SimuSCoP User Guide

Zhenhua Yu

qasim0208@163.com

1. Introduction

SimuSCoP is a tool designed for emulating next-generation sequencing (NGS) data by integrating the position and context dependent profiles into the simulation process. It consists of two modules: 1) “seqToProfile” utility is developed for inference of base substitution probabilities, base quality distributions and GC-content bias from aligned NGS reads, users can use this utility to build their own models from sequencing data; 2) “simuReads” utility is used for simulation of complex NGS data by incorporating the learned sequencing profiles.

2. Requirements

- ✓ Linux systems.
- ✓ CMake2.8+.
- ✓ C++11
- ✓ g++.

3. Installation

To build binary, do as follows:

```
tar -zxvf SimuSCoP.tar.gz
cd SimuSCoP
cmake .
make
```

After the installation, the main programs of SimuSCoP are generated in “bin” directory.

4. Usage

4.1. Use of “seqToProfile” utility

Users can use “seqToProfile” program to build a model from sequencing data generated from specific instrument. To successfully run “seqToProfile” you will need to obtain or create these items:

- ✓ A **BAM** file of normal sample [required].
- ✓ A **VCF** file generated from the normal BAM [required].
- ✓ A **FASTA** file of the genome sequence to which the reads were aligned [required].
- ✓ A **BED** file defining the target regions if whole-exome sequencing was used [optional].
- ✓ **SAMtools** software [required].

To extract germline SNP positions from the normal BAM, users need to install GATK software in their machines. The VCF file can be obtained by using following command:

```
java –jar gatk.jar HaplotypeCaller -I <normal>.bam -O <normal>.vcf -R < reference>.fasta
```

A BED file is needed if the normal BAM was produced using whole-exome sequencing strategy.

You should now have all the files to run “seqToProfile” program.

Usage:

./bin/ seqToProfile [options]

Option	**Description**	**Default value**
-h, --help	give help information	--
-b, --bam	normal BAM file	null
-t, --target	target file (.bed) for whole-exome sequencing	null
-v, --vcf	the VCF file produced from the normal BAM	null
-r, --ref	genome reference file (.fasta) to which the reads were aligned	null
-o, --output	output file	standard output
-s, --samtools	the path of samtools	samtools
-k, --kmer	the length of kmer sequence	3
-B, --bins	the number of bins into which bases of read are grouped	50

Example:

./bin/ seqToProfile -b <normal>.bam -t <target>.bed -v <normal>.vcf -r < reference>.fasta –k 4 –B 100 > <normal>.profile
./bin/ seqToProfile -b <normal>.bam -v <normal>.vcf -r < reference>.fasta -o <normal>.profile –s /path/to/samtools

4.2. Use of “simuReads” utility

A config file is needed to successfully run “simuReads” program. Users should create the config file according to the following instructions.

Usage:

./bin/simuReads <config file>

The config file define the parameters used to generate sequencing data, and detailed description of each parameter is as follows:

Parameter	**Description**	**Possible values**
ref	[required] FASTA reference file from which reads will be sampled	Ex: /path/to/hg19.fa
profile	[required] a profile file produced by " seqToProfile " utility	Ex: /path/to/sample.profile
variation	[optional] a file defining the variations (indel, SNV and CNV) to simulate	Ex: /path/to/variation.txt Default: null
snp	[optional] a file defining the SNPs to simulate	Ex: /path/to/ hg19_snp138.txt Default: null

target	[optional] a file defining target regions for sequencing. If the argument is not specified, sequencing data of all chromosomes defined in the reference will be produced	Ex: /path/to/ targetRegions.txt Default: null
name	[required] population names (comma-separated)	Ex: tumor, normal
abundance	[optional] abundance file (the argument is only effective if there are at least two populations defined in "name")	Ex: /path/to/abundance.txt Default: null
output	[required] output directory to save results	Ex: /path/to/results
layout	[optional] sequence layout	“layout=SE” for single-end, “layout=PE” for paired-end Default: SE
threads	[optional] the number of threads to use	Ex: threads=8 Default: 1
verbose	[optional] mode of information display	“verbose=1” will print the intermediate information Default: 1
coverage	[required] sequencing coverage	Ex: coverage=30
insertSize	[optional] insert size for paired end sequencing (only effective when the "layout" is set to "PE")	Ex: insertSize=300 Default: 350

Example:

./bin/simuReads configFiles/config_test_wes.txt

./bin/simuReads configFiles/config_test_wgs.txt

./bin/simuReads configFiles/config_test_tumor.txt

4.2.1 Definition of variation file

All the simulated variations should be defined in one file, and the columns of the variation file must be separated by tabs.

