Supplementary information

BRAT-nova: Fast and accurate mapping of bisulfite-treated reads

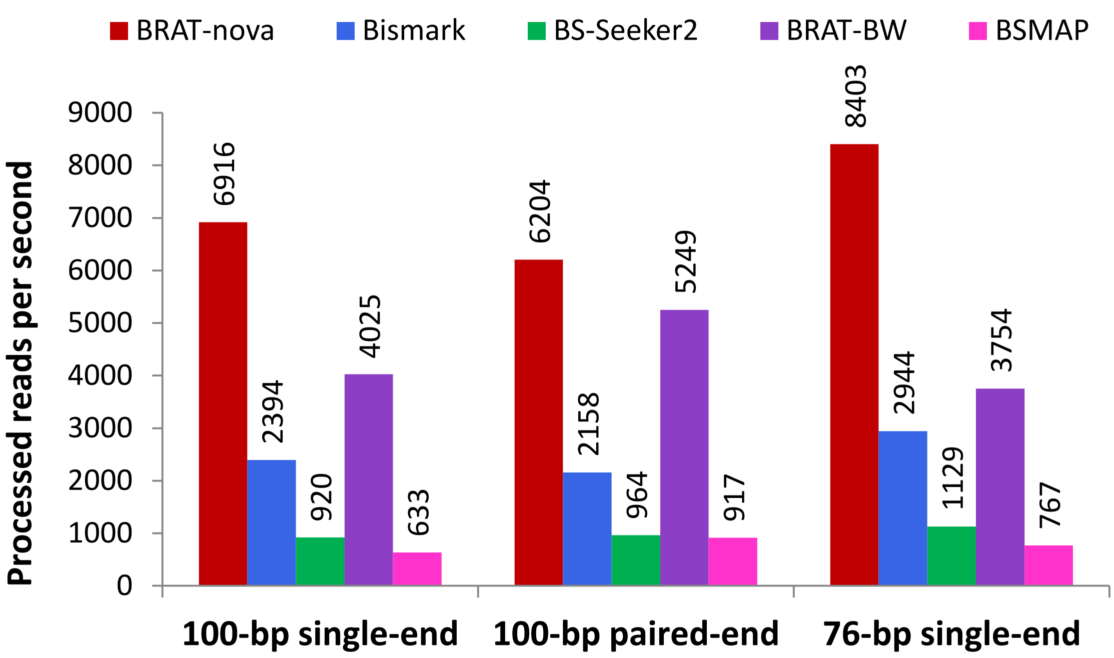
Elena Y. Harris[1,*], Rachid Ounit[2] and Stefano Lonardi[2]

[1]Department of Computer Science, California State University, Chico, CA 95929

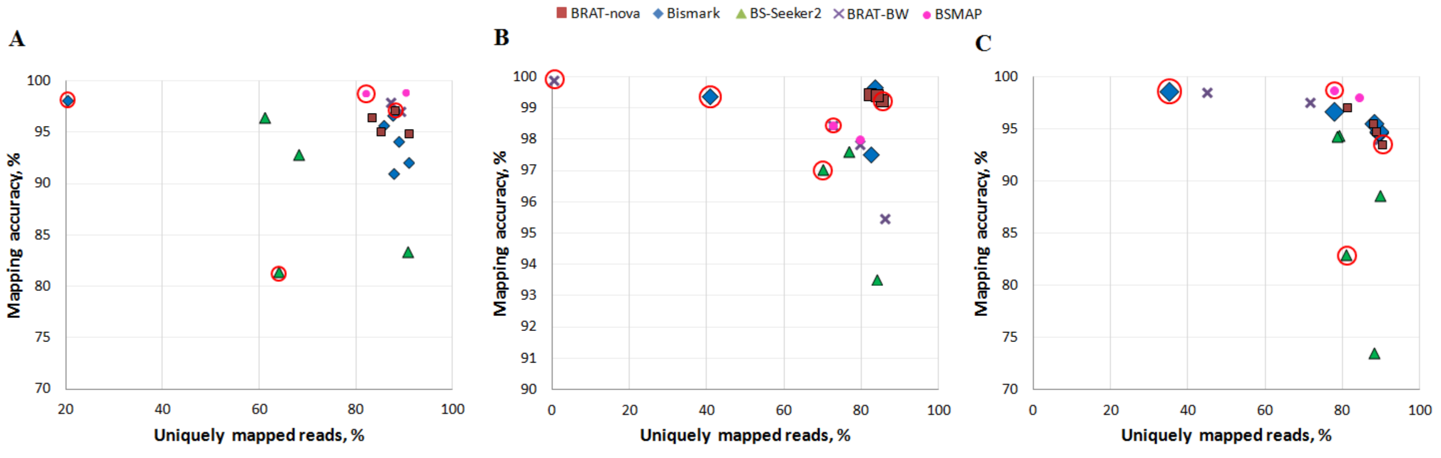
[2]Department of Computer Science, University of California, Riverside, CA 92521

[*]To whom correspondence should be addressed

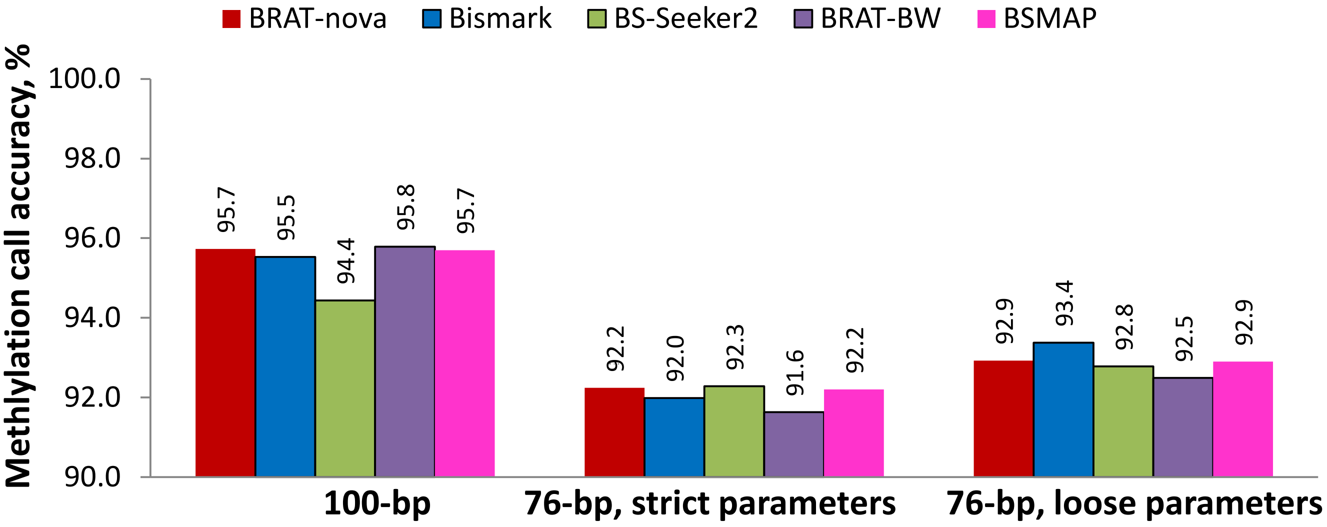
Supplemental Figure 1	Alignment speed for read data set
Supplemental Figure 2	Mapping accuracy as a function of the percentage of uniquely mapped reads
Supplemental Figure 3	Methylation call accuracy
Supplemental Figure 4	Methylation level accuracy
Supplemental Table 1	The results of mapping 10,633,033 real 101-bp long single-end reads from human genome, SRR306435, mate 1
Supplemental Table 2	The results of mapping 1M real 101-bp long paired-end reads from human genome, SRR306435
Supplemental Table 3	The results of mapping 1M real 76-bp long single-end reads from human genome, SRR306421
Supplemental Table 4	Mapping accuracy test: 1M synthetic 100bp single-end reads generated from human genome GRCh38 with 5% of sequencing errors and 2% of SNPs
Supplemental Table 5	Mapping accuracy test: 500,016 synthetic 100bp paired-end reads generated from human genome GRCh38 with 2% of sequencing errors, 1% of SNPs, and with a single indel introduced to 5% of reads, and adapter sequence up to 15bp long replaced 3'-end of 10% of the reads
Supplemental Table 6	Mapping accuracy test: 1M synthetic 76bp single-end reads with 3% of sequencing errors, 1% of SNPs, 10% of all reads having indels, and 30% of reads having an adapter sequence up to 15bp at the 3'-end of the reads
Supplemental Table 7	Methylation call and level accuracy test: 20M synthetic 100bp single-end reads from chromosome 21 of the human genome GRCh38 with 5% of sequencing errors and 2% of SNPs
Supplemental Table 8	Methylation call and level accuracy test: 20M synthetic 76bp single-end reads from chromosome 21 of the human genome GRCh38 with 5% of sequencing errors and 2% of SNPs, indels of length up to 10bp introduced to 10% of reads, adapter sequences up to 15bp replaced 3'-end of 10% of reads
Supplemental Table 9	Comparison of the performance of BRAT-nova on real data sets using local alignment versus full-length alignment
Supplemental Table 10	Feature comparison of BRAT-nova, Bismark, BS-Seeker2 and BRAT-BW
Supplemental Notes	BRAT-nova design and benchmarking



