SUPPLEMENTARY MATERIAL

lordFAST: sensitive and Fast Alignment Search Tool for LOng noisy Read sequencing Data

1 Data

1.1 Real data

Long reads for a human genome (CHM1) sequenced by PacBio RS II instrument using P5-C3 chemistry are available at Pacific Biosciences's Devnet repository:

`https://github.com/PacificBiosciences/DevNet/wiki/H_sapiens_54x_release`

We used long reads stored in a single fasta file (corresponding to one of the three files generated for a SMRT cell) for the real study which can be downloaded from:

`http://datasets.pacb.com/2013/Human10x/READS/2530572/0001/Analysis_Results/`
`m130929_024849_42213_c100518541910000001823079209281311_s1_p0.1.subreads.fasta`

After obtaining the fasta file, we filtered out reads shorter than 1000 base-pair so that we can focus more on the task of mapping longer reads which is the goal for long read mappers. This is motivated by the fact that more than 99% of the data is in reads longer than 1000 base-pair (Figure S1).

1.2 Synthetic data

We used PBSIM in order to generate synthetic data for the simulation study. PBSIM is a PacBio simulator that is capable of simulating long reads from a set of real reads. It uses real reads to infer the read length and error distribution. In addition to the generated long reads, PBSIM reports the true alignment between the generated long reads and the reference genome in MAF format. This enables us to evaluate tools based on their base-pair sensitivity and precision.

Here we explain the detailed instruction and commands for generating the synthetic data. First, the real fastq files used for generating simulated reads is obtained as the following:

```
mkdir simulated
cd simulated
wget http://datasets.pacb.com/2013/Human10x/READS/2530572/0001/Analysis_Results/
    m130929_024849_42213_c100518541910000001823079209281311_s1_p0.1.subreads.fastq
wget http://datasets.pacb.com/2013/Human10x/READS/2530572/0001/Analysis_Results/
    m130929_024849_42213_c100518541910000001823079209281311_s1_p0.2.subreads.fastq
wget http://datasets.pacb.com/2013/Human10x/READS/2530572/0001/Analysis_Results/
    m130929_024849_42213_c100518541910000001823079209281311_s1_p0.3.subreads.fastq
cat m130929_024849_42213_c100518541910000001823079209281311_s1_p0.1.subreads.fastq
    m130929_024849_42213_c100518541910000001823079209281311_s1_p0.2.subreads.fastq
    m130929_024849_42213_c100518541910000001823079209281311_s1_p0.3.subreads.fastq > real.fastq
```

Then we run PBSIM to generate the simulated reads:

```
pbsim --data-type CLR --depth 1 --length-min 1 --length-max 100000 --seed 0 --sample-fastq real.fastq
     hg38.fa
```

Then 25000 reads with minimum length of 1000 are sampled from the simulated reads as the synthetic dataset. The codes for sampling reads from the simulated reads is available at:

