rHAT: Fast aligning noisy long read with regional hashing

Supplementary Material

Bo Liu1, Dengfeng Guan1, Mingxiang Teng1 and Yadong Wang1

1Center for Bioinformatics, Harbin Institute of Technology, Harbin, Heilongjiang 150001, China

Contents

[Supplementary Fig. 1. A schematic illustration of the extension operation. 2](#_Toc428013366)

[Supplementary Fig. 2. A schematic illustration of the affection of false positive local matches on the sparse dynamic programming (SDP) heuristic. 3](#_Toc428013367)

[Supplementary Fig. 3. A case having large deletion near the beginning of the alignment 4](#_Toc428013368)

[Supplementary Table 1. The real and simulated datasets for evaluation. 5](#_Toc428013369)

[Supplementary Table 2. Evaluation of various aligners on real datasets. 6](#_Toc428013370)

[Supplementary Table 3. Evaluation of various aligners on simulated datasets. 7](#_Toc428013371)

[Supplementary Table 4. Evaluation of rHAT with various parameters on real datasets. 9](#_Toc428013372)

[Supplementary Table 5. Evaluation of rHAT with various settings on the local match length (-l) parameter for simulated datasets. 11](#_Toc428013373)

[Supplementary Table 6. Evaluation of rHAT with various settings on the max allowed window hits (-w) parameter for simulated datasets. 13](#_Toc428013374)

[Supplementary Table 7. Evaluation of rHAT with various settings on the number of candidates (-m) parameter for simulated datasets. 15](#_Toc428013375)

[Supplementary Table 8. Evaluation of rHAT with various settings on the k-mer size (-k) parameter for simulated datasets. 17](#_Toc428013376)

[Supplementary Table 9. Throughputs in multiple threads 19](#_Toc428013377)

[Supplementary Protocol 20](#_Toc428013378)

[1. Availability of real datasets. 20](#_Toc428013379)

[2. Read simulation with PBSim (Verison 1.0.3) 21](#_Toc428013380)

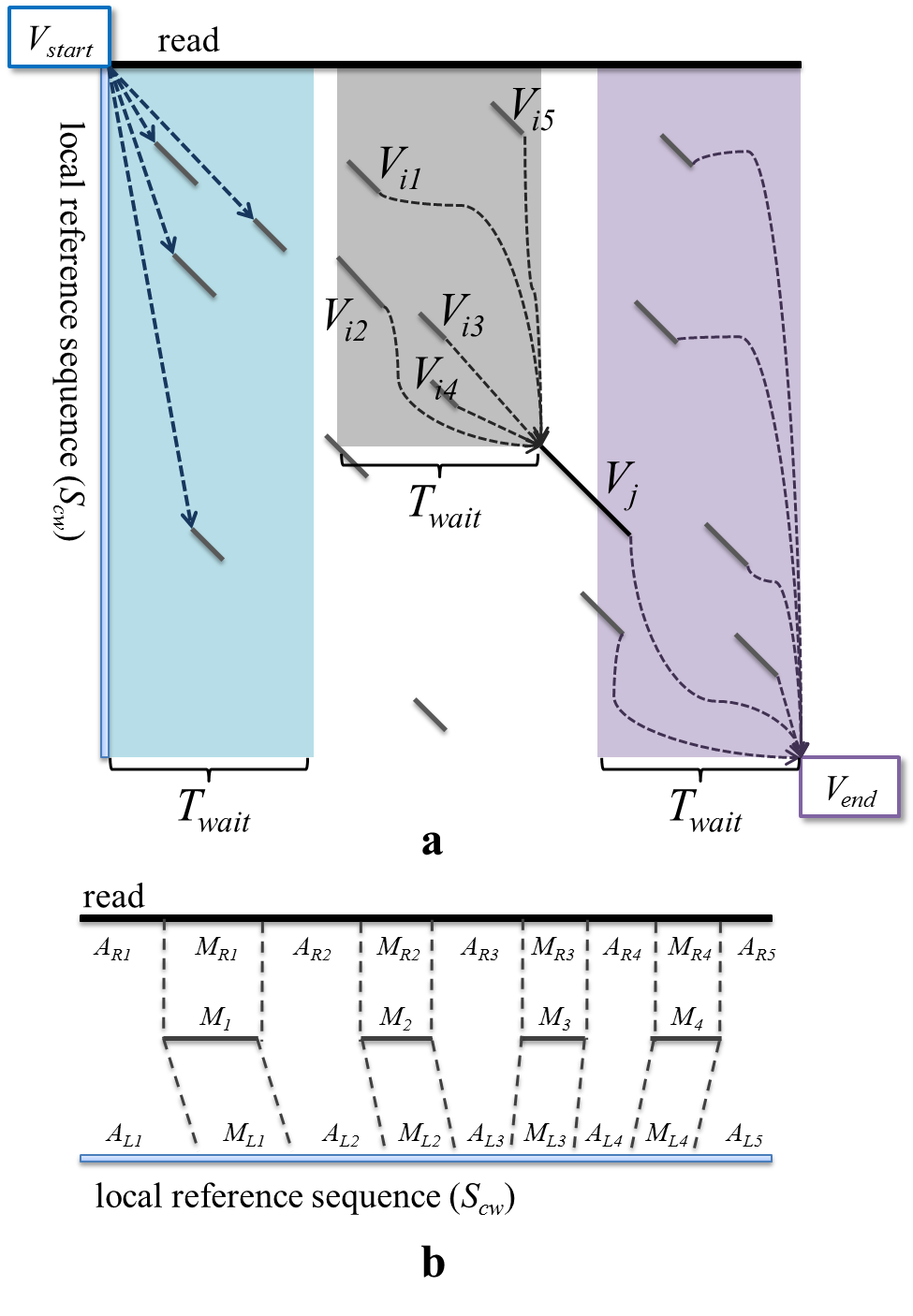
[3. Benchmarking for various aligners 22](#_Toc428013381)

[3.1 Benchmarking for BLASR (version 1.3.1) 22](#_Toc428013382)

[3.2 Benchmarking for BWA-MEM (version 0.7.12) 22](#_Toc428013383)

[3.3 Benchmarking for BWA-SW (version 0.7.12) 23](#_Toc428013384)

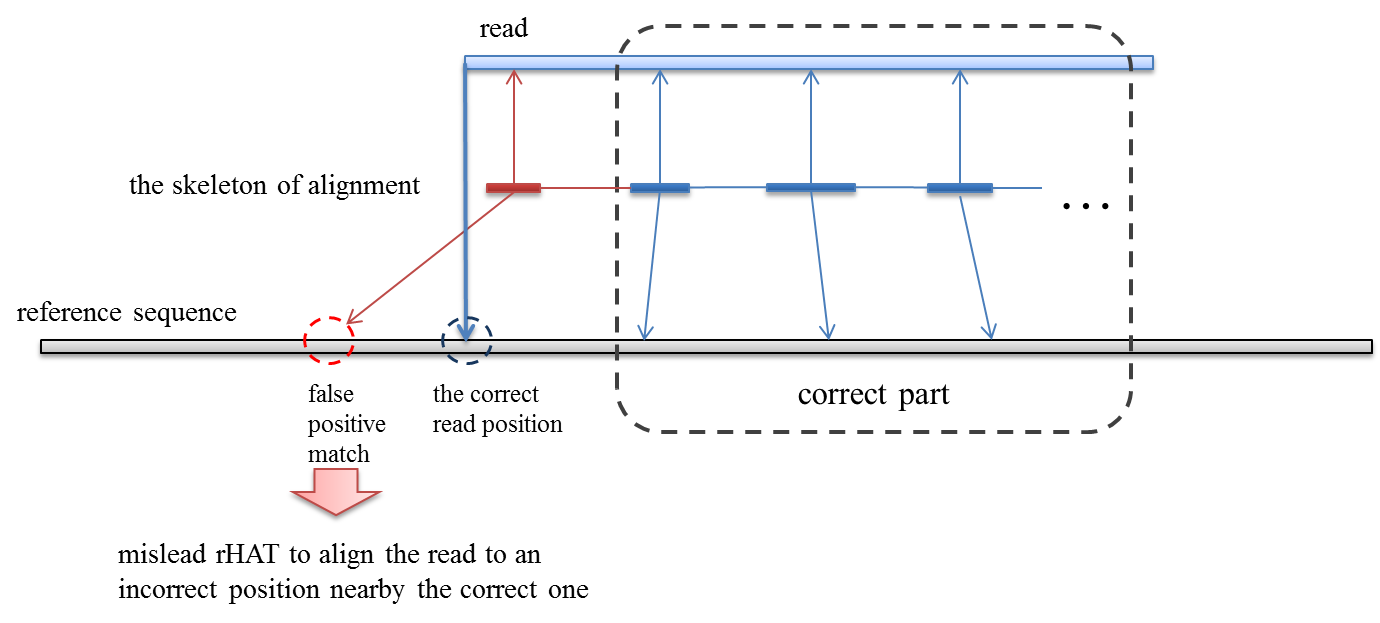
[3.4 Benchmarking for rHAT (version 0.1.1) 23](#_Toc428013385)



