simulate_illumina_reads_v3.0 说明文档

核酸序列构建单元 B　2011-8-3

程序路径：/ifs1/ST_ASMB/PMO/bin/Simulate_illumina_reads/Simulate_illumina_reads_v3.0

文档作者：袁剑颖 yuanjianying@genomics.org.cn; 史玉健 shiyujian@genomics.org.cn;

1. 程序简介

1.1 背景目的

针对 illumina（solexa）测序仪生产的数据而进行分析处理的一系列程序流程，包括数据处理、Contig 构建、Scaffold 构建、补洞程序等，需要使用高仿真的模拟数据来进行测评分析，从而较为准确地定义各程序流程的性能效果以及改进与否。

基于以上目的，使用本程序“simulate_illumina_reads_v3.0”模拟构造具有 Pairs-end 关系的 reads 文库，能较为真实地模拟 illumina 测序数据特点，包括插入片段长度分布、GC bias、测序错误分布、质量值等，同时，针对二倍体基因组，模拟杂合现象。

1.2 定义

a) read:测序读到的碱基序列段，测序的最小单位。
b) PE-reads: 对一定长度的 DNA 片段进行双末端测序得到的成对的 reads。
c) heterozygosis SNP：Single Nucleotide Polymorphisms，杂合现象中的单核苷酸多态性，是指个体间基因组 DNA 序列同一位置单个核苷酸改变所引起的多态性。
d) heterozygosis Indel：insertion-deletion, 杂合现象中的插入、缺失多态性。
e) GC bias:GC 偏离量，建库、测序过程中基因组不同 GC 含量的区域测序深度存在的偏向性。

1.3 Version3.0 更新改进说明

a) 此版本作为数据分析流程的其中一块并入数据分析流程包内，根据数据分析程序的统计结果进行模拟。数据统计流程见“illumia_reads_parameter_stator”程序包。
b) 模拟 SNP 时增加设置转换颠换比值。
c) 模拟结构变异时分为两种类型：1~6bp 的 small indel 以及较大的结构变异（见 2.3 使用注意事项 c）。
d) 规范化 indel, invertion 的坐标表示（见 3.4 文件格式说明）。
e) 测序错误、质量值以及 GC bias 按照输入的统计配置文件进行模拟。使模拟结果更接近于真实数据。
f) 增加两个参考序列处理程序 make_pure_ref_genome.pl 以及 split_scaffold.pl，用于处理带 N 的参考序列（见 2.3 使用注意事项 a）。

2. 程序使用说明

2.1 程序集架构

a) 程序主要包含两个可以独立使用的可执行文件：
   - simulate_snp_indel_seq ：用于模拟二倍体基因组，对输入基因组参考序列加入 Snp，Indel，invertion 位点，并输出模拟后的基因组参考序列。
   - simulate_illumina_reads：用于模拟 illumina 测序数据，输出 PE-reads 文件。可用于单倍体以及二倍体的 illumina 测序数据模拟。

b) 为方便用户使用， run_simulate_reads.pl 作为外部包装程序，自动调用以上两个可执行文件，协助模拟二倍体测序数据，当设置程序运行模式为任务提交模式时（即

qsub 模式），会调用 qsub-sge.pl 控制任务投递。

2.2 参数使用说明

perl run_simulate_reads.pl [option] <lib.lst>

【option】

-i <string> 输入基因组参考序列 *.fa 或者 *.fa.gz 格式
-f <string> 输入测序错误统计配置文件，默认值:（程序目录）/statistical_file_package/error_profile/ hum20110701.bwanosnp.count.mtarix
-F <string> 输入 GC 深度统计配置文件，默认值:（程序目录）/statistical_file_package/GC_depth_profile/ stat_100.dat.gc
-s <double> 设置杂合（SNP）比率，如 0.001 表示千分之一，默认值：0
-a <double> 设置转换颠换比值，默认值 2（表示转换：颠换=2:1）
-d <double> 设置插入缺失（Indel）比率，如 0.0001 表示万分之一，默认值：0
-v <double> 设置大的结构变异（big Indel/invertion）比率，如 0.000001 表示百万分之一，默认值：0
-e <double> 设置测序平均错误率，如 0.01 表示百分之一错误率，默认值：-1：测序错误配置文件的平均错误率
-g <int> 设置是否模拟 GC bias， 0:no, 1:yes, 默认值:1
-q <int> 设置是否模拟质量值, 0:no, 1:yes, 默认值:1
-j <int> 设置最大运行任务数, 默认值: 1
-o <string> 设置输出的 PE-reads 文件前缀，默认值：illumina
-c <int> 设置输出文件格式，0：文本输出，1：压缩输出，默认值：1
-r <int> 设置程序运行方式，1：直接运行（在测试节点）， 0:任务提交(qsub), 默认值:0
-m <int> 设置任务提交模式下申请的内存大小，单位为 G，默认值：0.5
-h 输出用户帮助信息