`https://github.com/vpc-ccg/lordfast-extra`

2 Analysis of the real data

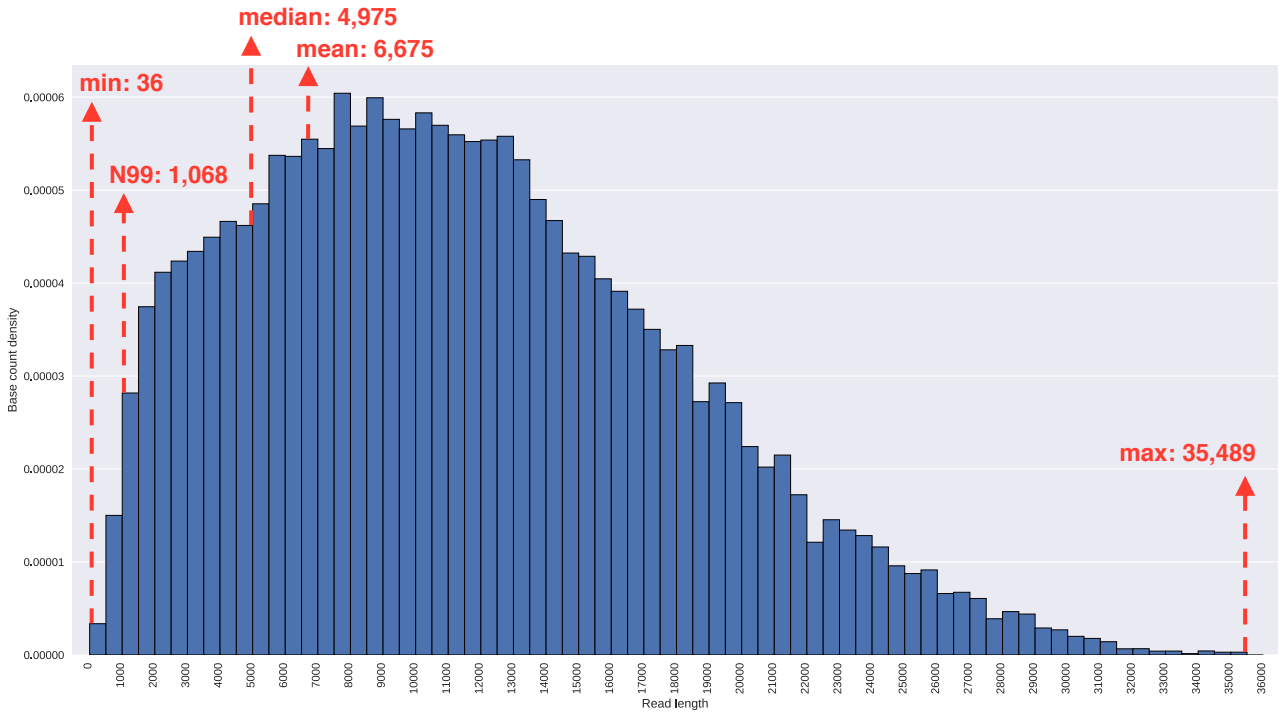

Fig. S1. The read length distribution of 72,708 real PacBio reads from a human genome (CHM1) dataset. The Y axis shows the number of bases in each bin rather than the number of reads. At least 99% of the bases are in the reads longer than 1000 bases.

3 Comparison between the standard FM index and our combined index for exact match searching

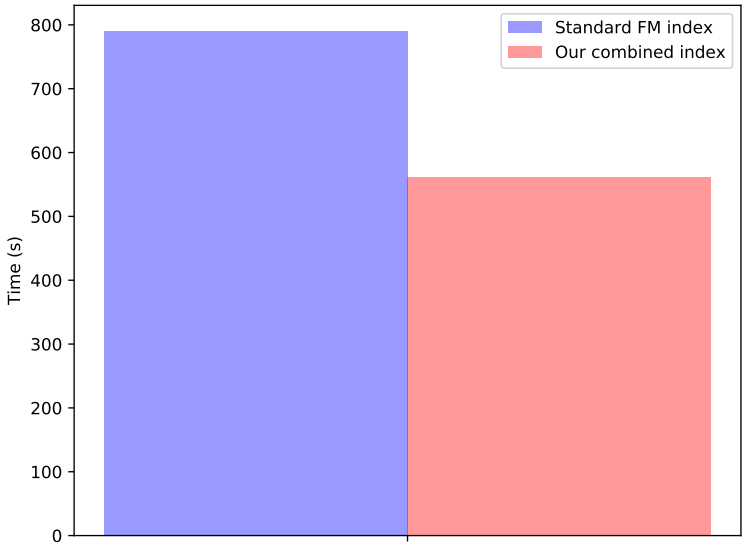

Fig. S2. The speed up when using our combined index for searching exact matches compared to the standard FM index. That is 29% speed up for finding all anchors in the first step.

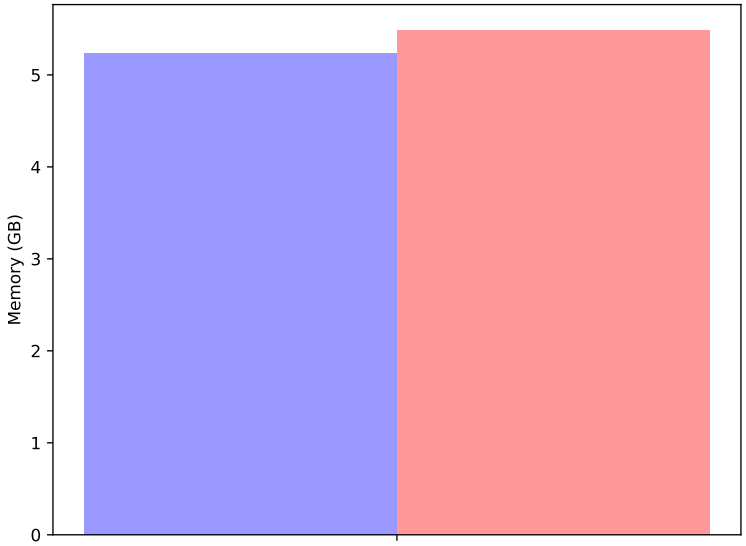

Fig. S3. The combined index uses only 0.25 GB more memory compared to the standard FM index.

