Supplementary Material

Alignments of simulated reads with up to 3 mismatches

We generated single-end and paired-end reads with 0 to 3 mismatches and without indels as shown in Tables S1 and S3. TopHat2 and MapSplice show the highest mapping sensitivity in read/pair and spliced read/pair alignments for both true mismatches (SNPs) and sequencing-error mismatches (Tables S2 and S4).

Table S1. The number of reads and spliced reads with up to 3 mismatches of 0 to 3.

Type	No. of total reads	No. of reads without mismatches (junction)	No. of reads with 1 mismatch (junction)	No. of reads with 2 mismatches (junction)	No. of reads with 3 mismatches (junction)
True mismatches	20,000,000	10,860,864 (4,654,864)	7,579,737 (2,289,006)	1,396,742 (428,136)	162,657 (46,185)
Sequencing-error mismatches	20,000,000	11,258,169 (4,010,662)	7,298,699 (2,610,246)	1,297,051 (462,717)	146,081 (52,481)

Table S2. The recall rates of read and spliced read alignments for true mismatches (SNPs) and sequencing-error mismatches.

Program	True mismatches								Sequencing-error mismatches							
	M0	M1	M2	M3	J0	J1	J2	J3	M0	M1	M2	M3	J0	J1	J2	J3
TopHat2 +Bowtie1	98.14	98.71	98.83	97.57	95.81	95.86	96.45	91.52	98.37	98.60	98.79	97.19	95.67	96.23	96.71	92.23
TopHat2 +Bowtie2	97.85	98.70	95.08	86.72	95.00	95.75	84.59	55.21	98.08	98.54	93.87	84.98	94.61	95.95	83.16	58.69
GSNAP	92.85	89.08	83.50	78.33	83.33	77.49	74.19	70.27	93.95	88.19	83.09	77.66	83.03	77.61	74.29	69.35
RUM	85.10	83.45	77.58	73.82	65.25	54.29	45.82	37.93	87.58	81.43	75.37	69.57	65.13	55.16	45.55	36.63
MapSplice	96.77	98.25	97.96	93.94	92.47	94.25	96.98	96.95	96.85	97.77	97.78	94.63	91.16	93.79	95.94	95.82
STAR	90.39	87.84	82.15	78.52	77.65	68.96	61.18	55.10	91.82	86.35	80.68	75.39	77.17	69.07	60.72	53.36

M0 is the sensitivity of read alignments with zero mismatches. M1 is the sensitivity of alignments with one mismatch. M2 and M3 are similarly defined. J0 is the sensitivity of spliced alignments with no mismatches. J1, J2, and J3 are similarly defined with mismatches of 1, 2, and 3, respectively, for spliced alignments. M0, M1, M2, and M3 also include spliced alignments as well as non-gapped alignments. Note that TopHat2 with Bowtie2 suffers a drop in performance compared to Bowtie1 when a single read has 3 mismatches (column J3). This occurs because TopHat2 splits reads into very short segments, 25 bp, when attempting to align across splice sites. TopHat2 then calls Bowtie1/2 to align these short segments. Bowtie2's default parameters are not designed for such short segments; however these can easily be modified by changing the parameters used to call Bowtie2 within TopHat2.

Table S3. The number of pairs and spliced pairs with mismatches of 0 to 3 for true mismatches (SNPs) and sequencing-error mismatches.

Type	No. of total pairs	No. of pairs without mismatches (junction)	No. of pairs with 1 mismatch (junction)	No. of pairs with 2 mismatches (junction)	No. of pairs with ≥ 3 mismatches (junction)
True mismatches	20,000,000	5,703,884 (3,809,739)	8,547,831 (4,031,062)	4,345,589 (1,779,403)	1,402,696 (557,825)
Sequencing-error mismatches	20,000,000	5,747,299 (2,818,790)	9,201,311 (4,553,143)	3,897,205 (1,923,580)	1,154,185 (568,559)

Note that each read can contain up to 3 mismatches, so it is possible that a pair can have more than 3 mismatches.

Table S4. The recall rates of pair and spliced pair alignments for true mismatches (SNPs) and sequencing-error mismatches.

