**Psim User Manual**

**Title**: Psim-a pilot study sequencing simulator

**Author**: Rachel Wu([rachelwu123@gmail.com](mailto:rachelwu123@gmail.com) ), BENM Feng([BinxiaoFeng@gmail.com](mailto:BinxiaoFeng@gmail.com) )

**Version**: 1.0.0

**Update**: 2013.06.16

**Depends**: perl, Math-Random

**Introduction**:

Psim is a pilot study sequencing simulator. It can simulate NGS sequencing data such as illuminaTM, 454 FLXTM , SOLiDTM , Ion TorrentTM or PacBioTM etc. It satisfied variety requires, using this tool to provide irregular or regular reads length, paired-end reads with hieratical insert span, single strand specified library, BAC liked loop sequences, pooling strategy, variation by randomly generated or custom introduced including single nucleotide polymorphism (SNP), small insertion and deletion (InDel) and structural variation (deletion, insertion, inversion, tandem repeat and translocation), DNA damage such as methylation or deamination, particular sequencing error models, library construction, pollution mixture, metagenome with multi-index or barcode, random passion/bimodal distribution, it also varied reference/ancestral sequence based on Bayesian probability models. Beyond these, Monte Carlo method is presented for pilot study by sequencing simulator PSim, with wide utility in generating DNA/mRNA sequencing with different sequencing data format.

The real sequencing library preparation process is simulated in Psim. The process include template DNA variation, \*DNA damage, \*DNA fragment pre-amplification, \*single-chain library or double-strain library, randomly fragmentation, add adapter or PE linker, generation of quality value and so on.

\*especially for damage DNA sequencing steps

**Obtaining and Installing:**

Psim is available from <https://sourceforge.net/projects/pilotsim/>. Psim depends on perl and Math-Random module.

Installation of perl:

1. Perl always has been pre-installed on Linux system. There is no need to reinstall;
2. Windows system user can download activeperl from <http://www.activestate.com/activeperl>.

Installation of Math-Ranodm:

1. Automatically install with CPAN (need to be root)

#Linux user

perl –MCPAN –e shell

cpan>install Math::Random

cpan>q

#Windows user

ppm

ppm>install Math::Random

1. Manual installation

Download Math-Random source code from <http://search.cpan.org/~grommel/Math-Random-0.70/Random.pm> and then type in a terminal:

tar zxvf Math-Random-0.70.tar.gz

cd Math-Random-0.70

perl Makefile.PL

make

make test

make install

#manually set the installation path if you don’t have root privilege

tar zxvf Math-Random-0.70.tar.gz

cd Math-Random-0.70

perl Makefile.PL PREFIX=YOURDIR

make

make test

make install

**Functions:**

1. **Randomly generate reference sequence**

Generate specified length or random length of Poisson distribution of sequence(s) which could be used ad reference sequence(s).

perl GenerateSeq.pl [-n] [-l] [-pre] [-auto] > OUTFILE

-n NUM number of sequence (default=1)

-l NUM or NUM,NUM... length of each sequence, separate different length by ","

-pre CHAR prefix of your sequence(s)

-auto randomly generate n sequences for Poisson distribution with average length of “–l”

Example:

#generate five sequences of 3000bp, 4000bp, 6000bp, 5000bp and 8000bp. Prefix of sequences is “test-1”. And the sequences file saved as “test-1.fa”

perl GenerateSeq.pl –n 5 –l 3000,4000,6000,5000,8000 –pre test-1 > test-1.fa

#generate five sequences of average length of 5000bp. Pre ix of sequences is “test-2”. And the sequences file saved as “test-2.fa”

perl GenerateSeq.pl –n 5 –l 5000 –pre test-2 –auto > test-2.fa

1. **Structure Variation**

Structural variation is the variation in structure of an organism's chromosome. It consists of many kinds of variation, such as deletion, insertion, tandem repeat, reversion and translocation. Typically a structure variation affects a sequence length about 1Kb to 3Mb. While, indel of short segment is also included in Psim. To specify variation location(s) and type(s), variation information file like /example/sv\_example.txt should be offered while running Psim.