# Supplementary Fig. 1. A schematic illustration of the extension operation.

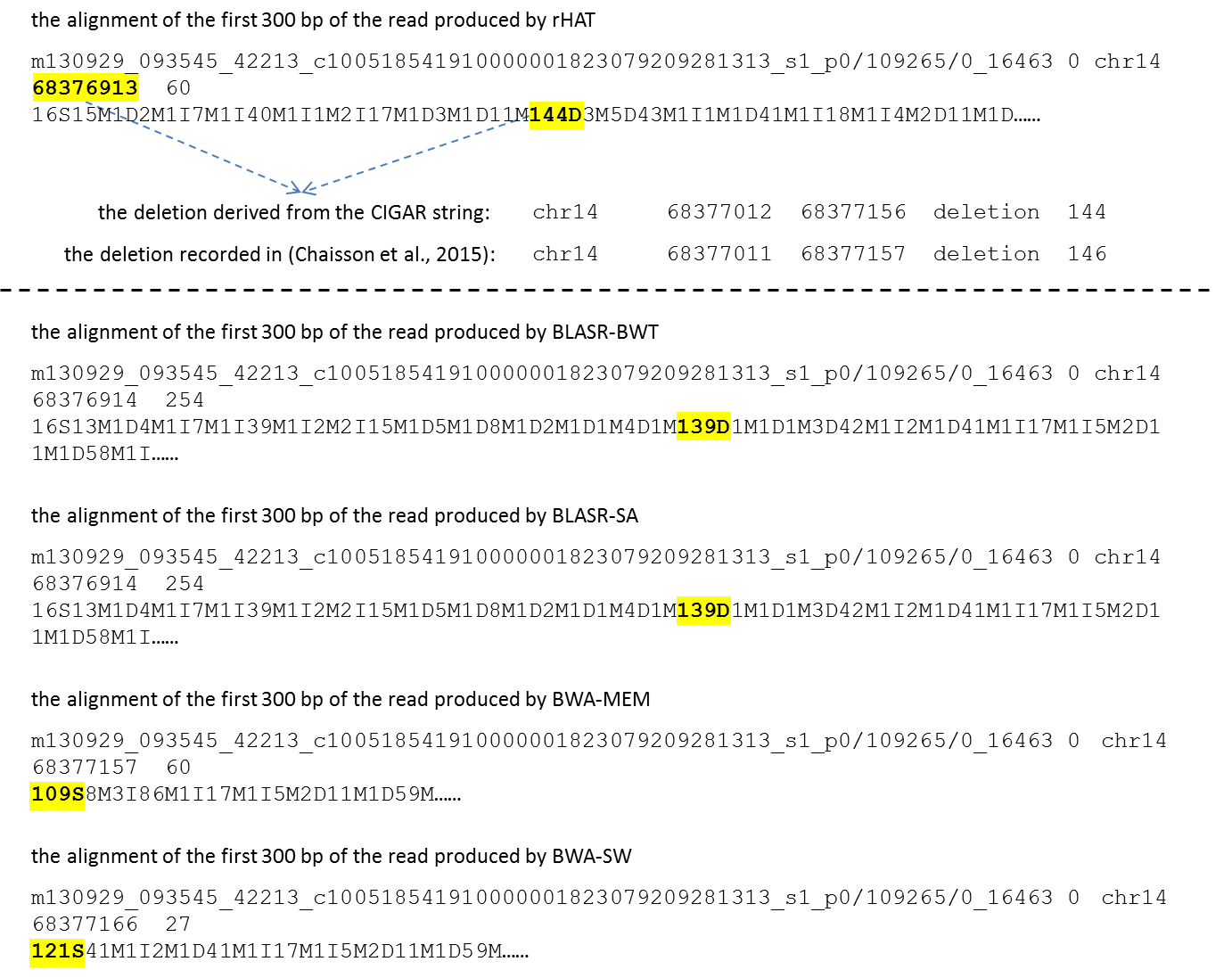
(a) An illustration of the DAG of the local matches used for the sparse dynamic programming (SDP)-based heuristic. The diagonal solid lines indicate the vertices of the DAG, i.e., the local matches between the read and the local reference sequence, and the two rectangles indicate the auxiliary *Vstart* and *Vend* vertices. For a certain vertex, *Vj* (marked by black solid line), all its precursors must be the matches ending within a neighborhood region (the gray area in the figure), which is defined by the connecting condition, i.e., and . In this case, there are five precursors (from *Vi1* to *Vi5* in the figure) for *Vj*. For the two auxiliary vertices, according to the same condition, the vertices starting within the aqua area depict the successors of *Vstart*, and the vertices ending within the purple area depict the precursors of *Vend*.

(b) An illustration of the partition of the read and local reference sequence by the skeleton of alignment. In this case, the skeleton (from *M1* to *M4*) partitions the read and the reference into 9 paired segments. Each pair of unaligned segments, i.e., (*ARi*, *ALi*), i=1, 2, 3, 4, 5, will be aligned to compose the whole read alignment.



# Supplementary Fig. 2. A schematic illustration of the affection of false positive local matches on the sparse dynamic programming (SDP) heuristic.

In the SDP heuristic, some vertices of the DAG could be false positive local matches. Such vertices may be mistakenly involved in the skeleton, since they can increase the total number of matched bases. The figure is an illustration about this issue. A false positive local match near the left end of the read (red bars) is involved in the skeleton due to the overmatch of rHAT. In this case, the read will be mapped to an incorrect position nearby the correct one. However, most parts of the skeleton can still be correct, since the SDP process can better prune the false positive local matches of the short tokens in the inner part of the read. Thus, the errors will only affect a portion of the bases, and other bases away from the false positive match can still be appropriately aligned (the “correct part” in the figure).



# Supplementary Fig. 3. A case having large deletion near the beginning of the alignment

The SAM record of the read (“m130929\_093545\_42213\_c1005185419100000018230792

09281313\_s1\_ p0/109265/0\_16463”, 16463 bp long) generated by rHAT indicates that it is aligned to the position “chr14, 68376913”. The CIGAR string corresponding to the first 300 bp of the read is shown in the figure. The “144D” in the CIGAR string indicates there is a 144 bp deletion in the beginning part of the alignment of the read. The interval of the deletion can be derived as [68377012, 68377156] of Chromosome 14. Referring to the results proposed in the study of (Chaisson et al., 2015), a similar deletion record, “chr14 68377011 68377157 deletion 146”, can be found, which suggested the alignment of this read is reasonable. BLASR also produced similar alignments with either SA or BWT-based index. However, both of BWA-MEM and BWA-SW clipped the read, i.e., over 100 bases of the read were not aligned.

# Supplementary Table 1. The real and simulated datasets for evaluation.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **no.** | **name** | **type** | **platform** | **# of Reads** | **# of Bases** | **reference genome** |
| 1 | *H. sapiens*-reala | real | PacBio RS II P5/C3 release | 290992 | 1810943188 | hg19 |
| 2 | *D. melanogaster*-reala | real | PacBio RS II P5/C3 release | 123234 | 1244028123 | DM5 |
| 3 | *E. coli*-simb | simulated | PBSim | 623 | 4938920 | *E.coli* Strain 536 |
| 4 | *S. cerevisiae*-simb | simulated | PBSim | 1520 | 12153653 | sacCer3 |
| 5 | *D. melanogaster*-simb | simulated | PBSim | 16291 | 129738789 | DM3 |
| 6 | *A. thaliana*-simb | simulated | PBSim | 14842 | 118558112 | TAIR10 |
| 7 | *H. sapiens*-simb | simulated | PBSim | 357226 | 2856372869 | hg19 |

a) Refer to **Supplementary Protocol** for the availability of real datasets.

b) Refer to **Supplementary Protocol** for the command lines of PBSim for simulating the reads.