Program	True mismatches								Sequencing-error mismatches							
	M0	M1	M2	M≥3	J0	J1	J2	J≥3	M0	M1	M2	M≥3	J0	J1	J2	J≥3
TopHat2 +Bowtie1	95.69	96.96	97.47	97.45	93.03	93.14	93.64	93.12	96.72	96.98	97.19	96.91	93.19	93.78	94.15	93.51
TopHat2 +Bowtie2	95.06	96.77	96.01	91.94	91.90	92.52	89.40	78.18	96.10	96.59	94.92	89.86	91.46	92.60	89.33	78.12
GSNAP	84.03	83.95	79.29	72.72	74.03	69.83	64.84	58.95	88.34	83.20	77.52	70.99	73.84	69.76	64.84	59.17
RUM	69.86	73.97	72.22	67.81	51.85	45.42	39.32	33.22	78.40	72.92	68.09	63.08	52.57	46.43	40.03	33.66
MapSplice	90.53	92.59	93.33	92.47	84.70	84.88	85.59	85.98	91.90	91.77	91.97	91.48	83.57	83.67	84.23	84.73
STAR	79.41	81.17	78.07	60.80	66.65	61.01	55.04	41.48	85.05	80.12	75.03	58.85	66.64	61.25	55.01	41.49

M0 is the sensitivity of read alignments with zero mismatches. M1 is the sensitivity of alignments with one mismatch. M2 and M3 are similarly defined with mismatches of 1, 2, and ≥3, respectively. J0 is the sensitivity of spliced alignments with zero mismatches. J1, J2, and J≥3 are similarly defined for spliced alignments.

Corresponding tables for Figures 2-4 in the main text.

Table S5. Table for Figure 2.

	Program	0	1	2	3
De novo alignment	TopHat2 realignment	54,956,129	77,364,055	87,355,369	93,265,424
	TopHat	50,422,413	73,228,140	84,633,702	92,396,448
	GSNAP	52,255,865	74,247,781	84,946,229	91,598,102
	MapSplice	48,896,741	70,032,327	81,847,468	90,360,661
	STAR	50,986,666	71,782,717	81,074,505	86,235,516
Alignment using annotation	TopHat2 realignment	55,634,580	77,988,848	88,370,540	94,752,200
	TopHat	55,225,852	77,447,497	87,992,406	94,596,600
	GSNAP	54,666,282	76,642,607	86,835,392	93,005,273

	RUM	54,949,609	76,963,699	87,157,875	93,352,293
	STAR	54,326,036	75,730,313	84,957,399	89,844,775

Table S6. Table for Figure 3.

Type		Program	0	1	2	3
Alignments whose splice sites correspond to gene annotation	De novo alignment	TopHat2 realignment	15,804,625	21,406,115	23,524,839	24,436,600
		TopHat	9,799,757	13,586,453	15,104,339	15,798,045
		GSNAP	13,549,591	18,438,736	20,759,433	22,175,182
		MapSplice	14,792,707	20,264,394	22,961,083	24,704,514
		STAR	11,568,529	15,338,930	16,918,024	17,714,913
	Alignment using annotation	TopHat2 realignment	17,372,910	23,531,960	26,340,120	27,982,780
		TopHat	17,368,853	23,530,365	26,353,413	28,018,284
		GSNAP	16,801,716	22,812,953	25,598,496	27,259,090
		RUM	16,516,786	22,263,594	24,839,636	26,331,306
		STAR	16,526,673	22,195,936	24,558,091	25,693,885
All spliced alignments including novel splice sites	De novo alignment	TopHat2 realignment	17,516,565	24,088,224	26,632,215	27,754,233
		TopHat	10,238,968	14,232,391	15,847,929	16,601,804
		GSNAP	13,864,319	18,899,654	21,302,999	22,777,308
		MapSplice	15,863,181	22,638,514	26,692,556	29,630,048
		STAR	11,994,236	15,936,866	17,600,134	18,445,153
	Alignment using annotation	TopHat2 realignment	18,932,114	25,985,178	29,191,692	31,039,091
		TopHat	17,779,753	24,112,605	27,019,281	28,752,182
		GSNAP	17,117,374	23,272,081	26,138,915	27,858,112
		RUM	16,823,909	22,716,678	25,399,661	27,009,138
		STAR	16,895,367	22,725,029	25,170,512	26,352,382

Table S7. Table for Figure 4.

