SELECTOR version 1.0

USER MANUAL

January 2016

Mathias Currat
mathias.currat@unige.ch

AGP lab
Department of Genetics and Evolution – Anthropology Unit
University of Geneva, Switzerland

Table of Contents

Program description

Computer simulation is a powerful tool to disentangle the effects of demography and natural section on the genetic diversity of populations. SELECTOR is a program that simulates the evolution of genetic loci with multiple alleles, with or without the effect of balancing selection, in a complex spatially-explicit population framework. It has been specifically designed to study MHC lineages but can be used on any other genetic loci with similar characteristics. Complex demographic scenarios can be simulated, incorporating demographic regulation, migration, intra-population competition and multiple population sources. All these parameters can be varied in space and time to account for geographical, environmental or cultural heterogeneity. SELECTOR may be used to generate allele frequency trajectories under complex demographic scenarios and can thus serve for hypothesis testing and parameters estimation. Resulting allele frequencies are outputted in a format compatible with the program ARLEQUIN (Excoffier and Lischer 2010), thereby allowing the user to perform ARLEQUIN-type analyses on simulated datasets. Moreover, SELECTOR has been designed to be easily integrated into an Approximate Bayesian Computation (ABC) framework (Beaumont et al. 2002) and can be used in conjunction with tools such as the ABCtoolbox (Wegmann et al. 2010). The main originality of SELECTOR is that it combines the simulation of population spatial demography and balancing or positive selection at a multi-allelic locus.

How to cite SELECTOR

M. Currat, P. Gerbault, D. Di, J.M. Nunes, A. Sanchez-Mazas (2016) Forward-in-Time, Spatially Explicit Modeling Software to Simulate Genetic Lineages Under Selection. ***Evolutionary Bioinformatics*** 2015:Suppl. 2 27-39.

Descriptive article download

http://www.la-press.com/forward-in-time-spatially-explicit-modeling-software-to-simulate-genet-article-a5417

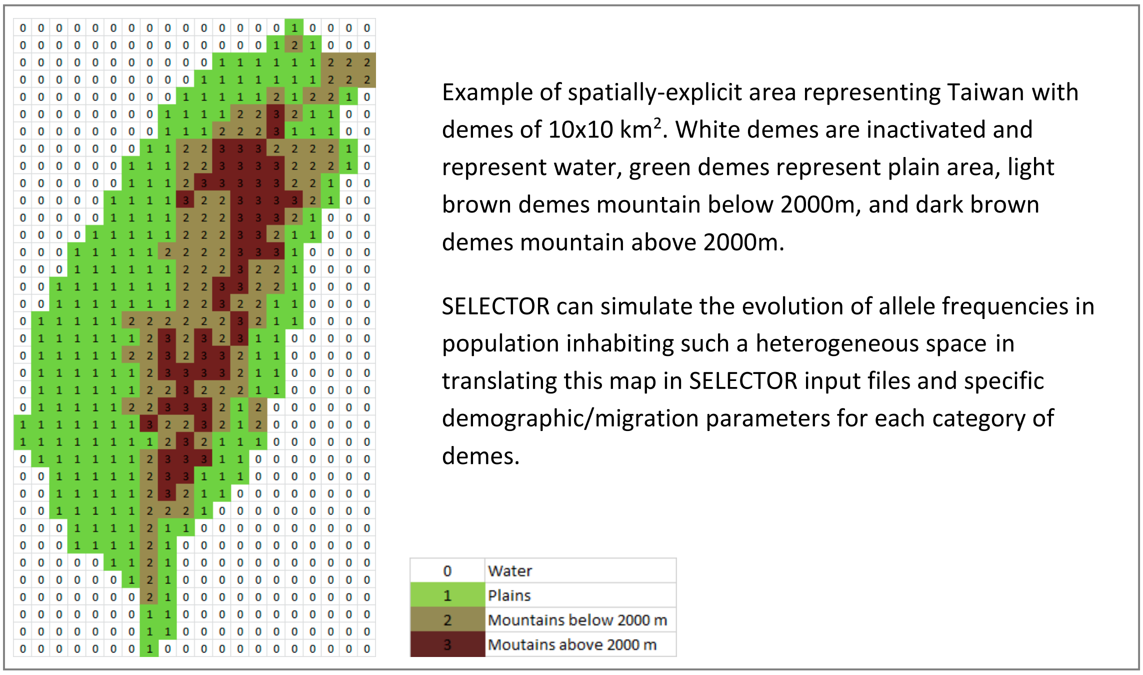
Program version and availability

SELECTOR is written in C++ and compiled for Windows and Linux operating systems (64bits). It can be freely downloaded at http://ua.unige.ch/en/agp/tools/selector/, together with two example datasets (BasicExampleFiles and ABCExampleFiles). Current version is SELECTOR 1.0 released in January 2016.

Methodological background

Spatially-explicit simulations

SELECTOR allows the forward-in-time simulation of diploid individuals, generation after generation (discrete generation) within a stepping-stone framework (Kimura 1953). The total population is subdivided into demes, where each has its own spatial coordinates and can exchange migrants with its neighbouring demes at each generation. The demographic and migration algorithms in SELECTOR are derived from those implemented in the program SPLATCHE (Ray et al. 2010, model n°1). The main difference between the two is that SELECTOR is a full forward simulator and therefore allows taking into account the effect of selection on genetic lineages/alleles (balancing and positive).



