PHYLOGEOSIM 2.0

>> manual (04-07-17)

S. Dellicour, C. Kastall, O.J. Hardy, P. Mardulyn

Evolutionary Biology and Ecology,
Université Libre de Bruxelles

PHYLOGEOSIM 2.0 is available free of charge from *ebe.ulb.ac.be/ebe/Software.html*
Contact and bugs report: Simon.Dellicour@ulb.ac.be

PHYLOGEOSIM (for "phylogeographic simulator") simulates the evolution of DNA sequences, microsatellites or SNPs under a model of coalescence on a 2-dimensional grid in which each cell is treated as one population. Going backward in time, populations can exchange gene copies and/or host coalescence events between two or more gene copies, at each generation. This geographic model of coalescence can be used to simulate data sets under evolutionary scenarios taking both demographic and geographic characteristics into account (e.g. isolation by distance, fragmentation, expansion, secondary contact, etc). In addition to generate data sets of genetic variation and gene genealogies, PHYLOGEOSIM also allows, in the case of DNA sequence data, the computation of several summary statistics, based on both genetic and geographic information. Users can use this set of statistics to compare different evolutionary scenarios or to estimate population genetic parameters. PHYLOGEOSIM is an open source software written in Java and can thus be run on any operating system on which a Java Virtual Machine is installed (e.g., Windows, Mac OS and Linux).

1. Model implemented in PHYLOGEOSIM

PHYLOGEOSIM simulates the evolution of DNA sequences, microsatellites or SNPs ("single nucleotide polymorphisms") under a demographic and geographic model of coalescence. The geographic structure of the model is based on a two-dimensional grid in which each cell corresponds to one population. Users have the opportunity to easily create a grid that overlays their own map (see example on the cover of this manual), and to determine the cells that are accessible at different time periods in the past. Throughout this manual, we will refer to a sampled DNA sequence, a microsatellite or a chromosome fragment in the case of SNPs, as a "gene copy".

1.1. Description of the algorithm implementing the coalescence simulation

A simulation begins at $t = 0$ (sampling time) and is finished when all gene copies of a given locus have coalesced. When more than one locus are simulated in parallel, the coalescence process for each locus is completely independent from that of other loci (assuming maximum recombination among loci). For the simulation of SNPs, recombination events within loci (i.e. chromosomes or fragments of chromosome; see below) are also considered.

Going backward in time, at each generation g, a given gene copy has the opportunity:

(1) to coalesce with another gene copy located in the same population. The probability of coalescence of a given gene copy located in the same population j is noted $P_c(j,g)$:

$$P_c(j,g) = \frac{n_j(g) - 1}{N_j(g)} \qquad \text{(eq. 1)}$$

with:

- N_j(g), effective size of population j (total number of gene copies) at $t = g$.
- n_j(g), the number of sampled gene copies located in population j at $t = g$.

(2) to recombine. At this stage, recombination events are only implemented for SNP data. The probability of recombination is defined by the recombination rate r of each gene copy, i.e. a recombination rate (per generation) defined by the length of the DNA fragment carrying the gene copy. The length of a DNA fragment is measured as the distance d between the positions of the two most distant SNPs present on this fragment. In PHYLOGEOSIM, the recombination rate r for each fragment is inferred from d_{max} (specified by the user), the distance on a chromosome above which the recombination rate r is maximal, i.e. $r = 0.5$. If the distance d between the two most distant SNPs on the DNA fragment is equal to or higher than d_{max}, the recombination rate for that fragment will be equal to 0.5. On the other hand, if d is smaller than d_{max}, this recombination rate will be equal to $(d/d_{max})*0.5$. Recombination events are implemented during the coalescence simulation using the algorithm of Hudson (Hudson, 1991), but with the two modifications proposed by Hein et al. (2005; p. 143-144).

(3) to migrate to one of the adjacent populations on the grid. The probability of a migration event equals the "backward" migration rate (see below) from the population of origin to the adjacent population. The probability of migration of a given gene copy located in population j is noted $P_m(j,g)$:

$$P_m(j,g) = \sum_{j'=1}^{K} m_{jj'}(g) \qquad \text{(eq. 2)}$$

with:

- $m_{jj'}(g)$, the backward migration from population j to population j' at

generation g, with $m_{jj'}(g) = 0$ if $j = j'$, that can be retrieved from the backward migration matrix.

- K, the total number of populations on the grid.

When two gene copies coalesce in a population, they merge into a single ancestral gene copy, and when a gene copy recombines, it splits into two ancestral gene copies. The simulation continues as long as more than one gene copy remains. Until the time to the most recent common ancestor (TMRCA), gene copies are free to move on the grid according to the different backward migration rates (see below). For further details on the algorithm used for the coalescence process (at least for the coalescence and migration events), see Dellicour *et al.* (2014).

At the end of the coalescence-migration(-recombination) simulation, the simulator builds a genealogy based on the recorded events of coalescence. In the case of **DNA sequences**, mutations are stochastically added on this genealogy according to a Jukes-Cantor model of DNA substitution (Jukes & Cantor, 1969). For each DNA sequence locus, the user has to specify: and (i) the number of mutation events OR a mutation rate (number of mutations per site and per generation) as well as (ii) the length of the sequence (number of nucleotides).

