

GenomePop2: Simulating SNPs

Antonio Carvajal-Rodríguez^{1*}

¹Departamento de Bioquímica, Genética e Inmunología, Facultad de Biología, Universidad de Vigo, 36310 Vigo, Spain

Abstract. The program GenomePop2 is a major update from a previous one, GenomePop, to handle SNPs under more flexible and useful settings. The program manages one chromosome with constant or variable (hotspots) recombination between sites. The initial metapopulation can be just one clone or an initial neutral equilibrium which can be computed theoretically or by simulation. Initial frequencies of SNPs at different positions can be predefined. Specific sites can undergo positive or negative selection and/or population expansion or bottlenecks in different populations during user-defined time periods. Any migration model is allowed. The output is given in GenePop 4.0 or Hudson ms program formats. GenomePop2 is available in the following web-link: <http://webs.uvigo.es/acraaj/GenomePop2.htm>.

Keywords: forward simulation, single nucleotide polymorphism.

1 Introduction

A previously published program GenomePop has been developed for simulating SNPs or DNA sequences under complex models of evolution and demography [1]. However, evolving sequences forward in time under complex evolution models as GTR is computationally very costly. Additionally, maintaining and evolving software that manages several different genetic models as DNA, codons and SNPs is hard. GenomePop2 discards previous DNA and codon models, focusing in two-allele models as SNPs with ancestral and derived alleles. It conserves previous powerful characteristics as migration models, scaling, population contraction-expansion scenarios and hot spot recombination. It also incorporates more output formats and selection and demographic episodes during user defined time periods.

Computer simulation of SNPs provides an efficient complement to experimental approaches in order to understand patterns of current DNA sequences in populations. For example, simulations can be used to check statistics about neutrality [2], the efficiency of selection detection methods [3,4], to study population linkage disequilibrium distributions [5], to test the accuracy of genomic artificial selection

* Corresponding author: acraaj@uvigo.es

methods [6] and so on. Henceforth, I will explain some basics about parameter input in GenomePop2 besides the new features added to the previous version [1].

2 Parameter input

The input file must be called GP2Input.txt and should be in the same directory as the executable file. GenomePop2 has a wide list of possible parameter inputs which are initialized with specific default values when possible. The complete list of default values can be viewed in the web page:

http://webs.uvigo.es/acraaj/GP2_Examples/GP2_DEFAULT_PARAMETERS.txt

The minimum input file for the program to work should include the following two lines:

chromsize	numchroms	N	Npops	maxgen	gmut	gRec
1000	1	1000	4	200	0.01	0.0

To identify the line the first word must be 'chromsize'. Below this line add the 7 values corresponding to: The chromosome size, number of chromosomes, initial population size, number of populations, number of generations, mutation rate per haploid genome and recombination rate per haploid genome. Thus, in the example, one run is defined with a one chromosome genome of 1000 positions, four populations with 1000 diploid individuals each. The simulation will run during 200 generations with a genome mutation rate of 0.01 and no recombination. The migration rate is 0 by default (see below).

More information about parameter input can be found in the following link:

<http://webs.uvigo.es/acraaj/GP2howto.htm>.

3 Demographic settings

Neutral mutation-drift equilibrium

An input file line beginning with the word *neutrleq* and the value 'true' below, indicates the program to compute mutation drift equilibrium [7], distributing the effective number of alleles, $\theta+1$, with $\theta = 4N\mu$ and μ the per site mutation rate, at frequency $q = 1/(\theta + 1)$ in the equilibrium population so that we will have q homozygotes and $\theta / (\theta+1)$ heterozygotes.

neutrleq

true

Simulated neutral equilibrium

Instead of computing the effective number of alleles we can simulate a scaled population under neutral conditions during $10N$ generations in order to reach the equilibrium. After the simulation, the program stores homozygotes and heterozygotes frequencies and will use such frequencies to begin the simulation process after equilibrium.

simneutraleq
true

The scaling can be used in order to improve the efficiency of the simulation. Therefore,

simscale
10

will evolve an equilibrium population of size $N/10$ during $t/10$ generations with mutation and recombination rates of 10μ and $10r$ respectively.

Migration

To set the migration rate the identifier “migration” with the desired value below must be used. Island or one-dimensional stepping stone models can be defined with the corresponding line in the input file. To set more complex migration models, the user should define an additional input file called MigrationModel.txt. Detailed explanation to define any migration model using this file can be found in the web link:

<http://webs.uvigo.es/acraaj/MigrationModels.htm>

Contraction-expansion scenarios

CEDS
1 1 20 2
1 250 350 2000

The settings above will define, under the CEDS identifier, a bottleneck of size $N = 2$ in population 1 from generation 1 to 20 and an expansion of $N = 2000$ in the same population from generation 250 to 350. After that, the original population size (N) will be recovered. The user can define as many lines as desired under the CEDS identifier. At each line, first item identifies the population, the next two define the generation period and the last one the desired population size.

4 Selection settings

In GenomePop2 the fitness scheme is multiplicative so that each site i contribute to the fitness with $1-s[i]h[i]-E$. By default, selection is not defined that is $s[i] = 0$ for every site, the same as epistasis E . The dominance coefficient has a default value of 0.5 under diploid models and is fixed to 1 under haploid ones. Selection can be modeled in two ways.