<lib.lst>

在 lib.lst 文件中，设置需要模拟的各文库信息，包括以下内容：

a) 插入片段长度期望值 <int>
b) read 长度 <int>
c) 测序乘数即碱基覆盖度 <double>
d) 插入片段长度标准差 <int> 一般情况下设置为：插入片段长度平均值/20
e) 设置是否环化处理 <int> 影响 PE-reads 方向, 0:read1 正向 read2 反向, 1:read1 反向，read2 正向, 默认值:0

用户必须按照以下参数顺序创建 lib.lst 文件内容，每一行为一个文库信息，并使用单个或多个空格隔开各数值

写入顺序：插入片段长度期望值/ read 长度/测序乘数/插入片段长度标准差/是否环化处理

例如：在 lib.lst 文件中写入以下信息：

170	100	20	10	0
500	100	20	25	0
800	100	10	40	0
2500	100	5	125	1

程序使用示例：

1. perl run_simulate_reads.pl -i ref_sequence.fa lib.lst
   说明：各参数按照默认值设置，注意-i 参数必须设置，以及 lib.lst 内容必须设置。
2. perl run_simulate_reads.pl -i ref_sequence.fa -g 1 -q 1 lib.lst
   说明：设置模拟 GC bias 以及质量值，当模拟质量值时，生成文件为*.fq 格式。
3. perl run_simulate_reads.pl -i ref_sequence.fa –e 0 –c 0 lib.lst
   说明：设置测序平均错误率为 0，以文本形式（非压缩）输出。
4. perl run_simulate_reads.pl –i ref_sequence.fa -j 2 -o test lib.lst
   说明：设置最大运行任务数为两个，并设置输出的 PE-reads 文件前缀名为 test。
5. perl run_simulate_reads.pl -i ref_sequence.fa -r 0 -m 0.8 lib.lst
   说明：设置运行模式为提交任务以及申请的内存大小为 0.8G。
6. perl run_simulate_reads.pl -i ref_sequence.fa -s 0.001 -d 0.0001 –a 2 –v 0.000001 lib.lst
   说明：二倍体基因组序列数据模拟：设置 snp 比率为千分之一，转换颠换比值为 2:1，indel 比率为万分之一，大结构变异比率为百万分之一。

2.3 使用注意事项

a) 输入参考序列为 fa 格式，如果带有“N”字符，可使用程序包内的 make_pure_ref_genome.pl 程序将“N”去除，得到纯基因碱基的序列，或者使用 split_scaffold.pl 程序，将带“N”的序列分割为多段 contig 序列。也可以不作处理，模拟程序在遇到“N”字符的序列区域时，自动跳过该区域不进行模拟，得出的 read 序列不会包含“N”字符。用户根据具体需求而选择策略。

b) 2.2 中使用说明为 illumina 测序仪 PE-read 数据模拟，如果用户只需要单独模拟二倍体基因组序列，给原始基因组序列加入 Snp 或者 Indel 位点并输出模拟后的基因组参考序列，则可以单独使用“simulate_snp_indel_seq”该执行文件，具体用法可以输入“(程序目录)/simulate_snp_indel_seq -h”查看。

c) 二倍体杂合是在染色体上随机的位点产生 snp、insertion、deletion、invertion，在模拟结构变异（SV）时，程序分两类情况进行设置模拟，一，small indel（-d 参数设置）： 插入缺失各占总比例的一半，indel 碱基为 1~6 个，其概率分布依照熊猫基因组而来：1bp-64.82%, 2bp-17.17%, 3bp-7.20%, 4bp-7.29%, 5bp-2.18%, 6bp-1.34%。二，big SV（-v 参数设置），包含插入、缺失、颠倒（invertion），各占总比例的三分之一，程序简单提供模拟一些较大的结构变异，其长度分布如下：100bp-70%, 200bp-20%, 500bp-7%, 1000bp-2%, 2000bp-1%。