In the case of **microsatellite loci**, a fictive ancestral microsatellite is stochastically modified through the genealogy following a TPM model ("two-phase model"; Di Rienzo *et al*, 1994). In this model, given that a mutation occurs, it has a probability p of being a one-step mutation and a probability (1-p) of being a multistep mutation. In the one-step phase, the descendent allele has an equal chance of being one repeat unit larger or smaller than its ancestor. In the multi-step phase, the change in the number of repeat units is drawn from a specified geometric distribution that allows for large changes in repeat number. For each microsatellite locus, the user has to specify (i) the number of different alleles OR the substitution rate, (ii) the probability p of the TPM model and (iii) the variance of the TPM geometric distribution defining the change in the number of repeat units in the multi-step phase.

For **SNP data**, two options are available regarding mutations: the user can provide (1) a uniform distribution of substitution rates or (2) *a priori* the proportion of SNPs with two, three and four alleles. With the first option, we assume that genotypes were generated via a RAD-seq- or GBS-type approach, i.e. that homologous DNA fragments of identical sizes (usually ± 100 pb), from different parts of the genome, were compared among individuals, to simultaneously identify and type polymorphic sites. The number of substitution events associated with each DNA fragment will then be drawn from a Poisson distribution of mean equal to μT (where μ is a substitution rate value drawn from the uniform distribution specified by the user and T is the length of the genealogy). The program assumes an infinite substitution model and creates a new SNP for a given fragment (associated with a single gene genealogy) for each substitution event. Therefore, some simulated fragments may be entirely monomorphic (no associated SNP) and others may include more than one SNP. It is trivial to infer the proportion of monomorphic DNA fragments and the average number of SNP per fragment from each simulated data set, two parameters that can also be extracted from the observed data. Comparison of these two parameters between simulated and observed SNP data, can help define a range of substitution rates for the simulations consistent with the observed data. Note that with this option, it is not possible to predict the exact number of SNP loci generated by a simulation (some DNA fragments will be entirely monomorphic, others may contain more than one SNP locus). With the second option, each simulated genealogy generates a single SNP, and the exact requested number of SNP loci will be generated. Use of this option should be made with caution however: it assumes that all simulated genealogies are sufficiently long that a mutation will necessarily occur at the associated site, an assumption that is unlikely to be realistic. Indeed, observed data sets contain only polymorphic sites (i.e. SNPs) as a result of an active selection process that has rejected monomorphic sites; monomorphic sites are on average associated with shorter genealogies (lower probability of mutation). For the simulations to be realistic, it should mimic

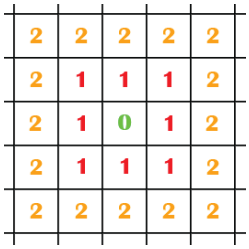
the process of rejecting sites associated with too short genealogies, which is not the case under this option (no filtering of genealogies). In addition, this option generates a single SNP per genealogy, and therefore does not allow generating more than one SNP per DNA fragment/genealogy, as we would expect under a Rad-seq-type approach.

To select the first mutation option, the user specify the five following parameters: (i) the total number of SNPs, (ii) the lower bound of the uniform distribution of substitution rates, (iii) the upper bound of the uniform distribution of substitution rates, (iv) the number of chromosomes and (v) the distance d_{max} ([0,1]) expressed as a proportion of total chromosome size. For the second option, an alternative line of five parameters must be defined: (i) the number of SNPs with two alleles, (ii) the number of SNPs with three alleles, (iii) the number of SNPs with four alleles, (iv) the number of chromosomes and (v) the distance d_{max} (i.e., parameters 4 and 5 are identical in both options). Note that for each simulation, the SNPs are randomly distributed between and along the different chromosomes. Also, after having been randomly distributed on each chromosome, groups of SNPs located on a same chromosome but separated by a distance higher than d_{max} from other groups will be treated as independent loci from the very beginning of the simulation process.

1.2. The need for a preliminary forward simulation to create a backward migration matrix

Coalescence simulations implemented in PHYLOGEOSIM use **backward migration rates.** In some cases, forward and backward evolution migration rates are not identical. For example, in the case of a geographic expansion, the probability that gene copies from a previously occupied cell A migrate in one generation to a newly colonized cell B on the grid could be set to 0.001, defining the *forward* evolution migration rate from A to B. However, to implement the coalescence simulation going backward in time, we need to know the *backward* evolution migration rate, i.e., the probability that gene copies are transferred from B to A, going backward in time. Because gene copies in B at the first generation of the geographic expansion (forward evolution) are all migrants from A, the *backward* migration rate from B to A is equal to 1.0. This *backward* migration rate will then decrease at each generation of *forward* evolution, as the population effective size of cell B increases (following reproduction and migration) until it reaches the maximum size set by the user. Because it would be extremely laborious to manually generate all necessary backward migration matrices for the coalescence simulation, and because there is a certain level of stochasticity associated with this process, PHYLOGEOSIM will generate these matrices automatically by performing a forward simulation.

