Here we provide a full list of our scripts together with the command lines used to benchmark BRAT-nova against Bismark, BS-Seeker2 and BRAT-BW. Please note that the order of arguments in the command lines is important.

| Script name<br><br>Brief description of purpose | Command Line<br><br>Options |
|---|---|
| ***generate_reads.cpp***<br><br>To compile:<br>g++ -o generate_reads generate_reads.cpp<br><br>Purpose:<br>generates bisulfite-treated single-end reads with sequencing errors, SNPs, indels, adapter contamination<br><br>Indels and adapter contamination<br>are introduced only if options -i and -a are used<br><br>Length of indels is random from 1 to 10<br><br>97% of all Cs are converted to Ts | ***./generate_reads -g references_names.txt -r reads.fastq -p positions.txt -l 100 -t 1000000 -e 5 -s 2 -i 7 -a 10 -m 30***<br><br>*references_names.txt* file contains the paths to the files with references (one reference per file): one path per line (option -g)<br><br>*reads.fastq* contains the resulting reads in FASTQ format (option -r, if not provided, output will be in file *reads.fastq*)<br><br>*positions.txt* contains the positions from which the reads have been generated in format: read_id, read, reference_name, strand, position starting with 1, mismatches, indel length, indel's position within the read (option -p, if not provided, output will be in file *positions.txt*)<br><br>100 is read length (default 100, option -*l*)<br><br>1000000 is the total number of reads generated (default 100, option -t)<br><br>5 is the percentage of sequencing errors of total number of read bases (default 2, option -e)<br><br>2 is the percentage of SNPs of total number of reads bases (default 1, option -s)<br><br>7 is the percentage of reads having an indel at a random position within a read; indel length is random in the range 1…10 (default value for the percentage of reads having an indel is 1, option -i)<br><br>10 is the percentage of reads with an adapter sequence of random length up to 15bp (next |

| | |
|---|---|
| | argument) replaced the 3'-end of the reads (default is 10%, option -a)<br><br>30 maximum length of adapter sequence (default is 15bp, option -m) |
| ***Generate_methylation_call.cpp***<br><br>To compile:<br>g++ -o Generate_methylation_call Generate_methylation_call.cpp<br><br>Purpose:<br>generates bisulfite-treated reads with sequencing errors, SNPs, indels, adapter contamination and records methylation<br>level of each cytocine derived from generated reads<br><br>Indels and adapter contamination<br>are introduced only if options -i and -a are used<br><br>97% of all Cs are converted to Ts | ***./Generate_methylation_call -g reference.fa -r reads.fastq -p positions.txt -l 100 -t 1000000 -e 5 -s 2 -i 7 -a 10 -m 30 -o methylation_level.txt***<br><br>*reference.fa* is a file with a reference in FASTA format (containing only a single reference, i.e. character ">" occurs only once in the file) (option -g)<br><br>*reads.fastq* contains the resulting reads in FASTQ format (option -r, if not provided, output will be in file *reads.fastq*)<br><br>*positions.txt* contains the positions from which the reads have been generated in format:<br>read_id, read, reference_name, strand, position starting with 1, mismatches, indel length, indel's position within the read (option -p, if not provided, output will be in file *positions.txt*)<br><br>100 is read length (default 100, option -*l*)<br><br>1000000 is the total number of reads generated (default 100, option -t)<br><br>5 is the percentage of sequencing errors of total number of read bases (default 2, option -e)<br><br>2 is the percentage of SNPs of total number of reads bases (default 1, option -s)<br><br>7 is the percentage of reads having an indel at a random position within a read; indel length equals to 3 (default value for the percentage of reads having an indel is 1, option -i)<br><br>10 is the percentage of reads with an adapter sequence of random length up to 15bp (next argument) replaced the 3'-end of the reads |