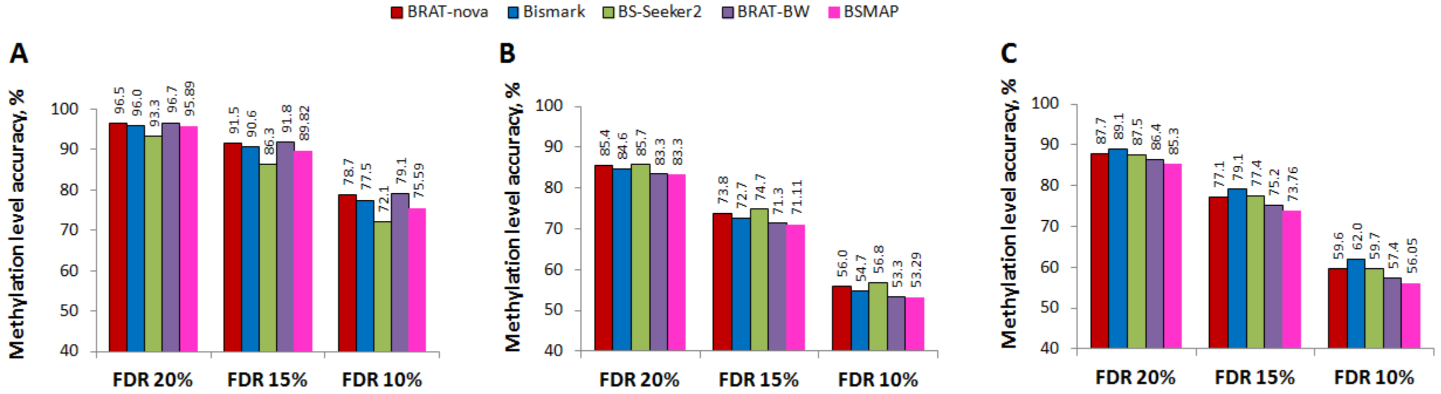
Supplemental Figure 1. Number of reads processed per second on the three real human datasets: 10.6M 100-bp single-end SRR306435 reads; 1M 100-bp SRR306435 pairs; and 1M 76-bp single-end SRR306421 reads. Combined results are provided for all experiments (described in Supplemental Tables 1-3 and Figure 1 of the manuscript): the total number of processed reads from all experiments divided by the total real time measured in all experiments. For BSMAP, only experiments run on a single CPU (thread) are shown. The parameters used with each tool are shown in Supplemental Tables 1-3. On 100-bp single-end reads, BRAT-nova was 2.9 times faster than Bismark, 7.5 times faster than BS-Seeker2, 1.7 times faster than BRAT-BW, and 10.9 times faster than BSMAP. On 100-bp paired end reads, BRAT-nova was 2.9 times faster than Bismark, 6.4 times faster than BS-Seeker2, 1.2 times faster than BRAT-BW and 6.8 times faster than BSMAP. On 76-bp single-end reads, BRAT-nova was 2.9 times faster than Bismark, 7.4 times faster than BS-Seeker2, 2.2 times faster than BRAT-BW and 11 times faster than BSMAP.



Supplemental Figure 2. Mapping accuracy as a function of the percentage of uniquely mapped reads. Mapping accuracy was measured as the ratio of unique reads mapped within 50bp of the original position (same strand, same chromosome) to the total number of uniquely mapped reads. The parameter settings used here are shown in Supplemental Tables 4-6. Circled results correspond to default parameters (not shown for BRAT-BW for A and C since default parameters for BRAT-BW is zero mismatches). (A) 1M 100-bp single-end reads generated from human genome GRCh38 with 5% sequencing error rate, 2% SNPs; (B) 500,000 100-bp paired-end reads (pairs) generated from human genome GRCh38 with 2% sequencing error rate and 1% SNPs, and with a single indel introduced in 5% of reads, and adapter sequence up to 15bp long replaced 3’-end for 10% of the reads; (C) 1M 76-bp single-end reads generated from human genome GRCh38 with 3% sequencing error rate, 1% SNPs, 10% of all reads having indels, and 30% of reads having an adapter sequence up to 15bp at the 3’-end. Dependent on different parameter settings, the percentage of uniquely mapped reads and mapping accuracy vary with all tools; some tools are more sensitive to parameter settings (BS-Seeker2) than other tools. BRAT-nova shows comparable results in terms of the percentage of uniquely mapped reads and mapping accuracy.

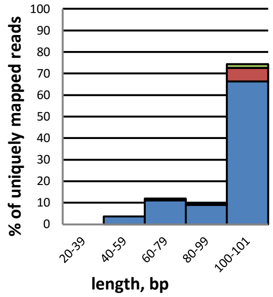
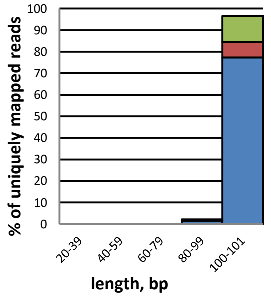
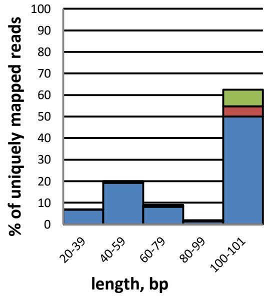
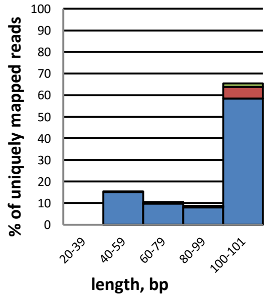


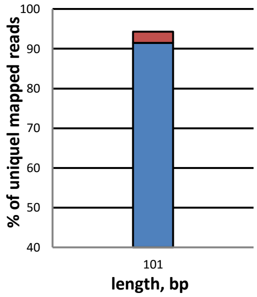
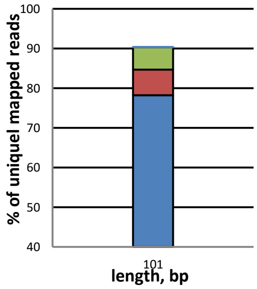
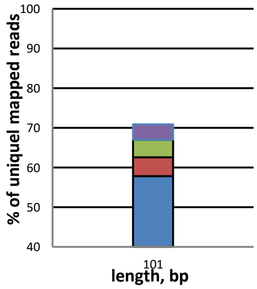
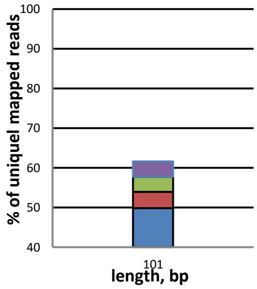
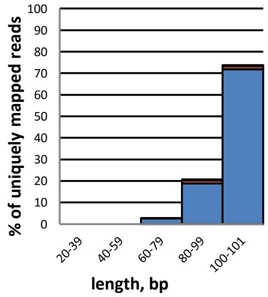
Supplemental Figure 3. Methylation call accuracy measured as the percentage of the cytosines whose methylation status was determined correctly, where a cytosine was considered methylated if its methylation level was at least 0.5, and unmethylated otherwise; all considered cytosines have to be covered by at least ten reads. The results are shown for two synthetic data sets each having a total of 20M reads: single-end 100bp reads and single-end 76bp reads generated from chromosome 21 of the human genome GRCh38. 100bp and 76bp reads have 5% sequencing error rate and 2% SNPs, while 10% of all 76bp reads have a single indel per read of length up to 10bp, and 10% of all 76bp reads have an adapter sequence of length up to 15bp at the 3'-end. The parameters of the tools were chosen so that the total number of uniquely mapped reads was similar among all tools, and 76bp reads were mapped using two different sets of parameters: strict parameters that ensured high mapping accuracy and loose parameters that ensured high percentage of unique reads mapped. All parameter settings are provided in Supplemental Tables 7 and 8. BRAT-nova shows comparable results in terms of methylation call accuracy.

