simulate_illumina_reads_v2.0 说明文档

核酸序列构建单元 B　2011-5-5

程序路径：/ifs1/ST_ASMB/USER/yuanjy/bin/DataProcessing/Simulate_illumina_reads_v2.0

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_v2.0”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 Version2.0 更新改进说明

a) 更正二倍体模拟 Snp、Indel 时，不同插入片段文库产生的 Snp&Indel 位点不一致的问题。
b) 使用 Box-muller 方法模拟插入片段长度正态分布。
c) 错误率分布曲线由原来的 f(x)=0.00001*x**3 调整为 f(x)=0.00001*x**4；
d) 加入测序错误偏向性模拟。
e) 加入 GC bias 模拟。
f) 加入质量值模拟。
g) 增设环化与否设置，即 direction, FR 或者 RF，用来解决 fosmid 模拟问题。
h) 支持读写压缩文件。
i) 重写程序集架构（见 2.1 说明）

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 以及 Indel 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
-s <float> 设置杂合（SNP）比率，如 0.001 表示千分之一，默认值：0
-d <float> 设置插入缺失（Indel）比率，如 0.0001 表示万分之一，默认值：0
-e <float> 设置测序平均错误率，如 0.01 表示百分之一错误率，默认值：0.01
-g <int> 设置是否模拟 GC bias， 0:no, 1:yes, 默认值:0
-q <int> 设置是否模拟质量值, 0:no, 1:yes, 默认值:0
-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 <float>
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	20	0
800	100	10	40	0
2500	100	5	100	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 0.01 –c 0 lib.lst
   说明：设置测序平均错误率为百分之一，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 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，indel 比率为万分之一。

2.3 使用注意事项

a) 输入参考序列为 fa 格式，不允许有“N”的序列（需过滤存在洞的序列）。
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，产生的 indel 分别有 1base、2base、3base，比例为 3:2:1。.
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 参数分别设置 snp 和 indel 比率，当模拟的是单倍体生物时，-s –d 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) 质量值模拟时，模型为每个 cycle 中正确碱基和错误碱基对应的质量值分布，当前 v2.0 统计模型在 100cycle 范围内，即 read 长度 100bp 以内，因此，当模拟 read 长度大于 100bp 时，不支持模拟质量值，请将-q 参数设为 0。

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 –d 0.0001 –o Human > simulate_seq.o 2> simulate _seq.e

结果输出：

a) Human.snp.indel.fa.gz ： 模拟加入 snp & indel 后的基因组序列。
b) Human_indel.lis ：模拟 indel 信息的列表。
c) Human_snp.lis：模拟 snp 信息的列表。
d) 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 0.01 -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 -e 0.02 -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.fa.gz ： 模拟加入 snp & indel 后的基因组序列。
b) Human_indel.lis ：模拟 indel 信息的列表。
c) Human_snp.lis ：模拟 snp 信息的列表。
d) simulate_seq.o 和 simulate _seq.e ：重定向输出文件，记录程序运行信息。
e) 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。

b) *_snp.lis
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，第三列表示原始参考序列碱基，第四列表示 snp 碱基。

c) *_indel.lis
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，第四列表示 indel 碱基个数，第五列表示 indel 的碱基。

