BSSim:

Bisulfite sequencing simulator for next-generation sequencing

Introduction

BSSim is implemented in the Python language and run in an operating system-independent manner. It can allow users to mimic various methylation level (total methylation level of cytosines, percentage of cytosines that methylated and methylation level of total methylcytosines) and bisulfite conversion rate in CpG, CHG and CHH context, respectively. It can also simulate genetic variations that are divergent from the reference sequence along with the sequencing error and quality distributions. In the output, both directional/non-directional, various read length, single/paired-end reads and alignment data in the SAM format can be generated. BSSim is a cross-platform BS-seq simulator offers output read datasets not only suitable for Illumina's Solexa, but also for Roche's 454 and Applied Biosystems' SOLiD.

Installation

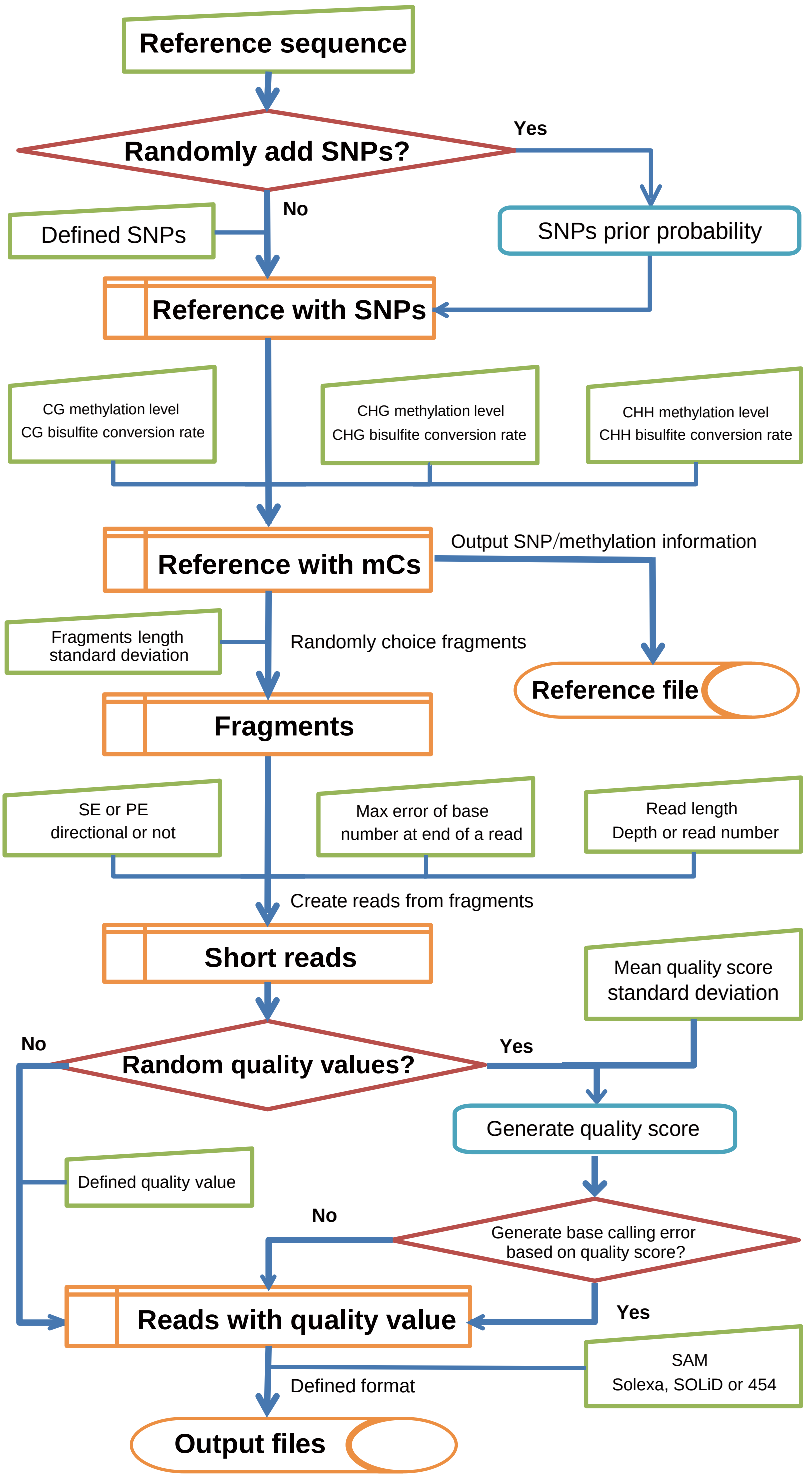
BSSim is implemented in Python2.6. It requires several freely Python packages: multiprocessing 2.6.2.1, nested_dict 1.0.9, nose-bisect 0.1.0, pyfasta 0.4.5 and wsgiserialize 0.3. Download and installation instructions of them are available at http://122.228.158.106/BSSim and packaged as Needed_pypi.tar.gz, which can also be downloaded from http://pypi.python.org/pypi.

