesite-outfile splicesites.txt <index> -1 <read_1> -2 <read_2>
		Second pass) hisat-align -p 3 --novel-splicesite-infile splicesites.txt <index> -1 <read_1> -2 <read_2>
TopHat2	2.0.11	tophat -p 3 --read-edit-dist 3 --no-sort-bam --read-realign-edit-dist 0 --keep-tmp <index> <read_1> <read_2>
OLego	1.1.2	Left read) olego --num-reads-batch 1024 -t 3 -M 3 -o out1.sam <index> <read_1>
		Right read) olego --num-reads-batch 1024 -t 3 -M 3 -o out2.sam <index> <read_2>
		Pairing alignments of left and right reads) mergePEsam.pl out1.sam out2.sam out.sam
GSNAP	2014-5-30	gsnap -A sam -t 3 --max-mismatches=3 -D . -N 1 -d <index> <read_1> <read_2>
STAR	2.4.0a August 2014	STAR --runThreadN 3 --genomeDir <index> --genomeLoad NoSharedMemory --readFilesIn <read_1> <read_2> --outFilterMismatchNmax 6
STARx2		First pass) STAR --runThreadN 3 --genomeDir <index> --genomeLoad NoSharedMemory --readFilesIn <read_1> <read_2> --outFilterMismatchNmax 6
		Indexing) STAR --genomeDir <new_index> --runMode genomeGenerate --genomeFastaFiles genome.fa --sjdbFileChrStartEnd SJ.out.tab.Pass1.sjdb --sjdbOverhang 99 --runThreadN 3
		Second pass) STAR --runThreadN 3 --genomeDir <new_index> --genomeLoad NoSharedMemory --alignSJDBoverhangMin 1 --readFilesIn <read_1> <read_2> --outFilterMismatchNmax 6

Program parameters for running simulated and real reads

Nature Methods: doi:10.1038/nmeth.3317

The last column (Parameters) shows specific parameters for each program that allow a pair to be aligned with edit distances of 0, 1, 2, and 3. For single-end reads, <read_2> field is not needed, "--outFilterMismatchNmax 3" is used in STAR/STARx2, and only "Left read" is needed in OLego. We ran the programs on Mac Pro with a 3.7 GHz Quad-Core (Intel Xeon E5 processor) and 64 GB of RAM (1866 MHz DDR3 ECC memory).

Supplementary Note

(1) Details on simulated data sets

Rather than precisely imitating real RNA-seq experiments, we generated reads specifically for the purpose of testing the alignment programs. To simulate reads, we used the transcript expression model from the Flux simulator [1]. First, we randomly ranked the transcripts from the protein coding genes found in the Ensembl human gene annotation (release 66). Then we modeled the expression levels of the transcripts as follows. The expression level y of a transcript is defined as $y = \left(\frac{x}{x_0}\right)^k e^{-\left(\frac{x}{x_1}\right)-\left(\frac{x}{x_1}\right)^2}$, where x is the rank number of a transcript, $x_0 = 5 \times 10^7$, $x_1 = 9500$, and $k = -0.6$. Fragment lengths are chosen according to a normal distribution (mean: 250 bp and s.d.: 40 bp) and fragments are generated from the transcripts with their 3' and 5' positions randomly selected according to a uniform distribution. Left and right reads (100-bp long) are generated from the fragments. To create reads with mismatches, we replaced each nucleotide of a read with a different base with a probability of 0.5%. The maximum number of mismatches allowed in each read is 3. All these procedures are implemented in the TuxSim simulation program, which was originally developed by Cole Trapnell and slightly modified by us to allow reads with mismatches.

To run TuxSim, download tuxsim-0.1.tar.gz from
http://www.ccb.jhu.edu/software/hisat/downloads/hisat-suppl/tuxsim-0.1.tar.gz and follow the instructions below.
(Version 1.38.0 or higher of the boost library, http://www.boost.org, is required.)

i) tar xvzf tuxsim-0.1.tar.gz
ii) cd tuxsim-0.1
iii) ./configure --with-boost=/path/to/boost_prefix_dir
iv) make

We used release GRCh37 of the human genome, available from many sites including
http://ccb.jhu.edu/software/hisat/downloads/hisat-suppl/genome.fa
The human gene annotation used for our simulation is available at
http://ccb.jhu.edu/software/hisat/downloads/hisat-suppl/genes.gtf

The 40 million error-free 100-bp reads (20 million pairs) are available for download at
http://ccb.jhu.edu/software/hisat/downloads/hisat-suppl/reads_perfect.tar.gz
Alternatively, these reads can be re-generated them as follows

i) tuxsim-0.1/src/tuxsim sim_perfect.cfg
(sim_perfect.cfg is available at
http://ccb.jhu.edu/software/hisat/downloads/hisat-suppl/sim_perfect.cfg)

ii) python tuxsim-0.1/src/extract_and_shuffle.py sim_20M 20000000

The 40 million 100-bp reads (20 million pairs) with mismatches at a rate of 0.5% are available for download at http://www.ccb.jhu.edu/software/hisat/downloads/hisat-suppl/reads_mismatch.tar.gz
This data set can be re-generated as follows:

i) tuxsim-0.1/src/tuxsim sim_mismatch.cfg
(sim_mismatch.cfg is available at
http://www.ccb.jhu.edu/software/hisat/downloads/hisat-suppl/sim_mismatch.cfg)

ii) python tuxsim-0.1/src/extract_and_shuffle.py sim_20M 20000000

(2) Downloading the real data sets used in this study

For the real data in the main text [GEO accession number: GSM981249], the reads are available to download at
ftp://hgdownload.cse.ucsc.edu/goldenPath/hg19/encodeDCC/wgEncodeCshlLongRnaSeq/wgEncodeCshlLongRnaSeqImr90CellPapFastqRd1Rep1.fastq.gz,
ftp://hgdownload.cse.ucsc.edu/goldenPath/hg19/encodeDCC/wgEncodeCshlLongRnaSeq/wgEncodeCshlLongRnaSeqImr90CellPapFastqRd1Rep2.fastq.gz