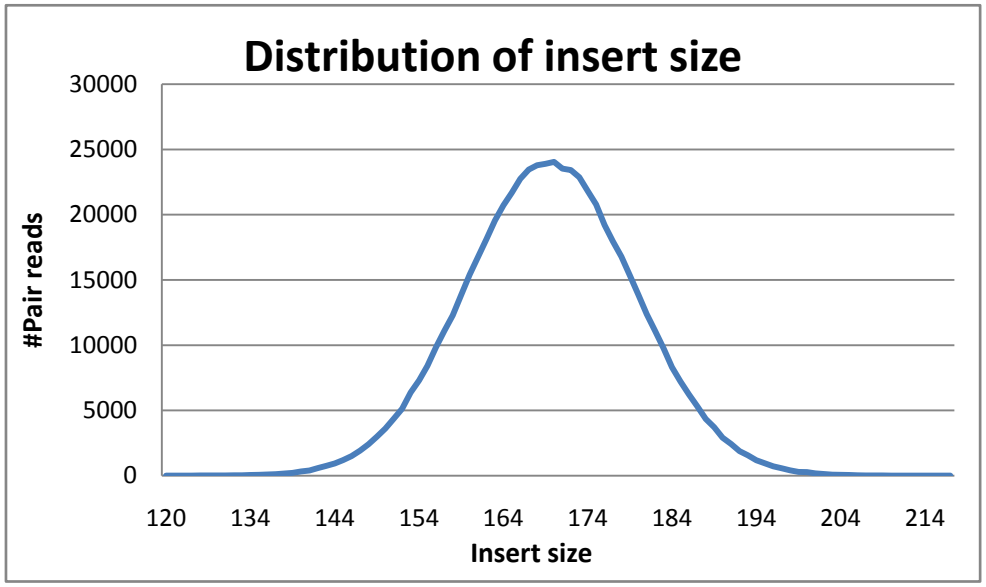
4. 测试分析

4.1 插入片段长度分布

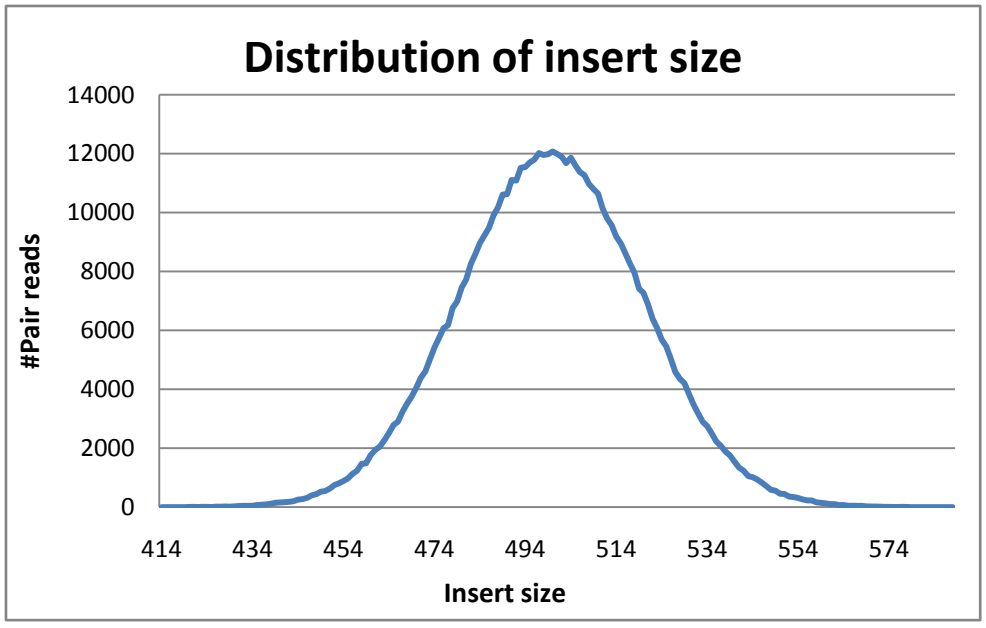
模型：

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

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



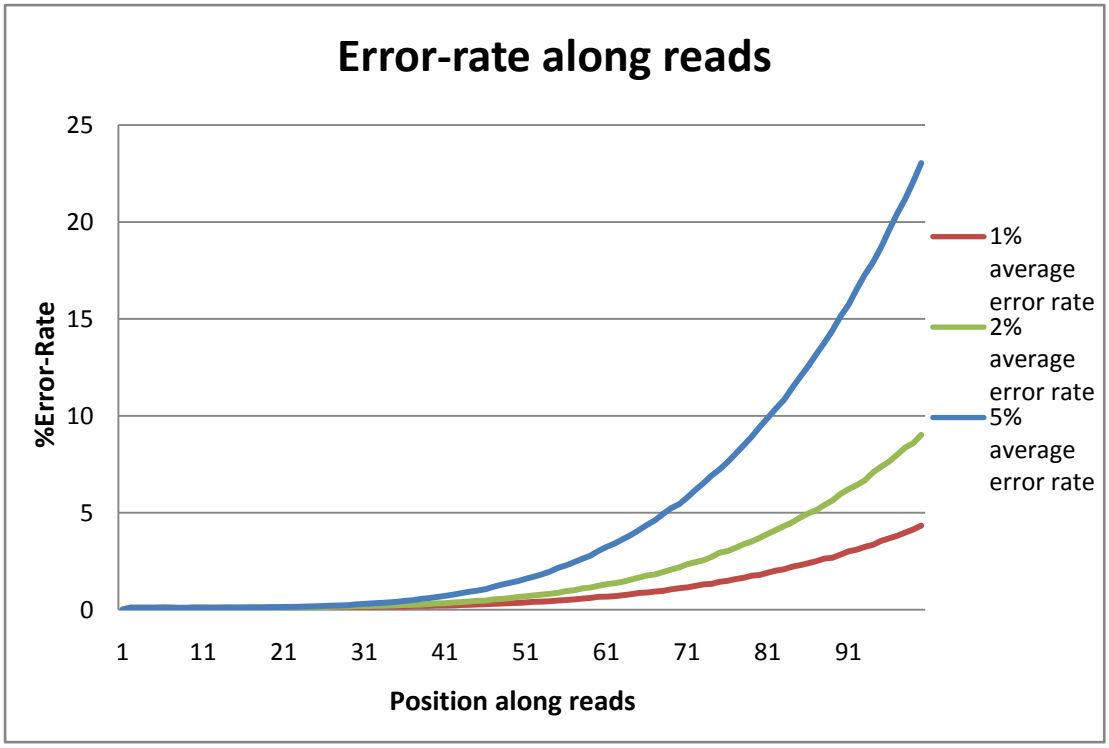
b) Saccharomyces; insertsize mean:500 insertsize-sd:20 coverage: 10X