# Supplementary Table 2. Evaluation of various aligners on real datasets.

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **aligner** | **parametersa** | **# of unaligned readsb** | **aligned basesc** | | **80% consecutively aligned readsd** | | **85% consecutively aligned readsd** | | **90% consecutively aligned readsd** | | **95% consecutively aligned readsd** | | **throughpute**  **(wall time)** |
| **#** | **%** | **#** | **%** | **#** | **%** | **#** | **%** | **#** | **%** |
| ***H. sapiens-real* dataset (4 SMRT cells, 290992 reads, 1810943188 bases)** | | | | | | | | | | | | | |
| BLASR | BWT | 6285 | 1682871435 | 92.93% | 250595 | 86.12% | 244212 | 83.92% | 234370 | 80.54% | 213940 | 73.52% | 86743 |
| BLASR | SA | 6110 | 1685814597 | 93.09% | 251516 | 86.43% | 243047 | 83.52% | 236684 | 81.34% | 219355 | 75.38% | 101629 |
| BWA-MEM | -x pacbio | 9517 | 1733371248 | 95.72% | 244986 | 84.19% | 240174 | 82.54% | 234231 | 80.49% | 224988 | 77.32% | 44660 |
| BWA-SW | −b5 -q2 -r1 -z20 | 7526 | 1616353770 | 89.25% | 70620 | 24.27% | 62991 | 21.65% | 55347 | 19.02% | 46450 | 15.96% | 137191 |
| rHAT | -l 11, -w 1000, -m 5, -k 13 | 220 | 1779033318 | 98.24% | 267931 | 92.08% | 265992 | 91.41% | 262822 | 90.32% | 255695 | 87.87% | 25259 |
| ***D. melanogaster-real* dataset (4 SMRT cells, 123234 reads, 1244028123 bases)** | | | | | | | | | | | | | |
| BLASR | BWT | 3685 | 1122398735 | 90.22% | 100503 | 81.55% | 98654 | 80.05% | 95922 | 77.84% | 90870 | 73.74% | 133850 |
| BLASR | SA | 3601 | 1123831639 | 90.34% | 100739 | 81.75% | 98954 | 80.30% | 96489 | 78.30% | 92160 | 74.78% | 67291 |
| BWA-MEM | -x Pacbio | 1902 | 1207026172 | 97.03% | 96965 | 78.68% | 95223 | 77.27% | 93232 | 75.65% | 90403 | 73.36% | 59993 |
| BWA-SW | −b5 -q2 -r1 -z20 | 1133 | 1176672298 | 94.59% | 11647 | 9.45% | 10147 | 8.23% | 8577 | 6.96% | 6704 | 5.44% | 1663047 |
| rHAT | -l 11, -w 1000, -m 5, -k 13 | 150 | 1227958298 | 98.71% | 102502 | 83.18% | 101750 | 82.57% | 100610 | 81.64% | 97907 | 79.45% | 8470 |

a) Refer to **Supplementary Protocol** for the command lines of various aligners for alignment.

b) A read is considered as unaligned if it is not covered by any alignment.

c) A base is considered as aligned only if it is covered by at least one alignment of the corresponding read.

d) A read is considered as “x% consecutively aligned” (x = 80, 85, 90, 95) only if there is at least one alignment which can cover at least x% of the bases of the read.

e) The throughput is defined as the wall time of the alignment (in seconds).

# Supplementary Table 3. Evaluation of various aligners on simulated datasets.

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **aligner** | **parametersa** | **# of unaligned readsb** | **aligned basesc** | | **correctly aligned readsd** | | **bases covered by position-correct alignmentse** | | **throughputf**  **(wall time)** |
| **#** | **%** | **#** | **%** | **#** | **%** |
| ***E. coli-sim* dataset (623 reads, 4938920 bases)** | | | | | | | | | |
| BLASR | BWT | 0 | 4922428 | 99.67% | 599 | 96.15% | 4728811 | 95.75% | 57 |
| BLASR | SA | 0 | 4931452 | 99.85% | 609 | 97.75% | 4824688 | 97.69% | 65 |
| BWA-MEM | -x pacbio | 0 | 4934344 | 99.91% | 622 | 99.84% | 4926709 | 99.75% | 186 |
| BWA-SW | −b5 -q2 -r1 -z20 | 0 | 4906428 | 99.34% | 0 | 0.00% | 4856812 | 98.34% | 78 |
| rHAT | -l 11, -w 1000, -m 5, -k 13 | 0 | 4935507 | 99.93% | 621 | 99.68% | 4927695 | 99.77% | 4 |
| ***S. cerevisiae-sim* dataset (1520 reads, 12153653 bases)** | | | | | | | | | |
| BLASR | BWT | 0 | 12116104 | 99.69% | 1475 | 97.04% | 11753001 | 96.70% | 192 |
| BLASR | SA | 0 | 12137124 | 99.86% | 1489 | 97.96% | 11900539 | 97.92% | 237 |
| BWA-MEM | -x pacbio | 0 | 12139901 | 99.89% | 1507 | 99.14% | 12044936 | 99.11% | 441 |
| BWA-SW | −b5 -q2 -r1 -z20 | 0 | 12055024 | 99.19% | 0 | 0.00% | 11729142 | 96.51% | 761 |
| rHAT | -l 11, -w 1000, -m 5, -k 13 | 0 | 12148592 | 99.96% | 1510 | 99.34% | 12087668 | 99.46% | 15 |
| ***D. melanogaster-sim* dataset (16291 reads, 129738789 bases)** | | | | | | | | | |
| BLASR | BWT | 1 | 129308789 | 99.67% | 15787 | 96.91% | 125421486 | 96.67% | 4611 |
| BLASR | SA | 1 | 129519341 | 99.83% | 15799 | 96.98% | 125694916 | 96.88% | 2941 |
| BWA-MEM | -x pacbio | 0 | 129560120 | 99.86% | 16166 | 99.23% | 128722062 | 99.22% | 5616 |
| BWA-SW | −b5 -q2 -r1 -z20 | 0 | 128449045 | 99.01% | 2 | 0.01% | 121157735 | 93.39% | 11149 |
| rHAT | -l 11, -w 1000, -m 5, -k 13 | 0 | 129655492 | 99.94% | 15963 | 97.99% | 127791474 | 98.50% | 203 |
| ***A. thaliana-sim* dataset (14842 reads, 118558112 bases)** | | | | | | | | | |
| BLASRg | BWT | 13662 | 9351107 | 7.89% | 1135 | 7.65% | 8985722 | 7.58% | N/A |
| BLASR | SA | 0 | 118354960 | 99.83% | 14381 | 96.89% | 114707111 | 96.75% | 2572 |
| BWA-MEM | -x pacbio | 0 | 118189490 | 99.69% | 14401 | 97.03% | 115352067 | 97.30% | 4181 |
| BWA-SW | −b5 -q2 -r1 -z20 | 0 | 117501295 | 99.11% | 0 | 0.00% | 115460198 | 97.39% | 7965 |
| rHAT | -l 11, -w 1000, -m 5, -k 13 | 0 | 118535018 | 99.98% | 14768 | 99.50% | 117951707 | 99.49% | 159 |
| ***H.sapiens-sim* dataset (357226 reads, 2856372869 bases)** | | | | | | | | | |
| BLASR | BWT | 0 | 2846018934 | 99.64% | 345267 | 96.65% | 2751461026 | 96.33% | 79041 |
| BLASR | SA | 0 | 2849922539 | 99.77% | 343784 | 96.24% | 2743245952 | 96.04% | 111169 |
| BWA-MEM | -x pacbio | 0 | 2848196757 | 99.71% | 350106 | 98.01% | 2805323377 | 98.21% | 111451 |
| BWA-SW | −b5 -q2 -r1 -z20 | 0 | 2798414909 | 97.97% | 1 | 0.00% | 2726246875 | 95.44% | 297301 |
| rHAT | -l 11, -w 1000, -m 5, -k 13 | 0 | 2855626329 | 99.97% | 354546 | 99.25% | 2837879055 | 99.35% | 17172 |
| **Average statisticsh** | | | | | | | | | |
| BLASR | BWT | -- | -- | 99.67% | -- | 96.69% | -- | 96.36% | -- |
| BLASR | SA | -- | -- | 99.83% | -- | 97.16% | -- | 97.06% | -- |
| BWA-MEM | -x pacbio | -- | -- | 99.81% | -- | 98.65% | -- | 98.72% | -- |
| BWA-SW | −b5 -q2 -r1 -z20 | -- | -- | 98.92% | -- | 0.00% | -- | 96.21% | -- |
| rHAT | -l 11, -w 1000, -m 5, -k 13 | -- | -- | 99.96% | -- | 99.15% | -- | 99.31% | -- |

a) Also refer to **Supplementary Protocol** for the command lines of various aligners for alignment.

