Supplementary information: Lacking alignments? The next-generation sequencing mapper segemehl revisited

Christian Otto, Peter F. Stadler, Steve Hoffmann

Supplementary Methods

Datasets for comparison of read aligners

Each artificial dataset consists of 100 000 single- or paired-end reads and was simulated using `Mason` v0.1.1 [1] from the Human genome (hg19), excluding haplotypes, random contigs, and 'non-chromosomal' sequences. For the single-end Illumina datasets, `Mason` was run in Illumina mode with parameters `-hn2`, `-sq`, and the read length (`-n`) set to `100` and `30` for long and short reads, respectively. For the paired-end dataset, additionally, the parameters `-mp`, `-ll 375`, and `-le 100` were specified and the read length (`-n`) was again set to 100. The artificial 454 dataset was simulated in 454 mode with the parameter `-hn 2`, `-sq`, `-k 0.3`, `-bm 0.4`, `-bs 0.2`, and `-nm 400` (analogously to Langmead & Salzberg [2]).

The real datasets were downloaded, converted to fastQ format, some of them post-processed, and down-sampled to 100 000 single- or paired-end reads. The Illumina DNA-seq dataset was used as both single-end (by only using the first read sequences) and paired-end data. For both Illumina mRNA-seq datasets, the post-processing involved removing reads that possibly overlapped exon-exon junctions. To achieve this, the entire dataset was mapped using `segemehl` (with `-S` option), `STAR` [3], `TopHat 2` [4], and `SOAPsplice` [5], and reads which were split-mapped by any of these tools were removed prior to down-sampling. In case of the paired-end mRNA-seq dataset, only paired-end reads were kept where both ends were not split-mapped by any of the tools. For Illumina shortRNA-seq, 3′-adapter contaminations on the read sequences were clipped using `fastx_clipper` (part of the `FASTX-Toolkit`) with the Illumina shortRNA-seq adapter (TCGTATGCCGTCTTCTGCTTGT). Before down-sampling, reads outside of the expected length range (19-25 nt) were discarded.

An overview of the benchmarking datasets, their sequencing platforms, library types, and average read lengths is given in Supplementary Table S1. To permit full reproducability, we have assembled an Electronic Supplement[1] comprising all data, custom scripts, and detailed information on how to re-run the benchmarks.

Performance evaluation of `lack`

For the simulated data, we sampled 10 000 isoforms from the ASTD data base [6], extracted their sequences from the Human genome (hg19). Using `Mason` v.0.1.1 [1], we simulated single-end reads in Illumina and 454 mode of length 100 nt and 400 nt, respectively, from the isoform sequences at 20-fold coverage. For both datasets, two haplotypes (`-hn 2`) and base qualities (`-sq`) were simulated with `Mason` and the read length were specified accordingly, i.e., `-n 100` and `-nm 400` for Illumina and 454 data, respectively. In terms of the error models, we used the default parameter values of `Mason` for both datasets. The resulting datasets consisted of 989 387 and 247 346 reads in case of Illumina and 454, respectively. For the real data, we used one Illumina (access no. SRR534289) and one 454 RNA-seq dataset (access no. GSM951482). Both datasets were downloaded, converted to fastQ format, and down-sampled to 1 000 000 and 250 000 single-end reads for Illumina and 454, respectively. In addition to Illumina and 454 data, we simulated single-end Ion Torrent reads of length 200 nt, from the sampled ASTD isoform sequences at 20-fold coverage using `Mason`. Due to the fact that there is no Ion Torrent mode in `Mason` but the 454 and Ion Torrent technology are similar in terms of the homopolymer issue, we used the 454 mode and altered the parameters of the error model. As reference, we used a recent publication of Bragg and colleagues [7] that analyzed the error characteristics of the Ion Torrent Personal Genome Machines (PGM) using different library preparation kits, one for 100 nt and two for 200 nt long reads. Even though it was not possible with `Mason` to mimick the exact error profiles of Ion Torrent data, we tested various different

[1] http://www.bioinf.uni-leipzig.de/publications/supplements/13-008

combinations of the 454-specific error parameters in Mason and obtained reads with an average insertion and deletion rate very similar to the ones reported for 200 nt long reads. In result, Mason was executed using the following parameters: -hn 2, -sq, -k 0.3, -bm 0.35, -bs 0.18.

All datasets were mapped using segemehl v.0.1.7 (with option -S), Blat v.35x1 [8], TopHat 2 v.2.0.9 [4], and STAR v.2.3.0e [3] under default parameters. Using custom scripts, the output of each split-read aligner was converted to segemehl's extended SAM format where custom SAM tags were added to expand the capabilities of the SAM format to flexibly and consistently represent single- or paired-end multi-split alignments. If not provided by the aligner, a file with unmapped reads was generated in a post-processing step. Subsequently, segemehl without split-read option (-S) was executed to add information on the best seed found. For each dataset and split-read aligner, lack v.0.1.7 was executed under default parameter values on the set of unmapped reads using the reported alignments by the aligner as split junction data base. The user time and peak virtual memory consumption were tracked during all alignment runs.

To assess the performance of lack, we first inspected the fraction of mapped and remapped reads, i.e. reads that were mapped by the split-read aligner and lack, respectively, as well as unmapped reads that could not be mapped by split-read aligner and lack. The mapped reads were further subdivided into unsplit-mapped reads, where a continuous alignment block to the reference was found, and split-mapped reads, otherwise. The distinction between unsplit- and split-mapped reads was important since reads obtained from RNA-seq which did not overlap an exon-exon junction and hence did not contain any splice junction information, cannot be recovered by lack by design. We defined the remapping rate of lack as the fraction of reads that were missed by the aligner but were rescued by lack and calculated it for each dataset and split-read aligner. Due to the fact that lack only searches for split-read alignments, we further calculated the split-read remapping rates of lack, i.e., the fraction of split-mapped reads missed by the aligner but rescued by lack. To decide whether an unmapped read was split-mapped, we used the total amount of split-read alignments reported by any aligner or by lack. An unmapped read was deemed split-mapped if a split-read alignment of it was reported by any aligner or by lack. Thus, remapping and split-read remapping rates constituted lower and upper bounds, respectively, on the performance of lack in the recovery of unmapped RNA-seq reads that may have emerged from splicing events. To assess the quality of lack alignments compared to the ones reported by split-read aligners, the accuracy over all split-mapped and remapped alignments was estimated on the artificial datasets where the true origin and hence true (split-read) alignment was known. The accuracy of an alignment was given by the relative number of correct alignment blocks by comparing mapping strand, start, and end position. Due to the issue of multiple optimal alignments, a small deviation from the true position was permitted. Note that the calculation of the accuracy via correct alignment blocks was rather strict.

Secondly, we inspected the benefit of using lack on splice junctions and number of split-reads overlapping them, termed support. Higher support increases the confidence in a splice junction and often splice junctions below a given minimum read support are discarded as untrustworthy. For all datasets and split-read aligners, the number of splice junction was calculated as function of minimal read support with and without use of lack. Given a minimal required support, the difference between both values represented the gain in confident splice junctions due to lack. To illustrate whether lack greatly increased the support of few junctions or minorly enhances the support of many junctions, we looked at the number of junctions as function of their gain in read support, i.e. the number of additional split-reads overlapping them. During transcript quantification (e.g. with Cufflinks[9]), splice junctions with greatly increased support will lead to more reliable abundance estimates.

In addition to the evaluation of the performance of lack itself, we compared it to STAR's second pass approach, described in Dobin *et al.*[3], and to TopHat 2 using the option -j. For each dataset, the initial mapping was done using STAR and TopHat 2 as described above. In case of STAR's second pass, it was necessary to generate a new index for every set of input splice junctions. Alongside the alignment file, STAR generally reported a junction file that was converted to a specific format and provided as option --sjdbFileChrStartEnd during index generation. In addition, the option --sjdbOverhang was set to the maximal read length. For each dataset, a new index was generated and STAR was run with it on the set of reads that were not aligned with STAR during its first pass. For TopHat 2, it was not necessary to build a new index structure. The junction file of TopHat 2's first pass was converted using bed_to_juncs, supplied by TopHat 2. Then, TopHat 2 was simply run on the set of unmapped reads with the option -j set to the converted junction file. The user time and peak memory consumption were tracked across all alignment runs including the index generation of STAR's second pass. For comparison, the number of remapped reads and the alignment accuracy of the remapping methods of STAR and TopHat 2 were determined.

