

SeqSIMLA user manual

Introduction

SeqSIMLA can simulate sequence data in families with multiple affected and unaffected siblings or unrelated case-control under different disease models.

Requirements

GENOME

SeqSIMLA relies on GENOME to simulate sequence data. The software needs to be downloaded and installed on the local machine before running SeqSIMLA. GENOME can be obtained from <http://www.sph.umich.edu/csg/liang/genome/>. Please read the GENOME user manual carefully as some parameters in GENOME are required for SeqSIMLA.

Java

SeqSIMLA is implemented in Java. Hence a Java platform is required to run SeqSIMLA.

Java JDK 7 or JRE 7 is required to run SeqSIMLA. You will get error messages if running SeqSIMLA with JDK or JRE 6.

Installation

On Unix, unzip the file using: `unzip SeqSIMLA.zip`. On Windows, unzip the file using programs such as Winzip or 7-ZIP. There will be a jar file SeqSIMLA.jar and a directory b36_probability.

Use `"java -jar SeqSIMLA.jar"` to run the program.

Parameters

-p: population number: Specify the number of subpopulations and the effective sizes simulated in the data. More details about the simulation model can be found in the GENOME manual. Ex: 2,1000,2000 will simulate 2 subpopulations; the first subpopulation has 1000 samples and the second subpopulation has 2000 samples. The default value is -p=1,10000.

-s: fixed number of SNPs on a chromosome: Specify the number of SNPs on a chromosome. Ex: -s=20 will simulate a chromosome with 20 markers. If -s=-1, the number of SNPs will not be fixed and will follow a Poisson distribution. The default value is -s=-1.

-rec: recombination rate: the recombination rate can be a fixed value between 0 and 0.5 or a set of chromosome numbers.

1. A fixed value: Ex: -rec=0.01. In this case, only ONE chromosome will be simulated. And

the recombination rates between pieces on the chromosome are 0.01.

2. A set of chromosome numbers: Ex: `-rec=1,6`. In this case, two chromosomes (chroms 1 and 6) will be simulated. The recombination rates will be based on the estimated rates from the HapMap project. More details of how we parsed the recombination rates can be found in FAQ. Also if this option is specified, the option `"-s"` will be set to be `-1`. So SeqSIMLA will simulate whole chromosomes based on the recombination rates. Simulations may be slow when multiple chromosomes are specified.

`-site`: selection of causal loci: There are three ways to select causal loci.

1. If only one chromosome is simulated, you can specify the locus numbers. Ex: `1,2,4,6`. This tells SeqSIMLA to use loci 1,2,4,6 as the disease loci.

2. If multiple chromosomes are simulated, you can specify the locus numbers by pieces and chromosomes. Ex: `1:1,2:5/3:7`. This tells SeqSIMLA to use locus 1 in the first piece on chromosome 1, locus 5 in the second piece on chromosome 1 and locus 7 in the third piece on chromosome 2 as the disease loci. Here `a:b` means the `b`-th locus in the `a`-th piece. `"/"` is used to separate chromosomes.

3. You can specify the disease loci by allele frequency thresholds. Ex: `1:0.2,2:0.1`. This tells SeqSIMLA to select variants with allele frequencies LESS than 0.2 and 0.1 in pieces 1 and 2, respectively. This is useful to select a large number of rare variants as the disease loci.

`-pieces`: The number of blocks that will be simulated. The default recombination rate between blocks is fixed at 0.0001. The user can specify the rates using the option `-rec`.

`--mode-grr`: if this flag is given in the command line, the GRR penetrance function will be used as the disease model. Without this flag, the logistic function will be used as the penetrance model. Note that if more than 10 variants are selected as the disease loci, SeqSIMLA automatically uses the GRR model, even if the flag is not given. If this flag is specified, `-par` and `-grr-f0` must also be specified. A more detailed description of the logistic function and GRR models can be found in the SeqSIMLA manuscript or the FAQ.

`-grr-p`: the proportion of variants with protective alleles. The value ranges between 0 and 100. For example, `-grr-p=30` mean 30% of the disease variants are protective.

`-par`: the population attributable risk.

`-grr-f0`: the baseline probability for the GRR penetrance model. If `--mode-grr` is specified,

`-grr-p` and `-grr-f0` must be specified. Please refer to the SeqSIMLA manuscript or the FAQ for

more details about the GRR penetrance model.

-b: the conditional odds ratios for the disease. If the logistic function is used as the penetrance model, this option must be specified. There are three ways to specify the odds ratios.

1. You can use the same format as -site. For example, 1.2,2,2.5 for three disease loci.

2. You can use a fixed odds ratio for all disease loci. For example, -b=1.2 means all disease loci have odds ratios of 1.2.

-pre: the prevalence of disease for --mode-beta.

-var: variance of the quantitative trait. Specify the total variance of the trait.