| | |
|---|---|
| | (default is 10%, option -a)<br><br>30 maximum length of adapter sequence (default is 15bp, option -m)<br><br>*methylation_level.txt* is a file with methylation level in format: reference name, strand, position starting with 1, count of Cs, count of Cs plus Ts |
| **Correctness_methCall_rand_bratbw.cpp**<br><br>To compile:<br>g++ -o Correctness_methCall_rand_bratbw Correctness_methCall_rand_bratbw.cpp<br><br>Purpose: it calculates methylation call accuracy and methylation level accuracy, given an output of BRAT-BW' and BRAT-nova' tool *acgt_count* and given the file *methylation_level.txt* (produced by running Generate_methylation_call). | ***./Correctness_methCall_rand_bratbw -g reference.fa -m methylation_level.txt -t meth_level_BRAT_nova.txt -r 10 -f 0.2 -c 0.5***<br><br>*reference.fa* is the same file used to run Generate_methylation_call (option -g)<br><br>*methylation_level.txt* is the file produced by running Generate_methylation_call (option -m)<br><br>*meth_level_BRAT_nova.txt* is the output of *acgt_count* of BRAT-nova/BRAT-BW (if using BRAT-BW, concatenate two output files) (option -t)<br><br>10 is the minimum read coverage of a cytosine to consider it for methylation call/level accuracy estimation (default is 10, option -r)<br><br>0.2 is the threshold on FDR (program will use an error corresponding to 20% of false discovery rate to estimate methylation level accuracy) (default is 0.2, option -f)<br><br>0.5 is the threshold on methylation level to estimate methylation call accuracy (default is 0.5, option -c) |
| Correctness_meth_call_BSseeker.cpp<br><br>To compile:<br>g++ -o Correctness_meth_call_BSseeker Correctness_meth_call_BSseeker.cpp<br><br>Purpose: it calculates methylation call accuracy and methylation level accuracy, given an output of BS- | ***./ Correctness_meth_call_BSseeker -g reference.fa -m methylation_level.txt -t meth_level_BSseeker -r 10 -f 0.2 -c 0.5***<br><br>All options are described above (***Correctness_methCall_rand_bratbw***)<br><br>***meth_level_BSseeker*** is the file produced by |

| | |
|---|---|
| Seeker2's tool *bs_seeker2-call_methylation.py* (file provided with --CGmap option) and given the file *methylation_level.txt* (produced by running Generate_methylation_call). | *bs_seeker2-call_methylation.py* (file provided with --CGmap option) in the format:<br>reference name, dummy, pos (starts with 1), dummy, dummy, methylation level, count of Cs, count of Cs and Ts<br><br>where "dummy" values were ignored. |
| Correctness_meth_call_bismark.cpp<br><br>To compile:<br>g++ -o Correctness_meth_call_bismark Correctness_meth_call_bismark.cpp<br><br>Purpose: it calculates methylation call accuracy and methylation level accuracy, given an output of Bismark's tool *bismark_methylation_extractor* , the file, concatenation of the output files:<br>cat CpG_O* Non_CpG_O* > meth_level_Bismark.txt | ***./ Correctness_meth_call_bismark -g reference.fa -m methylation_level.txt -t meth_level_Bismark.txt -r 10 -f 0.2 -c 0.5***<br><br>*meth_level_Bismark.txt is in format:*<br>read id, strand, reference name, position (starts with 1), dummy<br><br>The rest of parameter options as in ***Correctness_methCall_rand_bratbw*** |
| correctness_bratbw.cpp<br><br>To compile:<br>g++ -o correctness_bratbw correctness_bratbw.cpp<br><br>Purpose: calculates mapping accuracy of BRAT-BW, given the file *positions.txt* (results of running generate_reads) and given the output of BRAT-BW | ***./correctness_bratbw -r output_bratbw.txt -p positions.txt -t 1000000 -e 50***<br><br>1000000 is the total number of reads produced by ***generate_reads*** (attempted to map) (option –t)<br>50 is the threshold on position, a read is considered to be correctly mapped if it is mapped to the same chromosome, same strand to the position within 50bp of original position (option –e)<br><br>positions.txt is the file with original positions (produced by running ***generate_reads***) (option –p)<br><br>output_bratbw.txt is the output file with mapped reads by BRAT-BW (reads initially generated using ***generate_reads***) (option –r) |
| correctness_SAM.cpp<br><br>To compile:<br><br>g++ -o correctness_SAM correctness_SAM.cpp<br><br>Purpose: calculates mapping accuracy of Bismark, BS-Seeker2 and BRAT-nova on single-end reads, given the file *positions.txt* (results of running generate_reads) and given the output of a tool with | ***./correctness_SAM -r mapped_reads.sam -p positions.txt -t 1000000 -e 50***<br><br>1000000 is the total number of reads produced by ***generate_reads*** (attempted to map) (option –t)<br>50 is the threshold on position, a read is considered to be correctly mapped if it is mapped to the same chromosome, same strand to the position within 50bp of original position (option –e) |