d) 在 PE-read 数据模拟时，必须注意只有二倍体才会有杂合性状，可以通过设置 -s –d –v 参数分别设置 snp 和结构变异比率，当模拟的是单倍体生物时，-s –d -v 参数无需设置或者设置为 0。

e) –r 参数设置程序运行方式，当用户所在节点为测试节点时，设置-r 1 直接运行程序，

当所在节点为任务提交节点（登陆节点）时，无需设置-r 或者设置为 0。

f) Lib.lst 内容必须按照规定的顺序写入，不可省略部分参数设置，每行代表一个插入片段文库，各数值以单个或者多个空格符隔开。

g) 错误率、质量值以及 GCbias 模拟是根据配置文件进行操作的，必须保证输入的配置文件格式正确，用户可以使用程序包内 statistical_file_package 文件夹下的配置文件，也可以根据数据分析流程 “illumia_reads_parameter_stator”自行统计得出。

g) 模拟 read 的最大读长由配置文件统计的 Cycle 数决定，read_length 必须小于等于 Cycle/2。

h) 平均错误率（-e）的设置：默认的平均错误率为配置文件统计的平均错误率，当用户设置 read 的读长决定后，程序会根据用户输入读长从配置文件中读入相应 Cycle 的数据，并按照读入的 Cycle 数值决定平均错误率的大小。用户对-e 参数不进行设置或者设置为-1 时，按照默认的平均错误率模拟。在模拟零错误率时，用户可以设置-e 为 0，如果用户需要模拟一定量的平均错误率，程序会根据用户输入的错误率与配置文件的平均错误率作等比例地模拟。如：假设配置文件的平均错误率为 0.5%，用户输入-e 0.01 即 1%平均错误率，程序将按照配置文件中每个 Cycle 错误率的两倍模拟测序错误。

3. 结果文件说明

3.1 二倍体基因组序列模拟结果

如：./ simulate_snp_indel_seq –i Human_ref.fa –s 0.001 –a 2 –d 0.0001 –v 0.000001 –o Human > simulate_seq.o 2> simulate_seq.e

结果输出：

a) Human.snp.indel.invertion.fa.gz ： 模拟加入 snp & SV 后的基因组序列。

b) Human_indel.lst ：模拟 indel 信息的列表。

c) Human_snp.lst：模拟 snp 信息的列表。

d) Human_invertion.lst： 模拟 invertion 信息的列表。

e) simulate_seq.o 和 simulate_seq.e ：重定向输出文件，记录程序运行信息。

3.2 单倍体 illumina 测序数据模拟结果

如：

假设用户设置 lib.lst 文件中内容为：

170 100 20 10 0

500 100 10 20 0

800 100 10 40 0

程序设置为：

perl ./run_simulate_reads.pl -i Saccharomyces_ref.fa -e -1 -g 1 –q 1 –j 2 –r 1 –o Saccharomyces lib.lst >simulate_saccharomyces_reads.o 2> simulate_saccharomyces_reads.e

结果输出：

创建了三个文件目录：Saccharomyces_170_100_20_10_0 Saccharomyces_500_100_10_20_0 Saccharomyces_800_100_10_40_0

每个目录下分别包含两个*.fq.gz 文件，对应 PE 关系的 reads 文件，一个插入片段长度分布文件，一个错误率分布文件以及其他日志文件记录模拟信息和程序运行信息。

3.3 二倍体 illumina 测序数据模拟结果

如：

假设用户设置 lib.lst 文件中内容为：

170 100 20 10 0
500 100 10 20 0
800 100 10 40 0
2500 100 5 100 1

程序设置为：

perl ./run_simulate_reads.pl -i Human_ref.fa –s 0.001 –d 0.0001 –v 0.000001 -e 0 -g 1 –q 0 –j 2 –r 0 –m 1.4 –o Human lib.lst >simulate_Human_reads.o 2> simulate_Human_reads.e