Format of CNVs:

VarType PopuName Chrom StartPos EndPos CN majorCN
c tumor chr20 1 500000 1 1
c tumor chr20 10000000 15000000 4 3

The meaning of each column: *VarType*, the type of variation; *PopuName*, the population having the CNV; *Chrom*, chromosome name; *StartPos*, 1-based start position of the simulated CNV; *EndPos*, 1-based end position of the simulated CNV; *CN*, total copy number; *majorCN*, major allele copy number.

Format of SNVs:

VarType PopuName Chrom Pos Ref Alt Type
s tumor chr20 2000000 c T homo
s tumor chr20 4000000 a g het

The meaning of each column: *VarType*, the type of variation; *PopuName*, the population having the SNV; *Chrom*, chromosome name; *Pos*, 1-based position of the simulated SNV; *Ref*, reference allele; *Alt*, mutated allele; *Type*, type of the simulated SNV ("homo" denotes homozygous mutation and "het" denotes heterozygous mutation)

Format of insertions:

VarType PopuName Chrom Pos SeqContent Type
i tumor chr20 1300000 cgtccgtc het
i tumor chr20 2500000 tcgag homo

The meaning of each column: *VarType*, the type of variation; *PopuName*, the population having the insert; *Chrom*, chromosome name; *Pos*, 1-based position of the simulated insert; *SeqContent*, nucleotide sequence to insert; *Type*, type of the simulated insertion.

Format of deletions:

VarType PopuName Chrom Pos Length Type
d tumor chr20 5000000 10 homo
d tumor chr20 13500000 4 het

The meaning of each column: *VarType*, the type of variation; *PopuName*, the population having the deletion; *Chrom*, chromosome name; *Pos*, 1-based start position of the simulated deletion; *Length*, length of the simulated deletion; *Type*, type of the simulated deletion.

Here is an example of the variation file defining 2 CNVs, 2 SNVs, 2 inserts and 2 deletions (no header is required):

```
c tumor chr20 1 500000 1 1
c tumor chr20 10000000 15000000 4 3
s tumor chr20 2000000 c T homo
s tumor chr20 4000000 a g het
i tumor chr20 1300000 cgtccgtc het
i tumor chr20 2500000 tcgag homo
d tumor chr20 5000000 10 homo
d tumor chr20 13500000 4 het
```