Supplementary Figures

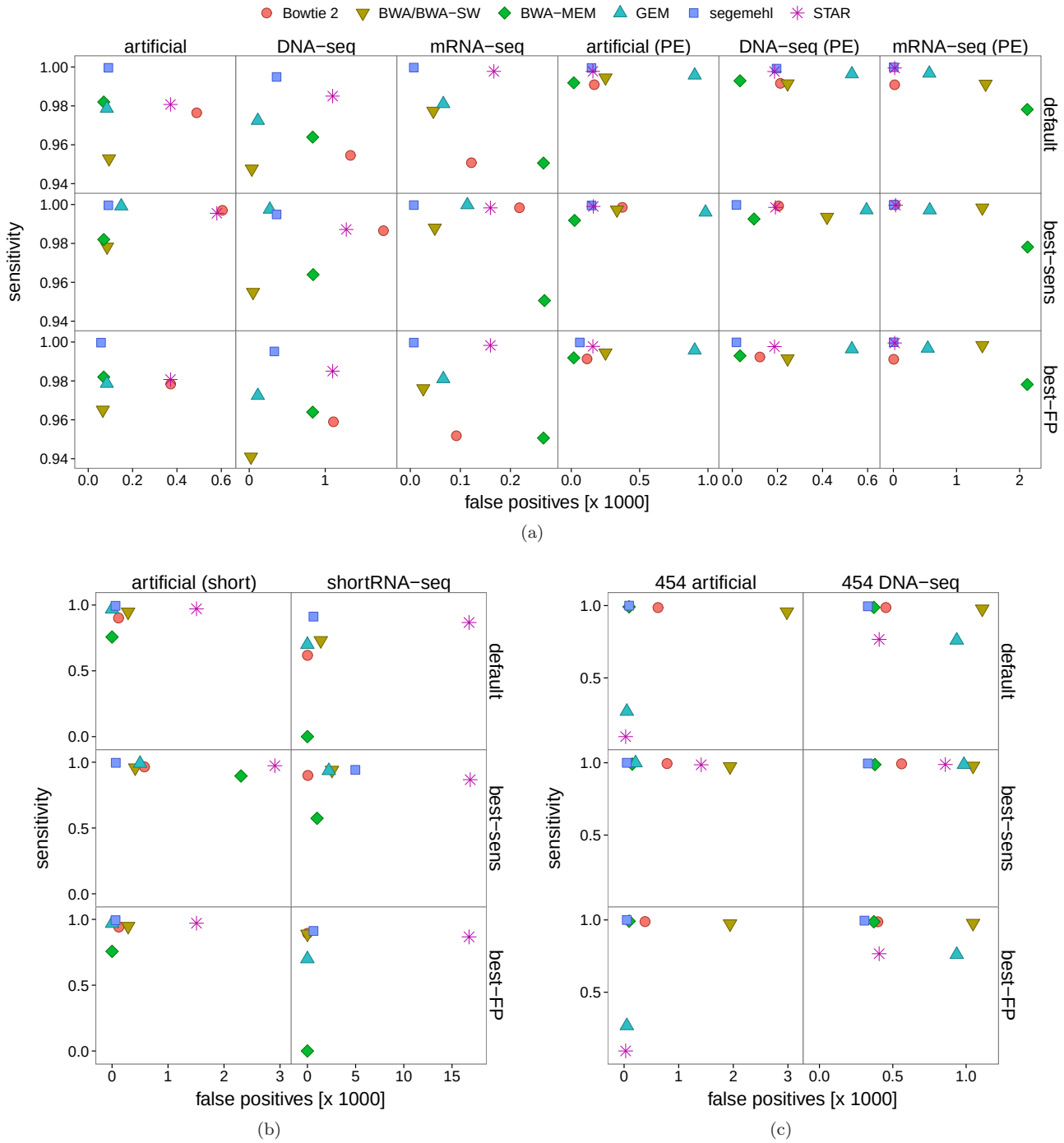


Figure S1: **Comparison of read aligners in the all-best benchmark.** The performance of different read aligners was assessed in terms of sensitivity and number of false positive alignments for (a) Illumina (long), (b) Illumina (short), and (c) 454 datasets. In addition to the default parameters, we evaluated a number of different parameter settings for each aligner to explore the trade off between sensitivity and number of false positive alignments. In such a way, best-sensitivity (best-sens) and best-false positive (best-FP) parameter settings were selected for each read aligner and dataset.

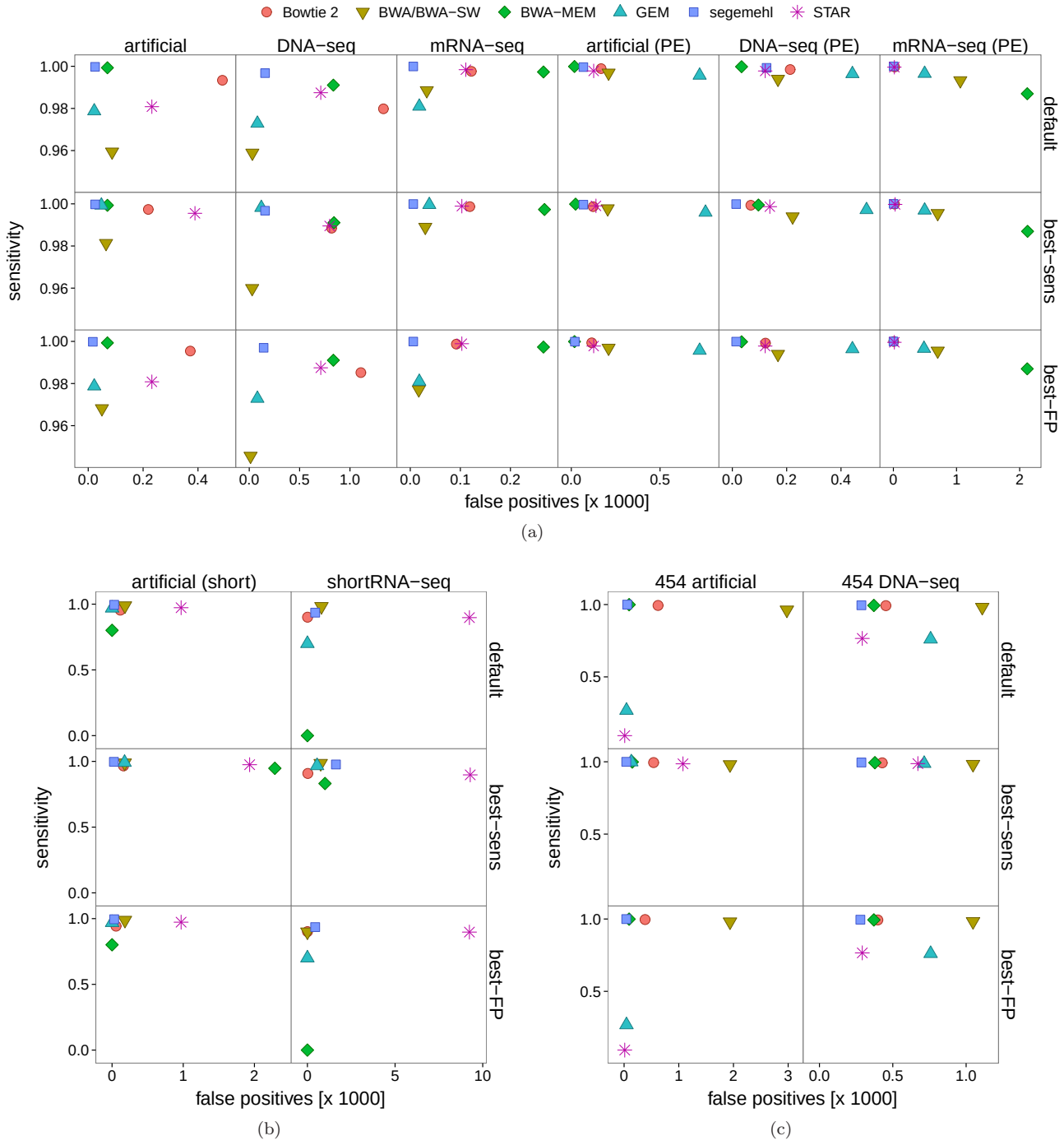


Figure S2: **Comparison of read aligners in the any-best benchmark.** The performance of different read aligners was assessed in terms of sensitivity and number of false positive alignments for (a) Illumina (long), (b) Illumina (short), and (c) 454 datasets. In addition to the default parameters, we evaluated a number of different parameter settings for each aligner to explore the trade off between sensitivity and number of false positive alignments. In such a way, best-sensitivity (best-sens) and best-false positive (best-FP) parameter settings were selected for each read aligner and dataset.