-vp: the proportions of variance explained by QTL. Specify the proportions of variance explained by QTL. Ex: -vp=0.1,0.1,0.1. This assumes three loci are selected in -site, and each site explains 10% of variance in the trait values.

--d: the dominant model. The default model is additive if --d or --r is not specified.

--r: the recessive model. The default model is additive if --d or --r is not specified.

-dt-f: number of families: Specify the number of extended families to be simulated.

SeqSIMLA simulates a three-generation extended pedigree for each family. The family structure is

-numaff: the minimum number of affected siblings in the third generation. By default, the ascertainment criteria is that at least one individual in individuals 7,8,9,10,11,12 are affected. If numaff is specified as n (the maximum of n is 3), at least n number of siblings will be affected in each family. Ex: -numaff=2 will generate families that at least 2 individuals in 7,8,9 are affected OR at least 2 individuals in 10,11,12 are affected.

-dt-cc: number of unrelated cases and controls: Specify the numbers of cases and controls to be simulated. SeqSIMLA can also simulate unrelated case-control samples. Ex:

-dt-cc=1000,500 will generate 1000 cases and 500 controls.

-batch: number of replicates of the simulated samples: Ex: -batch=500 will simulate 500 replicates of samples.

-seed: the seed of random number generator. If you would like to simulate multiple replicates

drawn from the same population, it is important to give a fixed value of seed. Ex: -seed=1234. Otherwise, each replicate of samples will come from a random population, determined by a random seed.

-gpath: the path to the directory where GENOME is installed.

-b36path: the path to the directory where the recombination files are located.

Optional files

Two files are required if the user would like to specify the recombination rates between blocks instead of using an uniform rate.

One file must be named as chrA, where A is a number that must match the number in -rec. See the example files in b36_probability.

For example, if -rec=5, the program will look for the file chr5.

The format of the file is the same as the recombination file in GENOME. For example:

Rec	Pos
0.044116479	1
0.01565426	2
0.035666338	3
0.037254565	4
0.075473212	5
0.058051934	6
0.066107947	7

The second file must be named as chr_length. For example:

chr1	92983	246124932	50
chr2	89272	242373921	1688
chr3	87348	198804087	1557
chr4	86240	191108845	1596
chr5	86835	180444000	1237

The first column specifies the chromosome id. The second and third columns specify the start and end positions. The last column specifies the number of blocks on the chromosome.

The number of blocks should match the number of rows in the chrA file.

Output files

_.parameter: the parameters specified.

*.map: A PLINK format map file, which contains the chromosome number, the marker name,

the genetic position, and the physical position.

*.ped: A PLINK format ped file, which contains six mandatory columns (Family ID, Individual ID, Father ID, Mother ID, Sex, Affection status), followed by allele codes (1 or 2).

Tutorial

```
java -jar SeqSIMLA.jar -gpath=c:\genome-0.2-Windows\ -p=1,1000 -s=300  
-site=67,95,26,52,24,198,1,42,10,46,116,56,74,57,2,199,68,44,7,53,5,106,107,126,114,102,1  
7,125,80,63 -dt-f=500 --mode-grr -grr-p=60 -par=0.0033 -grr-f0=0.1 -batch=1000  
-seed=123456
```

This command uses GENOME to generate one population with 1000 sequences (haplotypes). Assume GENOME is installed in c:\genome-0.2-Windows\. 300 variants (-s) will be in the output. 30 disease loci (-site) are selected out of the 300 variants. 500 families (-dt-f) will be generated with at least one affected sibling in the third generation. The GRR disease model (--mode-grr) will be used. 60% of the disease loci are risk variants (-grr-p) and the rest are protective variants. The overall population attributable risk (-par) is $0.0033 \times 30 = 0.1$. The baseline penetrance (grr-f0) is 0.1. 1000 replicates (-batch) will be generated. The random seed (-seed) is set to be 123456.

```
java -jar SeqSIMLA.jar -p=1,1000 -s=-1 -site=10,20,30,40,50 --mode-beta -b=2  
-dt-cc=1000,1000 -batch=100 -seed=123456 -pre=0.1  
-b36path=/home/rchung/NGS_algorithm/NGS7 -rec=1
```

This command uses GENOME to generate one population with 1000 sequences (haplotypes). The number of variants (-s) is not fixed and will follow a poisson distribution. 5 disease loci are selected. The logistic disease model (--mode-beta) will be used. Assume the odds ratios are all equal to 2 (-b). 1000 cases and 1000 controls will be generated (-dt-cc). The disease prevalence is set as 0.1 (-pre). The recombination rates saved in chr1 will be used (-rec=1). Note that if rec=A is specified, the recombination rate information must be saved in a file named as chrA. The chr1 file is saved in /home/rchung/NGS_algorithm/NGS7

Contact us

Suggestions and comments on the software are welcome. Also feel free to send us bug report. Please email: rchung@nhri.org.tw