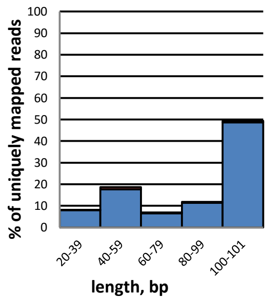
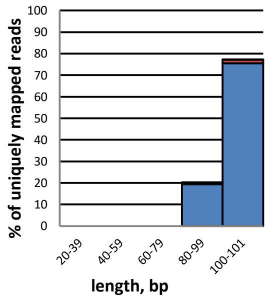
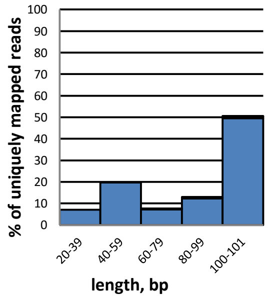
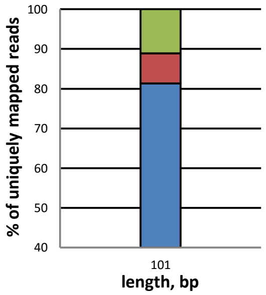
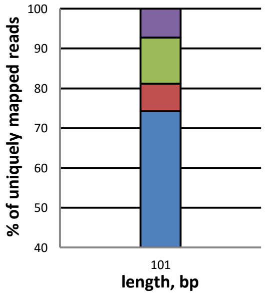
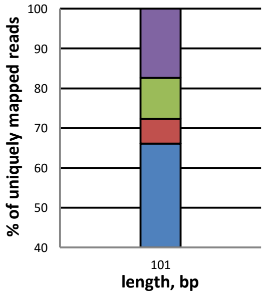


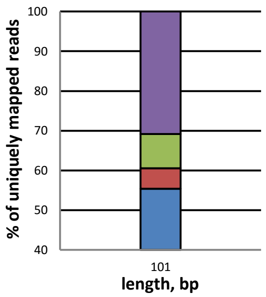
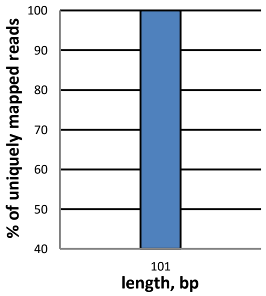
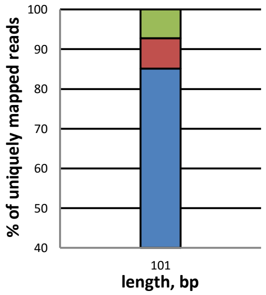
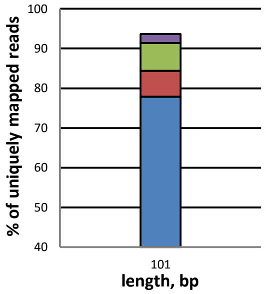
Supplemental Figure 4. Methylation level accuracy as defined in Supplemental Notes. (A) 20M synthetic single-end 100bp reads with 5% sequencing error rate and 2% SNPs; (B) 20M synthetic single-end 100bp reads with 5% sequencing error rate and 2% SNPs, 10% of the reads having indels up to 10bp long, and 10% of the reads having adapter contamination up to 15bp long at the 3'-end. (A) and (B) use "strict parameters" to ensure high mapping accuracy; (C) is the same dataset as (B), but all tools were run with looser parameters to increase mapping efficiency but at the cost of lower mapping accuracy. Parameters are provided in Supplemental Tables 7 and 8. All tools show comparable results in terms of methylation level accuracy.

Supplemental Table 1. The results of mapping 10,633,033 real 101-bp long single-end SRR306435 reads from the human genome. Figures in the last column show the distribution of uniquely mapped reads that do not have indels by length and by mismatches: reads mapped with 0-2 mismatches are shown in blue, with 3-5 mismatches in reddish-brown, with 6-11 mismatches in green and with greater than 11 mismatches in purple. The percentage of unique reads having indels is shown in the seventh column (previous to the last). Supplemental Figure 1 compares speed of BRAT-nova to that of other tools. The percentage of reads mapped by all tools varies dependent on the parameter settings. BRAT-nova shows a similar range of the percentage of mapped reads compared to other tools, but BRAT-nova is consistently faster than the other tools.

Aligner		**Options**	**Running time *real*/ *user***	**Uniquely mapped reads, %**	**RAM, GB**	**% unique reads with indels**	**Length and mismatch distribution of uniquely mapped reads** 🟦 0-2mism 🟥 3-5mism 🟩 6-11mism 🟪 >11mism
BRAT-nova	1	K 9, L, G q 95, l 60	28m26s 27m48s	7027269 66.09%	6.0	0.23%	
BRAT-nova	2	K 9, L, G q 90, l 90	25m35s 24m32s	6026381 56.68%	6.0	1.30%	
BRAT-nova	3	Default: K 9, L, G q 90, l 30	26m58s 26m25s	9323955 87.69%	6.0	0%	
BRAT-nova	4	K 9, L, G q 95, l 50	21m40s 21m2s	7985149 75.09%	6.0	0.06%	

Bismark	1	Default	68m45s 155m17s	5089248 47.9%	9.4	5.65%	
Bismark	2	--score-min L,0,-0.65 rdg 7,6; rfg 7,6	69m23s 160m57s	5967863 56.1%	9.4	9.61%	
Bismark	3	--score-min L,0,-1.8 rdg 11,6; rdg 11,6	78m56s 189m51s	8069199 75.9%	9.4	29.12%	
Bismark	4	--score-min L,0,-2.3 rdg 11,6; rdg 11,6	78m56s 191m8s	9368367 88.1%	9.4	38.35%	
BS-Seeker2	1	--score-min G,11,13 ma 1; mp 3,3; m 10	195m26s 315m41s	6264250 58.91%	6.4	2.95%	
BS-Seeker2	2	Default	244m24s 298m56s	9202526 86.55%	6.4	5.01%	

							
BS-Seeker2	3	--score-min G,1,17.3 D15; R 2 ma 1; mp 3,3 m 0.1	126m45s 198m11s	5958866 56.04%	6.4	2.69%	
BS-Seeker2	4	--score-min G,0,6 D15; R 2 ma 1; mp 3,3 m 0.05	203m39s 268m23s	9064614 85.25%	6.4	2.25%	
BRAT-BW	1	K 9, m 10	49m25s 48m52s	5740439 53.99%	5.4	0%	
BRAT-BW	2	K 9, m 20	52m 33s 51m 57s	6286917 59.13%	5.4	0%	
BRAT-BW	3	K 9, m 30	55m14s 54m36s	7062382 66.42%	5.4	0%	

BRAT-BW	4	K 9, m 40	48m1s 47m18s	8433027 79.31%	5.4	0%	
BRAT-BW	5	Default: K 1, m 0	14m55s 14m40s	3152121 29.64%	5.4	0%	
BSMAP	1	Default: r 0, p 1	280m7s 247m45s	5517052 51.9%	7.7	0%	
BSMAP	2	r 0, p 8, v 15, g 3	509m34s	6023517 56.6%	7.7	6.3%	

Supplemental Table 2. The results of mapping 1M real 101bp paired-end SRR306435 reads. Insert size between mates of a pair was kept the same for all tools, 0-800bp. Here, BRAT-nova showed comparable performance in terms of the number of unique reads mapped and RAM usage, but was faster than other tools.

