

SNPsim (current version is 0.8)

© 2002–2004 David Posada (dposada@uvigo.es) and Carsten Wiuf.

WWW: <http://darwin.uvigo.es/software/snpssim.html>

Disclaimer

This program is free software; you can redistribute it and/or modify it under the terms of the GNU General Public License as published by the Free Software Foundation; either version 2 of the License, or (at your option) any later version. This program is distributed in the hope that it will be useful, but WITHOUT ANY WARRANTY; without even the implied warranty of MERCHANTABILITY or FITNESS FOR A PARTICULAR PURPOSE. See the GNU General Public License for more details. You should have received a copy of the GNU General Public License along with this program; if not, write to the Free Software Foundation, Inc., 59 Temple Place – Suite 330, Boston, MA 02111–1307, USA.

Credits

This program was started at the Department of Zoology of BYU (Provo, UT) but mainly developed at Variagenics (Cambridge, MA).

Citation

Posada D and Wiuf C. 2003. Simulating haplotype blocks in the human genome. *Bioinformatics* 19: 289–290.

History*Version 0.1*

- Simulates genealogies under the coalescent with recombination
- Assumes SNPs are biallelic (0,1)
- Mutation model is JC (Jukes–Cantor) or ISM (Infinite Sites Model)
- Simulates SNPs under theta or generates a fixed number
- Simulate homogeneous recombination or recombination hotspots
- Allows for several demographies
- Exponential growth
- Several periods

Version 0.2 (8 October 2002)

- Read arguments as float although use some of them as integers (this way we can input $-e1e4$, for example)
- Expected and fixed number of hotspots as decimals
- Added heterogeneity of hotspot recombination rate (with gamma distribution)
- Added Uniform hotspot (until now there were always Normal)
- Added hotspot interference

Version 0.3 (18 October 2002)

- Bug in the hotspot interference implementation
- Cleaned some code that I was not using anymore

Version 0.4 (11 November 2002)

- First version released to the web
- Counting only recombination events that happen in ancestral material that did not found its MRCA (X material). Before we counted $X|X$, $X|0$, and $0|X$. Now we only count $X|X$.
- New option for not simulating data (just do coalescent)
- Redefine the interference interval instead of $m-1$ as m (so now instead of $-z3$ we use $-z4$)

Version 0.5 (30 January 2003)

- Small bug fixed reading options $-n$ and $-x$ on command line (thanks to Clara Singer)
- Change N_e , μ and noisy defaults (now $-e10e4$ $-u1e-8$ $-y1$)
- By default print SNPs genotypes and haplotypes
- Print memory available

Version 0.6 (7 April 2003)

- Error in the implementation of the infinite site mutation (ISM) model. We need to update the totalBranchSum after evolving each site to conform to expectations. (thanks to Jack Liu)

- Potential exception (array out of bounds: hotspot[poissonNumber-1]) fixed (thanks to Jack Liu)
- Debugging: When checking the distribution of mutations use only datasets with SNPs (CHECK_MUT_DISTRIBUTION functions)

Version 0.7 (20 October 2003)

- Made sure that all parameters are well parsed : read them as doubles or integers or cast them afterwards.
- Change seed argument symbol from "*" to "#". Now the seed argument should be passed as in -#34742, for example.
- Fixed: It was accepting demographics from command line

Version 0.8 (6 May 2004)

- Added PrintDefaults() function
- Set seed to time(null) at the very beginning, just for having it in the defaults.
- Comment main variables in the defaults section in the code.
- Count reps without SNPs.

Purpose

SNPsim is a population genetic simulator that generates samples of SNP (Single Nucleotide Polymorphisms) haplotypes and diploid biallelic genotypes. It is based on the coalescent with recombination (Hudson 1983) modified by Wiuf and Posada (in press) to result in heterogeneous recombination rates along the chromosomes simulated. SNPsim also allows for the specification of demographic periods and different mutation models.