The program will thus perform a preliminary forward simulation to estimate the different backward migration rates and effective population sizes occurring at each generation. For this purpose, PHYLOGEOSIM requires that users specify an initial matrix of ancestral effective population sizes, one or more matrices of maximal effective population sizes (and generations at which they change), a rate of reproduction, two short-distance and one long-distance **forward migration rates**. The two short-distance forward migration rates fm_1 and fm_2 correspond to two different short-distance migration levels, i.e. between adjacent populations and between populations separated by two cells on the grid:



- fm_1 (e.g. 0.001): first level of short-distance migration, migration between cell **0** (origin) and cells **1** (destinations).
- fm_2 (e.g. 0.0001): second level of short-distance migration, migration between cell **0** (origin) and cells **2** (destinations).

Figure 1: illustration of the two scales of migration on the grid.

In addition to the two levels of short-distance migration, PHYLOGEOSIM also allows the possibility of long-distance migration events. Such events are specified with the following parameters: (i) a LDD (long-distance dispersion) rate fm_{LDD}, (ii) an upper distance limit LDD_{max} expressed as an Euclidian distance in number of cells, (iii) the name of the statistical distribution of the LDD distances ("uniform" or "lognormal) and, in the case of a "lognormal" distribution, (iv) the value of the "scale" parameter of this lognormal distribution (the "location" parameter of this distribution being automatically set to zero). Note that the lognormal distribution is truncated based on the maximal LDD distance LDD_{max} also used to define the limit of the "uniform" distribution ([0, LDD_{max}]).

The model therefore assumes that the dispersal ability and reproduction rate of the studied organism is constant across the entire grid. Only the carrying capacity of each population is allowed to change by defining the maximum effective size of each cell separately. Note that the presence of a barrier to migration (e.g. a mountain range or a river) can be modelled by assigning a small (or null) maximal effective size to one or more squares, thereby reducing the probability of migration through them. The preliminary forward simulation is performed as followed: starting with the most ancestral matrix of effective size, at a time defined by the user, the simulation proceeds forward one generation at a time, until t = 0. At each generation:

(1) all the effective population sizes are recorded.
(2) migration events among populations are simulated, using the two pre-defined forward migration rates and the effective population sizes of the preceding generation. These migration events are recorded and used to estimate the backward migration rates at this generation.
(3) the effective size of each population increases as individuals reproduce and new migrants are brought in.
(4) if effective population sizes have exceeded their maximal value, these sizes are reduced to these maximal values.

When two populations have reached their maximal effective size, instead of being estimated by simulating an exchange of migrants, the two backward migration rates connecting them are directly estimated according to the following deterministic formula:

$$m_{jj'}(g) = \frac{N_{j'}(g-1)\cdot M_{j'j}}{\sum_{p=1}^{P}\left(N_p(g-1)\cdot M_{pj}\right)}$$

1th deterministic formula (eq. 3)

with:

- $m_{jj'}(g)$, the backward migration rate from population j to population j' at generation g.
- $N_j(g)$, the effective size (haploid case) of population j at generation g.
- $M_{jj'}$, the forward migration rate from population j to population j'. If $j = j'$:

$$M_{jj'} = 1 - \sum_{p \neq j} M_{jp}$$

- P, the total number of populations (cells) on the grid.

When all the effective sizes have reached their maximum, the forward simulation stops until the next change of maximal effective sizes matrix or until the end of the forward simulation. Until this point, all the backward migration rates are fixed values estimated with this 1th deterministic formula. The use of a deterministic formula to estimate backward migration rates when effective population sizes are constant allows a significant increase of the forward simulation speed.

At each generation, estimated effective population sizes and backward migration rates are recorded and will be used for the backward simulations. Since this preliminary forward simulation is **stochastic** (migration events occur according to **probabilities** defined by forward migration rates), the program can renew the forward simulation every X backward simulations to account for this stochasticity (X being defined by the user).

Note that the preliminary forward simulation does not necessarily cover the entire range of generations that will be spanned during the main backward simulation. Indeed, when the backward simulation reaches the most ancestral generation of the forward simulation, the program simply uses the most ancestral effective population sizes and backward migration rates are estimated according to the following deterministic formula:

$$m_{a,jj'} = \frac{N_{a,j'} \cdot M_{j'j}}{\sum_{p=1}^{P} \left(N_{a,p} \cdot M_{pj} \right)}$$

2th deterministic formula (eq. 4)

with:

- $m_{a,jj'}$, the backward migration rate from population j to population j' estimated during the most ancestral generation of the forward simulation.
- $N_{a,j}$, the effective size of population j (haploid case) estimated during the most ancestral generation of the forward simulation.