b) A read is considered as unaligned if it is not covered by any alignment.

c) A base is considered as aligned only if it is covered by at least one alignment of the corresponding read.

d) A read is considered as correctly aligned only if there is an alignment covering at least 80% of the bases of the read, and its leftmost position is within 50 bps of the grand truth deducting the clipped parts.

e) A base is considered as covered by position-correct alignment only if it is covered by at least one alignment whose leftmost position is within 50 bps of the grand truth deducting the clipped parts, regardless of the number of bases the alignment covered.

f) The throughput is defined as the wall time of the alignment (in seconds).

g) When running BLASR on the *A. thaliana*-sim dataset with the BWT-based index, the program corrupted. We recorded the alignments it generated, and evaluated the partial result.

h) The average proportions of “aligned bases”, “correctly aligned reads” and “bases covered by position-correct alignments” for multiple datasets. For BLASR-BWT, the values are calculated based on the four datasets other than “*A. thaliana*-sim”, as it corrupts on that dataset. For other aligners, the values are calculated by averaging the corresponding statistics on all the five datasets.

# Supplementary Table 4. Evaluation of rHAT with various parameters on real datasets.

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **aligner** | **parametersa** | **# of unaligned readsb** | **aligned basesc** | | **80% consecutively aligned readsd** | | **85% consecutively aligned readsd** | | **90% consecutively aligned readsd** | | **95% consecutively aligned readsd** | | **throughpute**  **(wall time)** |
| **#** | **%** | **#** | **%** | **#** | **%** | **#** | **%** | **#** | **%** |
| ***H. sapiens-real* dataset (4 SMRT cells, 290992 reads, 1810943188 bases)** | | | | | | | | | | | | | |
| rHAT | -l 11, -w 1000, -m 5, -k 13  **(default setting)** | 220 | 1779033318 | 98.24% | 267931 | 92.08% | 265992 | 91.41% | 262822 | 90.32% | 255695 | 87.87% | 25259 |
| rHAT | **-l 8**, -w 1000, -m 5, -k 13 | 14 | 1804752959 | 99.66% | 289119 | 99.36% | 288492 | 99.14% | 287206 | 98.70% | 282431 | 97.06% | 27320 |
| rHAT | **-l 9**, -w 1000, -m 5, -k 13 | 102 | 1785636329 | 98.60% | 278674 | 95.77% | 277441 | 95.34% | 274939 | 94.48% | 267823 | 92.04% | 30142 |
| rHAT | **-l 10**, -w 1000, -m 5, -k 13 | 242 | 1776927821 | 98.12% | 270118 | 92.83% | 268390 | 92.23% | 265375 | 91.20% | 258409 | 88.80% | 26630 |
| rHAT | -l 11, **-w 500**, -m 5, -k 13 | 243 | 1778390522 | 98.20% | 267668 | 91.98% | 265699 | 91.31% | 262553 | 90.23% | 255439 | 87.78% | 22667 |
| rHAT | -l 11, **-w 2000**, -m 5, -k 13 | 218 | 1779438029 | 98.26% | 268058 | 92.12% | 266101 | 91.45% | 262941 | 90.36% | 255779 | 87.90% | 28192 |
| rHAT | -l 11, -w 1000, **-m 1**, -k 13 | 694 | 1759551737 | 97.16% | 263775 | 90.65% | 261966 | 90.03% | 258815 | 88.94% | 251200 | 86.33% | 21296 |
| rHAT | -l 11, -w 1000, **-m 2**, -k 13 | 533 | 1769920210 | 97.73% | 265499 | 91.24% | 263661 | 90.61% | 260578 | 89.55% | 253295 | 87.05% | 22456 |
| rHAT | -l 11, -w 1000, **-m 10**, -k 13 | 99 | 1784039808 | 98.51% | 269186 | 92.51% | 267118 | 91.80% | 263849 | 90.67% | 256665 | 88.20% | 30578 |
| rHAT | -l 11, -w 1000, -m 5, **-k 10** | 8499 | 1533461706 | 84.68% | 175821 | 60.42% | 173002 | 59.45% | 169873 | 58.38% | 165275 | 56.80% | 22384 |
| rHAT | -l 11, -w 1000, -m 5, **-k 11** | 249 | 1775233679 | 98.03% | 264502 | 90.90% | 262436 | 90.19% | 259321 | 89.12% | 252498 | 86.77% | 42217 |
| rHAT | -l 11, -w 1000, -m 5, **-k 12** | 126 | 1781403558 | 98.37% | 268187 | 92.16% | 266239 | 91.49% | 263118 | 90.42% | 255916 | 87.95% | 51093 |
| rHAT | -l 11, -w 1000, -m 5, **-k 14** | 318 | 1776393952 | 98.09% | 267136 | 91.80% | 265183 | 91.13% | 262047 | 90.05% | 254966 | 87.62% | 14756 |
| rHAT | -l 11, -w 1000, -m 5, **-k 15** | 416 | 1773491543 | 97.93% | 266050 | 91.43% | 264085 | 90.75% | 260942 | 89.67% | 253963 | 87.27% | 11997 |
| ***D. melanogaster-real* dataset (4 SMRT cells, 123234 reads, 1244028123 bases)** | | | | | | | | | | | | | |
| rHAT | -l 11, -w 1000, -m 5, -k 13  **(default setting)** | 150 | 1227958298 | 98.71% | 102502 | 83.18% | 101750 | 82.57% | 100610 | 81.64% | 97907 | 79.45% | 8470 |
| HAT | **-l 8**, -w 1000, -m 5, -k 13 | 6 | 1236286592 | 99.38% | 117511 | 95.36% | 117317 | 95.20% | 116856 | 94.82% | 114842 | 93.19% | 13832 |
| rHAT | **-l 9**, -w 1000, -m 5, -k 13 | 150 | 1226387737 | 98.58% | 106393 | 86.33% | 105935 | 85.96% | 104997 | 85.20% | 102273 | 82.99% | 11349 |
| rHAT | **-l 10**, -w 1000, -m 5, -k 13 | 183 | 1225922608 | 98.54% | 102978 | 83.56% | 102326 | 83.03% | 101305 | 82.21% | 98696 | 80.09% | 9683 |
| rHAT | -l 11, **-w 500**, -m 5, -k 13 | 148 | 1227939053 | 98.71% | 102413 | 83.10% | 101641 | 82.48% | 100491 | 81.54% | 97779 | 79.34% | 7928 |
| rHAT | -l 11, **-w 2000**, -m 5, -k 13 | 146 | 1227942126 | 98.71% | 102511 | 83.18% | 101760 | 82.57% | 100632 | 81.66% | 97939 | 79.47% | 8557 |
| rHAT | -l 11, -w 1000, **-m 1**, -k 13 | 306 | 1207312533 | 97.05% | 98098 | 79.60% | 97341 | 78.99% | 96193 | 78.06% | 93691 | 76.03% | 3777 |
| rHAT | -l 11, -w 1000, **-m 2**, -k 13 | 242 | 1221611784 | 98.20% | 100320 | 81.41% | 99582 | 80.81% | 98408 | 79.85% | 95745 | 77.69% | 4951 |
| rHAT | -l 11, -w 1000, **-m 10**, -k 13 | 103 | 1230271824 | 98.89% | 103776 | 84.21% | 103061 | 83.63% | 101959 | 82.74% | 99290 | 80.57% | 13473 |
| rHAT | -l 11, -w 1000, -m 5, **-k 10** | 25 | 1231717050 | 99.01% | 103012 | 83.59% | 102274 | 82.99% | 101193 | 82.11% | 98566 | 79.98% | 26960 |
| rHAT | -l 11, -w 1000, -m 5, **-k 11** | 33 | 1231119860 | 98.96% | 103059 | 83.63% | 102338 | 83.04% | 101208 | 82.13% | 98569 | 79.99% | 16511 |
| rHAT | -l 11, -w 1000, -m 5, **-k 12** | 96 | 1229333696 | 98.82% | 102751 | 83.38% | 102015 | 82.78% | 100883 | 81.86% | 98214 | 79.70% | 10916 |
| rHAT | -l 11, -w 1000, -m 5, **-k 14** | 185 | 1226958980 | 98.63% | 102247 | 82.97% | 101499 | 82.36% | 100364 | 81.44% | 97672 | 79.26% | 7884 |
| rHAT | -l 11, -w 1000, -m 5, **-k 15** | 215 | 1225894670 | 98.54% | 101958 | 82.74% | 101214 | 82.13% | 100063 | 81.20% | 97410 | 79.04% | 7755 |