Haplotype blocks

Recent studies suggest that most single nucleotide polymorphisms (SNPs) in the human genome are organized in regions with high levels of linkage disequilibrium and little haplotype diversity (Daly et al. 2001; Jeffreys, Kauppi, and Neumann 2001; Patil et al. 2001; Gabriel et al. 2002). These regions, called haplotype blocks, are believed to be the result of population history and especially, non-homogeneous recombination. The interest in this structure is in that such blocks can be tagged with a few SNPs, enormously facilitating association studies searching for genes underlying complex diseases (Murphy et al. 2001). Indeed, a haplotype project has been started to describe the haplotype blocks and to select best tagging SNPs along the human genome. We (Wiuf and Posada, in press) have devised a population genetic model of recombination hotspots, based on the coalescent with recombination, that stochastically generates genetic samples with "haplotype block" structure. Such models are expected to be very useful to understand and interpret genomic variation within populations.

The coalescent with hotspot recombination

The model implemented in SNPsim (Posada and Wiuf 2003; Wiuf and Posada 2003) is an extension of the coalescent with recombination based on the neutral Wright-Fisher model of genetic variation (Hudson 1983; Hudson 2002). Given an expected number of recombination hotspots, a background homogenous recombination rate and a hotspot recombination rate, SNPsim starts by choosing the position and number of recombination hotspots for a particular sample. Adding recombination events around the hotspot center results in the specification a probability distribution for the recombination rate along the region of interest. Other recombination events coming from recombinational hotspots centered outside the region of interest are also considered. This fast simulation results in different recombination rates for different sites along the region (hotspots and coldspots). Given these recombination rates and other parameters like the effective population size (N) and growth rate, random genealogies are produced that describe the history of different portions of the region of interest. Complex demographic histories can be implemented by defining demographic periods in which population sizes augment, reduce, or remain constant. Time is scaled in units of 2N generations. Mutations can be placed upon the genealogies under a biallelic infinite-sites mutation model or under a biallelic Jukes-Cantor mutation model (JC) (Jukes and Cantor 1969) that allows for recurrent mutations. In a given sample, different sites can evolve under one of these models by setting up a parameter that represents the expected proportion of sites under each model. Different mutation rates can be specified for each model. In addition, there is the possibility of conditioning the simulations to produce samples with a fixed number of SNPs (Hudson 2002). The result is a sample of haplotypes, which are then randomly combined to form diploid genotypes. Next this process is described in more detail.

Scaling time

SNPsim is scaled in units of $2N$ generations. For a given site, and under constant population size, the time to the most recent common ancestor (*TMRC*A) in generations is:

$$E(TMRC A) = 2 \left(1 - \frac{1}{s} \right)$$

$$Var(TMRC A) = \sum_{i=2}^s \frac{4}{i^2(i-1)^2}$$

where s is the sample size.

Hotspot recombination model

The hotspots recombination model aims to represent the idea that some sites in a chromosome are more likely to recombine than others ("recombination hotspots"). This general model is composed of two basic recombination rates and a set of hotspots sites. The *background recombination rate* (r_B) is the per generation recombination rate at any site in the chromosome of length L , while the *hotspot recombination rate* (r_H) is the additional per generation recombination rate at the recombination hotspot. X is the number of hotspots.

$$\text{Background recombination rate } (R_B) = 4Nr_BL$$

$$\text{Hotspot recombination rate } (R_H) = 4Nr_HX$$

$$\text{Global recombination rate } (R_G) = R_B + R_H$$

In real life we do not expect the recombination hotspot to be always restricted to the same single site, to have always the same intensity, or to occur independently from other hotspots. The model described above can be generalized to include these relevant biological features. When the hotspot is not restricted to a single site, and under constant population size, the expected number of recombination events is

$$E(\text{number of recombination events}) = R_G \sum_{i=1}^{s-1} \frac{1}{i}$$

Hotspot location: interference

We will assume that the distance between hotspots is Gamma distributed, $\Gamma(m, \lambda)$, $m > 0$. If $m=1$, we have a Poisson process with intensity λ . Allowing $m \neq 1$ introduces *interference*. If $m > 1$ hotspots are pushed away from each other; if $0 < m < 1$ they tend to be clustered (although SNPsim will only accept integer values for m). The average number of hotspots in the gene is λ/m (see Figure 1).

