SHERMAN - Bisulfite-Read Simulator v0.1.6

---

09 September 2013

BASIC ATTRIBUTES:

-l/--length
The length of all sequences to be generated (between 1 and 300 bp). Has to be specified.

-n/--number_of_seqs
The number of sequences to be generated. Default: 1000000.

--genome_folder <path/to/genome/folder>
Enter the genome folder you wish to use to extract sequences from. The genomic coordinates are being printed into the read-ID field in addition to the read count number. Accepted formats are FastA files ending with '.fa' or '.fasta'. Default path: /data/public/Genomes/Mouse/NCBIM37/. Default: ON.

--random
The sequences will be generated with entirely random base composition instead extracting sequences from a real genome. This is a much quicker option for testing purposes.

-q/--quality
The default quality for all positions if error rate is set to 0% or if SNPs are to be introduced. Default: 40.

-pe/--paired_end
Will create paired-end read files with the names simulated_1.fastq and simulated_2.fastq. The minimum and maximum fragment sizes can be adjusted with the options -I/--minfrag or -X/--maxfrag. Default: OFF.

-I/--minfrag <int>
The minimum size for paired-end fragments. Default: 70.

-X/--maxfrag <int>
The maximum size for paired-end fragments. Default: 400.

-cr/--conversion_rate <float>

A uniform bisulfite conversion rate of <float> %. This value can be anything between 0 (no bisulfite conversion at all) and 100% (all cytosines will be converted into thymines). To simulate standard genomic sequences for experiments other than BS-Seq simply use -cr 0.

-CG/--CG_conversion <float>
Bisulfite conversion rate for cytosines in CG-context as <float> %. This value can be anything between 0 (no bisulfite conversion at all) and 100% (all CG-cytosines will be converted into thymines). Requires -CH/--CH_conversion to be specified as well.

-CH/--CH_conversion <float>
Bisulfite conversion rate for cytosines in CH-context as <float> %. This value can be anything between 0 (no bisulfite conversion at all) and 100% (all CH-cytosines will be converted into thymines). Requires -CG/--CG_conversion to be specified as well.

--colorspace
Using this option will print out all sequences in color space as well as in base space. The output will consist of 2 files: simulated.csfasta and simulated_QV.qual, whereby the qualities are a Phred values separated by spaces (e.g. 40 40 39 39 38 37 ...). Paired-end files will carry _1 or _2 in the filenames. Note that the conversion of base space to color space takes place before any quality values or errors are introduced.

--non_directional
The reads can orignate from any of the four possible strands produced by bisulfite conversion and PCR amplification. Default: OFF.

CONTAMINANTS:

-s/--snps <int>
The number of SNPs to be introduced. This value can be anything between 1 and the total sequence length. Default: 0. Introducing SNPs will always assume an error rate of 0%, the default quality for all bases can be specified with (-q/--quality).

-e/--error_rate <float>
The error rate in %. This can be anything between 0 and 60%. If the error rate is selected as 0%, no sequencing errors will be introduced (even though a Phred score of 40 formally translates into an error rate of 0.01%). The error rate will be a mean error rate per bp whereby the error curve follows an exponential decay model. This means that an error rate of 0.1% will - overall - introduce sequencing errors roughly every 1 in 1000 bases, whereby the 5' end of a read is much less likely to harbour errors than bases towards the 3' end.

--fixed_length_adapter <int>
Replaces the most 3' <int> bp of each read with Illumina adapter sequence. Sherman is currently using the Illumina Paired End PCR Primer 2 as adapter sequence, however this can be modified in the script if necessary (locate the following line to change it:

my $adapter_sequence = 'CAAGCAGAAGACGGCATACGAGATCGGTCTCGGCATTCCTGCTGAACCGCTCTTCCGATCT').

--variable_length_adapter <int>

For this contamination we simulate a normal distribution of fragment sizes for a mean insert size specified as <int> bp and replace a variable portion at the 3' end of reads with a adapter sequence if the fragment size is smaller than the read length. A normal distribution of fragment sizes will be modelled using the specified <int> as mean (mu) and a variance (sigma) of 60 (this is a fixed value which was determined empirically). The adapter sequence is the same as described for --fixed_length_adapter <int>.