Aligner	**Options**	**Running time, *real***	**Uniquely mapped concordant pairs, %**	**RAM, GB**
BRAT-nova	K 9, L, G, i 0, a 800, pe, q90, l 30	5m25s	78.07%	6.0
BRAT-nova	K 9, L, G, i 0, a 800, pe, q95, l 40	4m51s	75.17%	6.0
BRAT-nova	K 9, L, G, i 0, a 800, pe, q95, l 50	5m51s	68.65%	6.0
BS-Seeker2	--ignore-quals, p 1, X 800 --no-mixed, --no-discordant Default	53m46s	79.12%	6.4
BS-Seeker2	--ignore-quals, p 1, X 800 --no-mixed, --no-discordant m 0.1, --ma 1, --mp 3,3 -D 15, -R 2, --score-min G,1,17.3	18m48s	49.35%	6.4
BS-Seeker2	--ignore-quals, p 1, X 800 --no-mixed, --no-discordant m 0.05, --ma 1, --mp 3,3 -D 15, -R 2, --score-min G,0,10	26m45s	60.47%	6.4
BS-Seeker2	--ignore-quals, p 1, X 800 --no-mixed, --no-discordant m 0.05, --ma 1, --mp 3,3 -D 30, -R 2, --score-min G,0,6	38m56s	74.46%	6.4
Bismark	-X 800, --non_bs_mm Default	9m15s	43.5%	9.4
Bismark	-X 800, --non_bs_mm --rdg 11,6; --rfg 11,6 --score_min L,1,-1.2	11m36s	49.8%	9.4
Bismark	-X 800, --non_bs_mm --rdg 11,6; --rfg 11,6 --score_min L,1,-1.8	15m23s	54.7%	9.4
Bismark	-X 800, --non_bs_mm --rdg 11,6; --rfg 11,6 --score_min L,1,-2.5	25m32s	76%	9.4
BRAT-BW	-K 9, -i 0, -a 800, -pe, -m 10	4m28s	18.15%	5.4
BRAT-BW	-K 9, -i 0, -a 800, -pe, -m50	8m14s	45.25%	5.4

BSMAP	Default: r 0, p 1, m 0, x 800	36m19s	50.9%	7.7
BSMAP	r 0, p 1, m 0, x 800, v 0.7	45m40s	56.9%	7.7

Supplemental Table 3. The results of mapping of 1M real 76bp single-end SRR306421 reads. Results on this real human data set are consistent with previous results on 100-bp real human reads: BRAT-nova shows comparable results in terms of the percentage of uniquely mapped reads and RAM, and is faster than other tools (Supplemental Figure 1 shows exact numbers for speed up).

Aligner	**Options**	**Running time, *real***	**Uniquely mapped reads, %**	**RAM, GB**
BRAT-nova	K 5, L, G, q 90, l 30	2m14s	878697 87.87%	6.0
BRAT-nova	K 5, L, G, q 90, l 90	1m57s	725926 72.59%	6.0
BRAT-nova	K 5, L, G, q 95, l 90	1m46s	687577 68.76%	6.0
BS-Seeker2	--ignore-quals, p 1 Default	20m40s	865317 86.53%	6.4
BS-Seeker2	--ignore-quals, p 1 m 0.1, -D 15, -R 2 --ma 1, --mp 3,3 --score-min G,0,14	12m37s	719947 71.99%	6.4
BS-Seeker2	--ignore-quals, p 1 m 0.05, -D 15, -R 2 --ma 1, --mp 3,3 --score-min G,0,3.2	21m48s	840932 84.1%	6.4
Bismark	--non_bs_mms, Default	5m1s	605102 60.5%	9.4
Bismark	--non_bs_mm --rdg 11,6; --rfg 11,6 --score_min L,0,-2.3	6m1s	869582 87.0%	9.4
Bismark	--non_bs_mm --rdg 11,6; --rfg 11,6 --score_min L,0,-0.8	5m57s	738539 73.9%	9.4
BRAT-BW	Default	1m37s	385444 38.54%	5.4
BRAT- BW	-K 5, -m 8	5m49s	693431 69.34%	5.4
BRAT- BW	-K 5, -m 15	5m53s	778112	5.4

			77.81%	
BSMAP	Default: r 0, p 1	21m45s	666229 66.62%	7.7
BSMAP	r 0, p 1, v 0.2	28m 33s	775252 77.53%	7.7
BSMAP	r 0, p 1, v0.7	20m1s	775253 77.53%	7.7

Supplemental Table 4. Mapping accuracy test. Total of 1M single-end 100bp reads generated from human genome GRCh38 with 5% sequencing error rate, 2% SNPs and 97% of bisulfite-conversion rate. These results are also shown in Supplemental Figure 2. Tools were tested using different parameter settings. We measured two types of the mapping accuracy defined as the ratio of uniquely mapped reads aligned within 50bp and 0bp of the original positions (same chromosome, same strand) to the total number of uniquely mapped reads. The second type of mapping accuracy is stricter and requires that a read is mapped exactly to the original position from which it was extracted; the results for this type of mapping accuracy are shown in blue. The measures of two types of mapping accuracy differ due to indels or adapter contamination. As shown in this table, all tools except for Bismark show the same results for mapping accuracy of type "within 50 bp" and "start position equals to original position". Bismark tends to give preference to the alignments with indels when options ***rdg*** and ***rfg*** are kept default; this is why here we observe a drastic difference between the two measures for two types of mapping accuracy. BRAT-nova was 3-16 times faster in these experiments and showed comparable results in terms of mapping accuracy (94.92%-97.18%).

Aligner	**Options**	**Uniquely mapped reads, count**	**Ambiguously mapped reads, count**	**Mapping accuracy, start pos within 50bp, %**	**Mapping accuracy, start pos equals to original pos, %**	**Running time, *real/user***	**RAM, GB**
Bismark	--score-min L,0,-1.2 D=7,R=1	877432	81756	91.04	86.72	6m18s 15m3s	9.4
	--score-min L,0,-1.2 R=3	887879	89792	94.14	89.65	12m17s 27m1s	9.4
	--score-min L,0,-0.6	857884	84107	95.75	91.28	11m14s 24m36s	9.4
	Default	202354	19202	98.21	96.99	9m42s 18m21s	9.4
	--score-min L,0,-0.6 D=50	875652	88159	96.72	92.21	19m57s 41m35s	9.4
	--score-min L,1,-2.6	908518	90929	92.14	87.75	14m58s 30m52s	9.4
BS-Seeker2	--ignore-quals, p=1 Default	639175	3	81.44	80.31	20m16s 27m0s	6.4
	--ignore-quals, p=1 -D 15, -R 2 --ma 1, --mp=3,3 --score-min G,0,5.85 -m 0.1	907243	4	83.42	83.37	19m44s 28m15s	6.4
	--ignore-quals, p=1 -D 15, -R 2 --ma 1, --mp=3,3 --score-min G,0,17.58	609437	4	96.46	96.43	14m12s 24m2s	6.4

	-m 0.1						
	--ignore-quals, p=1 -D 15, -R 2 --ma 1, --mp=3,3 --score-min G,0,10.2 -m 0.05	680917	3	92.90	92.87	17m36s 27m56s	
BRAT-nova	K5, q90, L, l=90	831629	67751	96.74	96.74	1m98s 1m76s	6.0
	K 9, q 90, L l 90, G	878951	82860	97.18	97.18	3m45s 3m37s	6.0
	K 9, L, G q 95, l 50	850009	82836	95.11	95.10	4m2s 3m54s	6.0
	K9, L, G q 90, l 30 Default	907872	88527	94.92	94.92	4m33s 4m25s	6.0
BAT-BW	K9, m 20	892256	92155	97.08	97.08	7m46s 7m38s	5.4
	K 9, m 10	871071	86840	97.96	97.96	9m40s 9m31s	5.4
BSMAP	-r 0, -p 1, -H	818839	88325	98.87	98.87	32m45s 32m45s	7.7
	-r 0, -p 1, -H, -v 0.2	901538	96546	98.90	98.90	48m46s 41m29s	7.7

Supplemental Table 5. Mapping accuracy test for 500,016 paired-end 100bp reads generated from human genome GRCh38 with 2% sequencing error rate and 1% SNPs, and with a single indel introduced in 5% of reads, and adapter sequence up to 15bp long replaced 3'-end for 10% of the reads. Here we also show two types of mapping accuracy defined as the ratio of uniquely mapped reads aligned within 50bp and 0bp of the original positions (same chromosome, same strand) to the total number of uniquely mapped reads. The results of the stricter mapping accuracy (shown in blue) for paired-end reads are slightly lower than for the other type of mapping accuracy and this tendency is observed for all tools. BRAT-nova was 9 times faster than BSMAP in these experiments. BRAT-nova was 2-10 times faster in these experiments compared to the rest of the tools and showed comparable mapping accuracy for the type "within 50bp" (99.27%-99.46%) and for the type "within 0bp" (93.17%-95.84%).