Hotspot imprecision

We will consider that the hotspot location actually represents the center of the hotspot, and where recombination is more likely, but that there are also some sites around where recombination occurs with some frequency that decreases with increasing distance from the hotspot center. This can be represented by a Normal or a Uniform distribution for the location of the recombination. When the Normal distribution is used, this has a mean equal to the location of the hotspot center and variance called *hotspot imprecision* (σ^2) (see Figure 2). When the variance is small the hotspot tends to be narrow. If the variance is 0, the hotspot is 1 bp wide. When the Uniform distribution is used, recombination events occur with the same probability along a given width for the hotspot. In addition there is the possibility of recombination events coming from hotspots located outside the region of interest of length L . To implement this idea we can extend the region of interest by a number of sites K at each end. In the Normal distribution an arbitrary, but seemingly reasonable value for K that assures that wide hotspots outside L are taking into account is

$$K = 10\sqrt{\sigma^2}$$

If the uniform distribution is used, K equals half the width of the hotspot.

Hotspot heterogeneity

In addition, not all hotspots have to be equally “hot”. We can model this heterogeneity of recombination rates using a gamma distribution with a mean of 1. The shape of this distribution (α) will determine the strength of this hotspot heterogeneity. The smaller α , the bigger the heterogeneity of recombination rates at the hotspots.

SNPsim implementation

A nice feature of the hotspot model is that it allows for the construction of a distribution of recombination rate along the chromosome (\Re) (Figure 1) for every sample.

SNPsim starts by building \Re . The first step is to set up the number of hotspot centers (X) along the extended region $L + 2K$. The average number of hotspots in the gene is λ/m , and when $m > 1$ we use a thinning algorithm to locate these hotspots (Figure 1). If $m=1$ we distributed the hotspots according to a Poisson distribution with intensity λ .

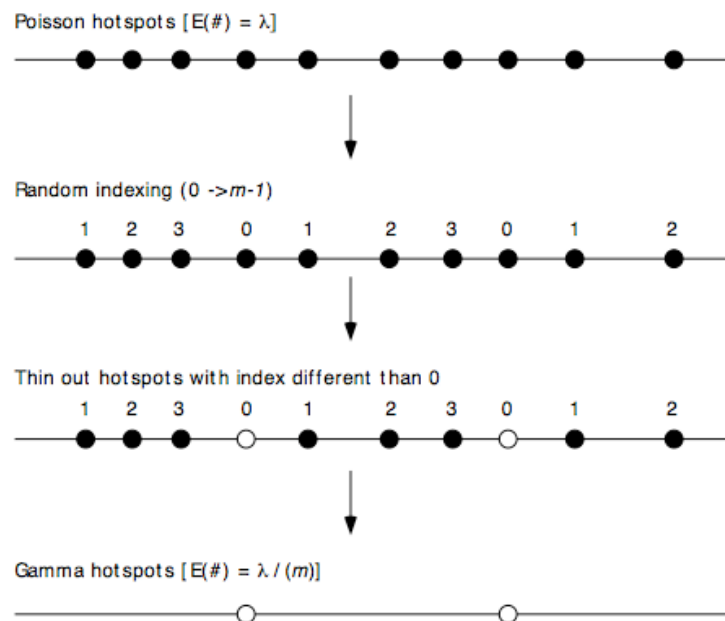


Figure 1. Thinning algorithm for the location of hotspots with interference ($m=4$). The index for the first hotspot site is chosen uniformly between 0 and $m-1$.

Alternatively the user can specify a fixed number of hotspots, that will be located uniformly (see option `-q` below).