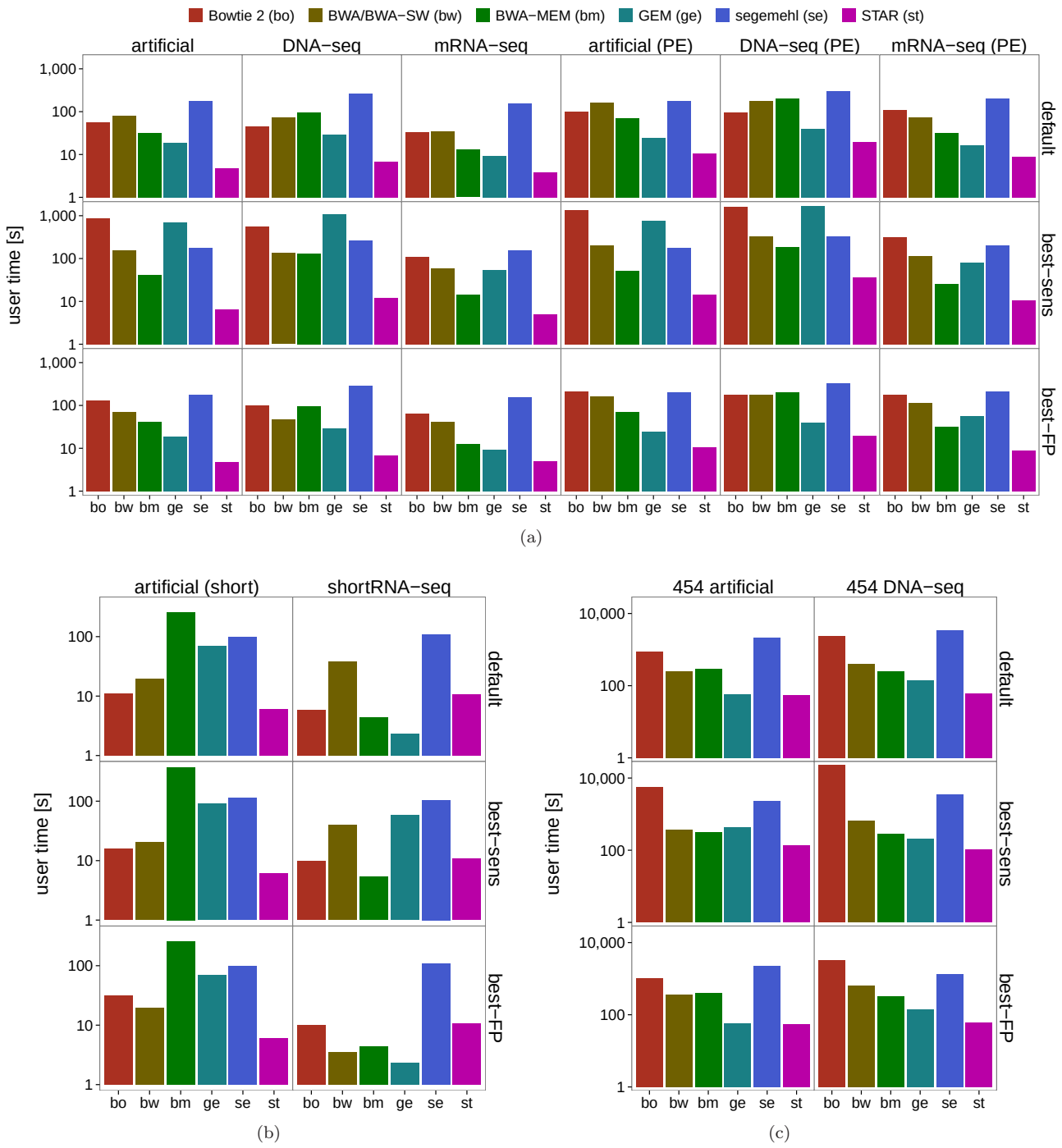
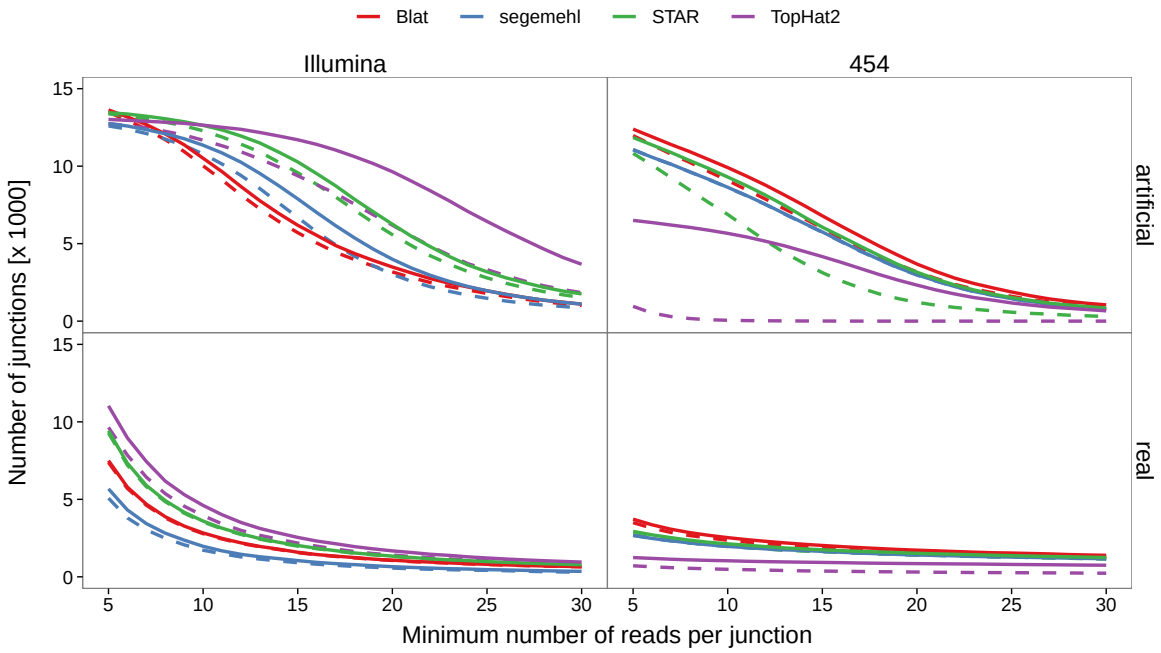
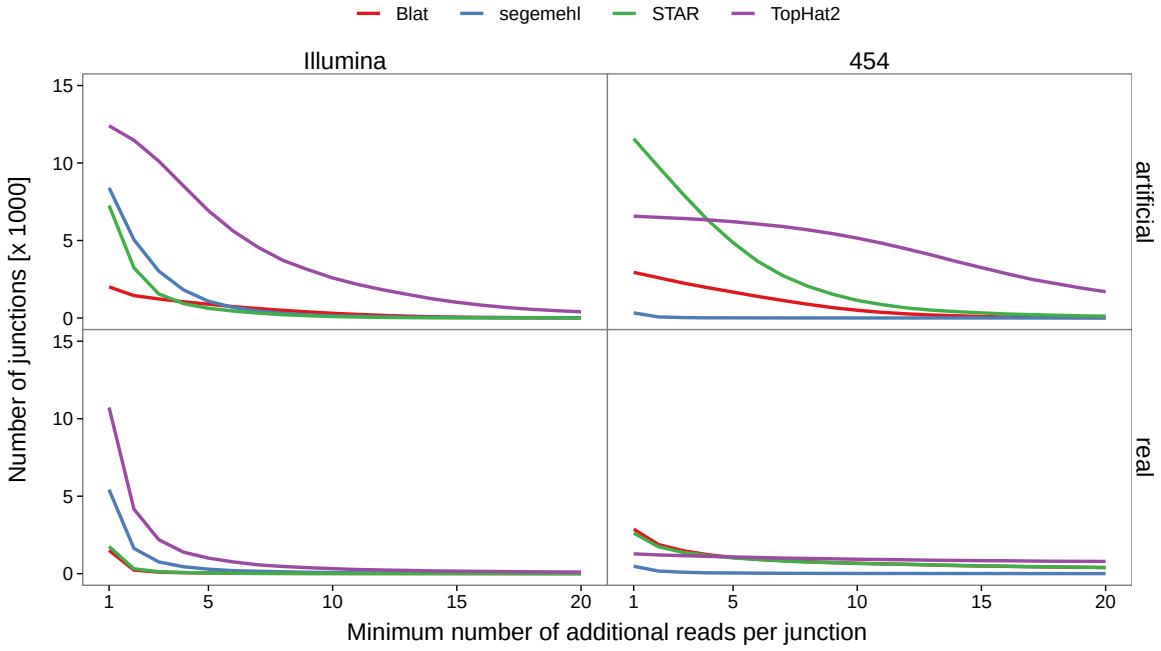


Figure S3: **Runtime comparison of read aligners.** The user time of different read aligners was compared for (a) Illumina (long), (b) Illumina (short), and (c) 454 datasets. In addition to the default parameters, we evaluated a number of different parameter settings for each aligner to explore the trade off between sensitivity and number of false positive alignments. In such a way, best-sensitivity (best-sens) and best-false positive (best-FP) parameter settings were selected for each read aligner and dataset.