Aligner	**Options**	**Uniquely mapped concordant pairs**	**Mapping accuracy, start pos within 50bp, %**	**Mapping accuracy, start pos equals to original pos, %**	**Running time, *real***	**RAM, GB**
Bismark	Default	203764 40.75%	99.37%	98.24%	7m33s	9.4
	--score-min L,0,-1.2 --rdg 11,6 --rfg 11,6	412183 82.4%	97.51%	94.03%	9m16s	9.4
	--score-min L,0,-1.8 --rdg 11,6 --rfg 11,6	416863 83.37%	99.66%	93.21%	10m37s	9.4
BS-Seeker2	--ignore-quals, p 1 --no-mixed, X 800 --no-discordant Default	350244 70.05%	97.09%	94.02%	31m27s	6.4
	--ignore-quals, p 1 --no-mixed, X 800 --no-discordant D 15, R2, m 0.1 --ma 1, --mp 3,3 --score-min G,1, 17.3	383666 76.73%	97.62%	95.23%	14m45s	6.4
	--ignore-quals, p 1 --no-mixed, X 800 --no-discordant --ma 1, --mp 3,3 D 15, R2, m 0.1 --score-min G,0,6	426506 83.85%	93.54%	90.26%	20m22s	6.4
BRAT-nova	Default: K 9, L, G, q 90, l 30	425627 85.12%	99.27%	93.17%	3m2s	6.0
	K 9, L, G, q 90, l 90	410680 82.13%	99.46%	95.84%	3m2s	6.0
	K 9, L, G, q 90, l 70	420367 84.07%	99.42%	94.28%	3m12s	6.0
BRAT-BW	Default	1727	99.97%	99.91%	2m1s	5.4

	i 0, a 800	0.34%				
	K 9, m 10 i 0, a 800	397071 79.41%	99.99%	97.85%	6m31s	5.4
	K 9, m 50, i 0, a 800	429558 85.91%	99.99%	95.49%	6m48s	5.4
BSMAP	Default, r 0, p 1, m 0, x 800	362496 72.5%	98.46%	96.73%	31m36s	7.7

Supplemental Table 6. Mapping accuracy test using 1M single-end 76bp reads with 3% sequencing error rate, 1% SNPs, 10% of all reads having indels, and 30% of reads having an adapter sequence up to 15bp at the 3'-end. The stricter mapping accuracy requiring mapping reads to the exact position, strand and chromosome is shown in blue. In this data set, indels and adapters were introduced; both of these cases may cause the starting position of the read to differ from its exact original position. Therefore, the stricter mapping accuracy (shown in blue) is lower by up to 10%. BSMAP showed slightly better results for mapping accuracy of type "position is within 50bp of original position" than other tools, but had similar mapping accuracy of type "exact position as original position". BRAT-nova was 2-11 times faster than other tools and showed comparable results in terms of mapping accuracy of type "within 50bp" (93.61%-97.12%) and of type "within 0bp" (84.25%-91.07%).

Aligner	**Options**	**Uniquely mapped reads, %**	**Mapping accuracy, start pos within 50bp, %**	**Mapping accuracy, start pos equals to original pos, %**	**Running time, *real***	**RAM, GB**
Bismark	Default	34.95	98.64	96.40	8m26s	9.4
	--score-min L,0,-4.42 --rfg 11,6; --rdg 11,6	89.3	94.71	85.86	10m14s	9.4
	--score-min L,0,-3.32 --rfg 11,6; --rdg 11,6	89.3	94.72	85.87	9m44s	9.4
	--score-min L,0,-1.5 --rfg 11,6; --rdg 11,6	88.15	95.53	86.64	8m16s	9.4
	--score-min L,0,-0.65 --rfg 11,6; --rdg 11,6	77.86	96.69	89.92	7m34s	9.4
BS-Seeker2	--ignore-quals, p 1, Default	80.59	82.91	76.81	19m33s	6.4
	--ignore-quals, p 1, D 15, R 2, m 0.1 --ma 1, --mp 3,3 --score-min G,0, 4.6	87.95	73.50	68.07	20m59s	6.4
	--ignore-quals, p 1, D 15, R 2, m 0.1 --ma 1, --mp 3,3 --score-min G,0, 14	78.95	94.45	89.21	15m31s	6.4
	--ignore-quals, p 1,	89.50	88.63	81.29	18m30s	6.4

	D 15, R 2, m 0.1 --ma 1, --mp 3,3 --score-min G,0, 7.85					
	--ignore-quals, p 1, D 15, R 2, m 0.05 --ma 1, --mp 3,3 --score-min G,0,13.16	78.47	94.34	88.31	15m54s	6.4
BRAT-nova	Default, K 5, L, G q 90, l 30	90.12	93.61	84.25	2m26s	6.0
	K 5, L, G q 90, l 50	87.70	95.58	86.27	2m17s	6.0
	K 5, L, G q 95, l 80	81.09	97.12	90.67	1m47s	6.0
	K 5, L, G q 90, l 90	80.19	97.02	91.07	2m1s	6.0
	K 5, L, G q 90, l 40	88.87	94.86	85.39	1m47s	6.0
	K 5, L, G q 95, l 30	89.76	93.81	84.42	1m48s	6.0
BRAT-BW	K 5, m 3	44.8	98.58	97.25	4m54s	5.4
	K 5, m 8	71.45	97.59	93.05	3m51s	5.4
	K 5, m 38	89.4	94.04	84.46	4m35s	5.4
BSMAP	Default r 0, A, p 1	77.6	98.74	87.34	22m3s	7.7

Supplemental Table 7. Methylation call accuracy and methylation level accuracy test. 20M synthetic single-end 100bp reads from human chromosome 21 GRCh38 with 5% sequencing error rate and 2% SNPs. Methylation call accuracy was measured for all cytosines covered by at least ten reads using a threshold for methylation level of 0.5 to determine methylated status of each cytosine. A cytosine with methylation level of 0.5 or higher was considered to be methylated; unmethylated otherwise. Methylation call accuracy was defined as the ratio of the cytosines whose methylated status was called correctly to the total number of the cytosines covered by at least ten reads. BRAT-nova's parameters used (-K 9, -q 90, -l 30, -L, -G); Bismark's parameters (--score_min L,0,-4.38, rdg 11,6, non_bs_mm, rfg 11,6), BS-Seeker2's parameters (score-min G,0,5.86, bt2, ma 1, mp 3,3, m 0.1, -D 15, -R 2, -p 1); BSMAP's parameters (r 0, p 1, v 0.1). The results of this experiment are also depicted in Supplemental Figures 3 and 4. BRAT-nova shows comparable results in terms of methylation call and methylation level accuracy. BRAT-nova was about 2-5 times faster than other tools.

Aligner	**Uniquely mapped reads, %**	**Mapping accuracy, pos within 50bp, %**	**Indels, % of reads having indels**	**Time to map reads, real/ user**	**Time to run methylation extractor program**	**Meth call accuracy, %**	**Total cytosines covered with at least 10 reads, x 10^6**	**% of Correct meth level, FDR 20%**	**% of Correct meth level, FDR 15%**	**% of Correct meth level, FDR 10%**
Brat-nova	83.47	95.53	0	94m38s 93m21s	1m17s	95.73	13.567	96.47	91.46	78.74
Bismark	83.21	93.00	3.74	220m41s 508m20s	37m16s	95.53	13.178	95.95	90.58	77.49
BS-Seeker2	82.81	85.29	1.41	421m52s 576m52s	136m23s	94.44	12.106	93.28	86.30	72.05
BRAT-BW	82.28	97.46	0	178m59s 177m50s	1m20s	95.79	13.634	96.67	91.75	79.11
BSMAP	81.3	99.38	0	509m24s 498m33s	11m54s	95.65	13.114	95.89	89.82	75.59

Supplemental Table 8. Methylation call accuracy and methylation level accuracy test. 20M 76bp long synthetic single-end reads from human genome GRCh38. 5% sequencing error rate and 2% SNPs, indels of length up to 10bp introduced to 10% of reads, adapter sequences up to 15bp replaced 3'-end of 10% of reads. Each tool was run on two parameter settings: strict and loose. With strict parameters, all tools mapped fewer reads with higher mapping accuracy. However, despite decreased mapping accuracy with loose parameters, methylation level accuracy was 2-4% higher for all tools compared to strict parameters. In these experiments, BRAT-nova was 2-8 times faster than Bismark, BS-Seeker2, BSMAP and BRAT-BW.