4.2 错误率分布

分布曲线模型：

$$F(x) = 0.00001x^4$$



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，模拟二倍体基因组时，只模拟 snp 的内存峰值约为最长序列的 5 倍，如果有 indel 模拟，内存峰值约为最长序列的 7 倍大小。模拟 illumina 测序数据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，则模拟包含 snp 的二倍体或者 illumina 测序数据，使用的内存峰值约为：1G 内存。模拟包含 indel 的二倍体使用的内存峰值约为：1.4G。

a) 二倍体基因组序列模拟：

- 测试参考序列为人基因组：

/ifs1/GAG/assemble/fanw/Assembly-2011/Reference_genomes/nonGap_genomes/Homo_sapiens.NCBI36.52.dna.toplevel.genome.chr.clean.fa.gz

模拟 snp:0.001 indel:0.0001

注：输入输出均为压缩形式

时间：14min

内存峰值：1.3G

内存平均值：420M

- 测试序列为拟南芥基因组：

/ifs1/GAG/assemble/fanw/Assembly-2011/Reference_genomes/nonGap_genomes/Arabidopsis_thaliana_TIGR5.clean.fa.gz

模拟 snp:0.001 indel：0.0001

注：输入输出均为压缩形式

时间：40second
内存峰值：160M
内存平均值：72M

- 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/GAG/assemble/liubinghang/bin/Train/genome/Saccharomyces_cerevisiae.SGD1.01.52.dna.toplevel.chr.clean.fa
模拟 snp:0.001 indel:0.0001
注：输入为文本格式，输出为压缩格式。
时间：4 second
内存峰值：12M
内存平均值：8M

b) Illumine 测序数据模拟：

- 测试参考序列为人基因组：

/ifs1/GAG/assemble/fanw/Assembly-2011/Reference_genomes/nonGap_genomes/Homo_sapiens.NCBI36.52.dna.toplevel.genome.chr.clean.fa.gz
模拟 error_rate:0.01; coverage:3X;
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： 56min
内存峰值：900M
内存平均值：600M

- 测试参考序列为拟南芥基因组：

/ifs1/GAG/assemble/fanw/Assembly-2011/Reference_genomes/nonGap_genomes/Arabidopsis_thaliana_TIGR5.clean.fa.gz
模拟：error_rate 0.01; 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：8min
内存峰值：133M
内存平均值：100M

- 测试参考序列为酵母（Yeast）基因组：

/ifs1/GAG/assemble/fanw/Assembly-2011/Reference_genomes/nonGap_genomes/Homo_sapiens.NCBI36.52.dna.toplevel.genome.chr.clean.fa.gz
模拟 error_rate:0.01; coverage:20X;
注：输入为文本格式，输出为压缩形式，单倍体基因组模拟（输入单条参考序列）
时间：1.5min
内存峰值：8.4M
内存平均值：6M

5. 结论

5.1 能力

“simulate_illumina_reads_v2.0”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，测序错误偏向性模型，GC bias 模型以及二倍体的杂合现象。

5.2 限制

质量值模拟依然是难点所在，模型复杂度有待提高。

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。

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）：由樊伟发起指导，史玉健以及胡雪松负责数据分析统计、建模，并由袁剑颖完成代码的修订以及测试。