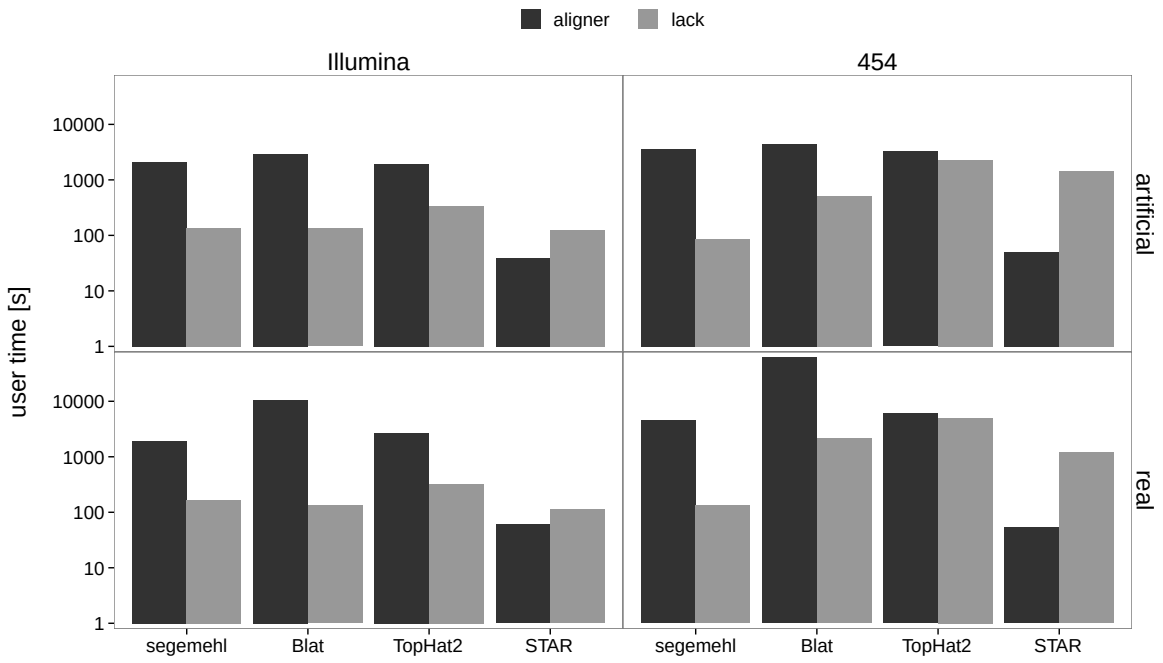


(a)

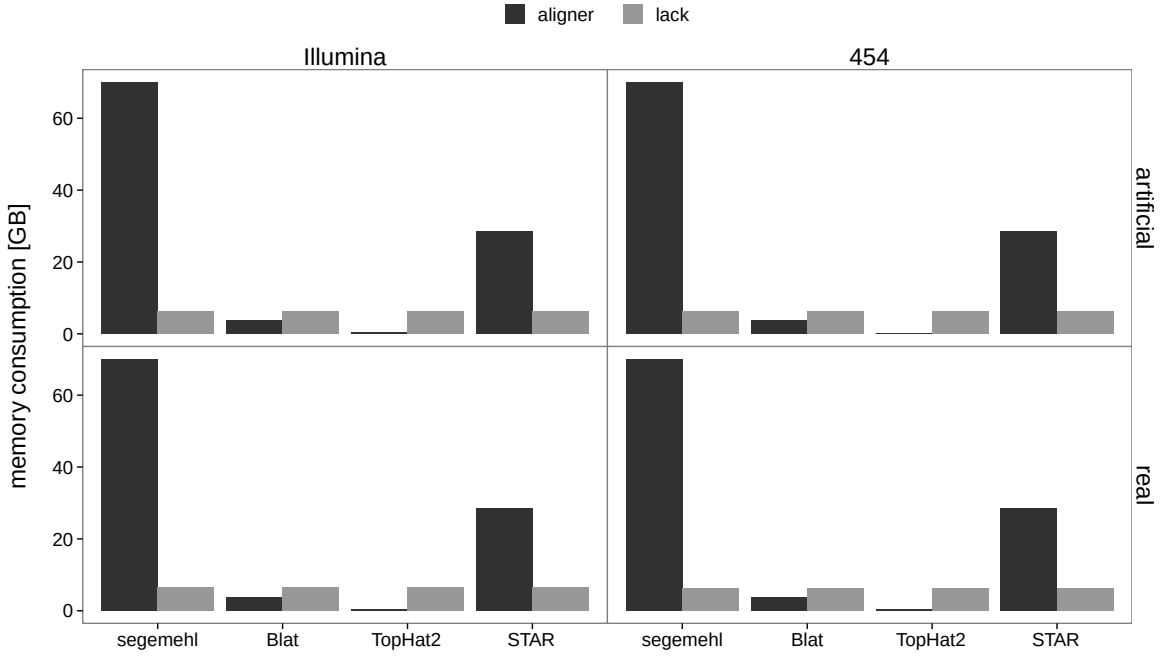


(b)

Figure S4: **Benefit of using `lack` on splice junctions.** (a) Number of potential splice junctions of different split-read aligners with (solid) and without (dashed) `lack`. (b) Number of potential splice junctions as function of minimum additional read support by `lack`. All split-read aligners as well as `lack` were executed with default parameters.

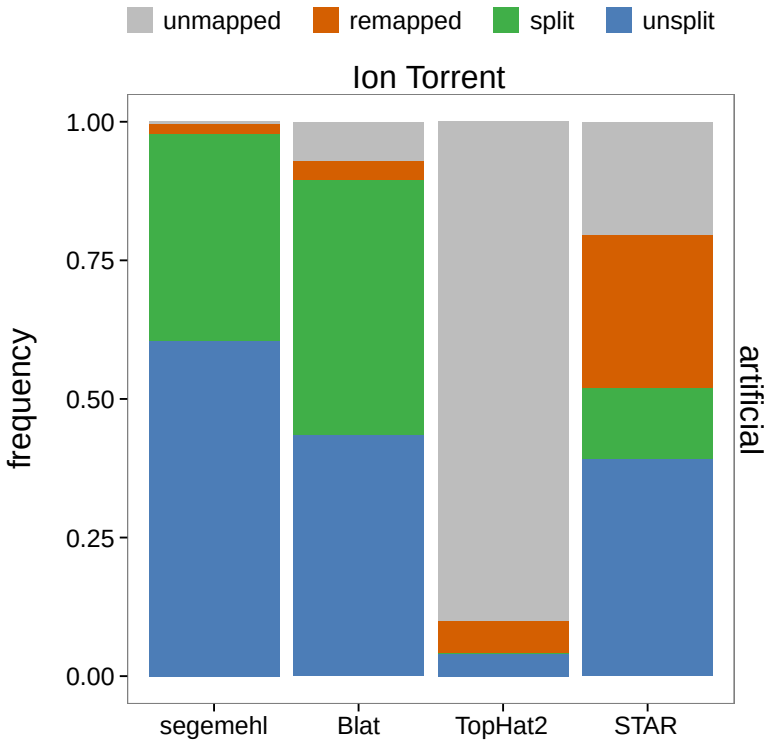


(a)



(b)

Figure S5: **Performance comparison between split-read aligners and `lack`.** Comparison of (a) running times and (b) memory consumptions of mapping with different split-read aligners and `lack`. All split-read aligners as well as `lack` were executed with default parameters.



(a)

Figure S6: **Performance comparison between split-read aligners and `lack` on simulated Ion Torrent data.** Frequency of unsplit-mapped and split-mapped reads of different split-read aligners as well as initially unmapped reads recovered by `lack`. Reads that were not mapped by the aligner and `lack` are termed unmapped. With an average of 44.7%, the remapping rates of `lack` were 82.1% for `segemehl`, 33.2% for `Blat`, 5.9% for `TopHat 2`, and 57.5% for `STAR`. The split-read remapping rates of `lack` were 92.2% for `segemehl`, 94.2% for `Blat`, 10.2% for `TopHat 2`, and 79.8% for `STAR`, i.e., on average 69.1%.