4 Analysis of the results for the simulated dataset

Table S1. Comparison between different tools capable of mapping PacBio long reads on the simulated human dataset. This dataset contains 25,000 reads and 183.61 million bases. Best results are marked with bold typeface.

Minimum overlap (p)	Mapper	Correctly mapped	Correct bases (Mb)	Incorrect bases (Mb)	Unmapped bases (Kb)	Sensitivity[a] (%)	Precision[b] (%)
1 bp	BLASR	24,642	171.74	11.17	698.22	93.53	93.89
	BWA-MEM	24,603	171.76	11.36	525.11	93.53	93.80
	GraphMap	24,161	177.33	3.98	2,297.27	96.58	97.81
	LAMSA	24,458	177.65	5.75	282.15	96.72	96.87
	rHAT	24,409	177.87	5.35	391.52	96.87	97.08
	NGMLR	24,194	172.83	6.53	4,246.51	94.13	96.36
	Minimap2	24,745	181.56	**1.84**	223.46	98.88	99.00
	minialign	24,567	179.68	3.31	621.60	97.86	98.19
	lordFAST	**24,751**	**181.74**	**1.84**	**29.35**	**98.98**	**99.00**
90%	BLASR	24,563	171.66	11.27	675.95	93.50	93.84
	BWA-MEM	24,485	171.37	11.85	417.84	93.32	93.53
	GraphMap	24,161	177.33	3.98	2,297.27	96.58	97.81
	LAMSA	24,371	177.52	5.94	208.22	96.65	96.76
	rHAT	24,372	177.82	5.71	80.98	96.85	96.89
	NGMLR	23,769	171.99	8.11	3,508.56	93.67	95.50
	Minimap2	24,740	181.53	**1.85**	223.20	98.87	98.99
	minialign	24,469	179,27	4.11	233.74	97.64	97.76
	lordFAST	**24,747**	**181.73**	**1.85**	**29.10**	**98.98**	**98.99**

A read is considered to be mapped correctly if its aligned subsequence in the reference overlaps with the "correct" mapping subsequence by at least p bases. On the other hand, **a base in a read is considered to be correctly mapped if the read is correctly mapped and the mapping location of the base is within a 50 bp vicinity of the correct alignment locus of the base.** [a] The sensitivity is defined as the number of correctly mapped bases / the total number of bases. [b] The precision is defined as the number of correctly mapped bases / the number of mapped bases.

Table S2. Comparison between different tools capable of mapping PacBio long reads on the simulated human dataset. This dataset contains 25,000 reads and 183.61 million bases. Best results are marked with bold typeface.

Minimum overlap (p)	Mapper	Correctly mapped	Correct bases (Mb)	Incorrect bases (Mb)	Unmapped bases (Mb)	Sensitivity[a] (%)	Precision[b] (%)
1 bp	BLASR	24,642	136.73	46.18	698.22	74.47	74.75
	BWA-MEM	24,603	164.35	18.78	525.11	89.49	89.75
	GraphMap	24,161	175.48	5.83	2,297.27	95.57	96.79
	LAMSA	24,458	163.97	19.42	282.15	89.27	89.41
	rHAT	24,409	172.63	10.59	391.52	94.02	94.22
	NGMLR	24,194	151.75	27.61	4,246.51	82.65	84.60
	Minimap2	24,745	160.90	22.50	223.46	87.62	87.73
	minialign	24,567	159.37	23.62	621.60	86.80	87.09
	lordFAST	**24,751**	**180.91**	**2.67**	**29.35**	**98.53**	**98.54**
90%	BLASR	24,563	136.67	46.26	675.95	74.43	74.71
	BWA-MEM	24,485	163.95	19.26	417.84	89.28	89.49
	GraphMap	24,161	175.48	5.83	2,297.27	95.57	96.79
	LAMSA	24,371	163.84	19.62	208.22	89.21	89.31
	rHAT	24,372	172.60	10.93	80.98	94.00	94.04
	NGMLR	23,769	150.91	29.19	3,508.56	82.19	83.79
	Minimap2	24,740	160.87	22.52	223.20	87.61	87.72
	minialign	24,469	158.95	24.42	233.74	86.57	86.68
	lordFAST	**24,747**	**180.91**	**2.67**	**29.10**	**98.53**	**98.54**

A read is considered to be mapped correctly if its aligned subsequence in the reference overlaps with the "correct" mapping subsequence by at least p bases. On the other hand, **a base in a read is considered to be correctly mapped if the read is correctly mapped and the mapping location of the base is within a 5 bp vicinity of the correct alignment locus of the base**. [a] The sensitivity is defined as the number of correctly mapped bases / the total number of bases. [b] The precision is defined as the number of correctly mapped bases / the number of mapped bases.