Aligner	**Uniquely mapped reads, %**	**Mapping accuracy, pos within 50bp, %**	**Indels, % of reads having indels**	**Time to map reads**	**Time to run methylation extractor program**	**Meth call accuracy, %**	**Total cytosines covered with at least 10 reads, x 10^6**	**% of Correct meth level, FDR 20%**	**% of Correct meth level, FDR 15%**	**% of Correct meth level, FDR 10%**
Brat-nova Strict K 5,L,G q 90, l 90	68.46	96.44	0.6	39m55s 39m3s	1m12s	92.21	11.797	85.27	73.6	55.71
Brat-nova Loose K 5,L,G q 90, l 30	83.71	86.18	0	41m55s 40m54s	1m13s	92.91	12.099	87.71	77.01	59.45
Bismark Strict L,0,-0.56; rdg 11,6 rfg 11,6	65.3	96.16	3.79	120m25s 300m16s	23m1s	91.99	11.283	84.57	72.66	54.67
Bismark Loose L,0,-4.42; rdg 11,6 rfg 11,6	81.2	89.71	15.63	178m15s 418m19s	33m31s	93.38	12.069	89.06	79.05	61.97
BS-Seeker2 Strict G,2,12 m 0.1 ma 1	71.85	91.58	1.73	284m5s 430m43s	92m40s	92.28	11.134	85.65	74.74	56.77

mp 3,3 p 1, R 2 D 15										
BS-Seeker2 Loose G,0,6.5 m 0.1 ma 1 mp 3,3 p 1, R 2 D 15	79.94	84.50	5.33	339m7s 474m22s	118m59s	92.78	10.830	87.47	77.43	59.71
BRAT-BW Strict K 5, m 8	64.9	96.75	0	76m46s 75m54s	1m14s	91.63	11.656	83.34	71.30	53.31
BRAT-BW Loose K 5, m 20	75.46	93.36	0	80m15s 79m17s	1m11s	92.49	12.261	86.39	75.18	57.39
BSMAP Strict r 0, v 0.1, p 1, A	69.6	98.55	0	190m40s 183m58s	8m55s	92.18	12.341	83.30	71.11	53.29
BSMAP Loose r 0, v 0.2 g 3, p 8, A	77.8	98.08	5.9	199m27s 1567m33s	9m34s	92.95	12.710	85.30	73.76	56.05

Supplemental Table 9. Comparison of the performance of BRAT-nova on real data sets using local alignment versus full-length alignment. With local alignment, BRAT-nova shows very similar time performance, but maps 3%-38% more reads.

Data set	**Options**	**Local alignment L, G**	**Full-length alignment**	**Local alignment L, G**	**Full-length alignment**	**Difference of uniquely mapped reads between local and full-length alignments, %**	**RAM, GB**
		Running time, *real*	**Running time, *real***	**Uniquely Mapped, %**	**Uniquely Mapped, %**		
Real Single-end 101bp	K 5, q 95, l 60	28m26s	16m32s	7027269 66.09%	5820358 54.73%	11.36%	6.0
	K 5, q 90, l 90	25m35s	16m46s	6026381 56.68%	5820358 54.73%	2.95%	6.0
	K 5, q 90, l 30	26m58s	17m23s	9323955 87.69%	5219793 49.09%	38.6%	6.0
	K 5, q 95, l 50	21m40s	16m48s	7985149 75.09%	5219793 49.09%	25.97%	6.0
Real Paired-end 101bp	K 5, q 90, l 30	5m25s	3m48s	780699 78.07%	511527 51.15%	26.92%	6.0
	K 5, q 95, l 40	4m51s	3m49s	751703 75.17%	457054 45.71%	29.46%	6.0
	K 5, q 95, l 50	5m51s	3m47s	686539 68.65%	457054 45.71%	12.94%	6.0
Real Single-end 76bp	K 5, q 90, l 30	2m14s	2m7s	878697 87.87%	690845 69.08%	18.79%	6.0
	K 5, q 90, l 90	1m57s	1m46s	725926 72.59%	690845 69.08%	3.51%	6.0
	K 5, q 95, l 50	1m46s	2m4s	687577 68.76%	623541 62.35%	6.41%	6.0

Supplemental Table 10. Feature comparison of BRAT-nova, Bismark, BS-Seeker2, BRAT-BW and BSMAP.

Feature	**BRAT-nova**	**BRAT-BW**	**Bismark**	**BS-Seeker2**	**BSMAP**
Number of FM-instances	1	2	2	2	N/A
Paired-end (PE) support	Yes	Yes	Yes	Yes	Yes
Variable read length	Yes	Yes	Yes	Yes	Yes
Adjustable insert size (PE)	Yes	Yes	Yes	Yes	Yes
Uses basecall qualities for FastQ mapping	No	No	Yes	No	Yes
Supports typeI/typeII bisulfite libraries (directional/unidirectional)	Yes/No	Yes/Yes	Yes/Yes	Yes/Yes	Yes/Yes
Supports alignment score as a function of the read length	Yes	No (fixed number of mismatches)	Yes	Yes	Yes
Supports indels	Yes One per read Length is unlimited	No	Yes Unlimited	Yes Unlimited	Yes One per read Up to 3bp
Supports alignment type	Full-length Local	Full-length only	Full-length only	Local only	Full-length only

Supplemental Notes

1. Design

Building the FM-index. In the FM-index used in BRAT-nova for a given reference genome S we first replace all Cs with Ts, then compute the reverse S^R of the resulting string; for example, if S is CCATG, then S^R is GTATT. Then, we take the reverse complement of S, replace all Cs with Ts, and compute the reverse S^C of the resulting string; for example, if S = CCATG, then S^C = GGTAT. In BRAT-nova the FM-index is built on the concatenation of the strings S^C and S^R, *i.e.*, on the string S^CS^R; for example, if S = CCATG, then S^CS^R = GGTATGTATT.

Read alignment overview. In order to align a read to the reference genome, BRAT-nova (1) determines the exact occurrences of k-mers (*seeds*) from the read using the FM-index; (2) aligns the read to the genomic regions sharing the same seeds using a binary representations of the genome and reads and taking advantage of binary operations that support a T in a read matching a C in the genome; (3) runs a linear time dynamic programming algorithm to find the best-score local alignment; and (4) if a read has two perfectly aligned seeds within a specified distance from each other, it attempts to find an alignment with a single indel using the linear time dynamic programming algorithm.

To improve the accuracy of the alignment, BRAT-nova uses a similar strategy as in BRAT-BW (Harris *et al.*, 2012). It prompts to align a flexible-length seed of a read using the FM-index, where seeds start at equal distance from each other with the first seed at the beginning of the 5'-end of the read, and with a user-defined total number of seeds. Starting at a specific position in a read, BRAT-nova aligns a seed character by character until a unique position in the genome matches the seed or until there are no positions in the genome that match the seed (in this case, BRAT-nova uses the last valid alignment of the seed).

For each seed, and each genomic position sharing the seed, BRAT-nova performs a full-length alignment of the read to the genomic position using a binary representation of the genome and a binary representation of the reads. The full length alignment is implemented using binary operations as it was done with BRAT-BW and BRAT (Harris *et al.*, 2010). The total number of mismatches is calculated, where a T in a read matching to a C in the genome is not considered as a mismatch. If the total number of mismatches is higher than a user-defined threshold, then BRAT-nova does not look for a local alignment or alignment with indels for this read and reports the best-score full-length alignment. Otherwise, BRAT-nova attempts to find a local alignment or an alignment with an indel, and reports the best-score alignment.

To control the quality of an alignment, BRAT-nova uses two user-defined thresholds: one that controls the alignment length (l) and another that determines the alignment quality (q). The threshold for an alignment score is calculated as a function of the read length, r, using the formula

$Score\_Threshold = \lfloor l \cdot q \cdot r \rfloor$

For example, for a read of length 100bp, alignment length l = 0.9, and alignment quality q = 0.95, the threshold on the alignment score is 85. Moreover, the alignment quality q re-enforces the alignment to have q% of bases matched perfectly (as a function of the aligned portion of the read). For our example, the read may be aligned with at most 5 mismatches using full-length alignment, or the contiguous 85bp of the read may be aligned with 0 mismatches using local alignment. To clarify, BRAT-nova uses the score threshold together with the quality alignment threshold to control the quality of the alignment for the aligned portion of the read, where the stretch is contiguous. The alignment score is calculated as follows

$$Alignment\_Score = matches - (mismatches \cdot mismatch\_penalty)$$
$$-(gap\_opening + indel\_length \cdot gap\_extension)$$

where *mismatch_penalty*, *gap_opening* and *gap_extension* are user-defined parameters.