Supplementary Tables

Table S1: **Overview of the datasets used for read aligner comparison.** For each dataset, the accession number (in case of real data), the sequencing platform, library type (SE = single-end or PE = paired-end), the median read length, and the number of reads (in total and used for evaluation). The accession number is not available (n/a) in case of simulated data. Note that library type indicates whether the dataset consists of single- or paired-end reads which does not necessarily match the library type of the original real dataset. For evaluation of the sensitivity and false positive alignments, we only used reads with an error rate of less than or equal to 10% as reported by `RazerS 3`. For these reads, `RazerS 3` guarantees to find all optimal alignments. In addition, reads with more than 10 optimal alignments were disregarded. For more details, please refer to the Methods section.

dataset	accession number	sequencing platform	library type	median read length [nt]	number of reads in total	number of reads for evaluation
Illumina artificial	n/a	Illumina	SE	100	100 000	99 472
Illumina DNA-seq	ERR037900	Illumina	SE	100	100 000	94 114
Illumina mRNA-seq	SRR534289	Illumina	SE	100	100 000	99 517
Illumina artificial (PE)	n/a	Illumina	PE	100	100 000	93 163
Illumina DNA-seq (PE)	ERR037900	Illumina	PE	100	100 000	87 623
Illumina mRNA-seq (PE)	SRR534289	Illumina	PE	100	100 000	81 405
Illumina artificial (short)	n/a	Illumina	SE	30	100 000	90 729
Illumina shortRNA-seq	GSM450598	Illumina	SE	22	100 000	93 962
454 artificial	n/a	454	SE	407	100 000	99 817
454 DNA-seq	SRR003161	454	SE	524	100 000	59 716

Table S2: **Overview of parameter settings of each read aligner.** To determine best-sensitivity and best-false positive settings, a number of different parameter settings were evaluated for each read aligner and dataset to explore the trade off between sensitivity and number of false positive alignments. In case of `BWA`, the groups of parameters were applied to the `aln` and `samse/sampe` command, respectively. The default parameter setting is marked by an asterisk. For better visualization, the parameter names of `STAR` are abbreviated: `--outFilterMismatchNmax` (=A), `--outFilterScoreMinOverLread` (=B), `--outFilterMatchNminOverLread` (=C), `--seedSearchStartLmax` (=D).

read aligner	parameter setting

`Bowtie 2`	`-k 100 --very-fast`
	`-k 10 --very-fast`
	`-k 100 --sensitive`
	`-k 10 --sensitive`
	`-k 10 --very-sensitive`
	`--sensitive`*
	`--very-sensitive`

`Bowtie 2` (local)	`--local --sensitive-local`*
	`--local --very-sensitive-local`
	`--local -k 10 --sensitive-local`
	`--local -k 10 --very-sensitive-local`
	`--local -k 10 --very-fast-local`
	`--local -k 100 --very-fast-local`
	`--local -k 100 --sensitive-local`

`BWA`	`-n 0.04 -l 24 -o 3` and `-n 10`
	`-n 0.04 -l 28 -o 3` and `-n 10`
	`-n 0.04 -l 32 -o 1` and `-n 10`
	`-n 0.04 -l 32 -o 3` and `-n 10`
	`-n 0.04 -l 32 -o 3` and `-n 3`*
	`-n 0.1 -l 32 -o 3` and `-n 10`
	`-n 0.1 -l 28 -o 3` and `-n 10`
	`-n 0.1 -l 24 -o 3` and `-n 10`

`BWA-SW`	`-z 2 -s 1`
	`-z 1 -s 3`*
	`-z 3 -s 1`
	`-z 2 -s 2`
	`-z 2 -s 3`
	`-z 3 -s 2`

`BWA-MEM`	`-T 10 -a -k 19 -r 5`
	`-T 10 -a -k 19 -r 1.5`
	`-T 10 -a -k 21 -r 1.5`
	`-T 10 -a -k 21 -r 5`

Table S2: **Overview of parameter settings of each read aligner.** To determine best-sensitivity and best-false positive settings, a number of different parameter settings were evaluated for each read aligner and dataset to explore the trade off between sensitivity and number of false positive alignments. In case of `BWA`, the groups of parameters were applied to the `aln` and `samse/sampe` command, respectively. The default parameter setting is marked by an asterisk. For better visualization, the parameter names of `STAR` are abbreviated: `--outFilterMismatchNmax` (=A), `--outFilterScoreMinOverLread` (=B), `--outFilterMatchNminOverLread` (=C), `--seedSearchStartLmax` (=D).

read aligner	parameter setting
	`-T 10 -k 19 -r 1.5`
	`-T 10 -k 19 -r 5`
	`-T 10 -k 21 -r 1.5`
	`-T 10 -k 21 -r 5`
	`-T 30 -k 19 -r 1.5`*
`GEM`	`-e 0.1 -m 0.04`
	`-e 0.1 -m 0.08`
	`-e 0.1 -m 0.10`
	`-e 0.04 -m 0.04`*
	`-e 0.04 -m 0.08`
	`-e 0.04 -m 0.10`
`segemehl`	`-D 0 -J 0 -E 5 -M 100`
	`-D 0 -J 0 -E 10 -M 200`
	`-D 0 -J 0 -E 50 -M 500`
	`-D 0 -J 1 -E 5 -M 100`
	`-D 0 -J 1 -E 50 -M 500`
	`-D 1 -J 0 -E 5 -M 100`*
	`-D 1 -J 0 -E 10 -M 200`
	`-D 1 -J 0 -E 50 -M 500`
`STAR`	`-A 10 -B 0.66 -C 0.66 -D 50`*
	`-A 100 -B 0.66 -C 0.66 -D 50`
	`-A 100 -B 0 -C 0.66 -D 50`
	`-A 100 -B 0 -C 0.1 -D 50`
	`-A 100 -B 0 -C 0.1 -D 30`

Table S3: **Comparison of read aligners with default parameters.** The performance of different read aligners was assessed in terms of number of mapped reads/pairs, sensitivity, number of false positive alignments, running time, and peak virtual memory consumption. The sensitivities and false positives have been calculated based on the set of reads with optimal alignments (Supplementary Table S1). For the number of mapped reads, only reads with at least one alignment with $\leq 10\%$ mismatches, indels, and clipped bases reported by the aligner were counted. This step was necessary to ensure comparability of local and semi-global alignments.

				all-best		any-best	
	time [s]	memory [GB]	mapped	sensitivity	FP	sensitivity	FP
Illumina (artificial)							
Bowtie 2	56	3.76	99809	0.976	489	0.993	489
BWA	80	4.19	96037	0.953	94	0.959	87
BWA-MEM	31	5.66	99915	0.982	70	0.999	70
GEM	18	4.68	97863	0.979	84	0.979	22
segemehl	174	70.05	99876	1.000	91	1.000	25
STAR	5	28.12	94272	0.981	371	0.981	232
Illumina (DNA-seq)							
Bowtie 2	45	3.76	93976	0.955	1332	0.980	1332
BWA	72	3.85	91223	0.948	34	0.959	31
BWA-MEM	98	5.68	93253	0.964	835	0.991	835
GEM	30	4.68	92165	0.972	113	0.973	81
segemehl	267	70.05	94954	0.995	360	0.997	157
STAR	7	28.12	92719	0.985	1098	0.987	710
Illumina (mRNA-seq)							
Bowtie 2	33	3.76	99489	0.951	122	0.998	122
BWA	34	3.70	98507	0.977	46	0.989	33
BWA-MEM	13	5.66	98888	0.951	266	0.997	266
GEM	9	4.68	97735	0.981	66	0.981	18
segemehl	158	70.05	99598	1.000	7	1.000	6
STAR	4	28.18	99301	0.998	167	0.998	111
Illumina (paired-end artificial)							
Bowtie 2	102	3.77	99804	0.991	168	0.999	168
BWA	160	3.99	99243	0.994	250	0.997	210
BWA-MEM	71	5.70	99843	0.992	18	1.000	18
GEM	24	4.71	99310	0.996	903	0.996	723
segemehl	179	70.05	99923	1.000	148	1.000	68
STAR	10	28.18	89294	0.998	158	0.998	126
Illumina (paired-end DNA-seq)							
Bowtie 2	95	3.77	90654	0.992	213	0.998	213
BWA	179	3.85	85781	0.991	245	0.994	168
BWA-MEM	204	5.75	85878	0.993	34	1.000	34
GEM	39	4.74	89716	0.996	530	0.996	442
segemehl	308	70.05	91883	0.999	196	0.999	126
STAR	19	28.18	87444	0.998	187	0.998	121
Illumina (paired-end RNA-seq)							
Bowtie 2	110	3.77	99399	0.991	20	1.000	20
BWA	74	3.72	98677	0.991	1461	0.993	1060
BWA-MEM	31	5.70	98533	0.978	2120	0.987	2120
GEM	16	4.72	98357	0.997	569	0.997	495
segemehl	200	70.05	99476	1.000	6	1.000	4
STAR	9	28.25	97810	1.000	21	1.000	16
Illumina (short artificial)							
Bowtie 2	11	3.76	95399	0.902	116	0.955	116
BWA	20	3.53	98603	0.948	287	0.988	181
BWA-MEM	259	6.06	79650	0.757	0	0.801	0
GEM	70	5.39	93849	0.969	1	0.969	1
segemehl	98	70.05	95173	0.994	62	0.995	30
STAR	6	28.12	86654	0.971	1508	0.973	970
Illumina (short RNA-seq)							
Bowtie 2	6	3.76	85310	0.618	8	0.901	8