To install BSSim, please simply download the BSSim.py and run the programs from the command line interface (eg, Terminal in Linux). Typing –h after the program name will give you basic usage instructions (more detailed instructions can be found later in this manual). For instance, typing:

```
./BSSim.py   – h
```

Workflow

The workflow of BSSim to simulate real BS-seq data is described by the following diagram:



Usage:

```
./BSSim.py [options]
```

General Options

-h Help.

-i Input reference sequence in the fasta format.

-d Sequencing depth (>0). Default: 30.

-U The max number of processes core (>0). Default: 2.

-l Read length (>0). Default: 90 bp.

-s Single-end pattern.

-p Paired-end pattern (default).

-t Sequencing platform: Solexa/SOLiD/454. Default: Solexa.

-f Fragment length (library size) (>0). Default: 300 bp.

--FR Standard deviation of -f (0~(f/2.58)). Default: 20 bp.

-n Number of reads to be generated (>0).

-q Quality score (mean value of quality score) [0~1]. Default: 0.95 (95% of highest score).

-e Number of max error base at the end of a reads [0~l]. Default: 0.

-N The number of bases to be read into RAM one time (>0). Default is 1000000. This option to control the memory of the program.

-o Prefix of output file. Default is set by name of input file. It contains the information about: fragment length, depth, total methylation level of CpG, CHG and CHH context.

-D Directional: reads 1 is same direction with reference sequencing (Watson strand) and read 2 is from Crick strand. Default: non-directional.

-P Output position information into the output file. Default is not.

-A Output alignment result in SAM format. Default is not.

-V Version information.

-R Output the reference methylation information. Default is not.

format is:

Chromosome	Position	ref_genome	ref_A	Methylation Pattern	default methylation rate(ignore it if ref is A,T)	Ref_B(homologous chromosome)	methylation pattern	default methylation rate
chr10	114	G	G	CHH	0.002	G	CHH	0.002
chr10	115	C	C	CG	0.972702	C	CG	0.951118

DNA methylation

--ML Total methylation level of cytosines (overall DNA methylation level) (0~1). Default: 0.0612.

--CL CG methylation level (0~1). Default: 0.8653.

--GL CHG methylation level (0~1). Default: 0.0203.

--HL CHH methylationlevel (0~1). Default: 0.0227.

--MR All mC/C rate (the ratio of total methylcytosines/total cytosines) (0~1). Default: 0.073.

--CR mCG/CG rate (0~1). Default: 0.852.

--GR mCHG/CHG rate (0~1). Default: 0.019.

--HR mCHH/CHH rate (0~1). Default: 0.025.

--MM Methylation level of total methylcytosines. Default: 0.8529.

--CM mCG methylation level (0~1). Default: 0.8529.

--GM mCHG methylation level (0~1). Default: 0.0837.

--HM mCHH methylation level (0~1). Default: 0.9091*0.0994+(1-0.9091)*0.8965.

--MCS Standard deviation of --MM (0~(1-MM)*MM). Default: 0.01.