The distribution \Re can be constructed according to a Normal (x, σ^2), or a Uniform (*hotspot width*) distribution. If there is hotspot recombination, a random gamma variable will scale the recombination events at each hotspot. The next figure represents a realization of this process:

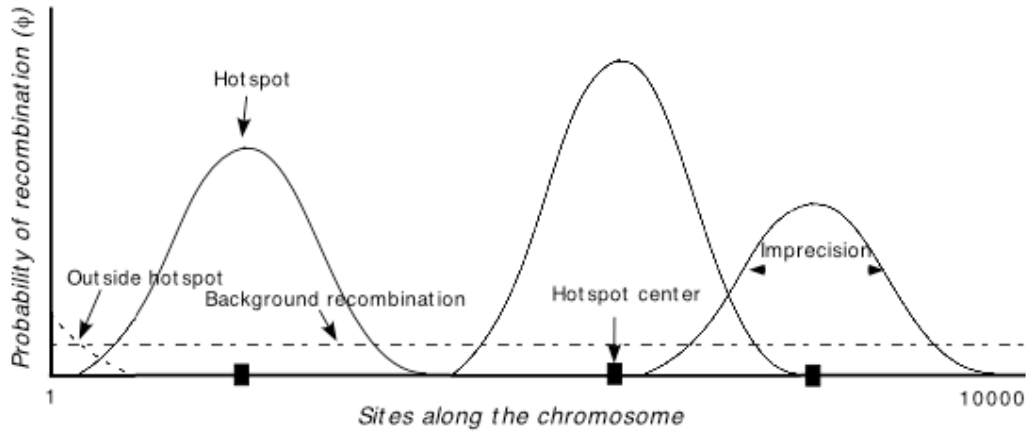


Figure 2. Schematic representation of \Re in the case of three Normal hotspots for the region of interest ($L = 10000$). In this case the hotspot imprecision is quite big. Note that some recombination probability is contributed by a hotspot outside the region of interest. The background recombination is the same for all sites. In this case there is hotspot heterogeneity.

We can use now \Re now to set the global recombination rate per each site i :

$$r_{Gi} = r_{Bi} + r_{Hi}\varphi_i$$

where φ_i is the probability of block recombination at site i .

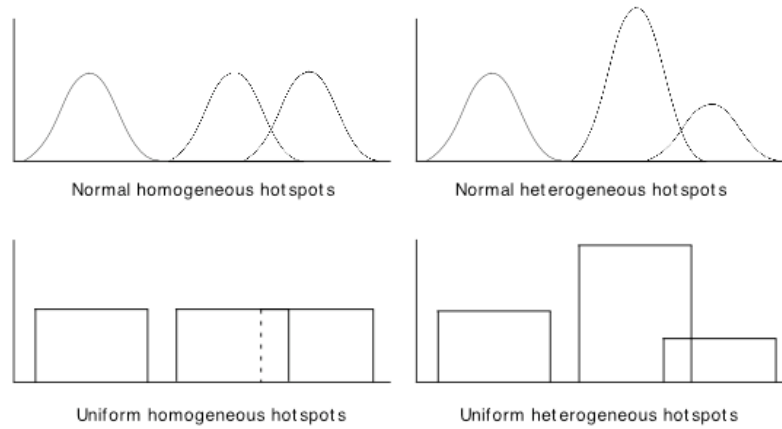


Figure 3. Schematic representation of different \Re .

Now the coalescent proceeds backwards starting from the sample of s gametes. The times to a coalescence (CA) or recombination (RE) event are exponentially distributed, and under constant population size:

$$\text{Time to CA} \propto \text{Exp} \left[\frac{k(k-1)}{2} \right] \cdot 2N$$

$$\text{Time to RE} \propto \text{Exp} (2NG) \cdot 2N$$

where k is the current number of gametes and G is the total recombination rate at all valid recombination sites.

$$G = \sum_{j=1}^k \sum_{i=1}^L r_{Gi}$$

where i is valid recombination site. A *valid* recombination site has to have at both sides ancestral material that has not found its MRCA yet.

The next event will be a coalescent if the time to coalescence is shorter than the time to a recombination, and a recombination if it is larger. Given that a recombination occurs, a gamete is chosen according to the total rate at potential recombining sites in that gamete. Breakpoint sites are chosen according to the recombination probabilities per site (r_G).

Exponential growth

SNPsim allows for the specification of the exponential population growth rate. The only modification concerns to the expected time to a coalescence:

$$Time\ to\ CA \sim \frac{\log \left[e^{\beta t} + \beta \exp \left[\frac{k(k-1)}{2} \right] 2N \right]}{\beta} - t$$

where t is the current time.