				all-best		any-best	
	time [s]	memory [GB]	mapped	sensitivity	FP	sensitivity	FP
BWA	38	3.52	95706	0.731	1388	0.983	801
BWA-MEM	4	5.64	0	0.000	0	0.000	0
GEM	2	4.66	66204	0.700	0	0.700	0
segemehl	107	70.05	89248	0.912	630	0.935	442
STAR	11	28.12	83941	0.867	16768	0.897	9241
454 (artificial)							
Bowtie 2	863	3.91	99762	0.986	621	0.994	621
BWA-SW	254	5.71	32279	0.957	2981	0.963	2977
BWA-MEM	299	5.71	98558	0.991	92	0.999	92
GEM	59	4.80	26512	0.268	50	0.268	44
segemehl	2162	70.05	100000	0.999	93	0.999	58
STAR	54	28.24	2545	0.094	29	0.094	12
454 (DNA-seq)							
Bowtie 2	2410	7.68	54828	0.986	453	0.992	453
BWA-SW	391	5.75	24634	0.977	1110	0.981	1110
BWA-MEM	252	5.75	46464	0.987	371	0.994	371
GEM	142	5.12	33322	0.760	936	0.761	758
segemehl	3491	70.05	73275	0.995	330	0.995	286
STAR	61	28.31	20261	0.766	407	0.766	292

Table S4: **Comparison of read aligners with best-sensitivity parameters.** To determine best-sensitivity settings, a number of different parameter settings were evaluated for each read aligner. For each read aligner and dataset, we selected the parameter sets achieving the highest sensitivity. The performance of different read aligners was assessed in terms of number of mapped reads/pairs, sensitivity, number of false positive alignments, running time, and peak virtual memory consumption. The sensitivities and false positives have been calculated based on the set of reads with optimal alignments (Supplementary Table S1). For the number of mapped reads, only reads with at least one alignment with $\leq 10\%$ mismatches, indels, and clipped bases reported by the aligner were counted. This step was necessary to ensure comparability of local and semi-global alignments. The default parameter setting is marked by an asterisk. For better visualization, the parameter names of `STAR` are abbreviated: `--outFilterMismatchNmax` (=A), `--outFilterScoreMinOverLread` (=B), `--outFilterMatchNminOverLread` (=C), `--seedSearchStartLmax` (=D).

					all-best		any-best	
		time [s]	memory [GB]	mapped	sensitivity	FP	sensitivity	FP
Illumina (artificial)								
`Bowtie 2`	`-k 100 --sensitive`	859	3.77	99985	0.997	605	0.997	219
`BWA`	`-n 0.04 -l 24 -o 3 -n 10`	158	5.22	98195	0.978	85	0.981	65
`BWA-MEM`	`-T 10 -a -k 19 -r 1.5`	41	5.68	99915	0.982	70	0.999	70
`GEM`	`-e 0.1 -m 0.10`	685	44.51	99947	0.999	149	0.999	48
`segemehl`	`-D 1 -J 0 -E 5 -M 100*`	174	70.05	99876	1.000	91	1.000	25
`STAR`	`-A 100 -B 0 -C 0.1 -D 30`	7	28.12	94403	0.996	580	0.996	389
Illumina (DNA-seq)								
`Bowtie 2`	`-k 100 --sensitive`	570	3.77	94496	0.987	1766	0.988	818
`BWA`	`-n 0.04 -l 24 -o 3 -n 10`	134	4.72	91320	0.955	50	0.960	29
`BWA-MEM`	`-T 10 -a -k 19 -r 1.5`	132	5.69	93253	0.964	841	0.991	841
`GEM`	`-e 0.1 -m 0.10`	1101	39.74	94739	0.997	271	0.998	120
`segemehl`	`-D 1 -J 0 -E 5 -M 100*`	267	70.05	94954	0.995	360	0.997	157
`STAR`	`-A 100 -B 0 -C 0.1 -D 30`	12	28.06	92799	0.987	1277	0.990	798
Illumina (mRNA-seq)								
`Bowtie 2`	`-k 10 --very-sensitive`	112	3.76	99599	0.998	218	0.999	119
`BWA`	`-n 0.04 -l 24 -o 3 -n 10`	58	3.99	98547	0.988	49	0.989	30
`BWA-MEM`	`-T 10 -a -k 19 -r 1.5`	14	5.66	98888	0.951	268	0.997	268
`GEM`	`-e 0.1 -m 0.10`	54	4.68	99609	1.000	114	1.000	38
`segemehl`	`-D 1 -J 0 -E 5 -M 100*`	158	70.05	99598	1.000	7	1.000	6
`STAR`	`-A 100 -B 0 -C 0.1 -D 30`	5	28.18	99307	0.998	160	0.999	103
Illumina (paired-end artificial)								
`Bowtie 2`	`-k 100 --very-fast`	1349	4.28	99794	0.999	373	0.999	123
`BWA`	`-n 0.1 -l 24 -o 3 -n 10`	203	3.93	99348	0.997	334	0.998	205
`BWA-MEM`	`-T 10 -a -k 19 -r 5`	52	5.70	99838	0.992	24	1.000	24
`GEM`	`-e 0.1 -m 0.10`	767	61.39	99392	0.996	982	0.996	756
`segemehl`	`-D 1 -J 0 -E 5 -M 100*`	179	70.05	99923	1.000	148	1.000	68
`STAR`	`-A 100 -B 0 -C 0.1 -D 30`	14	28.18	89544	0.999	161	0.999	139
Illumina (paired-end DNA-seq)								
`Bowtie 2`	`-k 100 --sensitive`	1590	4.28	90936	0.999	205	0.999	68
`BWA`	`-n 0.04 -l 24 -o 3 -n 10`	331	4.72	85877	0.994	420	0.994	222
`BWA-MEM`	`-T 10 -a -k 19 -r 5`	186	5.76	85795	0.993	96	0.999	96
`GEM`	`-e 0.1 -m 0.10`	1693	55.09	90571	0.997	595	0.997	494
`segemehl`	`-D 1 -J 0 -E 50 -M 500`	324	70.05	91901	1.000	18	1.000	14
`STAR`	`-A 100 -B 0 -C 0.1 -D 30`	37	28.18	88176	0.999	191	0.999	138
Illumina (paired-end RNA-seq)								
`Bowtie 2`	`-k 10 --very-sensitive`	316	3.77	99452	1.000	30	1.000	18
`BWA`	`-n 0.04 -l 24 -o 3 -n 10`	116	4.00	98709	0.998	1409	0.995	702
`BWA-MEM`	`-T 10 -a -k 19 -r 5`	25	5.70	98477	0.978	2125	0.987	2125
`GEM`	`-e 0.1 -m 0.10`	81	4.77	98846	0.997	576	0.997	495
`segemehl`	`-D 1 -J 0 -E 5 -M 100*`	200	70.05	99476	1.000	6	1.000	4
`STAR`	`-A 100 -B 0 -C 0.1 -D 30`	10	28.25	98309	1.000	35	1.000	28
Illumina (short artificial)								
`Bowtie 2`	`-k 10 --very-sensitive`	16	3.76	96476	0.964	579	0.966	161
`BWA`	`-n 0.1 -l 24 -o 3 -n 10`	20	3.53	98670	0.958	414	0.989	177
`BWA-MEM`	`-T 10 -a -k 19 -r 5`	369	7.12	93599	0.895	2303	0.947	2287
`GEM`	`-e 0.1 -m 0.10`	90	5.40	96209	0.991	500	0.992	176
`segemehl`	`-D 1 -J 0 -E 50 -M 500`	112	70.05	96603	0.995	67	0.997	26
`STAR`	`-A 100 -B 0 -C 0.1 -D 30`	6	28.12	86654	0.973	2908	0.975	1933