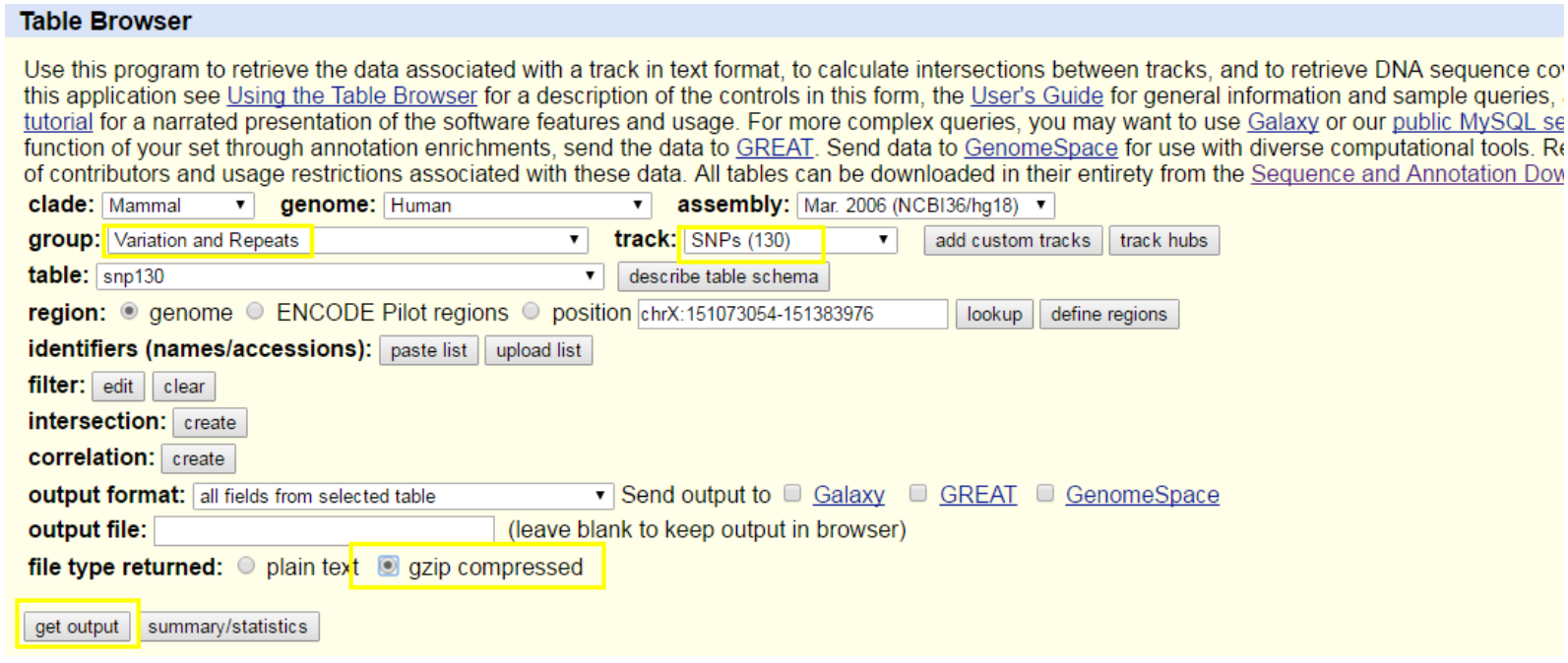
4.2.2 Definition of SNP file

The format of SNP file is as follows:

```
rs58108140 chr1 10583 A/G + G
rs189107123 chr1 10611 C/G + C
rs71252251 chr1 14976 C/T- G
```

The meaning of each column: 1) the name of the simulated SNP; 2) chromosome name; 3) 1-based position of the simulated SNP; 4) observed alleles; 5) the strand of the SNP; and 6) reference allele.

The SNP data can be downloaded at https://genome.ucsc.edu/cgi-bin/hgTables or manually defined by users.



The downloaded data should be further processed to only include the required columns as described above.

4.2.3 Definition of target file

Target file is used to define genomic regions for sequencing. If the file is not provided, sequencing data of all chromosomes defined in reference file will be produced. The format of the target file is as follows:

```
chr1    0    100000
chr1    200000    50000000
chr2    5000000    100000000
chr3    0    0
```

The meaning of each column: 1) chromosome name; 2) 0-based start position of the target region; 3) 1-based end position of the target region. Users can set start and end positions to 0 to select whole chromosome.

4.2.4 Definition of abundance file

Abundance file is used to specify the mixed proportion of each population defined in the “name” parameter. If there is only one population defined in “name” parameter, this argument will be ineffective. Suppose the “name” parameter is set to “name=tumor, normal”, then the abundance file will look like as follows:

```
0.7 0.3
0.85    0.15
0.3 0.7
```

Note that the sum of the proportions of each line is 1. The abundance file above defines 3 tumor samples, and corresponding tumor purities are 0.7, 0.85 and 0.3 respectively. The names of the generated FASTQ files will be tumor_0.700+normal_0.300, tumor_0.850+normal_0.150 and tumor_0.300+normal_0.700.

5. Contact

If you have any questions, please contact qasim0208@163.com.