If the growth rate is negative, the coalescent time may be infinite (i.e., coalescence does not happen), and SNPsim will stop and issue an error message.

Demographic periods

Under a specified demographic period i , β_i is the growth rate inferred for a demographic period i that goes from size N_{B_i} in the past to size N_{E_i} in l_i generations (Figure 2):

$$\beta_i = \frac{-\log \left(\frac{N_{B_i}}{N_{E_i}} \right)}{l_i}$$

The time to coalescence will be:

$$Time\ to\ CA \sim \frac{\log \left[\exp \left(\frac{k(k-1)}{2} \right) \beta_i 2N_{E_i} e^{-\beta_i(t-t_{i-1})} + 1 \right]}{\beta_i}$$

where t is the current time and t_i is the cumulative time from the present:

$$t_i = \sum_{j=0}^{j=i} l_j$$

We should realize that these parameters are looking back in time, so it is not a good idea to specify a strong negative growth rate for the last period, as the coalescent time could become infinite in the past. In such a case SNPsim will stop and issue an error message.

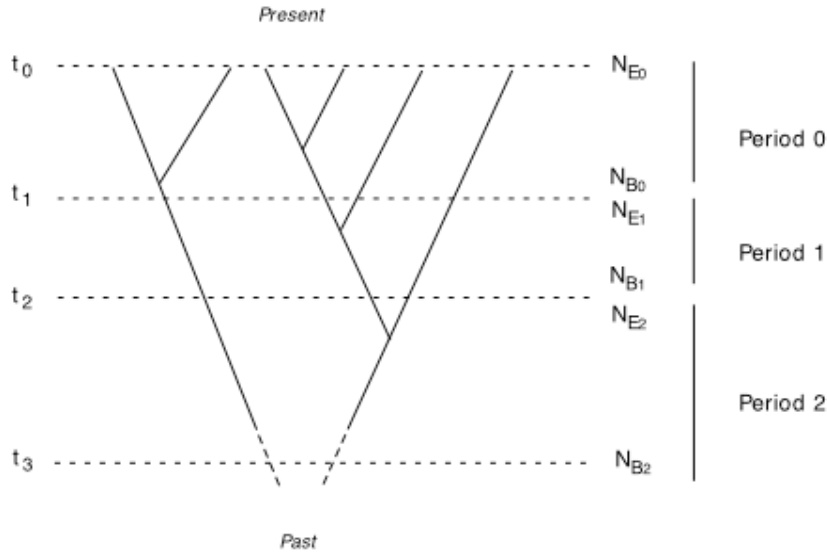


Figure 4. Representation of the demographic periods. The growth rate after the last period will be the same as the one implied by the last period.

Mutation models

SNPsim implements two mutation models that can be combined in a third mixed-model by setting up a proportion parameter (see below).

Infinite-sites mutation model (ISM)

Under this model any site can change only once ($0 \rightarrow 1$), so no recurrent mutations are allowed. The number of mutations (= number of SNPs) is distributed as a Poisson according to the total tree length over all sites, unless a fixed number of SNPs is requested (see option `-x` below). The specific sites where to place these mutations are selected according to the length of the trees at those sites, while the specific branch where this mutation occurs is selected according to branch lengths. Under the ISM (and under no recombination for the variance):

$$E(\text{Number of SNPs}) = \theta \sum_{i=1}^{i=n-1} \frac{1}{i}$$

$$\text{Var}(\text{Number of SNPs}) = \theta \sum_{i=1}^{i=n-1} \frac{1}{i} + \theta^2 \sum_{i=1}^{i=n-1} \frac{1}{i^2}$$

where $\theta = 4N\mu L$. L is the number of sites.

Jukes-Cantor model (JC)

The Jukes-Cantor (1969) model for 4 states is modified here to allow for 2 states only. Under this model recurrent mutations are possible. The probability of change under this model is:

$$\text{Pr}(\text{change}) = \frac{1}{4} - \frac{1}{4} e^{-4\mu}$$

Mixed ISM-JC

Both ISM and JC can be incorporated by setting up a proportion of sites expected to evolve under the JC model (p) (option `-o` below). Each site is assigned to each mutation category, JC or ISM, with probability p and $1-p$, respectively. Also, a different mutation rate can also be specified for the JC model (option `-w` below), allowing for the simulation of some fast homoplasious sites like CpG islands or homopolymeric sites.