5 Benchmarking of different methods using multiple threads

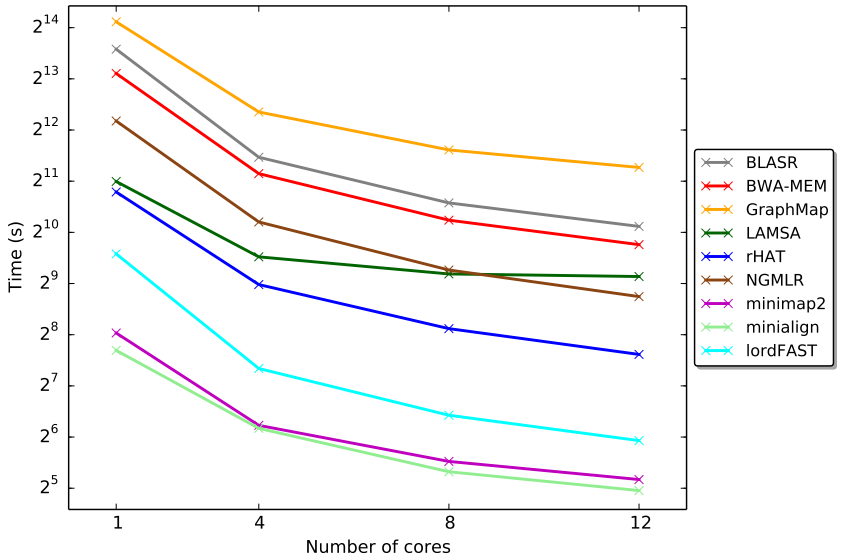

Fig. S4. Run-time comparison of different methods for mapping 23,155 real human reads using different threads. Note that the y-axis is in logarithmic scale.

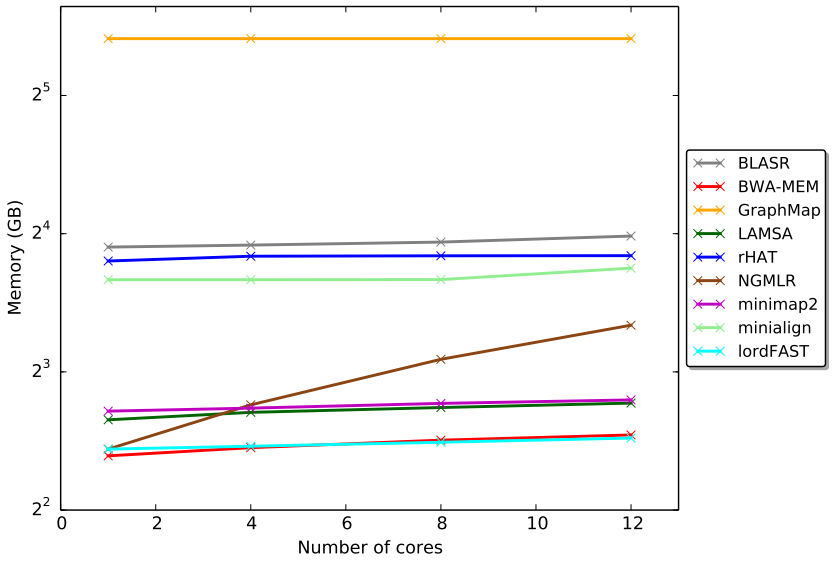

Fig. S5. Memory comparison of different methods for mapping 23,155 real human reads using different threads. Note that the y-axis is in logarithmic scale.