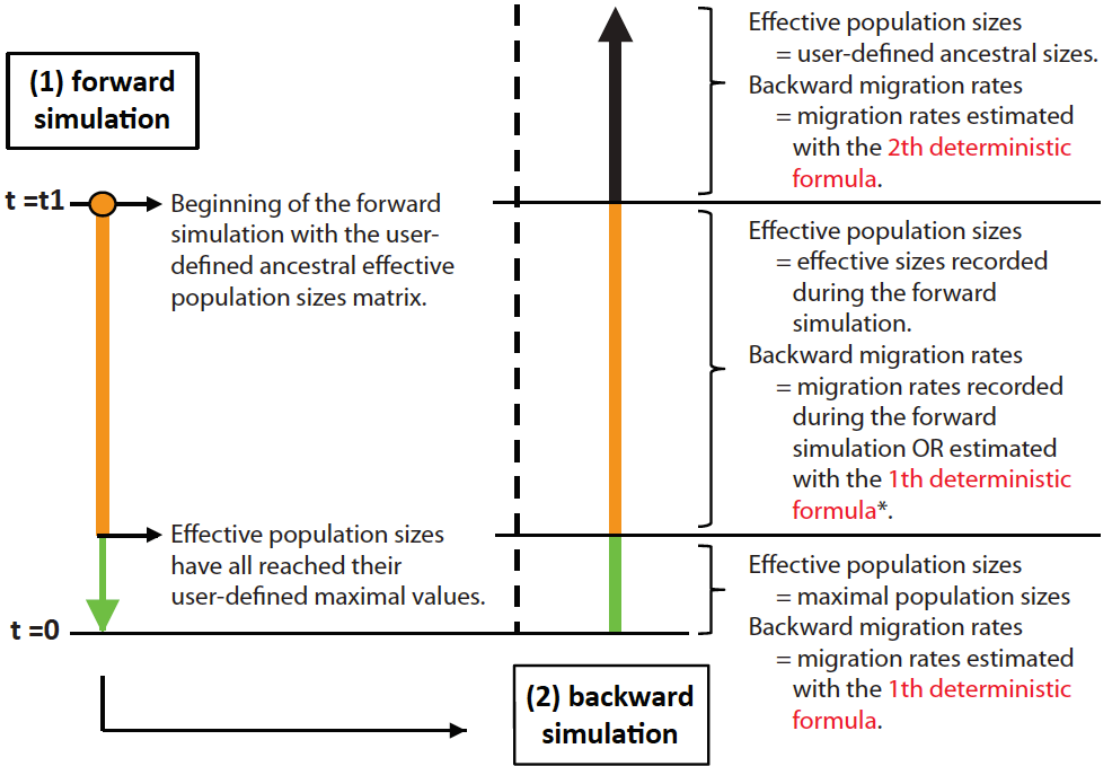


Figure 2: schematic program flow with the "input.txt" example file. (*) During the forward simulation, when two populations have reached their maximal effective size, the two backward migration rates connecting them are directly estimated using the 1th deterministic formula, instead of being estimated by simulating an exchange of migrants.

1.3. Simulation parameters

Below, the list of parameters that needs to be specified by the user:

- Number of populations (or number of cells on the grid).
- Number of independent backward simulations to perform.
- Rate at which the software will renew the forward simulation. The software thus uses backward migration rates and effective population sizes recorded on only one forward simulation but renew this simulation every X backward simulations (X being defined by the user).
- Reproduction rate t_R.
- Two short-distance forward migration rates fm_1 and fm_2, respectively between adjacent populations and populations separated by one grid cell.
- Number of user-defined groups. These groups of populations are specified in an additional input file ("groups.txt", cfr. 3. Input files) and are used as a basis for the computation of several summary statistics (cfr. 2. Computed summary statistics for DNA sequences).
- Long-distance forward migration rate fm_{LDD}, along with the name of the statistical distribution of the LDD distances ("uniform" or "lognormal"), the upper distance limit LDD_{max} (expressed as an Euclidian distance in number of cells), and, in the case of a "lognormal" distribution, the value of the "scale" parameter of this lognormal distribution.
- In the case of DNA sequences, for each locus: number of mutations that will occur on the final genealogy or the mutation rate (number of mutations per site and per generation), and the length the DNA sequence (i.e. number of nucleotides).
- In the case of microsatellites, for each locus: number of different alleles or the mutation rate, probability p of the TPM model and variance of the TPM geometric distribution defining the change in the number of repeat units in the multi-step phase.
- In the case of SNPs, there are two options: (1) total number of SNPs, lower and upper bounds of the uniform distribution of substitution rates, number of chromosomes and distance d_{max} expressed as a proportion of total chromosome size; (2) number of SNPs with two, three and four alleles, number of chromosomes and distance d_{max} expressed as a proportion of total chromosome size.
- A list of the different summary statistics to compute after each simulation (case of DNA sequences).
- A matrix with all population ID's, determining the position of each population on the grid and the dimensions of this grid.
- A matrix with the number of sampled individuals in each cell (population) of the grid. This matrix has to correspond to the matrix of population ID's.
- The maximal effective population sizes matrices and the generation at which they occur.
- The ancestral effective population sizes matrix.