结果输出：

a) Human.snp.indel.invertion.fa.gz ： 模拟加入 snp & SV 后的基因组序列。
b) Human_indel.lst ： 模拟 indel 信息的列表。
c) Human_snp.lst ： 模拟 snp 信息的列表。
d) Human_invertion.lst: 模拟 invertion 信息的列表。
e) simulate_seq.o 和 simulate_seq.e ： 重定向输出文件，记录程序运行信息。
f) 创建了三个文件目录：Human_170_100_20_10_0 Human_500_100_10_20_0 Human_800_100_10_40_0 Human_2500_100_5_100_1
每个目录下分别包含两个*.fa.gz 文件，对应 PE 关系的 reads 文件，一个插入片段长度分布文件，一个错误率分布文件，以及其他日志文件记录模拟信息和程序运行信息。

3.4 文件格式说明

a) *.fa/*.fa.gz
如：文件名*_100_500_1.fa 表示读长 100bp,平均插入片段长度 500bp, read1 文件内容：
>read_500_1/1 I 82473 100 505
TAGAAAAACCAGAGTGGT……..GTACGTTGGGGGCTCGTTTGTCGA
>read_500_2/1 I 9704 100 485 74,G;
CCTTTCATAACTAAACCAA……TGCAGAAATGTCATGGATACCATT
第一行中，“500”表示平均插入片段长度；接下来的“1”表示该文件第几条 read；最后一个“1”表示 read1；“I”表示参考序列 id，“82473”表示该 read 在参考序列上的起始位标；“100”表示 read 读长；“505”表示该 read 实际插入片段长度。
第三行中，“read_500_2/1 I 9704 100 485”对应以上解释，“74”表示错误碱基在该 read 的位标（以 1 为起始位标）；“G”表示对应位标的正确碱基应为 G。
b) *_snp.lst
如：

I 39540 G A
I 45342 C T
I 104775 C T

I 220818 C G

第一列表示参考序列 id; 第二列表示 snp 位标(以 1 为起始位标)，第三列表示原始参考序列碱基，第四列表示 snp 碱基。

如第一行表示：原始染色体 I 的第 39540 个碱基由 G 转换为 SNP 位点 A。

c) *_indel.lst

如：

I - 4280 1 C
I - 104003 1 G
I - 81587 1 G
I + 206841 2 TC

第一列表示参考序列 id，第二列表示缺失“-”或者插入“+”，第三列表示 indel 位标（以 1 为起始位标），注意 indel 位标为开始改变碱基的前一个碱基位标，如：第一行中表示在原参考序列上的第 4281 位标碱基缺失，最后一行表示在原参考序列上的 206841 位标碱基后加入两个碱基 TC。第四列表示 indel 碱基个数，第五列表示 indel 的碱基。

d) *_invertion.lst

如：

I 50191 100
I 948984 200

第一列表示参考序列 id，第二列表示 invertion 位标（以 1 为起始位标），注意 invertion 位标表示发生颠换碱基区域的前一个碱基位标，第三列表示发生颠换碱基的长度。