SNPsim Usage

SNPsim offers very flexible simulation schemes, and the user can specify many options in the command line (all OS except Macintosh) or in a parameter file (all OS).

Command line

At the command line, the program will first look for arguments. If the user specifies any argument, the program will use the values indicated for the specified values, and the default values (see below) for the parameters not included in the command line. If no argument is specified, the program will look for the parameter file *parameters*. If no arguments are specified in the command line, and there is no *parameters* file, SNPsim will stop and throw an error. It is important that the user checks in the screen info or in the resulting files that all the parameters have the desired values. Example command lines could be:

```
./snpsim0.8 -n100 -s18 -l500
```

or

```
./snpsim0.8 -n100 -s10 -x10 -e10e+4 -d2 1000 5000 200000 100 500 30000 -r1e-6 -k1.0e-5 -h2 -v1.0e+1 -u1.0e-07 -p0.5 -w1.5 -y1 -g1000
```

Parameter file

SNPsim can read its arguments from a file called *parameters*. In Macintosh, this is the only way of interacting with the program. This file should include only arguments, and anything within brackets will be ignored. It is quite helpful to include some text within brackets to remember the meaning of the arguments. An example parameter file is included that can be used as template.

Output files

All output files produced by the program are moved to the folder *Results*. Unless option -1 (do not simulate data) is invoked, the program will always output two data files:

- the file *SNPs* includes the SNP genotypes for those replicates that showed variation
- the file *haplotypes* includes the SNP haplotypes for those replicates that showed variation

Other outfiles can contain full information on genotypes (-f) or sequences (-a), or about the trees (-i), node times (-j) or mutation models (-o). For a detailed description of these options see below.

Arguments

Possible arguments for SNPsim are (# indicate a number):

-n# : Number of replicates

The number of samples to be generated. Each sample is an independent realization of the coalescent.

-s# : Sample size

The number of chromosomes to be generated for each sample (S). Because these chromosomes will be combined in individuals, this should be an even number; otherwise SNPsim will make it even by adding one chromosome. In such a case SNPsim will issue a warning message.

-l# : Number of sites

The total length (L), in base pairs or nucleotides, of the chromosomes. This length includes variable and non-variable sites.

-e# : Effective inbreeding population size

The effective inbreeding population size (N) of the population from which the sample was theoretically drawn.

-d# : Demographic periods

The number of periods (from present to past) and N during those periods. The first number specifies the number of periods. The for each period there should be three consecutive

numbers indicating the size N at the beginning and at the end of the period, and the duration of the period in generations.

The exponential growth rate during the period (positive or negative) will be deduced from the specified N at the beginning and at the end of the period. The growth rate derived for the last period will continue into the indefinite past. This implementation is borrowed from Hudson (2002). This option is incompatible with the exponential growth rate option ($-b$). **NOTE:** these parameters are looking back in time, so it is not a good idea to specify a negative growth rate for the last period, as the coalescent time could become infinite in the past.

$-b\#$: Exponential growth rate

Rate of exponential growth per individual per generation. This option is incompatible with the demographic periods ($-d$). **NOTE:** these parameters are looking back in time, so it is not a good idea to specify a negative growth rate for the last period, as the coalescent time could become infinite in the past.

$-r\#$: Homogeneous recombination rate

The background recombination rate per site per generation. All sites share this same rate, which is the standard recombination rate.

$-k\#$: Hotspot recombination rate

Expected recombination rate at the hotspots sites. If the hotspots are homogeneous (option $-t$ is not invoked) all the hotspots have the same rate.

$-h\#$: Expected number of hotspots

This is the expected number of hotspots for a given sample in the absence of interference. This parameter corresponds to the intensity parameter for a Poisson distribution from which the actual number of hotspots is drawn. For a given sample, the actual number of hotspots will change around this value. It does not have to be an integer. **NOTE:** When interference is specified we need to divide this number by the interference interval ($-z\#$) to obtain the expected number of hotspots. It does not have to be an integer.