The variation information file format as: the first column with reference name, the second column with variation type (include deletion、insertion、tandem\_repeat、reversion and translocation), the third column with location, length or other information. Such as

deletion site1, length1; site2, length2……

insertion insert\_site1,sequence1; insert\_site2,sequence2……

inversion site1, length1; site2, length2……

tandem\_repeat site1, length1, repeat\_times1 ;site2, length2, repeat\_times2……

translocation site1, length1, insert\_site1; site2, length2, insert\_site2……

Parameter:

--sv <NUM|FILE> structure variation rate and average length (default= 0.1:3000) or specific sv type and site file(format refer to ../example/sv\_example.txt)

Example:

#variation as SV information file (like DIR/Psim/example/sv\_example.txt)

--sv DIR/Psim/input/sv\_example.txt

#randomly generate SV of average length of 3000bp and coverage of 10%

--sv 0.1:3000

#no SV

--sv 0

1. **SNP**

To specify SNP site(s) and base(s), mutation information file like /example/ssnp\_example.txt should be offered while running Psim. The mutation information file format as: the first column with reference name, the second column with mutation site, the third column with possible mutation base(s) and the fourth column with mutation rate(S). Separate different sites and rates in column three and four by comma.

Parameter:

--snp <NUM|FILE> random snp rate (default=0.001) or specific snp site and rate file (format refer to ../example/snp\_examp le.txt)

Example:

#mutation as SNP information file DIR/Psim/example/snp\_example.txt

--snp DIR/Psim/example/snp\_example.txt

#randomly generate SNP of 10% coverage

--snp 0.001

#no SNP

--snp 0

1. **Library preparation**

Library preparation process includes random fragmentation of library sample and adapter or PE linker added. For fixed read length sequencing methods, most or all the lengths of fragments should be set above read length. And for non-fixed read length sequencing methods, the length of fragment is the read length.

In this section, user need to specify the average length of fragment (-fragmean), stander deviation (-fragsd) and length limit (-fraglim, optional). Adapter sequences are added by Illumina like sequencing methods and linker sequence is added by Roche 454 PE like sequencing methods.

Parameter:

--cov <NUM> sequencing coverage of reference sequence(s) (default=3)

--fragmean <INT> average length of library fragment (default=200)

--fragsd <INT> standard deviation of library fragment (default=10)

--fraglim <INT> limit length of fragment library ("20+" means must above 20nt, and "240-" means must shorter than 240nt,if(-damage) this parameter default=20+)

--adapter <CHAR> adapter sequence (default sequence refer to ../example/adapterIllumina\_example.txt, split two adapters by ":")

While adopting Roche 454 PE sequencing method:

--insert <INT> average insert size and sd(default=8000:30 if --pe)

--linker <CHAR> PE insert sequence (default=\"ATAACTTCGTATAATGTATGCT ATACGAAGTTAT\")

Example:

#simulating sequencing data of 150bp fragments average length, 6 stander deviation, 5 folds coverage and use the default adapter sequences

--cov 5 --fragmean 150 --fragsd 6

#simulating Roche 454 PE sequencing data of 8000bp insert size, 35 stander deviation of insert length, 4 folds coverage, 650bp fragments average length, 50 stander deviation of fragment length and use the default linker sequence

--pe --cov 4 --insert 8000:35 --fragmean 650 --fragsd 50

1. **RNA-seq**

Simulate RNA-seq based on given GFF (General Feature Format) file. Refer to Sanger website (<http://www.sanger.ac.uk/resources/software/gff/spec.html>) for GFF file format detail.