2. Computed summary statistics for DNA sequences

In the case of DNA sequences simulation, PHYLOGEOSIM proposes the computation of the following summary statistics after each simulation (this list is likely to be expanded as needed, suggestions are welcome)[1]:

- $N_H tot$, the total number of allelic types for the considered locus.
- global G_{ST} estimator (Pons & Petit, 1995) on sampled populations (sampled cells).
- global N_{ST} estimator (Pons & Petit, 1996) on sampled populations (sampled cells).
- AMOVA global Φ_{ST} estimator for K=1 (Excoffier *et al*, 1992) on sampled populations (sampled cells).
- *IBDSC*: isolation by distance slope coefficient. This is the slope coefficient of the linear regression estimated from y = f(ln(x)) with y = (Φ_{ST}/(1-Φ_{ST})) (Rousset, 1997).
- X_H: ratio between the number of haplotypes in a user-defined group of populations and the total number of haplotypes.
- π: nucleotidic diversity (Nei & Li, 1979) within each user-defined group of populations.

$$\pi = \frac{2!\,(n-2)!}{n!}\sum_{i=1}^{n-1}\sum_{i'=i+1}^{n} k_{ii'}$$

with:

- $k_{ii'}$, number of differences between sequence i and sequence i'.
- n, number of sequences in the considered user-defined group.

- π_R: computed for each user-defined group of population, this is the ratio between the nucleotidic diversity within a considered user-defined group of populations and the nucleotidic diversity within the virtual group formed by all the other populations which are not in this group (Mardulyn *et al*, 2009).
- A_R: estimator of allelic richness within each user-defined group of populations (El Mousadik & Petit, 1996).
- $X_H S$, πS, $\pi_R S$ and $A_R S$: ratios between each X_H, π, π_R or A_R and the surface of the corresponding user-defined group. This surface is the number of cells or the number of populations present in the group.
- Φ_{SC}, Φ_{ST}, Φ_{CT}: AMOVA Φ-statistics (Excoffier *et al*, 1992) computed for the population structure linked to the user-defined groups.
- pairwise Φ_{ST} (Excoffier *et al*, 1992) between user-defined groups of populations.
- $m\Phi_{ST}dgeo$: the average of ratios between Φ_{ST} estimators and geographical distances between all pairwise populations.

$$m\Phi_{ST}dgeo = \frac{2!\,(p-2)!}{p!}\sum_{j_1 \neq j_2}\left(\frac{\Phi_{ST\,j_1 j_2}}{d_{j_1 j_2}}\right)$$

with:

- p, the number of populations.
- Φ_{STj1j2}, Φ_{ST} between populations j_1 and j_2.
- d_{j1j2}, geographical distance between populations j_1 and j_2.

[1] These same summary statistics can be computed on real data sets using the software SPADS (as described at the end of this manual).

3. Input files

PHYLOGEOSIM requires two different input files:

(1) **the “input.txt” file** is the main input file containing all the simulation parameters. Here is an example of this file, designed for an expansion scenario on a 10x10 grid:

```
Input file for PhyloGeoSim 2.0

1 --------------------------
100 1000 0.01 1.5 0.001 0.0001 4 0.00001 lognormal 8 2

2 --------------------------
DNA 15 800 11 600 5 330

3 --------------------------
Nhtot Gst Nst Fst mFstdgeo IBDSC Xh Pi Pir Ar AMOVA groupsFst

4 --------------------------
1 2 3 4 5 6 7 8 9 10
11 12 13 14 15 16 17 18 19 20
21 22 23 24 25 26 27 28 29 30
31 32 33 34 35 36 37 38 39 40
41 42 43 44 45 46 47 48 49 50
51 52 53 54 55 56 57 58 59 60
61 62 63 64 65 66 67 68 69 70
71 72 73 74 75 76 77 78 79 80
81 82 83 84 85 86 87 88 89 90
91 92 93 94 95 96 97 98 99 100
```

Matrix of population ID’s

```
5 --------------------------
0 10 0 0 0 0 0 0 0 10
0 0 0 0 0 0 0 0 0 0
0 0 0 0 0 0 0 0 0 0
0 0 0 0 0 0 10 0 0 0
0 0 0 10 0 0 0 0 0 0
0 0 0 0 0 10 0 0 0 0
0 0 0 10 0 0 0 0 0 0
0 0 0 0 0 0 0 0 0 0
10 0 0 0 0 0 0 0 10 0
0 0 0 0 0 0 0 0 0 0
```

Number of sampled gene copies per population

```
6 --------------------------
1000
```

Times at which the maximal effective population sizes matrices occur. Here: only one matrix, between t=1000 (t1) and t=0.

```
7 --------------------------
5000 5000 5000 5000 5000 5000 5000 5000 5000 5000
5000 5000 5000 5000 5000 5000 5000 5000 5000 5000
5000 5000 5000 5000 5000 5000 5000 5000 5000 5000
5000 5000 5000 5000 5000 5000 5000 5000 5000 5000
5000 5000 5000 5000 5000 5000 5000 5000 5000 5000
5000 5000 5000 5000 5000 5000 5000 5000 5000 5000
5000 5000 5000 5000 5000 5000 5000 5000 5000 5000
5000 5000 5000 5000 5000 5000 5000 5000 5000 5000
5000 5000 5000 5000 5000 5000 5000 5000 5000 5000
5000 5000 5000 5000 5000 5000 5000 5000 5000 5000
```