	Type	0	1	2	3
Read alignments	Realignment 0	54,956,129	77,364,055	87,355,369	93,265,424
	Realignment 1	54,508,641	77,227,362	87,334,380	93,272,963
	Realignment 2	53,007,141	76,631,857	87,168,673	93,244,130
	No Realignment	50,422,413	73,228,140	84,633,702	92,396,448
Spliced read alignments	Realignment 0	17,516,565	24,088,224	26,632,215	27,754,233
	Realignment 1	14,179,269	19,895,371	22,278,929	23,389,758
	Realignment 2	12,755,976	17,578,938	19,577,384	20,558,593
	No Realignment	10,238,968	14,232,391	15,847,929	16,601,804

Alignment rates for reads of different lengths (error-free)

In addition to 100-bp simulated reads in the main text we also generated single and paired-end reads of different lengths (50, 150, 200 bp) in order to check how TopHat2 works compared to other alignment software for various read lengths. We used different fragment lengths 200, 250, 350, 450 bp for read lengths 50, 100, 150, 200 bp, respectively. Figure S1 shows that TopHat2 performs better than the other programs for different read lengths. TopHat2 also outputs much more accurate alignments for spliced reads and spliced reads with small anchors. These results suggest that TopHat2 may be the better choice for longer reads (≥150 bp) that will likely become prevalent in the near future, as well as for currently available reads (50 ~ 100 bp).

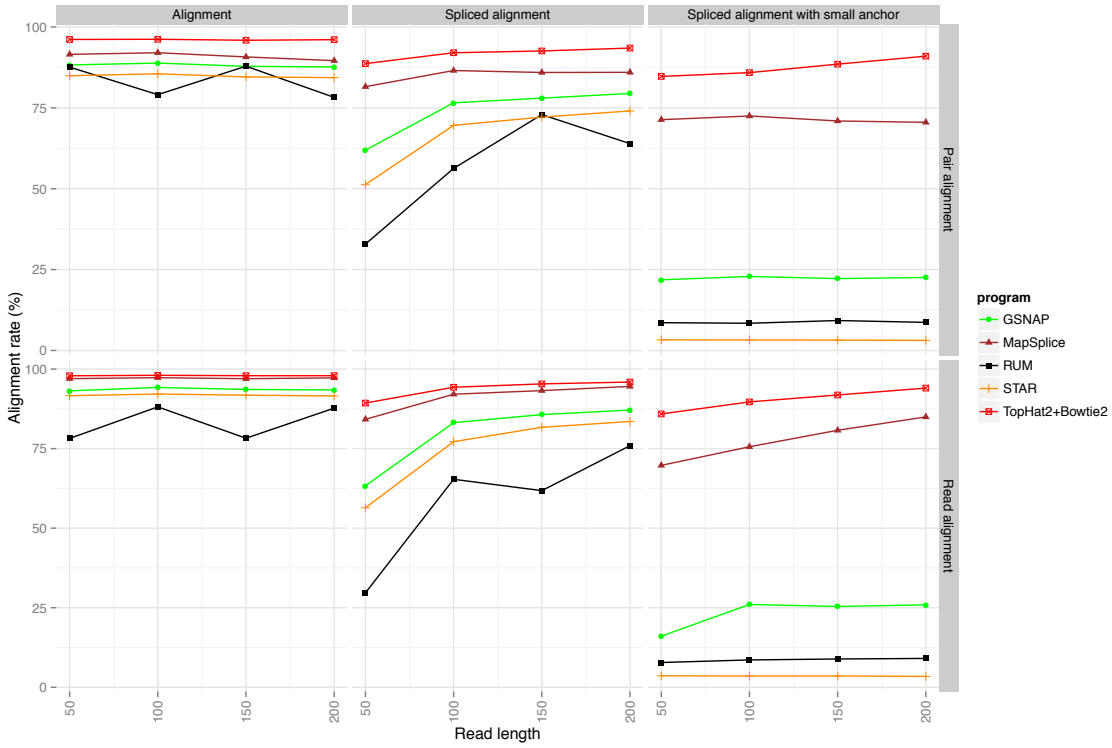

Figure S1. Mapping accuracy for single- and paired-end reads in different read lengths (20 million reads and 20 million pairs). Using simulated reads, the figure shows the ratio of correctly aligned reads (bottom) or pairs (top) for read alignment (the left column), spliced read alignment (the middle column), and spliced read alignment with small anchors (the right column).

Simulation of reads with indels and mismatches based on Ensembl human gene annotation (release 66)

We used the transcript expression model from the Flux simulator [1] to generate RNA-seq reads from the protein coding genes found in the Ensembl human gene annotation, release 66. First, the transcripts from the protein coding genes are randomly ranked. Then, the expression levels of the transcripts are modeled 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 $k = -0.6$.
Reads are simulated for the purpose of testing the alignment programs, as opposed to precisely mimicking real RNA-seq experiments. When generating reads with true indels, we include at most one indel per exon in a way that if the length of an exon *L* is greater than or equal to 1000 bp, we place either an insertion (50%) or a deletion (50%) into the exon at a random location. Otherwise an indel is introduced into a random location of the transcript with the chance of $\frac{L}{1000}$. Reads are generated from these transcripts so that the reads share any changes that their derivative transcripts have. For reads with true mismatches we change the nucleotides of each transcript in such a way that the average distance between two nearby mismatches is 150.5 bp and the distribution of the distance is uniform (1 to 300 bp). Reads are then generated from these modified transcripts. Reads with either indels or mismatches from sequencing errors are simulated in the same way except the transcript being used is changed every time a read is generated.