Linear time dynamic programming algorithm for local alignment with no indels. BRAT-nova uses a linear time dynamic programming algorithm to find a local alignment with no indels. Let read R be of length r, and let m_p be a mismatch penalty. Let G be a reference genome of length n. By default, BRAT-nova uses m_p=1 for a match penalty. Assume that the read is being aligned to a genomic position i. Let M[j], where $1 \leq j \leq r$, is the best alignment score to align $R[1\ldots j]$ to genomic location G[$i\ldots i$+j-1]. Then the recursive formula to calculate M for a local alignment with no indels is given by:

$$\begin{cases} M[1] = 1, & if\ R[1] = G[i] \\ M[1] = m_p, & if R[1] \neq G[i] \\ M[j] = \max \begin{cases} 1, & \text{if } \mathrm{R[j] = G[i + j - 1]} \\ m_p, & if\ R[j] \neq G[i + j - 1] \\ \mathrm{M[j - 1] + 1}, & \text{if } \mathrm{R[j] = G[i + j - 1]} \\ \mathrm{M[j - 1]} + m_p, & \text{if } \mathrm{R[j] \neq G[i + j - 1]} \end{cases} \end{cases}$$

Linear time dynamic programming algorithm for local alignment with an indel. Because a small proportion of the sequenced reads can contain indels, some aligners employ an adapted Smith-Waterman algorithm (Smith and Waterman, 1981) to allow alignment with gaps. For example, Bowtie-2 (Langmead *et al*., 2009; Langmead and Salzberg, 2012) uses a single-instruction multiple-data parallel processing strategy to accelerate the quadratic-time Smith-Waterman dynamic programming algorithm. BRAT-nova uses a different approach to handle indels based on the following observations: (1) only a very small fraction of the BS-Seq reads contains indels; (2) BS-Seq reads are relatively short (typically up to 100bp) and thus the occurrence of more than one indel in a single read is very unlikely; and (3) in most methylation studies the sequenced sample is comprised of a pool of donors to allow the study of genome-wide methylation across a population (hence, the contribution of individual indels to a general pattern of methylation is not representative). Thus, BRAT-nova employs a linear time dynamic programming algorithm to support local alignment with one indel per read. Since indels can potentially occur at any position within a read, BRAT-nova looks for the best-scored aligned portion of a read and clips off poorly aligned ends of the reads due to indels. This approach is also useful to trim adapter sequences and multiple sequencing errors at the end of the reads. If a read has two seeds perfectly aligned to two genomic positions at a user-defined distance from each other, BRAT-nova uses a linear time dynamic programming algorithm to find the best-score alignment with a single indel that occurs between the two seeds in the read. Similar to the local alignment with no indels, BRAT-nova calculates $M_{seed1}[1\ldots r]$ from the leftmost end of the read to the rightmost end of the read for the leftmost seed. Then BRAT-nova calculates $M_{seed2}[1\ldots r]$ from the rightmost end of the read to the leftmost end of the read for the rightmost seed (the formula for M above has to be slightly modified accordingly). Finally, using M_{seed1} and M_{seed2} BRAT-nova calculates $M^*[1\ldots r]$, where $M^*[j]$ is the best score of aligning $R[1\ldots j]$ bases of the read with genomic bases $G[i...i+j-1]$, then having an indel after R[j], and aligning the rest of the read bases.

2. Benchmarking

We benchmarked BRAT-nova against Bismark (release 0.14.3 with Bowtie2), BS-Seeker2 (release V2.0.9 with Bowtie2), BSMAP (release 2.90) and BRAT-BW (release 2.0.1). All experiments were run on a dual ten-core Intel Xeon Processor 3GHz, 512 GB of RAM, 1333MHz DDR3 RAM, and 216 TB of raw storage space running Linux Ubuntu.

The running time was measured using the Linux command *time* that provides *real* and *user* time. *User* time for Bismark and BS-Seeker2 is usually twice as much as *real* time because the tools run two threads simultaneously: one per each of the FM-indexes. RAM space was measured using the Linux command *top.* For all tools and in all experiments, we ran a single thread (Bowtie2 option -p 1). We used output files format SAM for Bismark and BS-Seeker2, and BRAT-nova. BRAT-BW produces a program-specific output format.

Building the FM-index. To benchmark BRAT-nova against state-of-the-art tools Bismark, BS-Seeker2, BSMAP and BRAT-BW, we used human genome GRCh38 (24 chromosomes). We removed contiguous stretches of Ns longer than 49bp (since the minimum length of the reads in BRAT-nova and BS-Seeker2 is 50bp). The resulting genome size was 2,934,896,319 bp. To make sure that comparison was fair, we used similar parameters that control the number of genomic positions stored with FM-index: with Bowtie2 the option *offrate* was kept default, and with BRAT-BW and BRAT-nova option *S* was set to 32 (this means that every 32nd genomic position was stored). The rest of parameters were the default for each tool.

Time and memory required to build the FM-index. BRAT-nova builds a single index on the concatenation of the genome and its reverse-complement (500m, 3.8GB), Bismark builds two indexes in parallel (144m, 5.1·2=10.2GB), BS-Seeker2 builds four indexes in parallel (168m, 5.1·2+4.3+4.7 = 19.2GB), and BRAT-BW builds one index at a time with total of two indexes that users can build simultaneously or separately (255m, 1.9·2=3.8GB).

Scripts and simulated data sets used in benchmarking. All simulated data sets and scripts used in benchmarking are available on BRAT-nova's website listed in the main manuscript. The file README.txt contains a full list of the scripts together with the command lines.

Mapping efficiency on real reads. We evaluated all tools using real human genome reads from dataset SRR306435 (Molaro *et al*., 2011) and dataset SRR306421 (Hodges *et al*., 2011). For datasets of real reads, we first pre-processed the reads by (1) removing adapter sequences with *cutadapt* (Martin, 2011), (2) trimming low-quality ends with BRAT-BW's *trim*, and (3) removing duplicate reads by sorting the reads and keeping one representative of each copy. All remaining 101bp-long reads from SRR306435 mate 1 (about 10.6M) composed the first dataset. For the second and third data sets, from pre-processed reads, we chose 1M paired-end 101bp reads from SRR306435, and 1M single-end 76bp reads from SRR306421 respectively.

We ran each tool using various parameter settings and measured the percentage of uniquely mapped reads (*i.e.,* reads mapped with the highest score to a single location), and the running time. The percentage of uniquely mapped reads and running time can vary significantly depending on the parameter settings. In terms of uniquely mapped reads, BRAT-nova showed a comparable range of performance as Bismark, BSMAP and BS-Seeker2, and it was 2-11 times faster (Supplemental Figure 1). Supplemental Tables 1-3 show detailed information for these experiments. For example, on single-end 101bp reads, the percentage of uniquely mapped reads with indels reported by BS-Seeker2 was 2-5%, by BSMAP was 6.3%, whereas Bismark reported 5-38% and BRAT-nova reported 0-1.3%. In addition, the percentage of uniquely mapped reads aligned by BRAT-nova using local alignment was 2-38% (i.e. reads mapped with trimmed ends due to indels or sequencing errors at the end of the reads, or adapter contamination).

Mapping accuracy evaluation on synthetic reads. On synthetic reads, we measured the *mapping accuracy* as the ratio of uniquely mapped reads aligned within 50bp and 0bp of the original positions (same chromosome, same strand) to the total number of uniquely mapped reads. Three different data sets were used to measure read mapping accuracy: (a) 1M single-end 100bp reads with 5% sequencing error rate and 2% SNPs; (b) 0.5M paired-end 100bp reads with 2% sequencing error rate, 1% of SNPs, 5% of indels, and 10% of adapter contamination; and (c) 1M single-end 76bp reads with 3% sequencing error rate, 1% SNPs, 10% of indels, and 30% of adapter contamination. Sequencing error rate and SNP rate were calculated as a fraction of the total number of bases in the reads, while the percentage of indels and adapter contamination was based on the total number of reads having indels and adapter sequences at the 3'-end. SNPs were introduced using a uniform random distribution. In order to appropriately model Illumina's sequencing error distribution, we used the "per base sequence quality score" graph generated by *cutadapt* for dataset SRR306435. We observed that about 2/3 of the length of the reads starting at 5'-end had a quality score of 40 (corresponding to an error probability of 1 in 10,000), the next 1/6 of the length had a quality score of 30 (corresponding to an error probability of 1 in 1000), and the remaining 1/6 of the length had the poorest score, not worse than 20 (corresponding to an error probability of 1 in 100).

We introduced sequencing errors according to this distribution. Indels of random length (up to 10bp, uniform distribution) were introduced to randomly selected reads and adapter sequences of random length up to 15bp replaced 3'-ends of randomly chosen reads. The bisulfite-conversion rate was set to 97%.

