**Psim 用户使用手册**

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

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

**版本：**1.0.0

**更新日期：**2013.06.16

**依存包：**perl，Math-Random模块

**简要描述：**

Psim是一个用来模拟NGS测序数据的实验软件包，用perl语言编写。它可以模拟固定读长长度（如illumina或SOLiD）或随机读长长度（如Roche 454、Ion Torrent或PacBio）、单端测序（single end sequencing）或者双端测序（pair-end sequencing）、常规样本测序或损伤DNA样本测序、基因组DNA或cDNA测序等。用户可以提供指定的模拟参考序列，或者由程序根据设定长度随机生成参考序列。模拟数据与参考序列之间可以存在结构变异（structure variation）或单核苷酸多态性（single nucleotide polymorphism），这些参数可指定位点、区域及相应概率生成或随机生成。

在Psim程序的流程设计中，模拟了真实样本的处理过程，包括：模板DNA与已知序列的变异、\*DNA片段损伤，\*样品预扩增，\*单链建库/双链建库，建库过程随机打断、加入测序特异性接头（adapter）或PE连接序列（linker）、测序错误的产生、生成测序质量值等。

\*标记特指损伤DNA测序步骤。

**获取及安装：**

Psim软件包可以在sourceforge网站（<https://sourceforge.net/projects/pilotsim/>）下载。

Psim依赖于perl语言，用户需在安装了perl软件的计算机上使用；程序中使用到了Math-Random的生成泊松分布数函数，用户需自行安装该模块。

Perl安装方式：

1. Linux系统中通常以预先安装了perl程序，无需重复安装；
2. Windows系统用户请在activestate网站（<http://www.activestate.com/activeperl>）下载并安装activeperl。

Math-Random模块安装方式:

1. 使用CPAN模块自动安装（需要root权限）：

#Linux用户

perl –MCPAN –e shell

#初次运行需要做一些设置，选择CPAN镜像站点

cpan>install Math::Random

cpan>q

#windows用户

ppm

ppm>install Math::Random

1. 手动安装

在<http://search.cpan.org/~grommel/Math-Random-0.70/Random.pm>网站下载模块源代码，然后在终端中输入：

tar zxvf Math-Random-0.70.tar.gz

cd Math-Random-0.70

perl Makefile.PL

make

make test

make install

#如果没有root权限，请手动设定安装路径

tar zxvf Math-Random-0.70.tar.gz

cd Math-Random-0.70

perl Makefile.PL PREFIX=YOURDIR

make

make test

make install

**功能：**

1. **生成随机序列**

生成n条指定长度或者n条长度符合泊松分布的序列，可用于生成测试的参考序列。

主程序参数：

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”

示例：

#生成5条长度分别为3000bp, 4000bp, 6000bp, 5000bp, 8000bp的序列，序列名称前缀为test-1，序列文件名称为test-1.fa

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

#生成5条平均长度为5000bp的序列, 序列名称前缀为test-2, 序列文件名称为test-2.fa

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

1. **结构变异**

结构变异是指基因组内大于1kb的DNA片段的插入（insertion）、缺失（deletion）、重复（tandem repeat）、倒位（reversion）及易位（translocation），在本程序中也包括短片段的插入和缺失。如设定指定位置变异，用户需在运行模拟程序前编写包含结构变异信息的文件，参见example文件夹中的sv\_example.txt文件，格式为：第一列是参考序列名称（与输入参考序列的名称应完全相同），第二列是结构变异类型（包括deletion、insertion、tandem\_repeat、reversion及translocation），第三列是位点、长度等信息，具体如下：

deletion 起始位点1，长度1；起始位点2，长度2；……

#不同信息之间以分号“;”隔开

insertion 插入位点1，序列1；插入位点2，序列2；……

inversion 起始位点1，长度1；起始位点2，长度2；……

tandem\_repeat 起始位点1，长度1，重复次数1；起始位点2，长度2，重复次数2；……

translocation 起始位点1，长度1，插入位点1；起始位点1，长度1，插入位点1；……

主程序参数：

--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)