--CS Standard deviation of --CM (0~(1-CM)*CM). Default:

(1-CM)*CM/2.0.

--GS Standard deviation of --GM (0~(1-GM)*GM). Default: (1-GM)*GM/2.0.

--HS Standard deviation of --HM (0~(1-HM)*HM). Default: (1-HM)*HM/2.0.

--BC All cytosines' bisulfite conversion rate [0~1]. Defualt is 0.998.

--CC CG conversion rate [0~1]. Default: 0.998.

--GC CHG conversion rate [0~1]. Default: 0.998.

--HC CHH conversion rate [0~1]. Default: 0.998.

SNP

-S SNP file with SNP information, specifying location and frequency of SNPs. format is:

Chromosome	position	strand	A	T	C	G
chr10	1	+	0	0.4	0	0.6
chr10	2	+	0.3	0.2	0.1	0.4

--DS Do not add SNP. Default is add (based on prior probability).

-G Polyploid type of reference sequencing (>0). Default: 2.

-Y The frequency of homozygous SNPs [0~(1-Z)]. Default: 0.0005.

-Z The frequency of heterozygous SNPs [0~(1-Y)]. Default: 0.001.

Read quality

-q Quality score (mean value of quality score). Default: 0.95 (95% of highest score).

--DQ Randomly introduce quality value. Default: uniform quality score.

--RE Randomly introduce sequencing errors by sequencing quality value (Q =-10*log10(e),Q is the sequencing quality value (Phred score), e is the error rate, Massingham, et al., 2012). Default is not.

--QS Standard deviation of -q (0~(1-q)*q). Default: (1-q)*q/2.

BSSim can also allow users to split every read into three parts (The head part, the end part and interval) to add different quality value along the read.

--FP (Lengh of the head part)/(total read length) (0~1). Default: 0.01 (1% of total read length).

--FPS Standard deviation of --FP (0~(1-FP)*FP). Default: (1-FP)*FP/2.

--FQ The mean quality value of the head part less than -q (0~q). Default: 0.1

--FS Standard deviation of --FQ (0~(1-(q-FQ))*(q-FQ)). Default: (1-(q-FQ))*(q-FQ)/8.

--EP (Lengh of the end part)/(total read length) (0~1). Default: 0.8 (80% of total read length).

--EPS Standard deviation of --EP (0~(1-EP)*EP). Default: (1-EP)*EP/2.

--EQ The mean quality value of the end part less than -q (0~q). Default: 0.2

--ES Standard deviation of --EQ (0~(1-(q-EQ))*(q-EQ)). Default: (1-(q-EQ))*(q-EQ)/4.

Examples:

1. Generate some (30×) Illumina Paired-end reads (90bp) from the reference sequence (test.fa) with the same quality value:

```
./BSSim.py -i test.fa
```

This will create two output files:

test-Fragment_length-300-depth-30-methylation_level_CG-0.73801437-CHG-0.00169911-CHH-0.003901140053.1.fq

test-Fragment_length-300-depth-30-methylation_level_CG-0.73801437-CHG-0.00169911-CHH-0.003901140053.2.fq

2. Use your own Prefix of output file:

```
./BSSim.py -i test.fa –o out
```

This will create two output files:

out.1.fq   out.2.fq

3. Generate Roche/454 reads:

```
./BSSim.py -i test.fa -t 454 –o out
```

This will create four output files:

out.1.fna   out.1.qual

```
out.2.fna       out.2.qual
```

4. Use 4 processes cores to run the program:

```
./BSSim.py -i test.fa -U 4 -o out
```

5. Make some single-end and directional Illumina reads with position information:

```
./BSSim.py -i test.fa -s -D –P -U 4 -o out
```

This will produce the output file out.1.fq.

This is part of out.1.fq:

```
@chr10.1862-1951.+.W:1:17:32028:35678#0/1
ATTTTGGTTTTTTATTATTATATTTTGGGGTGGTATAGTTTGGTTTTATATATTG
TGTTTTATTGGTAATGAAAAGAGTTTTTGTTTTTT
+
ffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffff
```

This reads is come from the Watson strand of chr10:1862-1951 PRC fragment from Watson strand.