Parameter:

--cdna RNA-SEQ sequencing

--gff <FILE> input gene gff file

--covcdna <INT> mean coverage of cDNA sequences(default=10 if --cdna)

Example:

#simulating RNA-seq result with 20-fold coverage, gff file at DIR/Psim/example/gff\_exampl e.txt

--cdna --gff DIR/Psim/example/gff\_example.txt –covcdna 20

1. **Damage DNA**

The end of damage DNA fragment cannot

1. **Sequencing**

For fixed read length sequencing methods (such as Illumina and SOLiD), user should set read length with the parameter of “-read”.

**Output Files:**

**output.fasta**--reads fasta file of Roche or SOLiD sequencing result.

>Psim\_Roche454.1

CCTAACCCTAACCCTAACCCTCGCG

**output.qual**--quality file of Roche or SOLiD sequencing result. The quality score is present by number 0 to 40.

>Psim\_Roche454.1

38 38 38 38 38 38 38 38 38 38 38 38 38 38 38

**output.fastq**--reads and quality file of Illumina sequencing result. The format is standard fastq file format.

@RachelWu:1:BENM:1:1:7572:10362 0:N:0:ATCACG

TCAAACATAAATGAGCAGGCAAGCTGGCTAGAAAACCAC

+

=>?@AABCDEEEEEEEEEEEEEEDDDDDDDDDDDDDDDD

**output-1.fastq** --reads and quality file of Illumina SE forward sequencing result. The format is standard fastq file format.

**output-2.fastq** -- reads and quality file of Illumina SE reverse sequencing result. The format is standard fastq file format.

**NameRecordByPsim.txt**--the relationship between reads name and their physical information. The first column displays reads name; the second column displays its site information include reference name, reads number and start site, end site and length of its corresponding fragment.

>RachelWu\_1\_1\_0 chromosomeGRCh37-1-1040010bp 0 start=744403 end=744493 length=91

**SNPReportByPsim.txt**--SNP

chromosomeGRCh37-1-1040010bp 389093 C A

**SVReportByPsim.txt**--

chromosomeGRCh37-1-1040010bp inversion 240285 2896

chromosomeGRCh37-1-1040010bp deletion 641559 2205

chromosomeGRCh37-1-1040010bp translocation 533138 2394 405275

chromosomeGRCh37-1-1040010bp inversion 565987 3117

chromosomeGRCh37-1-1040010bp inserion

**MutationByPsim.txt**--

chromosomeGRCh37-1-1040010bp 576-2 1 C T

chromosomeGRCh37-1-1040010bp 576-2 59 G C

chromosomeGRCh37-1-1040010bp 580-2 47 T G

**Examples:**

A. SV

1. Based on ancestral diploid reference(~1Mb), simulate a children sequence with SVs, including mentioned above.

2. Fragment to 200, 500, 1000bp hierarchical insert size library, insert span s.d. <35; 10 folds for each library.

3. Used these reads for denovo assembly to be a bunch of contigs or scaffolds (using SOAPdenovo or Phusion-meta or CA or Velvet)

4. Mapped these reads and contigs to reference;

5. Using SV detector to find SV;

6. Draw module and diagnostic depicted figures, and validation/evaluation (FP, FN, TP, TN, Sn, Sp, CC, ACP, AC)

B. Paleogenomics

1. Based on ancestral diploid reference(~1Mb), simulate a children sequence with SNPs/InDels with same most part of same locus as dbSNP, and parts of private diversity (~0.!%).

2. Fragment to 30~200bp with overhanding or only single strand DNA;

3. Add damage by nucleotide substitution according to training stochastic matrix;

4. PCR duplication, ~1000x.

5. Sequencing with sequencing errors (~1%) for each reads via stochastic probability.

6. Mapped these reads to reference;

5. Using GATE to called variation;

6. Find damage (I will do this part);

7. Draw figures for evaluating DNA damage, and their distribution compared to design.

#generate a random reference file length of 10000 bp

perl GenerateSeq.pl -l 10000 > MyNewGenerateRefFile.fa

#generate several(assume 3) references with average length of 10000 bp, and seq pre is \"example\"

perl GenerateSeq.pl -n 3 -l 10000 -auto -pre example > MyNewGenerateRefFile.fa

#generate several(assume 3) references with length of 10000 bp, 20000 bp and 16000 bp

perl GenerateSeq.pl -n 3 -l 10000,20000,16000 > MyNewGenerateRefFile.fa

#generate roche SE sequencing data with no SV or SNP.