$-q\#$: Fixed number of hotspots

This option fixes the number of hotspots inside the region of interest, so every sample will have the same number. In this case the hotspot locations are chosen from a uniform distribution. If the hotspots overlap, they will be displaced to the closest available location. Note that in this case no recombination events will originate from a hotspot located outside the region of interest.

$-v\#$: Hotspot imprecision

The hotspot imprecision corresponds to the variance of a Normal distribution for the specific site to recombine around the hotspot center (chosen by a Poisson process). The bigger the imprecision, the wider is the hotspot. If the imprecision is 0, all the recombination events happen exactly at the hotspot center. See figures 1 and 2.

$-m\#$: Hotspot width

This option specifies the width of the hotspots. In this case any site in the hotspot has the same probability of recombination. If the width is 1 all the recombination events happen exactly at the hotspot center. This parameters has to be bigger than 0. See figures 1 and 2.

$-t\#$: Hotspot heterogeneity

This parameter indicates that there is hotspot heterogeneity, that is, hotspots may have different recombination rates. This heterogeneity is accomplished through the use of the continuous gamma distribution. The shape parameter of this distribution (α) will control the strength of this heterogeneity. The smaller the shape the stronger the heterogeneity. This is similar to the application of Yang (Yang 1996).

$-z\#$: Hotspot interference

This parameter indicates whether the location of the hotspots is not independent of each other. If this parameter is 1 there is no interference, if it is between 0 and 1 hotspots tend to cluster, and if it is bigger than 1 hotspots will tend to be pushed away from each other.

$-u\#$: Mutation rate

Nucleotide mutation rate per site per generation. Because the aim here is to simulate SNPs, only two nucleotide states are permitted, 0 (ancestral) and 1 (derived). When this parameter is specified, the simulations are conditioned on the mutation population parameter θ ($= 4N\mu L$). This implies that the number of SNPs will change for each sample.

-x#: Simulate a particular number of SNPs

When this argument is present all the samples will have the specified number of SNPs regardless the sample coalescent branch lengths. This option is only compatible with the ISM.

-p#: Proportion of JC sites

Proportion of sites that will mutate according to the Jukes–Cantor (1969) mutation model for biallelic sites. Again this is an expectation, so different samples will differ in their proportions. Sites evolving under one or the other site can be printed a file the argument `-o` (see below) is present. See above for a description of the JC and ISM models.

-w#: Relative JC/ISM mutation rate

Relative rate for the JC sites compared to the ISM sites. If this parameter is > 1 the JC sites will be faster than the ISM of sites.

-1: Do not simulate data

If this options is specified the program does not produce any data. This option may be useful when we are interested only in coalescent parameters, like the variance of the number of recombination events or the observed number of hotspots, for example. This option is off by default.

-a: Print sequences

When this argument is specified both haplotypes for each diploid individual including all variable and invariable sites are printed to the file *sequences* in the *Results* folder. This file is the different than the default file *haplotypes*, which includes only variable sites. This option will slow down the simulations.

-f: Print genotypes

When this argument is specified all genotypes, including those at invariable sites, will be printed to the file *genotypes* in the *Results* folder. This file is the different than the default file *SNPs*, which only include genotypes at variable sites. Also in this file it is indicated which two haplotypes every individual is carrying. This option will slow down the simulations.

-i: Print trees

When this argument is specified all trees for each sample and site (in the case of recombination) will be printed to the file *trees* in the *Results* folder. This option will slow down the simulations.

-j: Print times

When this argument is specified the coalescent times for each tree for each sample and site (in the case of recombination), will be printed to the file *times* in the *Results* folder. This option will slow down the simulations.

-o: Print mutation model

When this argument is specified the corresponding mutation model (ISM or JC) assignment for each site will be printed to the file *mutations* in the *Results* folder. This option will slow down the simulations.

-g#: Memory allocation for tree nodes

Total number of tree nodes to allocate memory for. This option should be increased if SNPsim runs out of memory. This can happen when there are many chromosomes, sites, and recombination events.