6. Output alignment result in SAM format:

```
./BSSim.py -i test.fa -A -U 4 -o out
```

This will create four output files:

```
out.1.fq        out.2.fq
```

out.Watson.sam out.Crick.sam

7. Set the DNA methylation information (50% cytosines is methylated, the methylation level of mCG is 0.9, the methylation level of mCHG is 0.3， the methylation level of mCG is 0.2) and output the reference methylation information:

```
./BSSim.py -i test.fa --MR 0.5 --CM 0.8 --GM 0.3 --HM 0.2 -R -A -U 4
```

This will create five output files:

test-Fragment_length-300-depth-30-methylation_level_CG-0.4-CHG-0.15-CHH-0.1.ref

test-Fragment_length-300-depth-30-methylation_level_CG-0.4-CHG-0.15-CHH-0.1.Watson.sam

test-Fragment_length-300-depth-30-methylation_level_CG-0.4-CHG-0.15-CHH-0.1.Crick.sam

test-Fragment_length-300-depth-30-methylation_level_CG-0.4-CHG-0.15-CHH-0.1.1.fq

test-Fragment_length-300-depth-30-methylation_level_CG-0.4-CHG-0.15-CHH-0.1.2.fq

8. Use your own SNP points and output the information into ref file:

```
./BSSim.py -i test.fa -S snp_test.txt -R -U 4 -o out
```

This will create three output files:

out.1.fq   out.2.fq   out.ref

This is part of out.ref:

Chromosome	Position	ref_genome	ref_A	Methylation Pattern	default methylation rate(ignore it if ref is A,T)	A	T	C	G
chr10	1	c	C	CHH	0.002	0	0.4	0	0.6
chr10	2	a	A		1	0.3	0.2	0.1	0.4
chr10	3	t	T		1	0.2	0.1	0.2	0.5
chr10	4	t	T		1	0.5	0	0	0.5
chr10	5	t	T		1	0	0	1	0

9. Simulate a sequence of haploid and set the frequency SNPs:

```
./BSSim.py -i test.fa -Z 0.2 -G 1 -R -U 4 -o out
```

This will create three output files:

out.1.fq   out.2.fq   out.ref

10. Randomly introduce quality value and randomly introduce sequencing errors by sequencing quality value:

```
./BSSim.py -i test.fa --DQ --RE -R -U 4 -o out
```

This will create three output files:

```
out.1.fq        out.2.fq        out.ref
```

This is part of out.1.fq:

```
@FC61FL8AAXX:1:17:79367:31079#0/1
GGGTGTGTTGTTATTATAATGTGAGGAAGAGGGTTTTGTAATGTTTTTAGTT
GTTAGTAGGCGGCGTGTTATTATTATATTGTGAGTAAG
+
hhhBGe@hhhhJchehfffffffffffffffffffffffffffffffffffffffffffffffffffffffffffbAcX@_fPZ@@
bahh
@FC61FL8AAXX:1:17:43793:15339#0/1
ACAAGTCTCACCTTACAATCCAAAAATAACATTCCTAAGTATTTTGACAACT
ACTTTGATGTTATTTCCCATCAAAAGCTACCATGCAGT
+
hhhDhhhhghhhhhhh[Ahhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhh
hhhhhhhhhhhhhN@h]hFhhN
```

Version

Changes since Version 1.0:

1.2:

- Added support for multi process.
- Users now can control the distribution of quality value.
- Optimized the memory footprint (control the number of bases to be read into RAM one time).

1.1:

- Added support for CpG, CHG and CHH context.
- Added support for output alignment result in SAM format.
- Added support for output the reference methylation information.
- Added support for standard deviation of methylation level of total methylcytosines.

1.0:

- Added support for three sequencing platform: Solexa/SOLiD/454

- Added support for user-defined input SNP information.

- Added support for haploid and polyploid.

- Fixed several bugs.