a) Also refer to **Supplementary Protocol** for the command lines of various aligners for alignment.

b) A read is considered as unaligned if it is not covered by any alignment.

c) A base is considered as aligned only if it is covered by at least one alignment of the corresponding read.

d) A read is considered as “x% consecutively aligned” (x = 80, 85, 90, 95) only if there is at least one alignment which can cover at least x% of the bases of the read.

e) The throughput is defined as the wall time of the alignment (in seconds).

# Supplementary Table 5. Evaluation of rHAT with various settings on the local match length (-l) parameter for simulated datasets.

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **aligner** | **parametersa** | **# of unaligned readsb** | **aligned basesc** | | **correctly aligned readsd** | | **bases covered by position-correct alignmentse** | | **throughputf**  **(wall time)** |
| **#** | **%** | **#** | **%** | **#** | **%** |
| ***E. coli-sim* dataset (623 reads, 4938920 bases)** | | | | | | | | | |
| rHAT | **-l 8**, -w 1000, -m 5, -k 13 | 0 | 4937379 | 99.97% | 617 | 99.04% | 4895389 | 99.12% | 8 |
| rHAT | **-l 9**, -w 1000, -m 5, -k 13 | 0 | 4931301 | 99.85% | 616 | 98.88% | 4893684 | 99.08% | 7 |
| rHAT | **-l 10**, -w 1000, -m 5, -k 13 | 0 | 4935317 | 99.93% | 621 | 99.68% | 4926481 | 99.75% | 8 |
| rHAT | **-l 11**, -w 1000, -m 5, -k 13 | 0 | 4935507 | 99.93% | 621 | 99.68% | 4927695 | 99.77% | 4 |
| ***S. cerevisiae-sim* dataset (1520 reads, 12153653 bases)** | | | | | | | | | |
| rHAT | **-l 8**, -w 1000, -m 5, -k 13 | 0 | 12150325 | 99.97% | 1484 | 97.63% | 11835147 | 97.38% | 31 |
| rHAT | **-l 9**, -w 1000, -m 5, -k 13 | 0 | 12149608 | 99.97% | 1503 | 98.88% | 12016956 | 98.88% | 18 |
| rHAT | **-l 10**, -w 1000, -m 5, -k 13 | 0 | 12148561 | 99.96% | 1509 | 99.28% | 12083283 | 99.42% | 15 |
| rHAT | **-l 11**, -w 1000, -m 5, -k 13 | 0 | 12148592 | 99.96% | 1510 | 99.34% | 12087668 | 99.46% | 15 |
| ***D. melanogaster-sim* dataset (16291 reads, 129738789 bases)** | | | | | | | | | |
| rHAT | **-l 8**, -w 1000, -m 5, -k 13 | 0 | 129696695 | 99.97% | 15716 | 96.47% | 125081377 | 96.41% | 287 |
| rHAT | **-l 9**, -w 1000, -m 5, -k 13 | 0 | 129666842 | 99.94% | 15865 | 97.39% | 126601906 | 97.58% | 206 |
| rHAT | **-l 10**, -w 1000, -m 5, -k 13 | 0 | 129663238 | 99.94% | 15936 | 97.82% | 127560936 | 98.32% | 195 |
| rHAT | **-l 11**, -w 1000, -m 5, -k 13 | 0 | 129655492 | 99.94% | 15963 | 97.99% | 127791474 | 98.50% | 203 |
| ***A. thaliana-sim* dataset (14842 reads, 118558112 bases)** | | | | | | | | | |
| rHAT | **-l 8**, -w 1000, -m 5, -k 13 | 0 | 118528657 | 99.98% | 14430 | 97.22% | 115282845 | 97.24% | 259 |
| rHAT | **-l 9**, -w 1000, -m 5, -k 13 | 0 | 118531285 | 99.98% | 14636 | 98.61% | 116899854 | 98.60% | 176 |
| rHAT | **-l 10**, -w 1000, -m 5, -k 13 | 0 | 118533625 | 99.98% | 14711 | 99.12% | 117449655 | 99.07% | 163 |
| rHAT | **-l 11**, -w 1000, -m 5, -k 13 | 0 | 118535018 | 99.98% | 14768 | 99.50% | 117951707 | 99.49% | 159 |
| ***H.sapiens-sim* dataset (357226 reads, 2856372869 bases)** | | | | | | | | | |
| rHAT | **-l 8**, -w 1000, -m 5, -k 13 | 0 | 2855642567 | 99.97% | 347237 | 97.20% | 2775616976 | 97.17% | 19361 |
| rHAT | **-l 9**, -w 1000, -m 5, -k 13 | 0 | 2855580764 | 99.97% | 351473 | 98.39% | 2810839210 | 98.41% | 17584 |
| rHAT | **-l 10**, -w 1000, -m 5, -k 13 | 0 | 2855556784 | 99.97% | 353444 | 98.94% | 2828539377 | 99.03% | 17366 |
| rHAT | **-l 11**, -w 1000, -m 5, -k 13 | 0 | 2855626329 | 99.97% | 354546 | 99.25% | 2837879055 | 99.35% | 17172 |

a) Also refer to **Supplementary Protocol** for the command lines of various aligners for alignment.

b) A read is considered as unaligned if it is not covered by any alignment.

c) A base is considered as aligned only if it is covered by at least one alignment of the corresponding read.

d) A read is considered as correctly aligned only if there is an alignment covering at least 80% of the bases of the read, and its leftmost position is within 50 bps of the grand truth deducting the clipped parts.

e) A base is considered as covered by position-correct alignment only if it is covered by at least one alignment whose leftmost position is within 50 bps of the grand truth deducting the clipped parts, regardless of the number of bases the alignment covered.

f) The throughput is defined as the wall time of the alignment (in seconds).

# Supplementary Table 6. Evaluation of rHAT with various settings on the max allowed window hits (-w) parameter for simulated datasets.