如第一行表示：在参考序列 I 上，第 50192~50291 位标的碱基区域发生颠换。

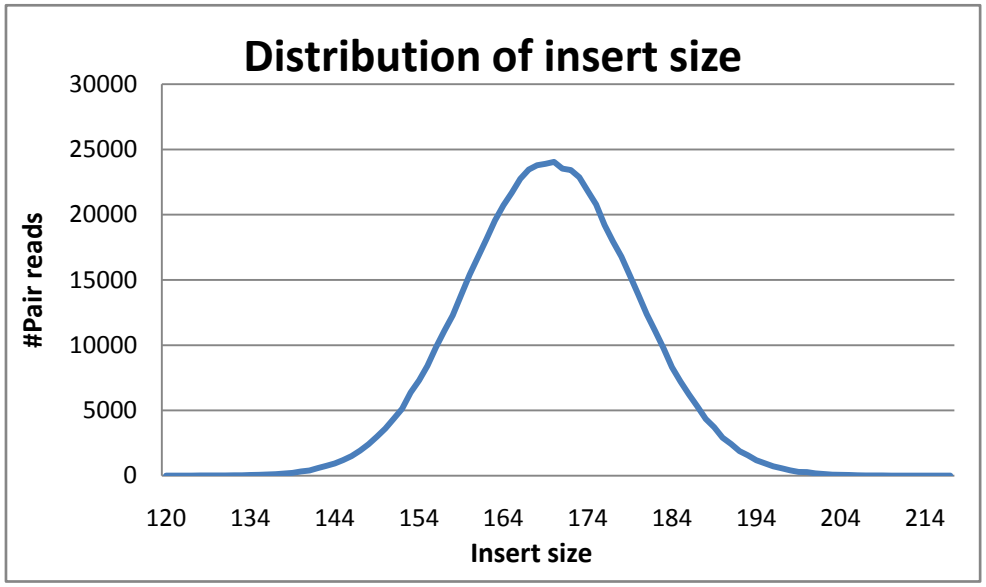
4. 测试分析

4.1 插入片段长度分布

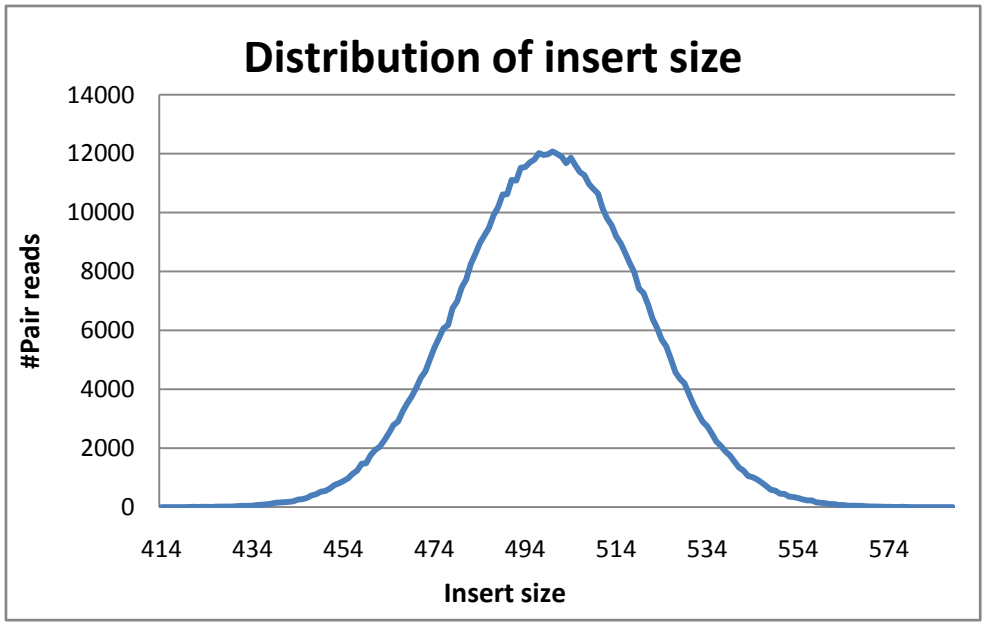
模型：

$$F(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}(\frac{x-\mu}{\sigma})^2}$$

