SCSsim User Guide

Zhenhua Yu

zhyu@nxu.edu.cn

1. Introduction

SCSsim is a tool designed for emulating single-cell genome sequencing (SCS) data. It consists of three modules: 1) "simuVars" module aims to generate single cell genome from a given reference sequence by inserting user-defined genomic variations into specific genomic loci; 2) "learnProfile" component is developed to infer sequencing platform dependent profiles from real data; 3) "genReads" utility is provided to mimic single-cell genome amplification and read generation procedures based on the results of "simuVars" and "learnProfile".

2. Requirements

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

3. Installation

To build binary, do as follows:

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

After the installation, the main programs of SCSsim are generated in "bin" directory.

4. Usage

4.1. Use of "simuVars" utility

Users can use "simuVars" program to simulate the genome sequence of single cell by defining various genomic variations. The types of variations include single nucleotide polymorphism (SNP), single nucleotide variation (SNV), short insert and deletion (indel), and copy number variation (CNV). The inputs of "simuVars" are listed as follows:

Parameter	Description	Possible values
-r, --ref	[required] FASTA reference file from which reads will be sampled	Ex: /path/to/hg19.fa
-v, --var	[optional] a file defining the variations (indel, SNV and CNV) of single cell to be simulated	Ex: /path/to/variation.txt Default: null
-s, --snp	[optional] a file defining the SNPs of single cell to be simulated	Ex: /path/to/hg19_snp138_chr1.txt Default: null
-o, --output	[required] output file (.fasta) to save generated sequences	Ex: /path/to/simu.fa

Example:

```
./bin/simuVars -r <reference>.fasta –s /path/to/snp.txt -v /path/to/variation.txt –o /path/to/output.fasta
```

The SNP file contains all the SNPs to be simulated, and the data is formatted as follows:

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

Each row gives the detailed information of one SNP, and the columns represent the name of the SNP, chromosome name, 1-based position of the SNP, observed alleles, the strand of the SNP and reference allele, respectively. The SNP data can be defined manually by users or downloaded from https://genome.ucsc.edu/cgi-bin/hgTables. The downloaded data should be further processed to only include the required columns.

All genomic variations to be simulated should be given in a single file. Here is an example of the variation file defining 2 CNVs, 2 SNVs, 2 inserts and 2 deletions (no header is required):

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

Each row defines a variation, and different types of variations are distinguished by the first column: *c* denotes CNV, *s* represents SNV, *i* refers insert, and *d* denotes deletion. A CNV contains these items: chromosome name, 1-based start position, 1-based end position, total copy number, and major allele copy number. A SNV is depicted by chromosome name, 1-based position, reference allele, mutated allele, and type of the SNV ("homo" refers homozygous mutation and "het" denotes heterozygous mutation). A short insert is described by chromosome name, 1-based position, nucleotide sequence to insert, and type of the insert ("homo" or "het"). Finally, a short deletion is specified using items including chromosome name, 1-based start position, length and type of the deletion ("homo" or "het").

4.2. Use of "learnProfile" utility

The "learnProfile" utility is designed to infer sequencing profiles from real sequencing data generated from Illumina instruments. In the current version, four profiles including indel error distributions, base substitution probabilities, Phred quality distributions and GC-content bias are measured. Users can build their own profiles from a given real dataset using this utility. The inputs of the program are described as follows:

Parameter	Description	Possible values
-b, --bam	[required] a non-tumor BAM file	Ex: /path/to/sample.bam
-t, --target	[optional] a BED file defining the target regions if whole-exome sequencing was used	Ex: /path/to/targets.bed Default: null
-v, --vcf	[required] a VCF file generated from the non-tumor BAM	Ex: /path/to/sample.vcf
-r, --ref	[required] FASTA reference file to which the reads were aligned	Ex: /path/to/hg19.fa
-w, --wsize	[optional] the length of windows used to infer GC-content bias	Ex: wsize=10000 Default: 1000
-k, --kmer	[required] the length of kmer sequence to infer base substitution probabilities	Ex: kmer=4 Default: 3
-s, --samtools	[optional] the path of samtools	Ex: /path/to/samtools Default: samtools
-o, --output	[optional] output file to save results	Ex: /path/to/profile.txt Default: null

Examples:

./bin/learnProfile -b <sample>.bam -t <targets>.bed -v <sample>.vcf -r <reference>.fasta > <sample>.profile
./bin/learnProfile -b <sample>.bam -v <sample>.vcf -r <reference>.fasta -o <sample>.profile –s /path/to/samtools

4.3. Use of "genReads" utility

The “genReads” program is developed to simulate single-end or paired-end reads based on the results of “simuVars” and “learnProfile”. Single-cell genome amplification and read generation procedures are implemented in this utility. The inputs of the program are listed as follows:

Parameter		Description	Possible values
	-i, --input	[required] FATTA Sequence file generated by “simuVars” program	Ex: /path/to/simu.fa
MALBAC options	-p, --primers	[optional] the number of each type of primer	Ex: primers=10000 Default: 100000
	-r, --gamma	[optional] proportion of the primers used in each amplification cycle	Ex: gamma=0.000001 Default: measured from data
Read generation options	-m, --model	[required] a file defining sequencing profiles inferred using “learnProfile” program	Ex: /path/to/profile.txt
	-l, --layout	[optional] read layout (SE for single end, PE for paired-end)	Ex: layout=PE Default: SE
	-c, --coverage	[optional] sequencing coverage	Ex: coverage=30 Default: 5
	-s, --isize	[optional] mean insert size for paired-end sequencing	Ex: isize=300 Default: 260
	-t, --threads	[optional] the number of threads to use	Ex: threads=8 Default: 1
	-o, --output	[required] the prefix of output file	Ex: /path/to/reads

Example:

./bin/simuReads -i /path/to/simu.fa -m /path/to/hiseq2500.profile -t 5 -o /path/to/reads

5. Contact

If you have any questions, please contact zhyu@nxu.edu.cn.