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **aligner** | **parametersa** | **# of unaligned readsb** | **aligned basesc** | | **correctly aligned readsd** | | **bases covered by position-correct alignmentse** | | **throughputf**  **(wall time)** |
| **#** | **%** | **#** | **%** | **#** | **%** |
| ***E. coli-sim* dataset (623 reads, 4938920 bases)** | | | | | | | | | |
| rHAT | -l 11, **-w 500**, -m 5, -k 13 | 0 | 4935507 | 99.93% | 621 | 99.68% | 4927695 | 99.77% | 4 |
| rHAT | -l 11, **-w 1000**, -m 5, -k 13 | 0 | 4935507 | 99.93% | 621 | 99.68% | 4927695 | 99.77% | 4 |
| rHAT | -l 11, **-w 2000**, -m 5, -k 13 | 0 | 4935507 | 99.93% | 621 | 99.68% | 4927695 | 99.77% | 4 |
| ***S. cerevisiae-sim* dataset (1520 reads, 12153653 bases)** | | | | | | | | | |
| rHAT | -l 11, **-w 500**, -m 5, -k 13 | 0 | 12148592 | 99.96% | 1510 | 99.34% | 12087668 | 99.46% | 11 |
| rHAT | -l 11, **-w 1000**, -m 5, -k 13 | 0 | 12148592 | 99.96% | 1510 | 99.34% | 12087668 | 99.46% | 15 |
| rHAT | -l 11, **-w 2000**, -m 5, -k 13 | 0 | 12148592 | 99.96% | 1510 | 99.34% | 12087668 | 99.46% | 11 |
| ***D. melanogaster-sim* dataset (16291 reads, 129738789 bases)** | | | | | | | | | |
| rHAT | -l 11, **-w 500**, -m 5, -k 13 | 0 | 129655495 | 99.94% | 15963 | 97.99% | 127793350 | 98.50% | 187 |
| rHAT | -l 11, **-w 1000**, -m 5, -k 13 | 0 | 129655492 | 99.94% | 15963 | 97.99% | 127791474 | 98.50% | 203 |
| rHAT | -l 11, **-w 2000**, -m 5, -k 13 | 0 | 129657812 | 99.94% | 15965 | 98.00% | 127817736 | 98.52% | 189 |
| ***A. thaliana-sim* dataset (14842 reads, 118558112 bases)** | | | | | | | | | |
| rHAT | -l 11, **-w 500**, -m 5, -k 13 | 0 | 118535018 | 99.98% | 14768 | 99.50% | 117951707 | 99.49% | 151 |
| rHAT | -l 11, **-w 1000**, -m 5, -k 13 | 0 | 118535018 | 99.98% | 14768 | 99.50% | 117951707 | 99.49% | 159 |
| rHAT | -l 11, **-w 2000**, -m 5, -k 13 | 0 | 118535018 | 99.98% | 14768 | 99.50% | 117951707 | 99.49% | 159 |
| ***H.sapiens-sim* dataset (357226 reads, 2856372869 bases)** | | | | | | | | | |
| rHAT | -l 11, **-w 500**, -m 5, -k 13 | 0 | 2855529983 | 99.97% | 354247 | 99.17% | 2836738697 | 99.31% | 14907 |
| rHAT | -l 11, **-w 1000**, -m 5, -k 13 | 0 | 2855626329 | 99.97% | 354546 | 99.25% | 2837879055 | 99.35% | 17172 |
| rHAT | -l 11, **-w 2000**, -m 5, -k 13 | 0 | 2855671237 | 99.98% | 354750 | 99.31% | 2838642935 | 99.38% | 19916 |

a) Also refer to **Supplementary Protocol** for the command lines of various aligners for alignment.

b) A read is considered as unaligned if it is not covered by any alignment.

c) A base is considered as aligned only if it is covered by at least one alignment of the corresponding read.

d) A read is considered as correctly aligned only if there is an alignment covering at least 80% of the bases of the read, and its leftmost position is within 50 bps of the grand truth deducting the clipped parts.

e) A base is considered as covered by position-correct alignment only if it is covered by at least one alignment whose leftmost position is within 50 bps of the grand truth deducting the clipped parts, regardless of the number of bases the alignment covered.

f) The throughput is defined as the wall time of the alignment (in seconds).

# Supplementary Table 7. Evaluation of rHAT with various settings on the number of candidates (-m) parameter for simulated datasets.

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **aligner** | **parametersa** | **# of unaligned readsb** | **aligned basesc** | | **correctly aligned readsd** | | **bases covered by position-correct alignmentse** | | **throughputf**  **(wall time)** |
| **#** | **%** | **#** | **%** | **#** | **%** |
| ***E. coli-sim* dataset (623 reads, 4938920 bases)** | | | | | | | | | |
| rHAT | -l 11, -w 1000, **-m 1**, -k 13 | 0 | 4934447 | 99.91% | 617 | 99.04% | 4908592 | 99.39% | 2 |
| rHAT | -l 11, -w 1000, **-m 2**, -k 13 | 0 | 4934457 | 99.91% | 618 | 99.20% | 4914539 | 99.51% | 4 |
| rHAT | -l 11, -w 1000, **-m 5**, -k 13 | 0 | 4935507 | 99.93% | 621 | 99.68% | 4927695 | 99.77% | 4 |
| rHAT | -l 11, -w 1000, **-m 10**, -k 13 | 0 | 4935511 | 99.93% | 621 | 99.68% | 4927699 | 99.77% | 6 |
| ***S. cerevisiae-sim* dataset (1520 reads, 12153653 bases)** | | | | | | | | | |
| rHAT | -l 11, -w 1000, **-m 1**, -k 13 | 0 | 12135762 | 99.85% | 1478 | 97.24% | 11896749 | 97.89% | 7 |
| rHAT | -l 11, -w 1000, **-m 2**, -k 13 | 0 | 12144457 | 99.92% | 1494 | 98.29% | 12013128 | 98.84% | 7 |
| rHAT | -l 11, -w 1000, **-m 5**, -k 13 | 0 | 12148592 | 99.96% | 1510 | 99.34% | 12087668 | 99.46% | 15 |
| rHAT | -l 11, -w 1000, **-m 10**, -k 13 | 0 | 12148622 | 99.96% | 1511 | 99.41% | 12088694 | 99.47% | 15 |
| ***D. melanogaster-sim* dataset (16291 reads, 129738789 bases)** | | | | | | | | | |
| rHAT | -l 11, -w 1000, **-m 1**, -k 13 | 0 | 129392156 | 99.73% | 15587 | 95.68% | 125848857 | 97.00% | 134 |
| rHAT | -l 11, -w 1000, **-m 2**, -k 13 | 0 | 129612180 | 99.90% | 15808 | 97.04% | 127008849 | 97.90% | 141 |
| rHAT | -l 11, -w 1000, **-m 5**, -k 13 | 0 | 129655492 | 99.94% | 15963 | 97.99% | 127791474 | 98.50% | 203 |
| rHAT | -l 11, -w 1000, **-m 10**, -k 13 | 0 | 129684954 | 99.96% | 16064 | 98.61% | 128272211 | 98.87% | 234 |
| ***A. thaliana-sim* dataset (14842 reads, 118558112 bases)** | | | | | | | | | |
| rHAT | -l 11, -w 1000, **-m 1**, -k 13 | 0 | 118507973 | 99.96% | 14592 | 98.32% | 116716612 | 98.45% | 109 |
| rHAT | -l 11, -w 1000, **-m 2**, -k 13 | 0 | 118531650 | 99.98% | 14705 | 99.08% | 117463080 | 99.08% | 118 |
| rHAT | -l 11, -w 1000, **-m 5**, -k 13 | 0 | 118535018 | 99.98% | 14768 | 99.50% | 117951707 | 99.49% | 159 |
| rHAT | -l 11, -w 1000, **-m 10**, -k 13 | 0 | 118535088 | 99.98% | 14776 | 99.56% | 117992740 | 99.52% | 196 |
| ***H.sapiens-sim* dataset (357226 reads, 2856372869 bases)** | | | | | | | | | |
| rHAT | -l 11, -w 1000, **-m 1**, -k 13 | 0 | 2854820181 | 99.95% | 349953 | 97.96% | 2804170119 | 98.17% | 16062 |
| rHAT | -l 11, -w 1000, **-m 2**, -k 13 | 0 | 2855263208 | 99.96% | 352840 | 98.77% | 2825643704 | 98.92% | 16332 |
| rHAT | -l 11, -w 1000, **-m 5**, -k 13 | 0 | 2855626329 | 99.97% | 354546 | 99.25% | 2837879055 | 99.35% | 17172 |
| rHAT | -l 11, -w 1000, **-m 10**, -k 13 | 0 | 2855677158 | 99.98% | 354800 | 99.32% | 2839084216 | 99.39% | 18118 |

a) Also refer to **Supplementary Protocol** for the command lines of various aligners for alignment.

b) A read is considered as unaligned if it is not covered by any alignment.

c) A base is considered as aligned only if it is covered by at least one alignment of the corresponding read.

d) A read is considered as correctly aligned only if there is an alignment covering at least 80% of the bases of the read, and its leftmost position is within 50 bps of the grand truth deducting the clipped parts.

e) A base is considered as covered by position-correct alignment only if it is covered by at least one alignment whose leftmost position is within 50 bps of the grand truth deducting the clipped parts, regardless of the number of bases the alignment covered.

f) The throughput is defined as the wall time of the alignment (in seconds).