Matrix of maximal effective population sizes

```
5000 5000 5000 5000 0 0 0 0 0 0
5000 5000 5000 5000 0 0 0 0 0 0
5000 5000 5000 5000 0 0 0 0 0 0
5000 5000 5000 5000 0 0 0 0 0 0
0 0 0 0 0 0 0 0 0 0
0 0 0 0 0 0 0 0 0 0
0 0 0 0 0 0 0 0 0 0
```

Matrix of ancestral population sizes

Description of each line in the input file:

- 1. (4th line) A list of parameters:
  (a) the number of populations (number of cells on the grid).
  (b) the number of independent backward simulations to perform.
  (c) the number of preliminary forward simulations to perform. This number has to be smaller than one and will be used as a rate at which the software performs a new forward simulation. In this example, there will be a new forward simulation every 10 backward simulations (1000*0.01 = 10).
  (d) the reproduction rate t_R (only used in the preliminary forward simulation).
  (e) the first level short-distance forward migration rate fm_1 (between adjacent cells; only used in the forward simulation).
  (f) the second level short-distance forward migration rate fm_2 (between adjacent cells; only used in the forward simulation).
  (g) the number of user-defined groups (only used for DNA sequence summary statistics estimation).
  (h) the long-distance forward migration rate fm_{LDD} (optional; only used in the forward simulation).
  (i) the name of the statistical distribution of the LDD distances ("uniform" or "lognormal") for the long-distance forward migration rate (optional).
  (j) the upper distance limit LDD_{max} (expressed as Euclidian distance in number of cells) for the long-distance forward migration rate (optional).
  (k) in the case of the LDD "lognormal" distribution, the value of the "scale" parameter of this lognormal distribution (optional).
- 2. (7th line) For each locus: the first element/keyword specifies the type of marker: "DNA", "mSAT" (or "STR"), or "SNP" (not case sensitive).
  If the type of marker is set to "DNA":
  (a) the number of mutations that will occur on the final genealogy of each locus OR the mutation rate of each locus (number of mutations/locus/generation).
  (b) the length (number of nucleotides) of each locus.
  If the type of marker is set to "mSAT" or "STR":
  (a) the number of different alleles OR the substitution rate.
  (b) the probability p of the TPM model.
  (c) the variance of the TPM geometric distribution defining the change in the number of repeat units in the multi-step phase.
  If the type of marker is set to "SNP":
  (a) the number of SNPs with two alleles OR the total number of SNPs (if a uniform distribution of substitution rates is defined).
  (b) the number of SNPs with three alleles OR the lower bound of the uniform distribution of substitution rates.
  (c) the number of SNPs with four alleles OR the upper bound of the uniform distribution of substitution rates.
  (d) the number of chromosomes.
  (e) the distance d_{max} ([0,1]) expressed as a proportion of total chromosome size.
  In this input file example, there are 3 DNA sequence loci: the first is 800 bp long and is associated to 15 substitutions, the second is 600 bp long with 11 substitutions and the third is 330 bp long with 5 substitutions.
- 3. (10th line) The list of the different summary statistics to compute after each simulation. The following keywords must be used to invoke them: "Nhtot", "Gst", "Nst", "Fst" (for Φ_{ST}), "mFstdgeo" (for $m\Phi_{ST}dgeo$), "S", "IBDSC", "Xh", "Pi" (for π), "Pir" (for π_R), "Ar", "AMOVA", "groupsFst" (for pairwise Φ_{ST}).
- 4. The matrix with all the populations ID's (numbers, from 1 to the total number of cells on the grid), determining the position of each population on the grid and the dimensions of this grid. The grid and this corresponding matrix do not have to be square-like, e.g. a grid with dimensions 20x30 can be created.

- 5. The matrix with the number of sampled individuals in each cell (population) of the grid. This matrix must correspond to the matrix of population ID's.
- 6. Times (in generations) until which the different maximal effective population sizes matrix occur. The last time will also correspond to the most ancestral effective population sizes matrix, and thus to the beginning of the preliminary forward simulations.
- 7. The different maximal effective population sizes matrices. These matrices correspond to the population ID's matrix. In this input file example, there will be only one population sizes change, occurring t = 1000 generations ago (corresponding to $t1$ in figure 2). The last matrix must be the ancestral population sizes matrix.

(2) **the "groups.txt" file**: this text file contains the different groups ID's followed by the ID's of populations that are in the group. These groups are thus user-defined and used for the computation of some summary statistics (X_H, π, π_R, and A_R). Here is an example:

```
groupA 1 2 3 4 5 11 12 13 14 15 21 22 23 24 25 31 32 33 34 35 41 42 43 44 45
groupB 6 7 8 9 10 16 17 18 19 20 26 27 28 29 30 36 37 38 39 40 46 47 48 49 50
groupC 51 52 53 54 55 61 62 63 64 65 71 72 73 74 75 81 82 83 84 85 91 92 93 94 95
groupD 56 57 58 59 60 66 67 68 69 70 76 77 78 79 80 86 87 88 89 90 96 97 98 99 100
```

Each line must correspond to a group and begin with the group ID followed by some populations ID's separated by single spaces. In this example, the second line refers to the group B which contains populations n°6, 7, 8, 9, 10, 16, 17, etc.