Figure S2 shows the proportions of reads spanning multiple exons, which increase approximately from 19% to 46% as the length of reads increases from 50 to 150 bp. On the other hand, as we may expect, the fragment length does not affect the proportions of spliced reads.

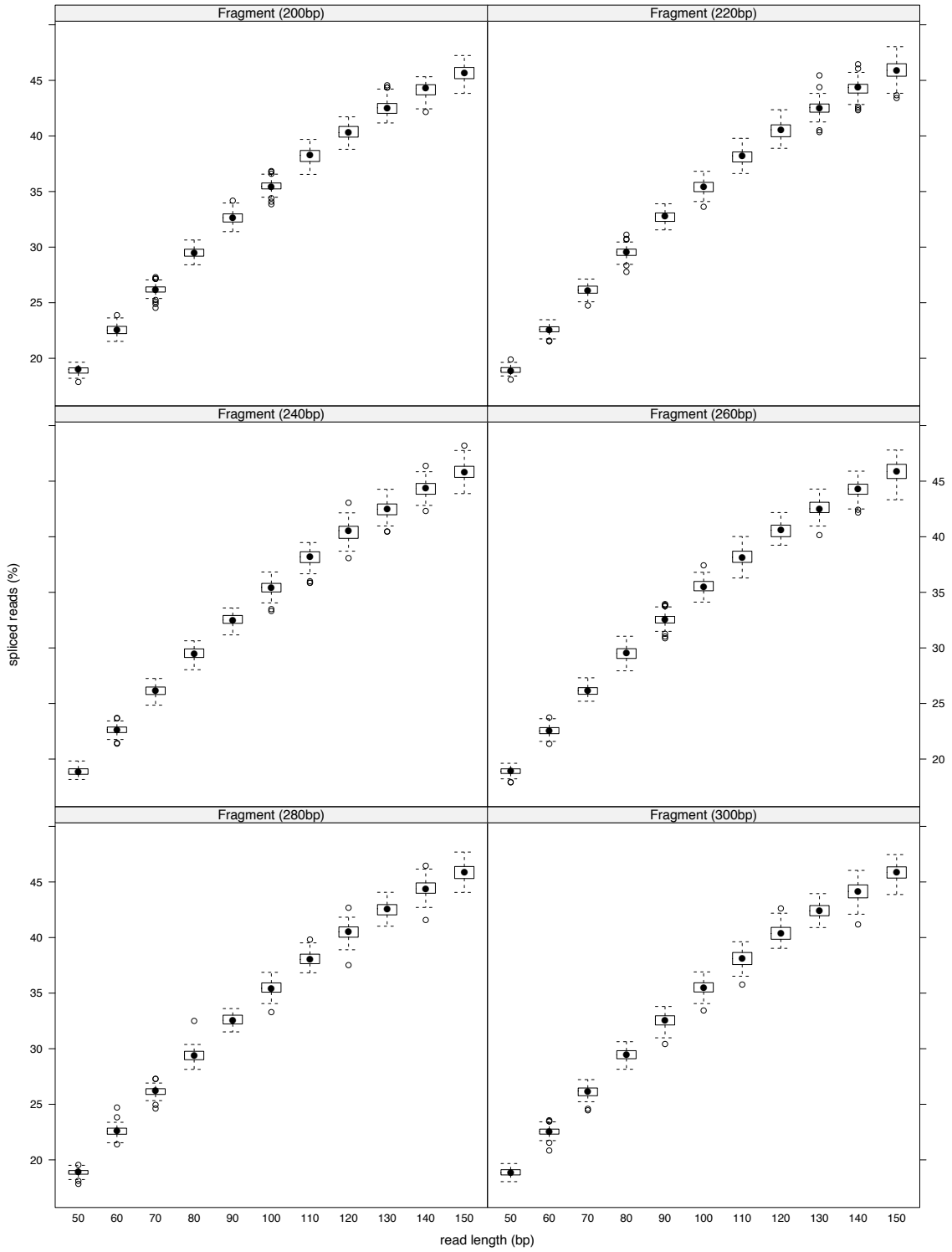

Figure S2. Proportions of spliced reads from different read lengths (50 to 150 bp) and fragment lengths (200 to 300 bp). For each fragment length (200, 220, 240, 260, 280, 300 bp), a whisker box plot shows 100 simulation results (the percentage of spliced reads) for each read length.