1) Gamma distribution

In the first way the user can set the parameters of the gamma distribution from where the selective coefficients will be sampled. The shape parameter is defined as β and the scale parameter, as s/β . The β parameter allows modeling the fitness effects distribution, e.g. a low value of β , e.g. 0.1, will sample many mutations with low effect and few with high. A β parameter of 1 corresponds to the exponential distribution. The lines

coefsel	coefdom	beta	Epistasis
0.001	0.5	0.5	0

define a gamma distribution with scale $\text{coefsel} / \text{beta}$ and expected mean of coefsel . The selective coefficient for each site i , $s[i]$, will be sampled from that distribution. The dominance coefficient h for each site i is computed as,

$$h[i] = U(0,1) * e^{(-k*s[i])}$$

where $k = \text{alpha} * [(2 * \text{coefdom})^{(-1.0/\text{beta})} - 1.0]$ and alpha is the rate parameter (1/scale). If beta is 0 then constant selection and dominance coefficients with value coefsel and coefdom are used.

2) SNPs under selection

We can alternatively define specific positions with derived alleles under positive or negative selection. For example:

selnuc	pop	position	s
	1	499999	-0.15
	2	499999	0.15

will define any derived allele in position 499999 as beneficial in population 1 but deleterious in population 2. Depending on the mutation rate it could occur that such position never mutate. The user can indeed force the existence of the mutation at such position (SNP) in the initial population, including the following lines:

isf	popid	position	freq
	1	499999	0.05
	2	499999	0.1

being “freq” the frequency of the heterozygotes carrying the derived allele at the given position.

If any equilibrium line is defined (*neutralseq* or *simneutralseq* identifiers) the “isf” information will be ignored, because the initial population will be the computed under equilibrium. If the user still wants to add some SNPs after the equilibrium, an identifier ‘isfposteq’ should be added. For example:

isfposteq	pop	pos	num indivs
	1	499999	1

will add a mutation at position 499999 in one individual after the equilibrium population was computed.

5 Output formats

There are two possible output formats. The default is the Hudson ms format which can produce two different files: First, GP2File_Run0.txt with the diploid (haploid) SNP haplotypes of the sampled individuals for each population. The SNPs positions are also given. A second file, gameteGP2File_Run0.txt, will be generated in the diploid case if the identifier “haplotypes” is set to true (by default). This latter file includes the sampled gametes from the individuals in the previous file. The user gets as many of such files as runs were defined at the input. The second output format is GenePop 4.0 output format [8] and includes all positions, not only SNPs. No file with gametes is produced.

6 Example: Divergent selection

In figure 1 we can appreciate the result of just one run of 100 generations of selection over a SNP occupying the central position in a 1Kb fragment. The derived allele is initially in heterozygous condition in 20% of the population 1 (gene frequency 0.1) and does not exist in population 2. In population 1 the derived allele has a selective coefficient $s = 0.1$, implying a loss of fitness of 1% and 0.5% in homozygous and heterozygous conditions respectively. On the contrary in population 2 the allele is favored by 2% and 1%, respectively ($s = -0.2$). Both populations are connected by

migration with $Nm = 10$. As expected, in the short term, the frequency diminishes in population 1 and increases in population 2 due to the combined effects of migration and selection. However, due to the high migration rate and the different selection intensities the derived allele will become fixed in both populations given an enough number of generations (result not shown). This occurs because the net selective advantage over the entire habitat is positive, implying the stable fixation of the favored allele given an adequate relationship between divergent selection and migration [9].



Fig. 1. Result of 100 generations of divergent selection between two populations connected by migration. Continuous line: population 2; Broken line: population 1.

7 Conclusions

GenomePop2 is software that allows the simulation of SNPs in reasonably complex evolutionary scenarios. The program is still under development. Suggestions, collaborations and improvements by potential and interested users are welcome.

Acknowledgments

I am grateful to M. Saura and A. Pérez-Figueroa for useful comments on the manuscript. I am currently funded by an Isidro Parga Pondal research fellowship from Xunta de Galicia (Spain).

References

1. Carvajal-Rodriguez A: GENOMEPOP: A program to simulate genomes in populations. *BMC Bioinformatics* 9: 223 (2008).
2. Achaz G: Frequency spectrum neutrality tests: one for all and all for one. *Genetics* 183: 249-258 (2009).
3. Hussin J, Nadeau P, Lefebvre JF, Labuda D: Haplotype allelic classes for detecting ongoing positive selection. *BMC Bioinformatics* 11: 65 (2010).
4. Huff CD, Harpending HC, Rogers AR: Detecting positive selection from genome scans of linkage disequilibrium. *BMC Genomics* 11: 8 (2010).
5. Mizuno H, Atwal G, Wang H, Levine AJ, Vazquez AEUE: Fine-scale detection of population-specific linkage disequilibrium using haplotype entropy in the human genome. *BMC Genet* 11: 27 (2010).
6. Calus MP, Meuwissen TH, de Roos AP, Veerkamp RF: Accuracy of genomic selection using different methods to define haplotypes. *Genetics* 178: 553-561 (2008).
7. Crow JF, Kimura M: *An Introduction to Population Genetics Theory*. New York: Harper & Row (1970).
8. Rousset F: genepop'007: a complete re-implementation of the genepop software for Windows and Linux. *Molecular Ecology Resources* 8: 103-106 (2008).
9. Roughgarden J: *Theory of Population Genetics and Evolutionary Ecology: An Introduction*. New York: MacMillan (1979).