a) Saccharomyces; insertsize mean:170 insertsize-sd:10 coverage: 10X

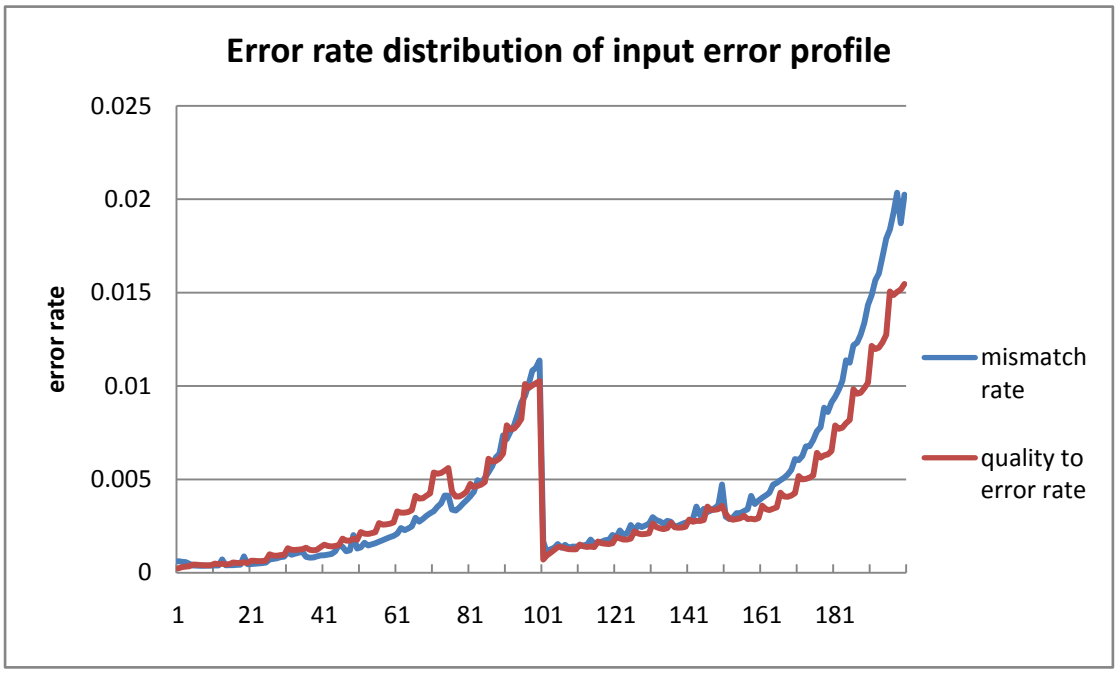


b) Saccharomyces; insertsize mean:500 insertsize-sd:20 coverage: 10X

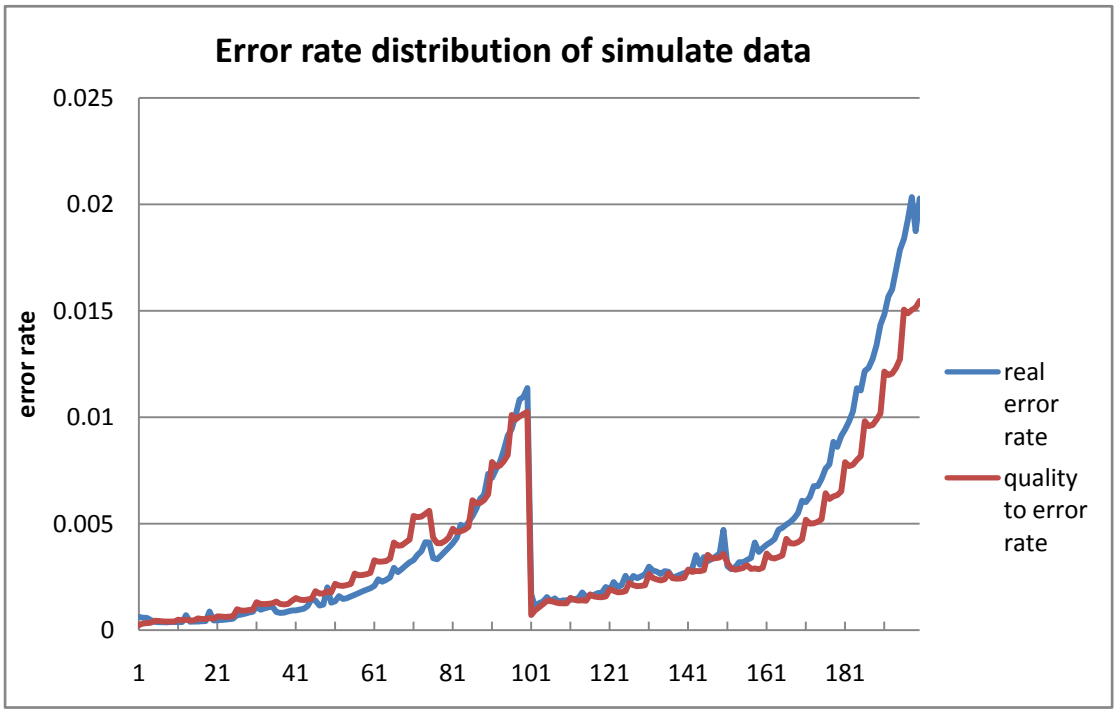


4.2 错误率分布

a) 配置文件的错误率分布



b) 实际模拟数据的错误率分布



说明： 输入配置文件的错误率分布与根据该配置文件模拟出的数据错误率分布一致。

4.3 时间内存使用

程序内存使用与基因组中最长的参考序列有关，内存峰值约为最长序列的 5 倍大小，如：人的全基因组中，最长染色体序列长度约为 237M bp，则内存峰值约为 1G 大小。实际内存与时间使用见如下测试：