6 Command details

6.1 BLASR (v5.3.4323a52)

Indexing:

```
sawriter hg38.fa.sa hg38.fa
```

Mapping:

```
blasr reads.fasta hg38.fa --sa hg38.fa.sa -m 5 --out map_blasr.m5 --nproc 1 --noSplitSubreads
```

6.2 BWA-MEM (v0.7.15-r1140)

Indexing:

```
bwa index hg38.fa
```

Mapping:

```
bwa mem -x pacbio -Y -t 1 hg38.fa reads.fasta > map_bwa.sam
```

6.3 GraphMap (v0.5.1)

Indexing:

```
graphmap align -I -r hg38.fa
```

Mapping:

```
graphmap align -r hg38.fa -d reads.fasta -o map_graphmap.sam -t 1
```

6.4 LAMSA (v1.0.0)

Indexing:

```
lamsa index hg38.fa
```

Mapping:

```
lamsa aln -t 1 -S -T pacbio -i 25 -l 50 hg38.fa reads.fasta > map_lamsa.sam
```

6.5 rHAT (v0.1.1)

Indexing:

```
rHAT-indexer . hg38.fa
```

Mapping:

```
rHAT-aligner . reads.fasta hg38.fa -t 1 > map_rhat.sam
```

6.6 NGMLR (v0.2.6)

Indexing:

When invoked for the first time, NGMLR generates the index and write it to the disk. It uses the saved index for next runs. Therefore, we ran it once to generate the index without including this run in comparisons.

Mapping:

```
ngmlr -r hg38.fa -q reads.fasta -t 1 -o map_ngmlr.sam
```

6.7 Minimap2 (v2.10-r761)

Indexing:

```
minimap2 -d hg38.fa.mmi hg38.fa
```

Mapping:

```
minimap2 -a -Y -x map-pb -t 1 hg38.fa.mmi reads.fasta > map_minimap2.sam
```

6.8 minialign (v0.5.3)

Indexing:

```
minialign -d hg38.fa.mai hg38.fa
```

Mapping:

```
minialign -x pacbio -t 1 -l hg38.fa.mai reads.fasta > map_minialign.sam
```

6.9 lordFAST (v0.0.9)

Indexing:

```
lordfast --index hg38.fa
```

Mapping:

```
lordfast --search hg38.fa --seq reads.fasta --thread 1 > map_lordfast.sam
```

7 Performace of lordFAST using fixed length anchors compared to extended anchors

Table S3. Comparison between the performance of lordFAST using fixed length anchors and extended anchors on the simulated human dataset. This dataset contains 25,000 reads and 183.61 million bases. This experiment is done using lordFAST v0.0.2. Best results are marked with bold typeface.

lordFAST's version	Correctly mapped	Correct bases (Mb)	Incorrect bases (Mb)	Unmapped bases (Kb)	Sensitivity[a] (%)	Precision[b] (%)
Variable length ≥ 14	24,748	181.96	1.61	35.14	99.10	99.12
Anchor length = 13	15,000	137.11	46.44	53.65	74.68	74.70
Anchor length = 14	15,408	137.56	46.00	40.25	74.92	74.94
Anchor length = 15	16,404	141.10	42.45	51.52	76.85	76.87
Anchor length = 16	17,188	143.87	39.69	48.01	78.36	78.38
Anchor length = 17	17,922	146.07	37.47	68.50	79.56	79.59
Anchor length = 18	18,826	149.20	34.32	86.42	81.26	81.30
Anchor length = 19	19,708	153.00	30.52	83.91	83.33	83.37
Anchor length = 20	20,712	157.21	26.31	93.40	85.62	85.66
Anchor length = 21	21,255	160.66	22.80	150.63	87.50	87.57
Anchor length = 22	21,791	164.23	19.14	243.86	89.44	89.56

A read is considered to be mapped correctly if its aligned subsequence in the reference overlaps with at least 90% of the bases of the "correct" mapping subsequence. On the other hand, a base in a read is considered to be correctly mapped if the read is correctly mapped and the mapping location of the base is within a 25 bp vicinity of the correct alignment locus of the base. [a] The sensitivity is defined as the number of correctly mapped bases / the total number of bases. [b] The precision is defined as the number of correctly mapped bases / the number of mapped bases.

8 Comparison on a large simulated dataset

We compared different mappers on a simulated dataset with 2x coverage to mimic a real low depth sequencing. As it can be seen, lordFAST performs best in finding the correct location of the reads with Minimap2 closely following. lordFAST shows the best sensitivity and precision. minialign is the fastest among all tools, however, it has higher number of unaligned/incorrectly aligned bases compared to lordFAST and Minimap2. The differences are clearer in the lower part of the table (which corresponds to a more stringent definition of "correct mapping").

Table S4. Comparison between different tools capable of mapping PacBio long reads on a simulated human dataset with 2x coverage. This dataset contains 843,500 reads and 6,178.30 million bases. Best results are marked with bold typeface.