Runtime and memory usage of TopHat2, GSNAP, RUM, MapSplice, and STAR for ~130 million paired-end reads from Chen et al. [2]

We ran each program using 8 threads on a Linux machine with memory of 256GB and 48 AMD processors (2.1GHz). Runtime (or wall time) and peak memory usage were measured using the GNU *time* program as shown in Table S5.

Table S8. Runtime and memory usage of RNA-seq alignment software (TopHat2, GSNAP, RUM, MapSplice, and STAR).

Program	Runtime (wall time)	Peak memory (GB)	Parameters
TopHat2 2.0.8 (Transcriptome only mapping)	8h 29m	4.9	-G --transcriptome-only --read-mismatches 3 --read-gap-length 3 --read-edit-dist 3 --mate-inner-dist 60 --mate-std-dev 60
TopHat2 2.0.8 (Default: genome and spliced mapping)	17h 1m	5.4	--read-mismatches 3 --read-gap-length 3 --read-edit-dist 3 --mate-inner-dist 60 --mate-std-dev 60
TopHat2 2.0.8 (With transcriptome mapping)	17h 31m	5.2	-G --read-mismatches 3 --read-gap-length 3 --read-edit-dist 3 --mate-inner-dist 60 --mate-std-dev 60
TopHat2 2.0.8 (Realignment with realignment edit distance of 0)	29h 55m	5.6	--read-mismatches 3 --read-gap-length 3 --read-edit-dist 3 --mate-inner-dist 60 --mate-std-dev 60 --read-realign-edit-dist 0
GSNAP 2013-01-23	55h 26m	7.6	--max-mismatches=3 -N 1
RUM 1.12_01	26h 34m	*36.4	
MapSplice 1.15.2	44h 50m	3.7	min_missed_seg = 0
STAR 2.3.0e	32m	27.8	--outFilterMatchNmin 97 --outFilterScoreMin 90 --outFilterMismatchNmax 3

Note the last column "Parameters" shows specific parameters in order for each program to allow a read to be aligned with edit distances of 0, 1, 2, and 3. Parameters for

specifying genome, gene annotation, RNA-seq read files, and the number of threads are not shown. The version of each program is shown in blue in the first column. Note that RUM uses separate processes, each of which consisted of Bowtie (2394MB) and BLAT (4660MB), requiring a total of 36.4GB memory when using 8 threads.

Specific program parameters for TopHat2, TopHat1, GSNAP, RUM, MapSplice, and STAR

Table S9. Parameters for specifying genome, gene annotation, RNA-seq read files, and the number of threads are not shown.

Test	Program	Reference genome	Gene annotation	Specific parameters
Alignments of simulated reads (error-free)	TopHat2 +Bowtie1	Whole human genome	No	--mate-inner-dist 50 --mate-std-dev 40 --bowtie1
	TopHat2 +Bowtie2			--mate-inner-dist 50 --mate-std-dev 40
	TopHat1.1.4			--mate-inner-dist 50 --mate-std-dev 40
	GSNAP			-N 1
	RUM		Yes	
	MapSplice			min_missed_seg = 0
	STAR			--outFilterMatchNmin 97 --outFilterScoreMin 90 --outFilterMismatchNmax 3
Alignments of simulated reads with short indels (1-3 bp)	TopHat2 +Bowtie1		No	--mate-inner-dist 50 --mate-std-dev 40 --read-mismatches 3 --read-gap-length 3 --read-edit-dist 3 --bowtie1
	TopHat2 +Bowtie2			--mate-inner-dist 50 --mate-std-dev 40 --read-mismatches 3 --read-gap-length 3 --read-edit-dist 3
	GSNAP			--max-mismatches=3 --indel-penalty=1 -N 1