For **Supplementary Fig. 6** and **Supplementary Table 5**, we used the same data set as in the main text, 217,498,662 paired-end reads (108,749,331 pairs) from fetal lung fibroblasts using Illumina HiSeq 2000 [GEO accession number: GSM981249]. All reads are 101 bp in length.

For **Supplementary Fig. 7** and **Supplementary Table 6**, we used RNA-seq reads gathered across a time course experiment reported by Chen et al. [2] [GEO accession number: GSM818581]. This data includes 125,642,456 million paired-end reads in 62,821,228 pairs. All reads are 101 bp in length.

Details on how HISAT handles alignment involving mismatches, indels, and short anchors on both ends of a read

Here we illustrate the HISAT algorithm in greater detail through the use of three examples: (a) a read with one mismatch, (b) a read with one indel, and (c) a read with short anchors on both ends (**Supplementary Fig. 11**). For case (b), it does not matter for the sake of discussion whether the indel is an insertion or a deletion. As with the examples given in the main text, reads are assumed to be 100-bp long. In addition to the global search, local search, and directed read extension strategies presented in the main text, HISAT also includes a gap closure operation, which combines two partial alignments and fills gaps if any exist. Also, the extension operation allows mismatches while extending the alignment of a read and, unlike index-based operations (global and local operations), the extension operation is bi-directional. We illustrate these relatively simple cases, providing some insight into how HISAT aligns more complicated cases. In the following description, we avoid excessive details so that we may convey key ideas about the core alignment algorithms of HISAT.

(a) *A read with one mismatch*. Similar to the examples in **Supplementary Fig. 8**, we first search the read from its right end using the global FM-index. Once we find a uniquely aligned anchor, we extend the alignment until we encounter a mismatch (**Supplementary Fig. 11a**). At this point, we might try either a local search (assuming an intron is present) or an extension operation to align the remaining part of the read. Suppose we use extension operation with one mismatch allowed. Since the read has only one mismatch and the read is included entirely within an exon (*e1*), the extension succeeds and sweeps across the rest of the read. Note that the sequence of operations applied to align this read is shown just below the operation arrows in the figure.

(b) *A read with an indel*. Suppose the indel in **Supplementary Fig. 11b** is 2 bp long. As in the previous case, we first use the global index to anchor the read and extend the alignment until we encounter a mismatch (due to the indel). We then try local search just after the mismatched base, but the local search fails because the indel is two bases long and the second base of the indel is included for the local search. We also try directed extension operation with one mismatch allowed, but the extension fails because the operation does not allow gaps. The next step is to skip a certain number of bases (e.g., 8 bp) and retry local search, producing a partial alignment on the left side of the indel. We now have two partial alignments on both sides of the indel, which we combine using the gap closure operation, producing the alignment with the indel as shown in the figure. The current version of HISAT allows only one indel per gap closure operation.

(c) *A read with short anchors (10-bp each) on both ends*.
Suppose the read in **Supplementary Fig. 11c** spans three exons (10 bp on the left exon, 80 bp on the middle one, and 10 bp on the right one). HISAT first searches the read from its right end, but fails to anchor it using the global index because a right segment of the read spans two exons. Note this failure will also happen if the segment includes

mismatches or indels. Next HISAT tries to anchor the read beginning at the 18th base (this parameter can be adjusted) and this time it succeeds in anchoring the read. HISAT then extends the anchored alignment in both directions as shown in the figure. The left extension and the right extension stop at 91st base and 10th base respectively, because of the exons (*e1* and *e3*) the read spans on its both ends. The left 10-bp segment of the read is aligned using local search. HISAT also uses the local index to align the right 10-bp segment. Note that the direction of the last local index search is right-to-left, aligning the small anchor from its right end.

HISAT provides several parameters with which users can customize its alignment strategy, including adjustable penalties for mismatches, indels, and non-canonical splice sites. The default penalty for a mismatch ranges from 2 (minimum: MN) to 6 (maximum: MX) depending on the quality score (Q) at the mismatched base, MN + floor((MX-MN) * MIN(Q, 40.0) / 40.0). The default gap opening and extension penalties are 3 and 5, respectively. The default penalty score for each non-canonical splice site is 12. The default minimum score for an alignment to be reported is -18. For example, if a read's alignment has one mismatch (assuming Q is 20) and involves one non-canonical splice site, the alignment score is -16, the sum of -4 (for the mismatch) and -12 (for the non-canonical splice site). Since the alignment score is greater than the default minimum score (-18), the alignment will be reported.

References

1. 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 research* 2012.
2. Chen R, Mias GI, Li-Pook-Than J, Jiang L, Lam HY, Miriami E, Karczewski KJ, Hariharan M, Dewey FE, Cheng Y, et al: **Personal omics profiling reveals dynamic molecular and medical phenotypes.** *Cell* 2012, **148:**1293-1307.