0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	2	1	0	0	0
0	0	0	0	0	0	0	0	0	0	0	1	1	1	1	1	1	2	2	2
0	0	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1	2	2	2
0	0	0	0	0	0	0	0	0	1	1	1	1	1	2	1	2	2	1	0
0	0	0	0	0	0	0	0	1	1	1	1	2	2	3	2	1	1	0	0
0	0	0	0	0	0	0	0	1	1	1	2	2	2	3	1	1	1	0	0
0	0	0	0	0	0	0	1	1	2	2	3	3	3	2	2	2	2	1	0
0	0	0	0	0	0	1	1	1	2	2	3	3	3	3	2	2	2	1	0
0	0	0	0	0	0	1	1	1	2	3	3	3	3	3	2	2	1	0	0
0	0	0	0	0	1	1	1	1	3	2	2	3	3	3	3	2	1	0	0
0	0	0	0	0	1	1	1	1	2	2	2	3	3	3	2	1	0	0	0
0	0	0	0	1	1	1	1	1	2	2	2	3	3	2	1	1	0	0	0
0	0	0	1	1	1	1	1	2	2	2	2	3	3	3	1	0	0	0	0
0	0	0	1	1	1	1	1	1	2	2	2	3	2	2	1	0	0	0	0
0	0	1	1	1	1	1	1	1	2	2	3	3	2	2	1	0	0	0	0
0	0	1	1	1	1	1	1	1	2	2	3	2	2	1	1	0	0	0	0
0	1	1	1	1	1	2	2	2	2	2	2	3	2	1	1	0	0	0	0
0	1	1	1	1	1	1	2	3	2	3	2	3	1	1	0	0	0	0	0
0	1	1	1	1	1	2	2	3	2	3	3	2	1	1	0	0	0	0	0
0	1	1	1	1	1	1	2	3	3	3	3	2	1	1	0	0	0	0	0
0	1	1	1	1	1	1	2	2	3	2	2	2	1	1	0	0	0	0	0
0	1	1	1	1	1	2	2	3	3	3	2	1	2	0	0	0	0	0	0
1	1	1	1	1	1	1	3	2	2	3	2	1	2	0	0	0	0	0	0
1	1	1	1	1	1	1	1	2	3	2	1	1	1	0	0	0	0	0	0
0	1	1	1	1	1	1	2	3	3	3	1	1	0	0	0	0	0	0	0
0	0	1	1	1	1	1	2	3	3	1	1	1	0	0	0	0	0	0	0
0	0	1	1	1	1	1	2	3	2	1	1	0	0	0	0	0	0	0	0
0	0	1	1	1	1	1	2	2	2	1	0	0	0	0	0	0	0	0	0
0	0	0	1	1	1	1	2	1	1	0	0	0	0	0	0	0	0	0	0
0	0	0	1	1	1	1	2	1	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	1	1	2	1	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	1	2	1	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	2	1	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0

Example of spatially-explicit area representing Taiwan with demes of 10x10 km^2. White demes are inactivated and represent water, green demes represent plain area, light brown demes mountain below 2000m, and dark brown demes mountain above 2000m.

SELECTOR can simulate the evolution of allele frequencies in population inhabiting such a heterogeneous space in translating this map in SELECTOR input files and specific demographic/migration parameters for each category of demes.

0	Water
1	Plains
2	Mountains below 2000 m
3	Moutains above 2000 m

Figure 1. Spatially-explicit framework which may be considered in SELECTOR.

Spatially-explicit area can be designed with precision with any GIS software able to generate ASCII grid files (ArcGIS, OpenGIS, etc…), which in turn can easily be translated into SELECTOR input files (see Figure 10 as example).

Demographic model

Each deme has its own demographic characteristics and is regulated independently using a logistic model as

Equation 1

$$N_i(t) = N_i(t-1)\left(1 + r_i \frac{K_i - N_i(t-1)}{K_i}\right)$$

where *N_i(t)* and *N_i (t-1)* are the number of individuals belonging to the current deme *i* at generation *t* and *t*-1, *r_i* is the growth rate by generation (i.e. the speed at which the population density increases or decreases) and *K_i* is the carrying capacity (the maximum number of individuals which may be supported by the resources of the deme). If *r_i* is equal to 0, the population does not grow while if *r_i* is equal to 1, it doubles during the earliest phases of growth, after which growth slows down due to effect of intra-deme competition ($\frac{K_i - N_i(t-1)}{K_i}$). If *K_i* is equal to 0, then the deme can never be occupied (no growth, no colonization) and if some individuals are added to it at the beginning of the simulation, they are not involved in any migration processes.