	RUM		Yes	
	MapSplice			min_missed_seg = 0
	STAR			--outFilterMatchNmin 97 --outFilterScoreMin 90 --outFilterMismatchNmax 3
Alignments of simulated reads with up to 3 mismatches	TopHat2 +Bowtie1	Whole human genome	No	--mate-inner-dist 50 --mate-std-dev 40 --read-mismatches 3 --read-gap-length 3 --read-edit-dist 3 --bowtie1
	TopHat2 +Bowtie2			--mate-inner-dist 50 --mate-std-dev 40 --read-mismatches 3 --read-gap-length 3 --read-edit-dist 3
	TopHat1.1.4			--mate-inner-dist 50 --mate-std-dev 40
	GSNAP			--max-mismatches=3 -N 1
	RUM		Yes	
	MapSplice			min_missed_seg = 0
	STAR		No	--outFilterMatchNmin 97 --outFilterScoreMin 90 --outFilterMismatchNmax 3
Alignments of a large set of real RNA-seq reads (Chen et al. [2])	TopHat2	Whole human genome	Yes/No	--read-mismatches 3 --read-gap-length 3 --read-edit-dist 3 --mate-inner-dist 60 --mate-std-dev 60
	TopHat2 realignment 0		Yes/No	--read-mismatches 3 --read-gap-length 3 --read-edit-dist 3 --mate-inner-dist 60 --mate-std-dev 60 --read-realign-edit-dist 0
	GSNAP		Yes/No	--max-mismatches=3 -N 1
	RUM		Yes	

	MapSplice	No	min_missed_seg = 0
	STAR	Yes/No	--outFilterMatchNmin 97 --outFilterScoreMin 90 --outFilterMismatchNmax 3

Note that a TopHat option "--read-realign-edit-dist" can be used to realign reads in the spliced alignment phase that are mapped against either transcriptome or genome.

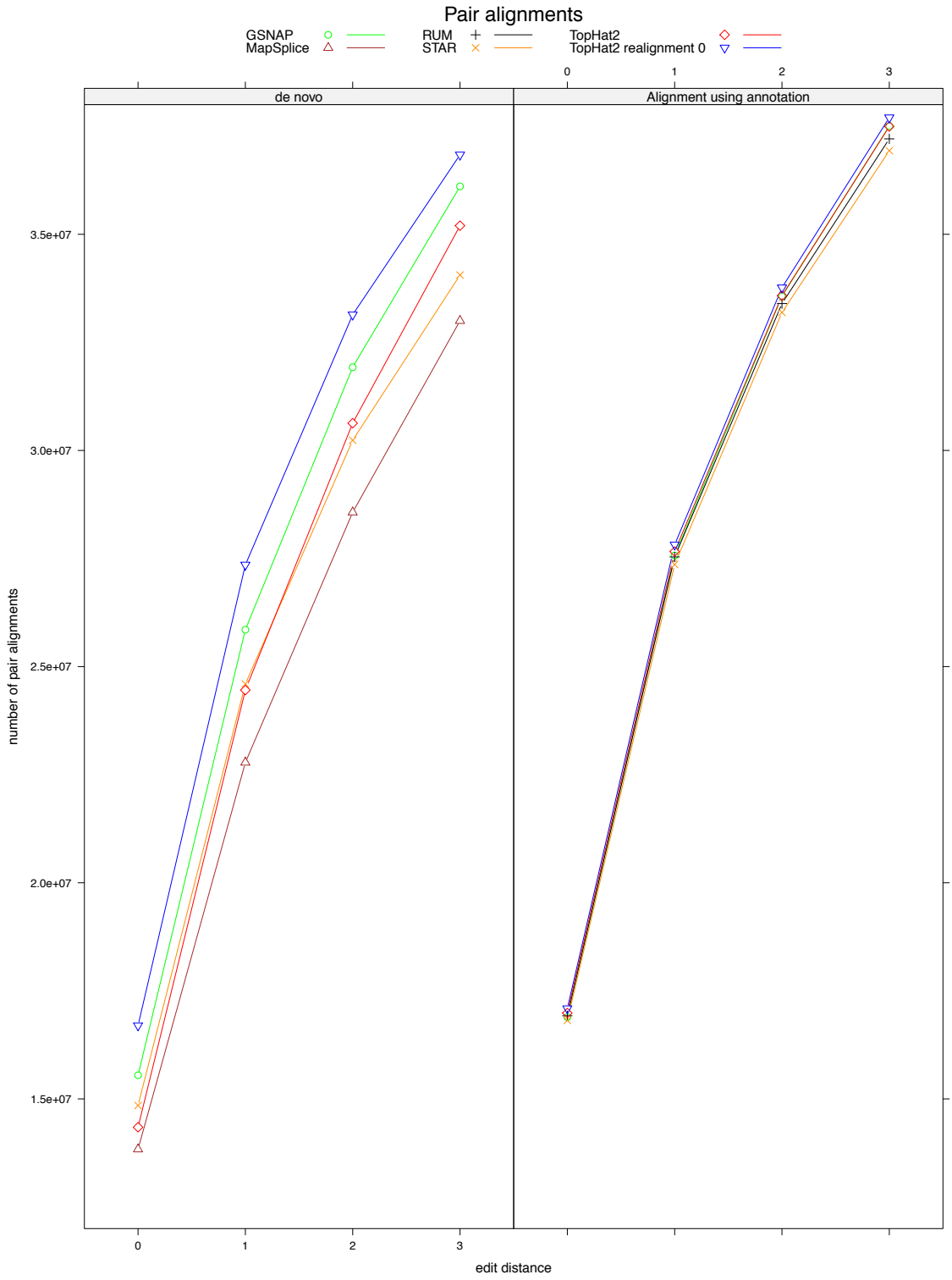

Figure S3. The number of pair alignments from TopHat2, GSNAP, RUM, MapSplice, and STAR (the RNA-seq reads are from Chen et al. [2]).