-y#: Noisy level

The level of information to be printed in the screen.

- 0 : nothing
- 1: run information summarizing the simulations
- 2: run settings for each sample
- 3: calculation status and event (recombinations and coalescences) information
- 4: ancestral status for each chromosome at each event + MRCA status

5: potential recombining locations (g and G vectors)

Note that noisy values bigger than 1 will slow down the simulations.

-# : Seed

Seed for the random number generator. If no seed is specified, the computer clock will be used.

Default settings

By default SNPsim will simulate 100 samples of 8 chromosomes with 200 sites, a constant effective size of 1000, constant size, no recombination, and a mutation rate of $1e-07$ under the ISM. 1000 nodes will be allocated and the noisy level is 1. SNPs will be printed to the *SNPs* file in the *Results* folder. To run the program with the equivalent arguments we would type:

```
./snpsim -n10 -s8 -l100 -e1000 -d0 -b0.0e+00 -r0.0e+00 -k0.0e+00 -h0 -v0.0e+00 -
u1.0e-07 -p0.0e+00 -w1 -y1 -g1000
```

Program benchmarking

SNPsim has been benchmarked against other programs (see Hudson 2002) and when possible, against different analytical and numerical expectations for the mean and variances of the number of SNPs, the number of recombination events, the position of the mutations in the chromosomes and in the trees and the TMRCA. However, this does not mean that the program is bug free.

References

- Daly, M. J., J. D. Rioux, S. F. Schaffner, T. J. Hudson, and E. S. Lander. 2001. High-resolution haplotype structure in the human genome. *Nat Genet* 29:229–232.
- Gabriel, S. B., S. F. Schaffner, H. Nguyen, J. M. Moore, J. Roy, B. Blumenstiel, J. Higgins, M. DeFelice, A. Lochner, M. Faggart, S. N. Liu–Cordero, C. Rotimi, A. Adeyemo, R. Cooper, R. Ward, E. S. Lander, M. J. Daly, and D. Altshuler. 2002. The structure of haplotype blocks in the human genome. *Science* 296:2225–2229.
- Hudson, R. R. 1983. Properties of a neutral allele model with intragenic recombination. *Theoretical Population Biology* 23:183–201.
- Hudson, R. R. 2002. Generating samples under a Wright–Fisher neutral model of genetic variation. *Bioinformatics* 18:337–338.
- Jeffreys, A. J., L. Kauppi, and R. Neumann. 2001. Intensely punctate meiotic recombination in the class II region of the major histocompatibility complex. *Nature Genetics* 29:217–222.
- Jukes, T. H., and C. R. Cantor. 1969. Evolution of protein molecules. Pp. 21–132 in H. M. Munro, ed. *Mammalian Protein Metabolism*. Academic Press, New York, NY.
- Murphy, W. J., E. Eizirik, W. E. Johnson, Y. P. Zhang, O. A. Ryder, and S. J. O'Brien. 2001. Molecular phylogenetics and the origins of placental mammals. *Nature (London)* 409:614–618.
- Patil, N., A. J. Berno, D. A. Hinds, W. A. Barrett, J. M. Doshi, C. R. Hacker, C. R. Kautzer, D. H. Lee, C. Marjoribanks, D. P. McDonough, B. T. N. Nguyen, M. C. Norris, J. B. Sheehan, N. Shen, D. Stern, R. P. Stokowski, D. J. Thomas, M. O. Trulson, K. R. Vyas, K. A. Frazer, S. P. A. Fodor, and D. R. Cox. 2001. Blocks of Limited Haplotype Diversity Revealed by High-Resolution Scanning of Human Chromosome 21 *Science* 294:1719–1723.
- Posada, D., and C. Wiuf. 2003. Simulating haplotype blocks in the human genome. *Bioinformatics* 19:289–290.
- Wiuf, C., and D. Posada. 2003. A coalescent model of recombination hotspots. *Genetics* 164:407–417.
- Yang, Z. 1996. Among-site rate variation and its impact on phylogenetic analysis. *Trends in Ecology and Evolution* 11:367–372.

David Posada, 9 May 2004