# Supplementary Table 8. Evaluation of rHAT with various settings on the k-mer size (-k) parameter for simulated datasets.

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **aligner** | **parametersa** | **# of unaligned readsb** | **aligned basesc** | | **correctly aligned readsd** | | **bases covered by position-correct alignmentse** | | **throughputf**  **(wall time)** |
| **#** | **%** | **#** | **%** | **#** | **%** |
| ***E. coli-sim* dataset (623 reads, 4938920 bases)** | | | | | | | | | |
| rHAT | -l 11, -w 1000, -m 5, **-k 10** | 0 | 4937960 | 99.98% | 622 | 99.84% | 4930148 | 99.82% | 6 |
| rHAT | -l 11, -w 1000, -m 5, **-k 11** | 0 | 4937960 | 99.98% | 622 | 99.84% | 4930148 | 99.82% | 5 |
| rHAT | -l 11, -w 1000, -m 5, **-k 12** | 0 | 4936267 | 99.95% | 621 | 99.68% | 4928100 | 99.78% | 4 |
| rHAT | -l 11, -w 1000, -m 5, **-k 13** | 0 | 4935507 | 99.93% | 621 | 99.68% | 4927695 | 99.77% | 4 |
| rHAT | -l 11, -w 1000, -m 5, **-k 14** | 0 | 4927879 | 99.78% | 617 | 99.04% | 4909185 | 99.40% | 5 |
| rHAT | -l 11, -w 1000, -m 5, **-k 15** | 0 | 4899197 | 99.20% | 609 | 97.75% | 4872751 | 98.66% | 27 |
| ***S. cerevisiae-sim* dataset (1520 reads, 12153653 bases)** | | | | | | | | | |
| rHAT | -l 11, -w 1000, -m 5, **-k 10** | 0 | 12151053 | 99.98% | 1509 | 99.28% | 12078730 | 99.38% | 22 |
| rHAT | -l 11, -w 1000, -m 5, **-k 11** | 0 | 12151219 | 99.98% | 1510 | 99.34% | 12084918 | 99.43% | 16 |
| rHAT | -l 11, -w 1000, -m 5, **-k 12** | 0 | 12151214 | 99.98% | 1512 | 99.47% | 12086963 | 99.45% | 16 |
| rHAT | -l 11, -w 1000, -m 5, **-k 13** | 0 | 12148592 | 99.96% | 1510 | 99.34% | 12087668 | 99.46% | 15 |
| rHAT | -l 11, -w 1000, -m 5, **-k 14** | 0 | 12144134 | 99.92% | 1505 | 99.01% | 12072087 | 99.33% | 44 |
| rHAT | -l 11, -w 1000, -m 5, **-k 15** | 0 | 12129521 | 99.80% | 1496 | 98.42% | 12040996 | 99.07% | 71 |
| ***D. melanogaster-sim* dataset (16291 reads, 129738789 bases)** | | | | | | | | | |
| rHAT | -l 11, -w 1000, -m 5, **-k 10** | 0 | 129672871 | 99.95% | 15985 | 98.12% | 127935992 | 98.61% | 1439 |
| rHAT | -l 11, -w 1000, -m 5, **-k 11** | 0 | 129659191 | 99.94% | 15975 | 98.06% | 127914572 | 98.59% | 631 |
| rHAT | -l 11, -w 1000, -m 5, **-k 12** | 0 | 129663482 | 99.94% | 15976 | 98.07% | 127888422 | 98.57% | 297 |
| rHAT | -l 11, -w 1000, -m 5, **-k 13** | 0 | 129655492 | 99.94% | 15963 | 97.99% | 127791474 | 98.50% | 203 |
| rHAT | -l 11, -w 1000, -m 5, **-k 14** | 0 | 129658651 | 99.94% | 15964 | 97.99% | 127810280 | 98.51% | 455 |
| rHAT | -l 11, -w 1000, -m 5, **-k 15** | 0 | 129653497 | 99.93% | 15945 | 97.88% | 127718434 | 98.44% | 186 |
| ***A. thaliana-sim* dataset (14842 reads, 118558112 bases)** | | | | | | | | | |
| rHAT | -l 11, -w 1000, -m 5, **-k 10** | 0 | 118534961 | 99.98% | 14765 | 99.48% | 117965311 | 99.50% | 985 |
| rHAT | -l 11, -w 1000, -m 5, **-k 11** | 0 | 118534931 | 99.98% | 14769 | 99.51% | 117959906 | 99.50% | 508 |
| rHAT | -l 11, -w 1000, -m 5, **-k 12** | 0 | 118535032 | 99.98% | 14768 | 99.50% | 117961330 | 99.50% | 255 |
| rHAT | -l 11, -w 1000, -m 5, **-k 13** | 0 | 118535018 | 99.98% | 14768 | 99.50% | 117951707 | 99.49% | 159 |
| rHAT | -l 11, -w 1000, -m 5, **-k 14** | 0 | 118532887 | 99.98% | 14764 | 99.47% | 117946503 | 99.48% | 155 |
| rHAT | -l 11, -w 1000, -m 5, **-k 15** | 0 | 118525227 | 99.97% | 14760 | 99.45% | 117938846 | 99.48% | 168 |
| ***H.sapiens-sim* dataset (357226 reads, 2856372869 bases)** | | | | | | | | | |
| rHAT | -l 11, -w 1000, -m 5, **-k 10** | 1115 | 2459057211 | 86.09% | 226577 | 63.43% | 2259004991 | 79.09% | 20401 |
| rHAT | -l 11, -w 1000, -m 5, **-k 11** | 0 | 2854815359 | 99.95% | 353076 | 98.84% | 2833035431 | 99.18% | 29989 |
| rHAT | -l 11, -w 1000, -m 5, **-k 12** | 0 | 2855616423 | 99.97% | 354538 | 99.25% | 2837870383 | 99.35% | 38088 |
| rHAT | -l 11, -w 1000, -m 5, **-k 13** | 0 | 2855626329 | 99.97% | 354546 | 99.25% | 2837879055 | 99.35% | 17172 |
| rHAT | -l 11, -w 1000, -m 5, **-k 14** | 0 | 2855576741 | 99.97% | 354344 | 99.19% | 2837157717 | 99.33% | 8352 |
| rHAT | -l 11, -w 1000, -m 5, **-k 15** | 0 | 2855511069 | 99.97% | 354080 | 99.12% | 2836165519 | 99.29% | 5166 |

a) Also refer to **Supplementary Protocol** for the command lines of various aligners for alignment.

b) A read is considered as unaligned if it is not covered by any alignment.

c) A base is considered as aligned only if it is covered by at least one alignment of the corresponding read.

d) A read is considered as correctly aligned only if there is an alignment covering at least 80% of the bases of the read, and its leftmost position is within 50 bps of the grand truth deducting the clipped parts.

e) A base is considered as covered by position-correct alignment only if it is covered by at least one alignment whose leftmost position is within 50 bps of the grand truth deducting the clipped parts, regardless of the number of bases the alignment covered.

f) The throughput is defined as the wall time of the alignment (in seconds).

# Supplementary Table 9. Throughputs in multiple threads

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **aligner** | **parametersa** | ***D. melanogaster-sim* dataset**  **(16291 reads, 129738789 bases)** | | | | ***A. thaliana-sim* dataset**  **(14842 reads, 118558112 bases)** | | | | ***H.sapiens-sim* dataset**  **(357226 reads, 2856372869 bases)** | | | |
| 1 thread | 2 threads | 4 threads | 8 threads | 1 thread | 2 threads | 4 threads | 8 threads | 1 thread | 2 threads | 4 threads | 8 threads |
| BLASR | BWT**b** | 4611 | 2195 | 1137 | 630 | N/A | N/A | N/A | N/A | 79041 | 41753 | 21315 | 11316 |
| BLASR | SA | 2941 | 1560 | 810 | 447 | 2572 | 1369 | 708 | 371 | 111169 | 69876 | 30988 | 17729 |
| BWA-MEM | -x pacbio | 5616 | 2824 | 1442 | 761 | 4181 | 2096 | 1095 | 555 | 111451 | 55597 | 28498 | 14683 |
| BWA-SW | −b5 -q2 -r1 -z20 | 11149 | 5959 | 3138 | 1762 | 7965 | 4273 | 2250 | 1249 | 297301 | 115303 | 61585 | 36760 |
| rHAT | -l 11, -w 1000, -m 5, -k 13 | 203 | 105 | 61 | 37 | 159 | 87 | 49 | 30 | 17172 | 8958 | 4762 | 2485 |

The throughputs of various aligners in multiple threads, defined as the wall time of the alignment (in seconds). The aligners were run on the three relative large simulated datasets, i.e., *D. melanogaster*-sim, *A. thaliana*-sim and *H.sapiens*-sim, each of them were run in 1, 2, 4 and 8 threads, implemented on a sever with 8 CPU cores (Supplementary Protocol).

a) Also refer to **Supplementary Protocol** for the command lines of various aligners for alignment.

b) When running BLASR on the *A. thaliana*-sim dataset with the BWT-based index, the program corrupted, so that the running time is not available.

