simulate_solexa_reads

鲁建亮 岳震

功能

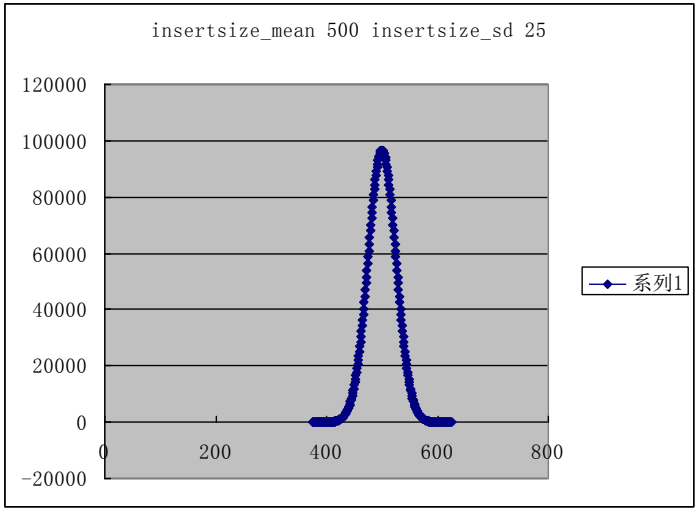
- 模拟Pair-End reads
- 模拟insertsize分布
- 模拟错误率
- 模拟二倍体杂合snp和indel

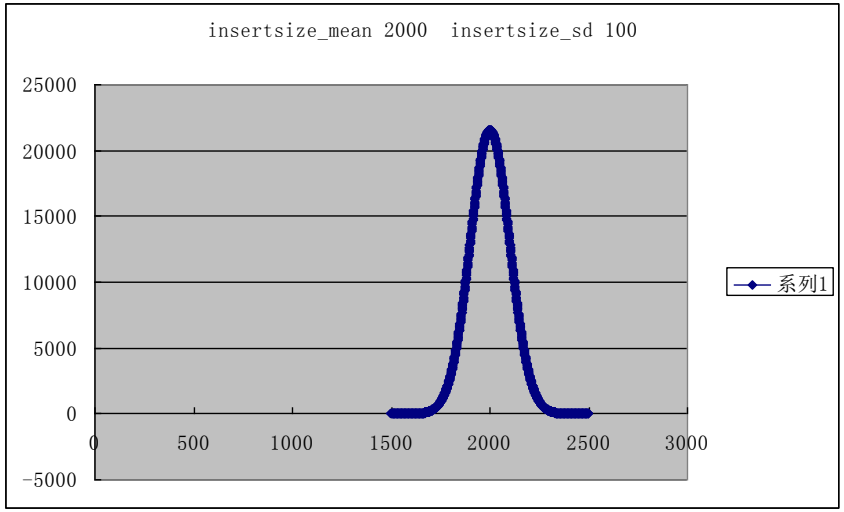
Insertsize分布

- 模型：

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

示例


insertsize_mean 500 insertsize_sd 25
120000
100000
80000
60000
40000
20000
0
-20000
0
200
400
600
800
系列1

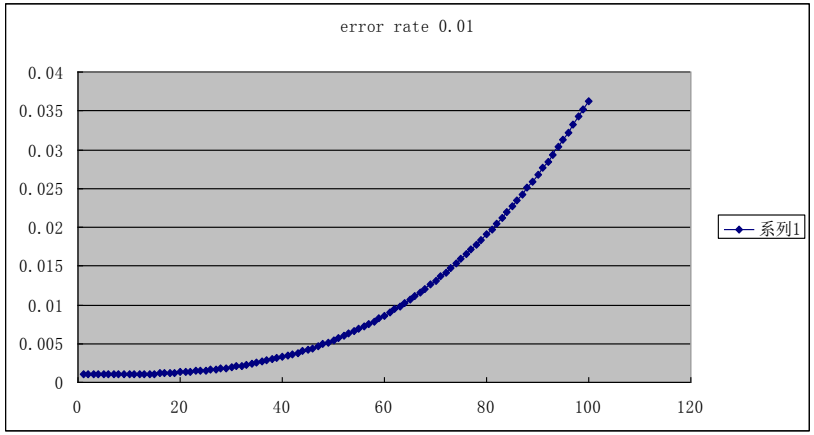

insertsize_mean 2000 insertsize_sd 100
25000
20000
15000
10000
5000
0
-5000
0
500
1000
1500
2000
2500
3000
系列1