WARNING:

- these two input files need to be formatted as text files with elements separated by single spaces or tabulations. One easy way to construct the "input.txt" file is to write it in Excel, and to save it into a "tab-delimited text file".
- the precise location of each line in the "input.txt" file is very important.

4. How to run the program

Input files ("input.txt" and "groups.txt" files) must be located in the same folder as the "PhyloGeoSim_2.0.jar" executable file. If the user wants to generate and saves the simulated genealogies and DNA sequence matrices, a folder precisely named "simulations_outputs" has to be created in the same folder as the executable file. A double click on the executable file will start the program. PHYLOGEOSIM was created as a command-line program, to facilitate its interaction with other programs. Here is the command to launch PHYLOGEOSIM from a terminal window:

java -jar PhyloGeoSim_2.0.jar

Note that if the user wants to specify a specific input file name and output prefix, the following command can be used to launch PHYLOGEOSIM from a terminal window:

java -jar PhyloGeoSim_2.0.jar input_file_name.txt output_file_prefix

5. Output files

PHYLOGEOSIM returns several output files:

(1) **simulated genealogies** (Newick format, ".tre" extension file) **and simulated molecular variation data (DNA sequences, SNPs or microsatellite genotypes).** These files will be generated only if a folder precisely named "simulations_outputs" is created in the same folder as the "PhyloGeoSim 2.0.jar" executable file. By not creating this folder, users can decide to avoid generating these files, e.g., in cases in which a high number of simulations are performed. If DNA sequences are simulated, the output file is generated in the Phylip format (Felsenstein, 2004) and has the ".phy" extension. When Newick and Phylip files are generated, they are labelled with the corresponding simulation number. If SNP data are simulated, a text file provides on each line the allele frequencies for a given SNP at all populations. The first text string in such a line identifies the SNP, locus and chromosome, then the number of each of the 4 possible alleles are given separately for each sampled population. Note that only polymorphic loci are included; since all simulated loci are serially numbered, it is easy to identify missing loci, i.e., monomorphic loci that were excluded from the output file, and thus to calculate the overall proportion of simulated monomorphic loci. Finally, if microsatellite data are simulated, the output file is generated in the SPAGeDi format (Hardy & Vekemans, 2002). This program can then be used to export the data file in other known formats, such as the GenePop format (Rousset, 2008).

(2) **A tab-delimited text file containing all the computed values of the chosen summary statistics (only if simulating DNA sequences)**. The name of this file is always "summary_stats_temp.txt". The name and the structure of this text file have been chosen in order to make it compatible with the package ABCtoolbox (Wegmann *et al*, 2010) (cfr. 6. ABCtoolbox compatibility). The first line contains the name of the different summary statistics chosen by the user. Below, each line corresponds to a distinct simulation and contains summary statistics values computed on the corresponding simulated data set. **A "messages_file.txt" containing several additional data**: the parameters values defined by the user and potential error messages.

6. *ABCtoolbox* compatibility

PHYLOGEOSIM 2.0 is compatible with the Approximate Bayesian Computation (ABC) package ABCtoolbox (Wegmann *et al*, 2010). This should make it possible to integrate this simulator in an ABC analysis. Note however that the amount of time needed to achieve one

simulation with PHYLOGEOSIM can be large. Since ABC analyses (even if coupled with an MCMC algorithm) often require more than 100,000 simulations, an ABC analysis may turn out to be unreasonably long. We thus advise users who want to use PHYLOGEOSIM in an ABC framework to perform exploratory analyses to check the speed of performing simulations and to identify useful summary statistics to compute.

7. Software limitation

In our view, PHYLOGEOSIM has one main practical limitation: a RAM memory limitation (or "Java heap space") can be reached during the forward simulations process because a lot of information has to be stored (migration events and population effective sizes).

This can be the case when:

- effective population sizes take time to reach their maximal sizes. Until this point, the forward simulation works generation by generation and thus saves one set of backward migration rates and effective sizes per generation. But when all effective population sizes reach their maximal value, the program uses a deterministic formula (1th deterministic formula) to estimate the only one set of backward migrations rates occurring until the next change of maximal effective population sizes matrix.
- user asks for a given number of preliminary forward simulations to perform before all the backward simulations. In this case, the software computes average backward migration rates and effective population sizes over all these recorded preliminary forward simulations. Since the program computes average backward migration rates and effective sizes based on all these preliminary forward simulations, it has to store a lot of information before computing average values.

These large amounts of information to save can cause the RAM memory limit of the Java virtual machine (JVM) to be reached. In this case, a solution to try again on the same computer is to increase the RAM memory allowed to the JVM by the operating system.