		time [s]	memory [GB]	mapped	all-best sensitivity	all-best FP	any-best sensitivity	any-best FP
Illumina (short RNA-seq)								
Bowtie 2	`-k 10 --very-sensitive`	10	3.76	86042	0.900	43	0.908	23
BWA	`-n 0.04 -l 24 -o 3 -n 10`	40	3.52	95897	0.940	2547	0.985	757
BWA-MEM	`-T 10 -a -k 19 -r 5`	5	5.64	78294	0.574	1002	0.832	1002
GEM	`-e 0.1 -m 0.10`	58	4.66	94127	0.935	2220	0.967	552
segemehl	`-D 1 -J 0 -E 50 -M 500`	105	70.05	94337	0.942	4971	0.977	1627
STAR	`-A 100 -B 0 -C 0.1 -D 30`	11	28.12	83941	0.867	16888	0.897	9287
454 (artificial)								
Bowtie 2	`--local -k 100 --sensitive-local`	5520	3.93	99772	0.994	786	0.995	540
BWA-SW	`-z 2 -s 3`	355	5.71	53040	0.974	1940	0.980	1937
BWA-MEM	`-T 10 -a -k 19 -r 5`	304	5.72	98540	0.991	150	0.998	150
GEM	`-e 0.1 -m 0.08`	422	4.85	99815	0.999	210	0.999	132
segemehl	`-D 1 -J 0 -E 50 -M 500`	2282	70.05	100000	1.000	50	1.000	40
STAR	`-A 100 -B 0 -C 0.1 -D 30`	136	28.25	15550	0.986	1412	0.986	1077
454 (DNA-seq)								
Bowtie 2	`--local -k 100 --sensitive-local`	22821	7.75	54849	0.993	560	0.993	428
BWA-SW	`-z 2 -s 3`	634	6.15	29344	0.977	1047	0.982	1047
BWA-MEM	`-T 10 -a -k 19 -r 5`	277	6.61	46462	0.987	378	0.994	378
GEM	`-e 0.1 -m 0.04`	207	5.12	67611	0.987	985	0.988	714
segemehl	`-D 1 -J 0 -E 5 -M 100*`	3491	70.05	73275	0.995	330	0.995	286
STAR	`-A 100 -B 0 -C 0.1 -D 50`	101	28.37	25347	0.988	858	0.987	671

Table S5: **Comparison of read aligners with best-false positive parameters.** To determine best-false positive settings, a number of different parameter settings were evaluated for each read aligner. For each read aligner and dataset, we selected the parameter sets achieving the lowest number of false positive alignments. The performance of different read aligners was assessed in terms of number of mapped reads/pairs, sensitivity, number of false positive alignments, running time, and peak virtual memory consumption. The sensitivities and false positives have been calculated based on the set of reads with optimal alignments (Supplementary Table S1). For the number of mapped reads, only reads with at least one alignment with $\leq 10\%$ mismatches, indels, and clipped bases reported by the aligner were counted. This step was necessary to ensure comparability of local and semi-global alignments. The default parameter setting is marked by an asterisk. For better visualization, the parameter names of `STAR` are abbreviated: `--outFilterMismatchNmax` (=A), `--outFilterScoreMinOverLread` (=B), `--outFilterMatchNminOverLread` (=C), `--seedSearchStartLmax` (=D).

					allbest		anybest	
		time [s]	memory [GB]	mapped	sensitivity	FP	sensitivity	FP
Illumina (artificial)								
Bowtie 2	`--very-sensitive`	133	3.76	99902	0.978	372	0.995	372
BWA	`-n 0.1 -l 32 -o 3 -n 10`	69	3.88	96871	0.965	66	0.968	50
BWA-MEM	`-T 10 -a -k 19 -r 1.5`	41	5.68	99915	0.982	70	0.999	70
GEM	`-e 0.04 -m 0.04`*	18	4.68	97863	0.979	84	0.979	22
segemehl	`-D 1 -J 0 -E 50 -M 500`	181	70.05	99962	1.000	58	1.000	17
STAR	`-A 10 -B 0.66 -C 0.66 -D 50`*	5	28.12	94272	0.981	371	0.981	232
Illumina (DNA-seq)								
Bowtie 2	`--very-sensitive`	102	3.76	94305	0.959	1108	0.985	1108
BWA	`-n 0.1 -l 28 -o 3 -n 10`	48	3.75	89933	0.941	23	0.946	13
BWA-MEM	`-T 30 -k 19 -r 1.5`*	98	5.68	93253	0.964	835	0.991	835
GEM	`-e 0.04 -m 0.04`*	30	4.68	92165	0.972	113	0.973	81
segemehl	`-D 1 -J 0 -E 50 -M 500`	285	70.05	94984	0.995	330	0.997	143
STAR	`-A 10 -B 0.66 -C 0.66 -D 50`*	7	28.12	92719	0.985	1098	0.987	710
Illumina (mRNA-seq)								
Bowtie 2	`--very-sensitive`	63	3.76	99572	0.952	92	0.999	92
BWA	`-n 0.1 -l 24 -o 3 -n 10`	42	3.70	97351	0.976	26	0.977	17
BWA-MEM	`-T 30 -k 19 -r 1.5`*	13	5.66	98888	0.951	266	0.997	266
GEM	`-e 0.04 -m 0.04`*	9	4.68	97735	0.981	66	0.981	18
segemehl	`-D 1 -J 0 -E 5 -M 100`*	158	70.05	99598	1.000	7	1.000	6
STAR	`-A 100 -B 0 -C 0.1 -D 30`	5	28.18	99307	0.998	160	0.999	103
Illumina (paired-end artificial)								
Bowtie 2	`--very-sensitive`	217	3.77	99844	0.991	115	0.999	115
BWA	`-n 0.04 -l 32 -o 1 -n 3`*	160	3.99	99243	0.994	250	0.997	210
BWA-MEM	`-T 10 -a -k 21 -r 1.5`	69	5.70	99843	0.992	18	1.000	18
GEM	`-e 0.04 -m 0.04`*	24	4.71	99310	0.996	903	0.996	723
segemehl	`-D 1 -J 0 -E 50 -M 500`	202	70.05	99945	1.000	62	1.000	20
STAR	`-A 10 -B 0.66 -C 0.66 -D 50`*	10	28.18	89294	0.998	158	0.998	126
Illumina (paired-end DNA-seq)								
Bowtie 2	`--very-sensitive`	180	3.77	90748	0.992	122	0.999	122
BWA	`-n 0.04 -l 32 -o 1 -n 3`*	179	3.85	85781	0.991	245	0.994	168
BWA-MEM	`-T 30 -k 19 -r 1.5`*	204	5.75	85878	0.993	34	1.000	34
GEM	`-e 0.04 -m 0.04`*	39	4.74	89716	0.996	530	0.996	442
segemehl	`-D 1 -J 0 -E 50 -M 500`	324	70.05	91901	1.000	18	1.000	14
STAR	`-A 10 -B 0.66 -C 0.66 -D 50`*	19	28.18	87444	0.998	187	0.998	121
Illumina (paired-end RNA-seq)								
Bowtie 2	`--very-sensitive`	177	3.77	99446	0.991	5	1.000	5
BWA	`-n 0.04 -l 24 -o 3 -n 10`	116	4.00	98709	0.998	1409	0.995	702
BWA-MEM	`-T 30 -k 19 -r 1.5`*	31	5.70	98533	0.978	2120	0.987	2120
GEM	`-e 0.04 -m 0.08`	56	4.75	98327	0.997	545	0.997	485
segemehl	`-D 1 -J 0 -E 50 -M 500`	215	70.05	99482	1.000	4	1.000	2
STAR	`-A 10 -B 0.66 -C 0.66 -D 50`*	9	28.25	97810	1.000	21	1.000	16
Illumina (short artificial)								
Bowtie 2	`-k 100 --very-fast`	31	3.77	94232	0.942	121	0.943	56
BWA	`-n 0.04 -l 32 -o 1 -n 3`*	20	3.53	98603	0.948	287	0.988	181
BWA-MEM	`-T 30 -k 19 -r 1.5`*	259	6.06	79650	0.757	0	0.801	0
GEM	`-e 0.04 -m 0.04`*	70	5.39	93849	0.969	1	0.969	1
segemehl	`-D 1 -J 0 -E 5 -M 100`*	98	70.05	95173	0.994	62	0.995	30