示例：

#根据预先设定的结构变异信息文件(如DIR/Psim/example/sv\_example.txt)对参考序列产生变异

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

#随机生成结构变异，变异序列长度占参考序列全长的10%，发生结构变异片段的平均长度是3000bp

--sv 0.1:3000

#无sv变异

--sv 0

1. **单核苷酸多态性**

如设定指定位点的突变，用户需在运行模拟程序前编写包含SNP变异信息的文件，参见example文件夹中的snp\_example.txt文件，格式为：第一列是参考序列名称（与输入参考序列的名称应完全相同），第二列是突变位置，第三列是可能突变的碱基，第四列是突变成不同碱基的概率。三四列不同碱基或数值之间用逗号隔开。

主程序参数：

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

示例：

#根据预先设定的SNP信息文件(如DIR/Psim/input/snp\_info.txt)对参考序列产生变异

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

#随机生成SNP突变，突变位点数目占参考序列全长的0.1%

--snp 0.001

#无SNP突变

--snp 0

1. **文库构建**

文库构建过程包括：样本的随机打断及连接接头序列。

DNA样本随机打断产生DNA短片段（fragment），对于固定读长的测序方法，应尽量保证全部fragment长度大于读长，对于非固定读长的测序方法，fragment长度即为读长。用户需设定fragment的平均长度（-fragmean）、标准差（-fragsd）及是否有长度限制（-fraglim）。类似Illumina的测序方法中，在fragment两端连接接头（adapter）片段；类似Roche 454 PE 测序方法中在fragment首位连接Linker序列。

主程序参数：

--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 ":")

在Roche 454 PE测序方法中：

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

--linker <CHAR> PE insert seq (default=\"ATAACTTCGTATAATGTATGCTATACG AAGTTAT\")

示例：

#模拟5倍覆盖度、fragment平均长度为150、标准差为6、使用默认接头序列

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

#Roche 454 PE，模拟4倍覆盖度、插入片段长度均值8kb、方差35、fragment平均长度为650、标准差为50、使用默认linker序列

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

1. **RNA-seq（cDNA测序）**

根据基因注释gff文件中给出的不同转录本的信息，以转录本为参考序列生成模拟测序数据。gff（General feature format）文件格式请参阅Sanger网站

<http://www.sanger.ac.uk/resources/software/gff/spec.html> 。

主程序参数：

--cdna RNA-SEQ sequencing

--gff <FILE> input gene gff file

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

示例：

#模拟RNA-seq测序，gff文件在DIR/Psim/example/gff\_example.txt，测序覆盖度为20倍

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

1. **损伤DNA测序**

DNA损伤模型为：损伤DNA的两条链末端不能够完全互补，形成粘性末端，粘性末端长度范围由参数—overhang设置；每条DNA片段不同位点都有一定的损伤（变异）概率，不同位点的损伤分布状态，需要用户输入用于训练模型的矩阵，参照DIR/Psim/example/mut site\_example. txt；某些区域会有定向突变（如某种碱基排布结构容易发生某种情况的变异），这一内容可以在SNP或SV的信息中提供。

损伤DNA测序的建库方式有单链建库和双链建库两种方式。单链建库时每条链的变异信息可以独立保留下来；双链建库中的补平末端操作（fill-in reaction）会将不能扩增补平的3’端剪切掉，损失部分的DNA及损伤信息，并且在两端都可扩增的情况双末端扩增概率可有参数--effic设置。在两种方法中DNA片段两条链都存在一定的丢失率，用户可通过参数--lost或--ds设置。

损伤DNA由于样本量小，建库前期会先进行预扩增扩大样本量，PCR扩增倍数可由参数--ampmean设置，扩增效率符合泊松分布。

主程序参数：

--damage the sequencing sample have overhanging and injured end

--overhang <INT> overhang length range of double strain (default= "3\_20")

--library <INT> library preparation type (1=single strain,2=double strain, default=1)

--ds <NUM> rate of lost one strain of double strain library (default=0.1)

--lost <NUM> lost rate single strain while library=1(default=0.5)

--ampmean <NUM> average amplification times (default=850)

--mutarray <FILE> mutation rate array file (format refer to ../examp le/mutationarray\_example, default 100% C->T)

--mutsite <FILE> mutation possibility along fragment site (format refer to ../example/mutationsite\_example, default 0.01)

--effic <NUM> efficiency of fill-in reaction (default=0.5)

示例：

#模拟损伤DNA测序，末端非互补区域长度范围是3nt~20nt，采用双链建库方式，建库过程单链损失率为30%，双末端扩增效率为30%，损伤模型位点矩阵文件是DIR/Psim/example e/mutsite\_example.txt, 损伤碱基转换矩阵文件是DIR/Psim/example/mutarray\_example.txt，PCR扩增倍数均值为800倍。