Table S10. Table for Figure S3.

	Program	0	1	2	3
De novo alignment	TopHat2 realignment	16,696,682	27,353,265	33,139,753	36,839,143
	TopHat	14,344,271	24,456,802	30,630,922	35,199,608
	GSNAP	15,546,886	25,853,039	31,925,593	36,108,336
	MapSplice	13,835,185	22,781,288	28,568,799	32,999,167
	STAR	14,847,145	24,598,381	30,235,116	34,057,210
Alignment using annotation	TopHat2 realignment	17,091,131	27,818,953	33,766,156	37,699,996
	TopHat	16,985,383	27,661,740	33,579,775	37,494,323
	GSNAP	16,890,487	27,569,140	33,566,349	37,503,456
	RUM	16,923,302	27,536,281	33,397,563	37,208,206
	STAR	16,815,984	27,361,365	33,191,954	36,933,241

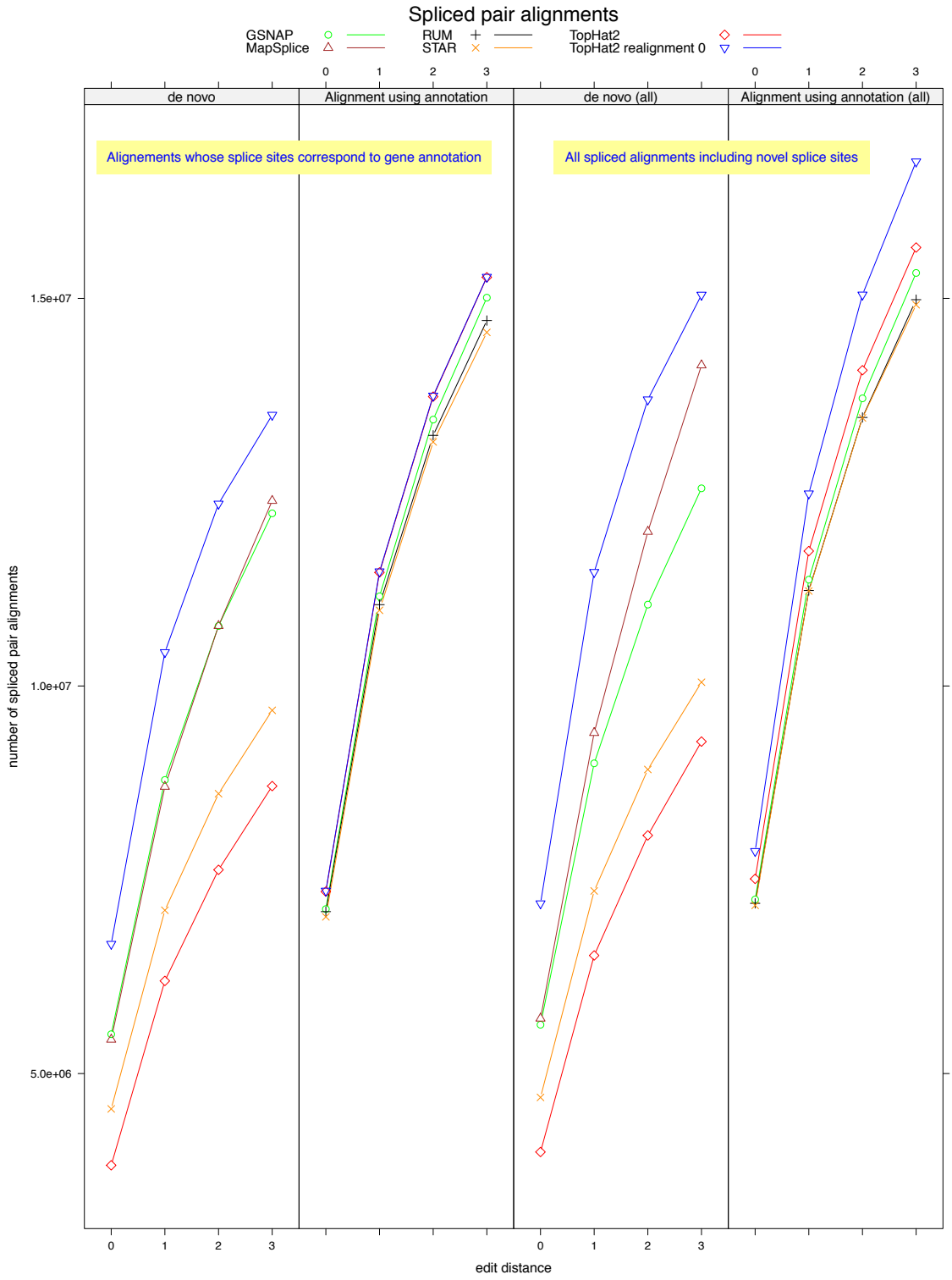

Figure S4. The number of spliced pair alignments from TopHat2, GSNAP, RUM, MapSplice, and STAR (the RNA-seq reads are from Chen et al. [2]).

Table S11. Table for Figure S4.

Type		Program	0	1	2	3
Alignments whose splice sites correspond to gene annotation	De novo alignment	TopHat2 realignment	6,670,997	10,434,104	12,349,897	13,496,341
		TopHat	3,816,460	6,195,116	7,628,348	8,708,508
		GSNAP	5,507,359	8,787,161	10,773,698	12,226,898
		MapSplice	5,438,391	8,701,358	10,775,190	12,389,538
		STAR	4,543,781	7,106,244	8,610,236	9,685,835
	Alignment using annotation	TopHat2 realignment	7,357,496	11,476,154	13,743,868	15,276,373
		TopHat	7,346,821	11,464,534	13,733,001	15,272,369
		GSNAP	7,121,858	11,156,844	13,436,485	15,009,697
		RUM	7,088,842	11,048,936	13,233,486	14,714,367
		STAR	7,021,511	10,975,663	13,147,910	14,561,264
All spliced alignments including novel splice sites	De novo alignment	TopHat2 realignment	7,193,604	11,468,318	13,694,621	15,045,756
		TopHat	3,988,139	6,523,309	8,072,162	9,282,468