8. PGSVIEWER 2.0

PGSVIEWER 2.0 allows to visualise the evolution of effective population size matrices during forward simulations performed in the same way as PHYLOGEOSIM. PGSVIEWER uses the same input files and stops at the end of the forward simulation. This can be useful before launching a big set of simulations with PHYLOGEOSIM in order to check if the forward simulation corresponds to the evolutionary scenario implemented by the user. PGSVIEWER 2.0 is a JavaScript application compiled from a R script with the "shiny" package. To use PGSVIEWER 2.0 and generate an animated GIF display the forward simulation, follow these steps:

(1) install ImageMagik® (it can be found on www.imagemagik.org)
(2) in R:

```
> install.packages("shiny")
> library(shiny)
> runGist("eded0ace2f52a1840451f9b91878d35c")
```

9. SPADS 1.0

SPADS 1.0 (for "Spatial and Population Analysis of DNA Sequences") is a population genetics software computing the same summary statistics as PHYLOGEOSIM but on real data sets. SPADS also implements two clustering algorithms (the SAMOVA and Monmonier algorithm) and several input file conversion. See SPADS manual for further details about this toolbox.

10. Software availability

PHYLOGEOSIM 2.0 is available free of charge at *ebe.ulb.ac.be/ebe/Software.html*. Java source code, example files and the software manuals can all be downloaded at this address. Questions and bug reports should be directed to simon.dellicour[at]ulb.ac.be.

11. References

Bandelt H.-J., Forster P., Röhl A. (1999). Median-joining networks for inferring intraspecific phylogenies. Mol Biol Evol **16**:37-48.

Dellicour S., Kastally C., Hardy O.J., Mardulyn P. (2014). Comparing phylogeographic hypotheses by simulating DNA sequences under a spatially explicit model of coalescence. *Molecular Biology and Evolution* **31**: 3359-3372.

Di Rienzo A., Peterson A.C., Garza J.C., Valdes A.M., Slatkin M., Freimer N.B. (1994). Mutational processes of simple-sequence repeat loci in human populations. *Proceedings of the National Academy of Sciences of the United States of America* **91**: 3166-3170.

El Mousadik A., Petit, R. J. (1996). Chloroplast DNA phylogeography of the argan tree of Morocco. *Molecular Ecology* **5**: 547–555.

Excoffier L., Smouse P.E., Quattro, J.M. (1992). Analysis of molecular variance inferred from metric distances among DNA haplotypes: application to human mitochondrial DNA restriction data. *Genetics* **131**: 479-491.

Excoffier L., Dupanloup I., Huerta-Sánchez E., Sousa V.C., Foll M. (2013) Robust demographic inference from genomic and SNP data. *PLoS Genetics* **9**: e1003905.

Felsenstein J. (2004). PHYLIP (PHYLogeny Inference Package) version 3.6a2, Department of Genome Sciences, University of Washington, Seattle (2004).

Hardy O.J., Vekemans X. (2002). SPAGeDi: a versatile computer program to analyse spatial genetic structure at the individual or population levels. *Molecular Ecology Notes* **2**: 618-620.

Hein J. (2005). Gene genealogies, variation and evolution: a primer in coalescent theory. Oxford, New York, Oxford University Press.

Hudson R. (1991). Gene genealogies and the coalescence process. In Oxford Surreys in Evolutionary Biology, eds D. Futuyama and J. Antonovics, Vol. 7, Oxford University Press, pp. 1-44.

Jukes T.H., Cantor C.R. (1969). Evolution of protein molecules. Mammalian protein metabolism, pp. 21-123. H.N. Munro, ed., Academic Press, New York.

Mardulyn P., Mikhailov Y., Pasteels J.-M. (2009). Testing phylogeographic hypotheses in a Euro-Siberian cold-adapted leaf beetle with coalescent simulations. *Evolution* **63**: 2717-2729.

Nei M., Li W.H. (1979). Mathematical model for studying genetic variation in terms of restriction endonucleases. *Proceedings of the National Academy of Sciences of the United States of America* **76**: 5269–5273.

Page, R. D. M. (1996). TREEVIEW: An application to display phylogenetic trees on personal computers. *Computer Applications in the Biosciences* **12**: 357-358.

Pons 0., Petit R. J. (1995). Estimation, variance and optimal sampling of genetic diversity. I. Haploid locus. *Theo'r. Appl. Genet.* **90**: 462-470.

Pons 0., Petit R. J. (1996). Measuring and testing genetic differentiation with ordered versus unordered alleles*. Genetics* **144**: 1237-1245.

Rousset F. (1997). Genetic differentiation and estimation of gene flow from F-statistics under isolation by distance. *Genetics* **145**: 1219–1228.

Rousset F. (2008). Genepop'007: a complete reimplementation of the Genepop software for Windows and Linux. *Molecular Ecology Resources* **8**: 103-106.

Wegmann D., Leuenberger C., Neuenschwander S., Excoffier L. (2010). ABCtoolbox: a versatile toolkit for approximate Bayesian computations. *BMC Bioinformatics* **11**: 116.