--damage --overhang 3\_20 --library 2 --ds 0.3 --ampmean 800 --mutarray DIR/Psim/example/mutarray\_example.txt --mutsite DIR/Psim/example/muts ite\_example.txt --effic 0.3

1. **测序**

对于固定读长的测序方法（如Illumina和SOLiD），需用户在--read参数中设定读长长度；对于SOLiD测序，还需要指定测序起始碱基类别。在测序过程中，同时产生每个碱基的质量值，值域为0~40，SOLiD和Roche 454测序结果中直接以数值表示，Illumina用ASCII码表示，根据不同的fastq格式有两种表示范围（Sanger/Illumina），详情参见<http://en.wikipedia.org/wiki/FASTQ_format>。

主程序参数：

--read <INT> average length of reads(default=100)

--error <NUM> sequencing error rate and possibilities of single base error, insert and deletion (default="0.0005:0.34:0.33:0.33")

--qtype <CHAR> quality type (!=Sanger format, @=illumina 1.3~1.8- format, default=!)

--qmean <NUM> peak value of quality score(default=37)

--qsd <NUM> standard deviation of quality score (default=1)

SOLiD

--header <CHAR> sequencing header base (default=G)

示例：

#读长长度202bp，测序错误率为0.001，三种类型（error，insert和deletion）的比重分别为0.5,0.25,0.25，平均质量值35，标准差1.5，质量值表示方式为Sanger类型。

--read 202 --error 0.001:0.5:0.25:0.25 --qtype ! --qmean 35 --qsd 1.5

#读长长度50bp，测序错误率为0.0002，三种类型（error，insert和deletion）的比重分别为0.5,0.25,0.25，平均质量值37，标准差1，SOLiD测序，起始碱基为A。

--read 50 --error 0.0002:0.5:0.25:0.25 --qmean 37 --qsd 1 --header A

**生成文件：**

**output.fasta**--Roche或SOLiD的测序reads文件，一行名称一行序列相间隔，序列名称前以“>”开头；

>Psim\_Roche454.1

CCTAACCCTAACCCTAACCCTCGCG

**output.qual**--Roche或SOLiD的测序质量文件，一行名称一行质量值相间隔，序列名称前以“>”开头，质量值范围0~40，用数字直接表示；

>Psim\_Roche454.1

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

**output.fastq**--llumina SE的测序reads及质量文件，标准fastq文件格式；

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

TCAAACATAAATGAGCAGGCAAGCTGGCTAGAAAACCAC

+

=>?@AABCDEEEEEEEEEEEEEEDDDDDDDDDDDDDDDD

**output-1.fastq** --llumina PE的测序正向reads及质量文件，标准fastq文件格式；

**output-2.fastq** --llumina PE的测序反向reads及质量文件，标准fastq文件格式；

**NameRecordByPsim.txt**--reads名称与其在参考序列上的位置、长度对应关系，第一列为reads名称，第二列是其在参考序列上的位置信息，包括：参考序列名称、reads编号、fragment在参考序列上的起始位点、终止位点、长度

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

**SNPReportByPsim.txt**--生成的SNP信息，每列信息分别是参考序列名称、SNP位点、突变前碱基、突变后碱基

chromosomeGRCh37-1-1040010bp 389093 C A

**SVReportByPsim.txt**--生成的SV信息，

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**--损伤DNA测序模拟生成的损伤信息，每列信息分别是参考序列名称、reads信息编号、突变位点、突变前碱基、突变后碱基

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

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

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

**示例：**

1）用默认参数生成Illumina测序数据，生成的文件存放于Psim/output文件夹中

perl Psim.pl illumina –-ref ../example/example.fa –-dir ../output

2）用默认参数生成Roche测序数据，生成的文件存放于Psim/output文件夹中

perl Psim.pl roche –-ref ../example/example.fa –-dir ../output

3）用默认参数生成SOLiD测序数据，生成的文件存放于Psim/output文件夹中

perl Psim.pl solid –-ref ../example/example.fa –-dir ../output

4）用示例参考序列生成15倍覆盖度、Illumina测序方法、PE测序、无SV及SNP变异、

perl Psim.pl illumina –ref ../example/example.fa –cov 15 –sv 0 –snp 0 –-dir ../output

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.