		GSNAP	5,630,093	9,002,188	11,049,842	12,550,023
		MapSplice	5,710,435	9,395,612	11,990,324	14,137,678
		STAR	4,692,109	7,355,651	8,922,567	10,048,254
	Alignment using annotation	TopHat2 realignment	7,868,376	12,481,943	15,047,081	16,764,777
		TopHat	7,511,707	11,740,351	14,073,199	15,656,508
		GSNAP	7,245,286	11,371,551	13,710,608	15,328,468
		RUM	7,195,805	11,231,525	13,463,953	14,983,419
		STAR	7,169,487	11,224,944	13,458,212	14,917,855

Table S12. The expression levels of genes with pseudogene copies from Illumina Body Map 2.0 data [3].

Number of pseudogene copies	Gene with pseudogene	Pair Count (%)	Ratio	Normalized count (%)	Normalized ratio
1	553 (1.7%)	4.66	x 2.73	7.33	x 4.30
2	113 (0.4%)	3.51	x 10.08	3.97	x 11.39
3	49 (0.2%)	0.62	x 4.13	1.05	x 6.96
4	27 (0.1%)	1.32	x 15.82	1.52	x 18.30
≥5	130 (0.4%)	3.61	x 9.01	5.23	x 13.04
Total (≥1)	872/32,439 (2.7%)	13.72	x 5.11	19.10	x 7.11

Columns are defined as in Table 5 in the main text.

Table S13. The expression levels of protein-coding genes with processed pseudogene copies from Chen et al. [2].

Number of pseudogene copies	Protein-coding gene with processed pseudogene	Pair Count (%)	Ratio	Normalized count (%)	Normalized ratio
1	267 (1.31%)	6.88	x 5.26	9.55	x 7.30
2	47 (0.23%)	6.31	x 27.42	6.07	x 26.39
3	21 (0.10%)	1.27	x 12.38	1.97	x 19.15
4	16 (0.08%)	0.84	x 10.73	1.02	x 13.02
≥5	40 (0.20%)	6.73	x 34.33	7.92	x 40.45
Total (≥1)	391/20,417 (1.92%)	22.03	x 11.50	26.54	x 13.86

Columns are defined as in Table 5 in the main text.

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.
3. **The Illumina Body Map 2.0 data** [http://www.ebi.ac.uk/arrayexpress/browse.html?keywords=E-MTAB-513&expandefo=on]