| | |
|---|---|
| mapped reads in SAM format<br><br>It also works for paired-end reads for the tools BRAT-nova and Bismark (for Bismark, first run "change_pairs_ids", described below, and only then run correctness_SAM) | positions.txt is the file with original positions (produced by running ***generate_reads***) (option –p)<br><br>mapped_reads.sam is the output file with mapped reads (reads initially generated using ***generate_reads***) (option –r) |
| ***generate_pairs.cpp***<br><br>To compile:<br>g++ -o generate_pairs generate_pairs.cpp<br><br>Purpose:<br>generates bisulfite-treated paired-end reads with sequencing errors, SNPs, indels, adapter contamination<br><br>Indels and adapter contamination<br>are introduced only if options -i and -a are used<br><br>Length of indels is random from 1 to 10<br><br>97% of all Cs are converted to Ts | ***./generate_pairs -g references_names.txt -r paired_end_reads -p positions.txt -l 100 -t 1000000 -e 5 -s 2 -i 7 -a 10 -m 30***<br><br>*references_names.txt* file contains the paths to the files with references (one reference per file): one path per line (option -g)<br><br>paired-end reads will be in two files: (1) paired_end_reads.mate1.fastq and (2) paired_end_reads.mate2.fastq in FASTQ format (thus, prefix of the file names is provided with option -r, if not provided, prefix is "paired_end_reads")<br><br>*positions.txt* contains the positions from which the reads have been generated in format: read_id, read, reference_name, strand, position starting with 1, mismatches, indel length, indel's position within the read (option -p, if not provided, output will be in file *positions.txt*)<br><br>100 is read length (default 100, option -*l*)<br><br>1000000 is the total number of reads generated (default 100, option -t)<br><br>5 is the percentage of sequencing errors of total number of read bases (default 2, option -e)<br><br>2 is the percentage of SNPs of total number of reads bases (default 1, option -s)<br><br>7 is the percentage of reads having an indel at a random position within a read; indel length is random in the range 1…10 (default value for the percentage of reads having an indel is 1, option - |

| | |
|---|---|
| | i)<br><br>10 is the percentage of reads with an adapter sequence of random length up to 15bp (next argument) replaced the 3'-end of the reads (default is 10%, option -a)<br><br>30 maximum length of adapter sequence (default is 15bp, option -m) |
| correctness_sam_paired_BSseeker.cpp<br><br>To compile:<br>g++ -o correctness_sam_paired_BSseeker correctness_sam_paired_BSseeker.spp<br><br>Purpose: calculates mapping accuracy for paired-end reads on BS-Seeker2 | ***./ correctness_sam_paired_BSseeker -r mapped_reads.sam -p positions.txt -t 1000000 -e 50***<br><br>All options are the same as described in **correctness_SAM**<br><br>1000000 here is the total number of attempted reads (count single reads, i.e. if there were 500,000 pairs, then we need to pass 1000000 = 2*500000 with opton -t) |
| change_pairs_ids.cpp<br><br>To compile:<br><br>g++ -o change_pairs_ids change_pairs_ids.cpp<br><br>Purpose: before calculating of mapping accuracy for paired-end reads on Bismark, run this script first, then use **correctness_SAM** | ***./change_pairs_ids output_paired_Bismark.sam***<br><br>output_paired_Bismark.sam is the output of Bismark mapping paired-end reads<br><br>The output with adjusted reads ids will be in ***output_paired_Bismark.sam.out*** (use this file to calculate mapping accuracy of Bismark on paired-end reads) |
| correctness_bratbw_paired.cpp<br><br>To compile:<br><br>g++ -o correctness_bratbw_paired correctness_bratbw_paired.cpp<br><br>Purpose: calculates mapping accuracy of BRAT-bw on paired-end reads | ***./ correctness_bratbw_paired -r output_bratbw.txt -p positions.txt -t 1000000 -e 50***<br><br>1000000 is the total number of reads produced by generate_pairs (attempted to map, total number of pairs multiplied by 2) (option –t)<br><br>50 is the threshold on position, a read is considered to be correctly mapped if it is mapped to the same chromosome, same strand to the position within 50bp of original position (option –e)<br><br>positions.txt is the file with original positions |

| | |
|---|---|
| | (produced by running generate_pairs) (option –p)<br><br>output_bratbw.txt is the output file with mapped reads by BRAT-BW (reads initially generated using generate_reads) (option –r) |
| To convert FASTQ reads to BRAT-BW's input reads format, use tool *trim* in BRAT-BW suit. | |