Supplemental Figure 2 and Supplemental Tables 4-6 report the results of the mapping accuracy tests. Observe again that the mapping accuracy can vary significantly depending on the parameter settings. All tools showed a higher mapping accuracy with stricter parameters at the expense of a smaller percentage of reads mapped. For example, on the synthetic data set with 76bp single-end reads (with 3% sequencing error rate, 1% SNPs, 10% of all reads having indels, and 30% of reads having an adapter sequence up to 15bp at the 3'-end), minimum and maximum percentage of uniquely mapped reads (together with the corresponding values of mapping accuracy of type "*within 50bp*" shown in parenthesis) was: for Bismark 34.95%-89.3% (98.64%-94.72%), for BS-Seeker2 78.47%-89.50% (94.34%-88.63%), for BSMAP 77.6%-84.1% (98.74%-98.09%), for BRAT-BW 44.80%-89.40% (98.58%-94.04%), and for BRAT-nova 81.09%-90.12% (97.12%-93.61%). Here, BRAT-nova showed comparable results in terms of the uniquely mapped reads (81.09%-90.12%) and BRAT-nova's performance in terms of mapping accuracy of type "within 50bp" (97.12%-93.61%) lies between BS-Seeker2's performance (lowest values) and BSMAP (highest values). BRAT-nova was 2-11 times faster in this experiment. Overall, in the experiments measured mapping accuracy (over all three data sets), BRAT-nova showed comparable results with other tools in terms of mapping accuracy of type "within 50bp" and of type "within 0bp", and was 2-16 times faster.

Methylation call accuracy and methylation level accuracy evaluation on synthetic reads. *Methylation call accuracy* was measured as the ratio of the cytosines whose methylation status was correctly identified (methylated or not methylated) to the total number of the cytosines covered by at least ten reads (in order to reduce noise, we only considered cytosines covered by at least ten reads). A cytosine was considered to be methylated if it had a methylation level of at least 0.5, and unmethylated otherwise.

To calculate the *methylation level accuracy*, we used the following randomized analysis. We compared the original methylation level (recorded during read generation) to the methylation level calculated by the tools. For each set of cytosines covered by at least ten reads, we used the same total number of Cs and Ts mapped to each cytosine by each tool, but randomly generated the number of Cs mapped to the cytosines. We calculated the error between the randomly-obtained methylation level and original methylation level (this procedure was repeated 100 times); then, we sorted all these errors and considered three thresholds on errors corresponding to 20%, 15% and 10% false discovery rates. For each error threshold, we defined the methylation level accuracy as the percentage of cytosines with methylation level calculated within the corresponding error threshold from the original methylation level. To measure methylation call accuracy and methylation level accuracy, we used two synthetic data sets of the reads generated from chromosome 21 of the human genome GRCh38: (a) 20M single-end 100bp reads with 5% sequencing error rate and 2% of SNPs, and (b) 20M single-end 76bp reads with 5% sequencing error rate and 2% SNPs, and with 10% of indels and 10% of adapter sequences. Initially, each cytosine of chromosome 21 was randomly assigned a methylation level of 0.2, 0.4, 0.5, 0.6, or 0.8. If a generated read covered a cytosine, the methylation status of that cytosine in the read was set randomly according to the methylation level of the genomic cytosine. For example, if the methylation level of a cytosine at a specific position was 0.2, then the methylation status of the corresponding cytosine in a read was set methylated, *i.e.,* C was unchanged to T, with probability of 0.2. To run these experiments, we chose the parameter settings for each tool so that the percentage of uniquely mapped reads across all tools was within 10% of one another (Supplemental Tables 7-8). With the dataset composed of 76bp reads, we used two different parameter settings (strict and loose) to measure how loose parameters affect methylation call and level accuracy. Supplemental Figure 3 and 4 and Tables 7-8 show the results for methylation call and methylation level accuracy tests. BRAT-nova showed comparable results with the other tools. With strict parameters, all tools mapped fewer reads with higher mapping accuracy. However, despite a decreased mapping accuracy with loose parameters, methylation level accuracy was 2-4% higher for all tools compared to strict parameters. In these experiments, BRAT-nova was 2-4 times faster than Bismark, 4-8 times faster than BS-Seeker2, 5 times faster than BSMAP and 2 times faster than BRAT-BW.

Options used with the tools in benchmarking. To make comparisons fair across all tools, in all experiments we ran all tools using one CPU. The exception was tests with indels with BSMAP because alignment of reads with indels on BSMAP is much slower than alignment without indels. Other options that ensured fairness for all the tools were the options controlling minimum and maximum insert size for paired-end reads (option *X* with Bismark and BS-Seeker2, options *i*/*a* with BRAT-bw and BRAT-nova and options *m/x* with BSMAP). For paired-end reads we also used options *no-discordant* and *no-mixed* with BS-Seeker2 (allowing only concordantly aligned pairs in the output).

Here we briefly explain the meaning of different options used with all tools in this benchmarking. BRAT-nova's options used were *K*, *L*, *G*, *l* and *q*. Options L and G allow local alignment and indels respectively (by default, BRAT-nova allows local alignment and indels). The option *K* is the number of shifts in multi-seed mapping, i.e. the total number of seeds in a read used in the alignment (each seed is 8bp apart from the previous seed). *K* was set to 9 for 100bp reads (optimal number, lower *K* might results in fewer reads mapped) and was set to 5 for 76bp reads (also optimal for this length). The options *l* and *q* are alignment length and alignment quality (described above) that affect the alignment score. Default parameters (*l* 30, *q* 90) allow mapping more reads. Setting higher values for *l* and *q* make these parameters *stricter* resulting in fewer reads mapped, but with higher mapping accuracy, but not necessarily with higher methylation call or methylation level accuracy.

With Bismark, we changed options *rdg* and *rfg* that set the read/reference gap open and extend penalties respectively (default 5,3 for both). Bismark tends to give preference to the alignments with indels when options *rdg* and *rfg* are kept default, so we set these options to higher values (11,6) to reinforce alignments with mismatches having preference over alignments with indels so that the results of alignments between Bismark and BRAT-nova are more consistent. The most important option used with Bismark is *score-min L, x, y,* which controls the alignment score, hence, the total number of uniquely mapped reads and affects mapping accuracy and running time. This option sets the minimum alignment score equal to *x* + *y**read-length, where *x* and *y* are decimal numbers (default is L,0,-0.2). The default parameter maps reads with high mapping accuracy, but at the cost of mapping significantly fewer reads. From our experiments, we found that the default setting for this parameter is unnecessarily strict, resulting in mapping fewer reads than with other settings of this parameter that also have high mapping accuracy (see Supplemental Tables 4-6 for different settings). On the other hand, mapping accuracy decreases when this parameter is set too loose. The other two options used were *D* and *R* (default values are 15 and 2 respectively). The definition of these parameters can be found on the Bowtie-2 website; in short, D is the number of consecutive seed extension attempts and R is the maximum number of times Bowtie-w re-seeds reads. These parameters might affect running time (the larger values of D and R, the greater the running time), but too small values might negatively affect mapping accuracy. From our experiments, the best parameter that effectively controls the number of mapped reads and mapping accuracy is *score-min*. By making this parameter less strict than the default setting, one can map more reads with slightly lower mapping accuracy and without much of an effect in running time.

With BS-Seeker2, we changed D and R parameters (described above) and alignment score function *score-min G, x, y* (option L with *score-min* does not work with BS-Seeker2). We kept D and R the same as default values for Bismark. G,*x*,*y* sets alignment score to *x* + *y**ln(read-length), where *x* and *y* are decimal numbers. The performance of BS-Seeker2 is sensitive to this parameter: tuning *x* and *y* helps to map more reads, and the mapping accuracy decreases with more reads mapped, as with Bismark. The other parameters used with BS-Seeker are *ma* (match bonus) and *mp* (mismatch penalty), which were kept to 1 and 3, consistent with the default settings of these parameters of BRAT-nova.

BSMAP has default settings that result in optimal running time; so we changed only option *v*, the number of mismatches allowed and the option that allowed indels (option *g*). BSMAP runs slower when the number of mismatches allowed increases and especially when indels are allowed. With a larger value of *v*, more reads are mapped without much of an effect on mapping accuracy, but the running time increases.

References

Langmead, B. *et al*. (2009) Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol., 10*, R25.

Langmead, B. *et al*. (2012) Fast gapped-read alignment with Bowtie 2. *Nature Methods*, 9, 357-359.

Martin, M. (2011) Cutadapt removes adapter sequences from high-throughput se-quencing reads. *EMBnet*, 17, 10-12.

Smith, T. and Waterman, M. (1981) Identification of Common Molecular Subsequences. *Journal of Molecular Biology*, 147, 195–197.