#save the output files to ../output/

perl Psim.pl roche --ref ../example/example.fa --snp 0 --sv 0 --dir ../output/

#generate illumina PE sequencing data.

#reference sequence is circle.

#sv information in the file of ../example/sv\_example.txt

#snp information in the file of ../example/snp\_example.txt

#coverage of reference is 10X

#Sequencing with sequencing errors (~0.5%) for each reads via stochastic probability.

#save the output files to ../output/

perl Psim.pl illumina --ref ../example/example.fa --snp ../example/snp\_example.txt --sv ../example/sv\_example.txt

--pe --circle 1 --cov 10 --error 0.005 --dir ../output/

#sv information

#one reference sequence named with \"ChromosomeOne\". It has a 2600bp deletion start from site 300, a 3300bp deletion start from site 900000, a 3000bp inversion start from site 37000, a 1200bp repeat start from stie 26000 and repeat time of 4, a 5000bp translocation with original start site of 43000 and new start site of 60000 and insert \"TTTTTTGGGGGGGGGCCCCA\" to the site of 16350.

#in the sv config file:

ChromosomeOne deletion 300,2600;900000,3300

ChromosomeOne inversion 37000,2600

ChromosomeOne tandem\_repeat 26000,1200,4

ChromosomeOne translocation 43000,5000,60000

ChromosomeOne insertion 16350,TTTTTTGGGGGGGGGCCCCA

#snp information

#one reference sequence named with \"ChromosomeOne\".

#The base at site 1462 has 30% possibility of mutation into \"A\" and 20% possibility of mutation into \"T\"

#The base at site 18209 has 40% possibility of mutation into \"C\"

#The base at site 2840 has 20% possibility of mutation into \"A\", 20% possibility of mutation into \"T\" and 30% possibility of mutation into \"G\"

#in the snp config file:

ChromosomeOne 1462 A,T 0.3,0.2

ChromosomeOne 18209 C 0.4

ChromosomeOne 2840 A,T,G 0.2,0.2,0.3

#damage1

#reference sequence is ../example/example.fa

#coverage of 10%

#SE sequencing

#average length of DNA fragments is 30~40 with the shortest length of 20. And the max length could be reach to about 200bp.

| ..

| . .

| . .

| . .

| .

| .

| .

---------------------------------------->

#overhang length of DNA double strain is 3bp to 20bp of random distribution.

#Single strain library preparation is adopted

#Add damage by nucleotide substitution according to training stochastic matrix.

#PCR duplication, ~1000x.

#lost rate single strain is 50%.

#efficiency of fill-in reaction is 50%

#WARNING:MAKE SURE THAT YOU HAVE ENOUGH MEMARY SPACE MORE THAN ReferenceSize\*Coverage\*Duplication\*3

perl Psim.pl illumina --damage --ref ../example/example.fa --cov 0.1 --fragmean 35 --fragsd 48 --fraglim 20+ --overhang 3\_20 --library 1 --lost 0.5 --ampmean 1000 --mutarray ../example/mutarray\_example.txt --mutsite ../example/mutsite\_example.txt --effic 0.5 --dir ../output/palo/

#RNA-SEQ

#Illumina PE sequencing

#randomly generate snp with the rate of 0.001

perl Psim.pl illumina --ref ../example/example.fa --cdna --gff ../example/gff\_example --covcdna 10 --snp 0.001 --dir ../output/