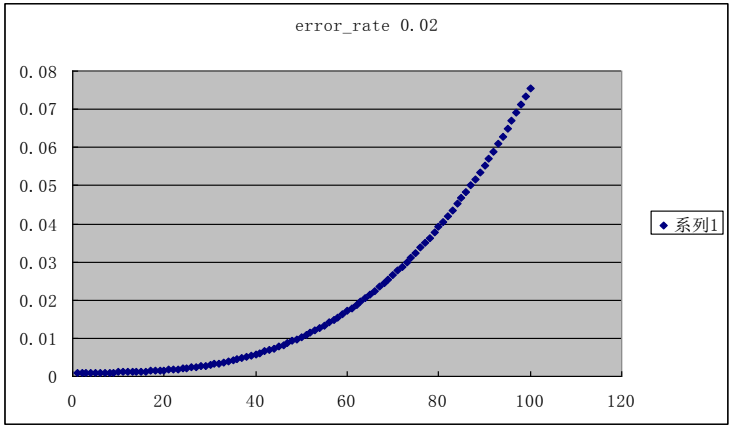
错误率

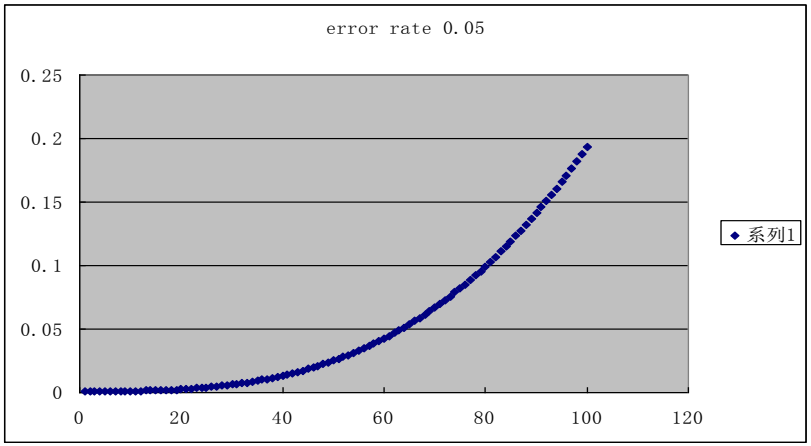
- 模型：

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

示例


error rate 0.01
0.04
0.035
0.03
0.025
0.02
0.015
0.01
0.005
0
0
20
40
60
80
100
120
系列1


error_rate 0.02
0.08
0.07
0.06
0.05
0.04
0.03
0.02
0.01
0
0
20
40
60
80
100
120
系列1


error rate 0.05
0.25
0.2
0.15
0.1
0.05
0
0
20
40
60
80
100
120
系列1

二倍体杂合

二倍体杂合是在染色体上随机的位点上产生snp、insertion、deletion

产生的indel分别为1base、2base、3base,比例为3:2:1。

用法

perl simulate_solexa_reads.pl

-i <string> input reference genome sequence

-l <int> set read length, default:100

-x <int> set the sequencing coverage,default:40

-m <int> set the average value of insert size,default:500

-v <int> set the standard deviation of insert sizes, default:25

-e <float> set the average error rate over all cycles,default:0.01

-s <float> set the heterozygous SNP rate ,default:0

-d <float> set the heterozygous indel rate ,default:0

-o <string> output file prefix default:solexa

./simulate_solexa_reads -i ref_sequence.fa -o humen –l 100 –x 20 –e 0.02 -s 0.01 -d 0.01

时间与内存

酵母	时间(min)	内存
I	44s	<3000bit
I，E	52s	<3000bit
I，E，H	1.11min	<3000bit

拟南芥（40X）	时间	内存
I	6.33min	30M
I，E	9.18min	30M
I，E，H	9.53min	30M

时间与内存

- 注：I表示只用了插入片段分布模块；I，E，表示用了插入片段模块与错误率模块；I，E，H表示插入片段模块，错误率模块，杂合模块
- 注：用人的全基因组模拟reads，5X，所用时间是30min，内存是230M。

谢谢