a) 二倍体基因组序列模拟：

- 测试参考序列为人基因组：

/ifs1/ST_ASMB/USER/yuanjy/data/noGap_genome/Homo_sapiens.NCBI36.52.dna.toplevel.genome.chr.clean.fa.gz
模拟 snp:0.001 indel:0.0001 big SV: 0.000001
注：输入输出均为压缩形式
时间：21min
内存峰值：1.1G
内存平均值：240M

- 测试序列为拟南芥基因组：

/ifs1/ST_ASMB/USER/yuanjy/data/noGap_genome/Arabidopsis_thaliana_TIGR5.clean.fa.gz
模拟 snp:0.001; small indel：0.0001; big SV: 0.000001
注：输入输出均为压缩形式
时间：60second
内存峰值：155M
内存平均值：47M

- 测试参考序列为酵母（Yeast）基因组：（酵母并非为二倍体基因组，此处模拟只用于测试）

/ifs1/ST_ASMB/USER/yuanjy/data/noGap_genome/Saccharomyces_cerevisiae.SGD1.01.52.dna.toplevel.chr.clean.fa.gz
模拟 snp:0.001; small indel：0.0001; big SV: 0.000001
注：输入输出均为压缩格式。
时间：6 second
内存峰值：10M
内存平均值：5.5M

b) Illumine 测序数据模拟：

- 测试参考序列为人基因组：

/ifs1/ST_ASMB/USER/yuanjy/data/noGap_genome/Homo_sapiens.NCBI36.52.dna.toplevel.genome.chr.clean.fa.gz
模拟 coverage:5X ;
注：输入输出均为压缩形式，二倍体基因组模拟（输入两条参考序列，模拟杂合后的基因组序列为 4.3 a）中输出的基因组序列，压缩形式）
时间： 3.16hour
内存峰值：870M
内存平均值：340M

- 测试参考序列为拟南芥基因组：

/ifs1/ST_ASMB/USER/yuanjy/data/noGap_genome/Arabidopsis_thaliana_TIGR5.clean

.fa.gz
模拟： coverage:10X;
注：输入输出均为压缩形式，二倍体基因组模拟（输入两条参考序列，模拟杂合后的基因组序列为 4.3 a）中输出的基因组序列，压缩形式）
时间：17min
内存峰值：120M
内存平均值：68M

- 测试参考序列为酵母（Yeast）基因组：
/ifs1/ST_ASMB/USER/yuanjy/data/noGap_genome/Saccharomyces_cerevisiae.SGD1.01.52.dna.toplevel.chr.clean.fa.gz
模拟 coverage:20X;
注：输入输出均为压缩形式，单倍体基因组模拟（输入单条参考序列）
时间：219second
内存峰值：8.3M
内存平均值：4.8M

5. 结论

5.1 能力

“simulate_illumina_reads_v3.0”程序能有效地模拟 illumina 测序数据特点，包括参考序列上随机位点的选取，插入片段长度的正态分布，由真实数据分析、统计构建出的测序错误分布模型，质量值分布模型，GC bias 模型以及二倍体的杂合现象。测序错误分布以及质量值模拟相对 V2.0 版本有了进一步的提升。

5.2 限制

由于数据模拟所考虑的因素有限，实际为按照某几种特定模型所产生的理想数据，其远小于真实数据模型的复杂度，并且随着测序实验技术的发展，模型也随之改变，因此，本程序需要对特定时间段的一定量真实数据进行统计分析，以便持续更新、改进模型，用户可按照数据分析流程对数据进行分析统计得出配置文件作为更新使用。

6. 参与人员

Version 1.0（2010-03-01 完成）：鲁建亮、岳震。
Version 2.0（2011-04-28 完成）：樊伟、史玉健、胡雪松、袁剑颖。
Version 3.0（2011-08-03 完成）：樊伟、史玉健、胡雪松、袁剑颖。