Minimum overlap (p)	Mapper	Correctly mapped	Correct bases (Mb)	Incorrect bases (Mb)	Unmapped bases (Mb)	Sensitivity[a] (%)	Precision[b] (%)	CPU hours[c]	Memory[c] (GB)
1 bp	BLASR	831,566	5,543.53	611.64	23.13	89.73	90.06	114.60	19.52
	BWA-MEM	830,867	5,761.41	399.79	17.98	93.24	93.51	71.53	5.70
	GraphMap	814,909	5,962.82	133.94	81.55	96.51	97.80	500.70	44.43
	LAMSA	824,897	5,957.13	215.48	9.18	96.37	96.51	22.28	7.13
	rHAT	823,443	5,971.55	193.99	12.77	96.65	96.85	15.03	14.75
	NGMLR	816,998	5,732.61	295.77	149.92	92.79	95.09	31.55	9.21
	Minimap2	834,990	6,058.30	112.98	7.67	98.05	98.17	1.42	8.10
	minialign	828,960	5,993.35	161.99	23.00	97.01	97.37	**0.43**	12.71
	lordFAST	**835,274**	**6,109.93**	**67.51**	**0.89**	**98.89**	**98.91**	5.43	**5.67**
90%	BLASR	828,176	5,540.60	615.66	22.04	89.68	90.00		
	BWA-MEM	826,833	5,749.43	415.94	13.47	93.05	93.25		
	GraphMap	814,888	5,962.80	133.95	81.55	96.51	97.80		
	LAMSA	822,276	5,952.50	221.71	6.94	96.30	96.41		
	rHAT	822,320	5,970.36	205.21	2.73	96.63	96.68		
	NGMLR	801,622	5,702.47	352.61	123.22	92.30	94.18		
	Minimap2	834,747	6,057.32	113.56	7.57	98.04	98.16		
	minialign	825,531	5,981.91	188.04	8.35	96.82	96.95		
	lordFAST	**835,181**	**6,109.86**	**67.59**	**0.87**	**98.89**	**98.91**		

A read is considered to be mapped correctly if its aligned subsequence in the reference overlaps with the "correct" mapping subsequence by at least p bases. On the other hand, a base in a read is considered to be correctly mapped if the read is correctly mapped and the mapping location of the base is within a 25 bp vicinity of the correct alignment locus of the base. [a] The sensitivity is defined as the number of correctly mapped bases / the total number of bases. [b] The precision is defined as the number of correctly mapped bases / the number of mapped bases. [c] This experiment is done using **8 threads**. The time and peak memory usage are measured using `/usr/bin/time -v` Unix command.

9 Simulation with structural variations

For this experiment, we performed simulation and SV calling as follows:

(i) We assigned SVs reported in DGV on chr1 of NA12878 individual into 3 groups based on their size (shorter than 500 bp, between 500 and 5000 bp, and longer than 5000 bp) and randomly selected 3 insertions, 3 deletions and 1 inversion from each group.
(ii) Selected SVs were inserted into the reference chr1 to get a simulated donor chromosome.
(iii) A set of long reads with 15x coverage were simulated from the donor chromosome using pbsim. pbsim was provided with a fastq file from a real human dataset to use its sample based model (–sample-fastq).
(iv) Long reads were mapped to the reference chr1 using different mappers. Sniffles requires MD tag in order to operate. Among different mappers, BLASR, LAMSA, and rHAT do not generate MD tag in the output sam file. Therefore, for these mappers, we used "samtools calmd" to calculate and add the MD tag. Minimap2 (version 2.10-r761; latest version at the time of writing this response) and minialign add MD tags with optional arguments. Other tools (including lordFAST) generate MD tags by default.
(v) For each mapper, a sorted bam file was generated from the sam file using "samtools sort".
(vi) Sniffles (version v1.0.8) was run with parameter "-s 4".

9.1 Command details

Blasr was run with parameters "--sam --bestn 1 --clipping subread --affineAlign --noSplitSubreads --nCandidates 20 --minPctSimilarity 75 --sdpTupleSize 6" and BWA-MEM was run with parameters "-x pacbio -MY" as mentioned in Sedlazeck *et al.* (2018). LAMSA was run with parameters "-T pacbio -i 25 -l 50 -S". minimap2 was run with parameters "-aY -x map-pb --MD". minialign was run with parameters "-x pacbio -T AS,XS,NM,NH,IH,SA,MD -P". rHAT, GraphMap, NGMLR, and lordFAST were run with default parameters.