# Supplementary Protocol

## 1. Availability of real datasets.

We download the *H. sapiens* and *D. melanogaster* datasets from Pacbio Devnet ([www.pacb.com/devnet/](file:///D:\Works\Molecular%20Biology\Alignment%20&%20Assembly\Project%20&%20Results\fast%20PacBio%20Alignment\Paper\www.pacb.com\devnet\)). The links to the introduction of the datasets are as following:

***H.Sapiens*:**

<https://github.com/PacificBiosciences/DevNet/wiki/H.-sapiens-10x-Sequence-Coverage-with-PacBio-data>

***D. melanogaster*:**

<https://github.com/PacificBiosciences/DevNet/wiki/Drosophila-sequence-and-assembly>

We respectively employed the first 4 SMRT cells from the *H. sapiens* and *D. melanogaster* datasets for evaluating the aligners. Each of the SMRT cells consists of 3 read files, i.e., 12 read files in total. We used the fastq format files as input for all the aligners in the benchmarking.

The IDs of the SMRT cells and links of the SMRT read files are as following.

***H.Sapiens*:**

SMRT cell 1: m130929\_024849\_42213\_c100518541910000001823079209281311\_s1\_p0

Data file links:

<https://s3.amazonaws.com/datasets.pacb.com/2013/Human10x/READS/2530572/0001/Analysis_Results/m130929_024849_42213_c100518541910000001823079209281311_s1_p0.1.subreads.fastq>

<https://s3.amazonaws.com/datasets.pacb.com/2013/Human10x/READS/2530572/0001/Analysis_Results/m130929_024849_42213_c100518541910000001823079209281311_s1_p0.2.subreads.fastq>

<https://s3.amazonaws.com/datasets.pacb.com/2013/Human10x/READS/2530572/0001/Analysis_Results/m130929_024849_42213_c100518541910000001823079209281311_s1_p0.3.subreads.fastq>

SMRT cell 2: m130929\_161837\_42213\_c100518541910000001823079209281315\_s1\_p0

Data file links:

<https://s3.amazonaws.com/datasets.pacb.com/2013/Human10x/READS/2530572/0002/Analysis_Results/m130929_161837_42213_c100518541910000001823079209281315_s1_p0.1.subreads.fastq>

<https://s3.amazonaws.com/datasets.pacb.com/2013/Human10x/READS/2530572/0002/Analysis_Results/m130929_161837_42213_c100518541910000001823079209281315_s1_p0.2.subreads.fastq>

<https://s3.amazonaws.com/datasets.pacb.com/2013/Human10x/READS/2530572/0002/Analysis_Results/m130929_161837_42213_c100518541910000001823079209281315_s1_p0.3.subreads.fastq>

SMRT cell 3: m130929\_093545\_42213\_c100518541910000001823079209281313\_s1\_p0

Data file links:

<https://s3.amazonaws.com/datasets.pacb.com/2013/Human10x/READS/2530572/0003/Analysis_Results/m130929_093545_42213_c100518541910000001823079209281313_s1_p0.1.subreads.fastq>

<https://s3.amazonaws.com/datasets.pacb.com/2013/Human10x/READS/2530572/0003/Analysis_Results/m130929_093545_42213_c100518541910000001823079209281313_s1_p0.2.subreads.fastq>

<https://s3.amazonaws.com/datasets.pacb.com/2013/Human10x/READS/2530572/0003/Analysis_Results/m130929_093545_42213_c100518541910000001823079209281313_s1_p0.3.subreads.fastq>

SMRT cell 4: m130928\_232712\_42213\_c100518541910000001823079209281310\_s1\_p0

Data file links:

<https://s3.amazonaws.com/datasets.pacb.com/2013/Human10x/READS/2530572/0004/Analysis_Results/m130928_232712_42213_c100518541910000001823079209281310_s1_p0.1.subreads.fastq>

<https://s3.amazonaws.com/datasets.pacb.com/2013/Human10x/READS/2530572/0004/Analysis_Results/m130928_232712_42213_c100518541910000001823079209281310_s1_p0.2.subreads.fastq>

<https://s3.amazonaws.com/datasets.pacb.com/2013/Human10x/READS/2530572/0004/Analysis_Results/m130928_232712_42213_c100518541910000001823079209281310_s1_p0.3.subreads.fastq>

***D. melanogaster*:**

SMRT cell 1: m131124\_190051\_42175\_c100583702550000001823087704281416\_s1\_p0

SMRT cell 2: m131124\_221952\_42175\_c100583702550000001823087704281417\_s1\_p0

SMRT cell 3: m131125\_013854\_42175\_c100583772550000001823087704281440\_s1\_p0

SMRT cell 4: m131125\_045830\_42175\_c100583772550000001823087704281441\_s1\_p0

Data file link:

https://s3.amazonaws.com/datasets.pacb.com/2014/Drosophila/raw/Dro1\_24NOV2013\_398.tgz

## 2. Read simulation with PBSim (Verison 1.0.3)

The P5/C3-like SMRT reads were simulated with the following command line. As PBSim may generate reads from the ambiguous regions of the genomes which cannot be aligned, such reads were filtered out in advance by an in-house python script.

pbsim --data-type CLR \

--model\_qc model\_qc\_clr \

--length-mean 8000 \

--difference-ratio 1:12:2 \

--accuracy-mean 0.85 \

--accuracy-min 0.8 \

--depth 1

The five simulated datasets used for benchmarking are available at:

<https://drive.google.com/folderview?id=0Bwibkj8plEJrZFlNOG1rd3hBRWM&usp=sharing>

## 3. Benchmarking for various aligners

The benchmarking was implemented on a server with two Intel E5640 CPUs at 2.67 GHz (four cores for each CPU), 24 Gigabytes RAM running Linux Ubuntu 10.04. The command lines of various aligners are as following.

### 3.1 Benchmarking for BLASR (version 1.3.1)

Genome indexing with BWT:

sawriter *SA\_File Genome\_File*

sa2bwt *Genome\_File SA\_File BWT\_File*

Genome indexing with suffix array:

sawriter *SA\_File Genome\_File*

Read alignment with BWT index:

blasr *Fastq\_File Genome\_File* -bwt *BWT\_File* -nproc *thread\_num* -sam > *Sam\_File*

Read alignment with suffix array:

blasr *Fastq\_File Genome\_File* -sa *SA\_File* -nproc *thread\_num* -sam > *Sam\_File*

### 3.2 Benchmarking for BWA-MEM (version 0.7.12)

Genome indexing:

bwa index *Genome\_File*

Read alignment:

bwa mem -x pacbio -t *thread\_num Genome\_Index Fastq\_File* > *Sam\_File*

### 3.3 Benchmarking for BWA-SW (version 0.7.12)

Genome indexing:

bwa index *Genome\_File*

Read alignment:

bwa bwasw –b 5 –z 20 –q 2 –r 1 -t *thread\_num Genome\_Index Fastq\_File* > *Sam\_File*

### 3.4 Benchmarking for rHAT (version 0.1.1)

Genome indexing:

rHAT-indexer *Genome\_Index\_Dir Genome\_File*

Read alignment (default setting):

rHAT-aligner *Genome\_Index\_Dir Fastq\_File Genome\_File* > *Sam\_File*

Read alignment (with various parameters):

rHAT-aligner –w *Max\_Num\_of\_Hits* –m *Num\_of\_Candidates* –k *Kmer\_Size* -t *thread\_num* *Genome\_Index\_Dir Fastq\_File Genome\_File* > *Sam\_File*