		time [s]	memory [GB]	mapped	allbest sensitivity	allbest FP	anybest sensitivity	anybest FP
STAR	-A 10 -B 0.66 -C 0.66 -D 50*	6	28.12	86654	0.971	1508	0.973	970
Illumina (short RNA-seq)								
Bowtie 2	-k 100 --very-fast	10	3.76	85310	0.894	8	0.901	3
BWA	-n 0.1 -l 24 -o 3 -n 10	4	2.89	85141	0.889	0	0.899	0
BWA-MEM	-T 30 -k 19 -r 1.5*	4	5.64	0	0.000	0	0.000	0
GEM	-e 0.04 -m 0.04*	2	4.66	66204	0.700	0	0.700	0
segemehl	-D 1 -J 0 -E 5 -M 100*	107	70.05	89248	0.912	630	0.935	442
STAR	-A 10 -B 0.66 -C 0.66 -D 50*	11	28.12	83941	0.867	16768	0.897	9241
454 (artificial)								
Bowtie 2	--local --very-sensitive-local	1050	3.91	99857	0.988	384	0.996	384
BWA-SW	-z 2 -s 3	355	5.71	53040	0.974	1940	0.980	1937
BWA-MEM	-T 10 -a -k 19 -r 1.5	406	5.73	98558	0.991	92	0.999	92
GEM	-e 0.04 -m 0.04*	59	4.80	26512	0.268	50	0.268	44
segemehl	-D 1 -J 0 -E 50 -M 500	2282	70.05	100000	1.000	50	1.000	40
STAR	-A 10 -B 0.66 -C 0.66 -D 50*	54	28.24	2545	0.094	29	0.094	12
454 (DNA-seq)								
Bowtie 2	--local --very-sensitive-local	3285	7.78	54839	0.987	398	0.993	398
BWA-SW	-z 2 -s 3	634	6.15	29344	0.977	1047	0.982	1047
BWA-MEM	-T 10 -a -k 19 -r 1.5	322	6.93	46464	0.987	371	0.994	371
GEM	-e 0.04 -m 0.04*	142	5.12	33322	0.760	936	0.761	758
segemehl	-D 0 -J 1 -E 50 -M 500	1349	70.05	73278	0.995	307	0.995	279
STAR	-A 10 -B 0.66 -C 0.66 -D 50*	61	28.31	20261	0.766	407	0.766	292

Table S6: **Remapping rates of `lack`.** The measure expresses the fraction of reads that were missed by the aligner but were rescued by `lack`. It was calculated for each dataset and split-read aligner. Averages of remapping rates per dataset (column-wise) and per split-read aligner (row-wise) are given in the last row and column, respectively. All split-read aligners as well as `lack` were executed with default parameters.

	Illumina		454		
	artificial	real	artificial	real	avg.
`segemehl`	0.954	0.249	0.775	0.390	0.592
`Blat`	0.313	0.072	0.632	0.712	0.433
`TopHat 2`	0.486	0.184	0.422	0.527	0.405
`STAR`	0.817	0.084	0.836	0.705	0.610
avg.	0.642	0.147	0.666	0.583	0.510

Table S7: **Split-read remapping rates of `lack`.** The measure expresses the fraction of split-mapped reads missed by the aligner but rescued by `lack`. To decide whether an unmapped read was split-mapped, we used the total amount of split-read alignments reported by any aligner or by `lack`. An unmapped read was deemed split-mapped if a split-read alignment of it was reported by any aligner or by `lack`. Averages of split-read remapping rates per dataset (column-wise) and per split-read aligner (row-wise) are given in the last row and column, respectively. All split-read aligners as well as `lack` were executed with default parameters.

	Illumina		454		
	artificial	real	artificial	real	avg.
`segemehl`	0.965	0.652	0.814	0.472	0.726
`Blat`	0.916	0.443	0.941	0.763	0.766
`TopHat 2`	0.834	0.471	0.585	0.556	0.612
`STAR`	0.924	0.307	0.935	0.734	0.725
avg.	0.910	0.468	0.819	0.631	0.707

Table S8: **Overview of alignment accuracy.** The accuracy of the alignments of split-mapped reads of different split-read aligners as well as alignment of remapped reads by `lack`. The measure was normalized by reads and was only available in case of artificial datasets where the simulated origin was known. All split-read aligners as well as `lack` were executed with default parameters.

	accuracy of	
	split-mapped	remapped
Illumina (artificial)		
`segemehl`	0.887	0.893
`Blat`	0.855	0.890
`TopHat 2`	0.714	0.879
`STAR`	0.899	0.898
454 (artificial)		
`segemehl`	0.866	0.792
`Blat`	0.833	0.877
`TopHat 2`	0.692	0.847
`STAR`	0.862	0.885

Table S9: **Comparison between `lack` and the remapping with `STAR`.** The performance of `lack` and remapping with `STAR` was assessed in terms of running time, memory consumption, and number of remapped reads. The datasets were initially mapped with `STAR`. `STAR`'s remapping method required a new index for every set of input splice junctions. The running time of the index construction is given in parenthesis and the maximal memory consumption of index construction and remapping is listed. `STAR` as well as `lack` were executed with default parameters.

	time [s]	memory [GB]	remapped reads
Illumina (artificial)			
`lack`	123	6.39	14432
`STAR`	2 (+ 8385)	32.13	11713
Illumina (RNA-seq)			
`lack`	117	6.42	2660
`STAR`	6 (+ 9352)	33.39	2464
454 (artificial)			
`lack`	1459	6.38	37736
`STAR`	29 (+12204)	33.39	23078
454 (HUVEC)			
`lack`	1189	6.37	20238
`STAR`	27 (+ 8932)	32.59	15897

Table S10: **Comparison between `lack` and remapping with `TopHat 2`.** The performance of `lack` and remapping with `TopHat 2` was assessed in terms of running time, memory consumption, and number of remapped reads. The datasets were initially mapped with `TopHat 2`. `TopHat 2` as well as `lack` were executed with default parameters.

	time [s]	memory [GB]	remapped reads
Illumina (artificial)			
`lack`	343	6.39	83199
`TopHat 2`	1075	0.32	283
Illumina (RNA-seq)			
`lack`	325	6.42	35536
`TopHat 2`	1367	0.33	350
454 (artificial)			
`lack`	2293	6.37	92733
`TopHat 2`	3908	0.27	41
454 (HUVEC)			
`lack`	4963	6.36	113593
`TopHat 2`	5554	0.31	40

References

[1] M Holtgrewe. Mason – a read simulator for second generation sequencing data. Technical Report TR-B-10-06, Department of Mathematics and Computer Science, FU Berlin, 2010.

[2] B Langmead and SL Salzberg. Fast gapped-read alignment with Bowtie 2. *Nat Methods*, 9(4):357–9, Mar 2012.

[3] A Dobin, CA Davis, F Schlesinger, J Drenkow, C Zaleski, S Jha, P Batut, M Chaisson, and TR Gingeras. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics*, 29:15–21, Jan 2013.

[4] D Kim, G Pertea, C Trapnell, H Pimentel, R Kelley, and SL Salzberg. TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biol*, 14:R36, Apr 2013.

[5] S Huang, J Zhang, R Li, W Zhang, Z He, TW Lam, Z Peng, and SM Yiu. SOAPsplice: Genome-Wide ab initio detection of splice junctions from RNA-Seq data. *Front Genet*, 2:46, 2011.

[6] G Koscielny, V Le Texier, C Gopalakrishnan, V Kumanduri, JJ Riethoven, F Nardone, E Stanley, C Fallsehr, O Hofmann, M Kull, E Harrington, S Boué, E Eyras, M Plass, F Lopez, W Ritchie, V Moucadel, T Ara, H Pospisil, A Herrmann, J G Reich, R Guigó, P Bork, M Doeberitz, J Vilo, W Hide, R Apweiler, TA Thanaraj, and D Gautheret. ASTD: The alternative splicing and transcript diversity database. *Genomics*, 93:213–20, Mar 2009.

[7] LM Bragg, G Stone, MK Butler, P Hugenholtz, and GW Tyson. Shining a light on dark sequencing: characterising errors in ion torrent PGM data. *PLoS Comput Biol*, 9:e1003031, Apr 2013.

[8] WJ Kent. BLAT–the BLAST-like alignment tool. *Genome Res*, 12:656–64, Apr 2002.

[9] C Trapnell, BA Williams, G Pertea, A Mortazavi, G Kwan, MJ van Baren, SL Salzberg, BJ Wold, and L Pachter. Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat Biotechnol*, 28:511–5, May 2010.