Migration model

In every deme, the number of emigrants *E* from deme *i* at each generation *t* and in each direction *j* is computed as

Equation 2 $$E_{ij}(t) = \left\lfloor \frac{m_{ij}N_i(t)}{B_i} + z_{ij}(t-1) \right\rfloor$$

where m_{ij} is the migration rate from *i* to *j*, $N_i(t)$ is the number of individuals in the current deme *i* at generation *t*, B_i is the number of neighbours in the current deme *i* and z_{ij} is the remaining migration fractional part from previous generations (i.e. t-1). *E* always corresponds to an integer (number of individuals) and is equal amongst all neighbouring demes, so that the remaining fractional part from *i* to *j* is recorded and affects the next generation as follows

Equation 3 $$z_{ij}(t) = \left(\frac{m_{ij}N_i(t)}{B_i} + z_{ij}(t-1) \right) - E_{ij}(t)$$

For example, if $N_i(t)$ = 70 and m_{ij}=0.2 and the current deme has 4 neighbours, $E_{ij}(t)$ will be 3, $z_{ij}(t)$=2, $E_{ij}(t+1)$=4, $z_{ij}(t+1)$=0, so that 3 emigrants will be sent to each neighbour at generation *t* while 4 will be sent in each direction at generation *t*+1; and the process continues like this, generation after generation. In this case, the absolute value of the number of emigrants at demographic equilibrium is *Nm*=14, but this is an average over generations (every two generations *Nm* is equal to 12 or 16, alternatively). Note that demes with K=0 cannot be considered as targets for migration and thus cannot receive migrants.

SELECTOR permits to modify some "routes" which correspond to the migration rate between a pair of demes (see description of "*selector_param.txt*" input file). The migration rate between two demes (m_{ij} in Equation 2 and Equation 3) is simply multiplied by the modifier which acts during the whole simulation. For example, the modified route "(1,2,0.5)" in the corresponding line of the "*selector_param.txt*" file, cuts the migration rate from deme n°1 to deme n°2 by half. This feature can be used to simulate partial or complete barriers to gene flow. Note that this modification affects the total number of emigrants from the current deme and thus the *Nm* coefficient of the deme.

Demographic processes

The order of demographic events is as follow:

1. Demographic regulation occurs in every demes *i* using Equation 1 leading to $N_i(t)$
2. The number of emigrants $E_{ij}(t)$ in all directions *j* is computed in every deme *i* using Equation 2 and Equation 3 and the total number of emigrant is computed as $E_i = \sum_1^{B_i} E_{ij}(t)$
3. The number of immigrants in each deme *i* is computed as $I_{i=} \sum_1^{B_i} E_{ji}(t)$
4. The density in every deme *i* is updated with the sum of emigrants and immigrants as

Equation 4 $$N_i^{'}(t) = N_i(t) - E_i + I_i$$

Simulated genetic data

SELECTOR simulates the evolution of allele frequencies in demes through the effect of genetic drift, migration, demographic variation and selection (see Figure 3). It records the two allelic variants of one genetic locus of each simulated individual. Those variants are simply coded "a*x*" where *x* goes from 1 to n_{max}, the maximum number of allelic variant in the system (defined by the user as input parameter). At the beginning of a simulation, alleles from 1 to n_{max} are randomly distributed in the first generation of individuals. This implies that, from one simulation to the next, initial allele frequencies vary and some alleles may not be represented in the source population. The larger the initial population size is and the smaller the probability that alleles are not represented. All initial populations can either be constituted from the same or form distinct ancestral allelelic pools (see details in the description of selector_param.txt).

Transmission of gametes

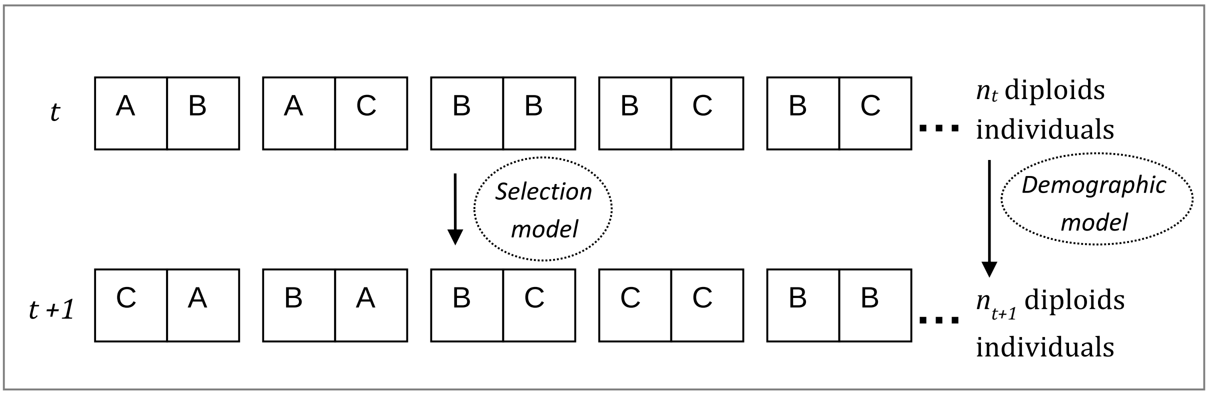


Figure 2. Schematic view of allele transmission from one generation (*t*) to the next (*t*+1).

Within each deme, the transmission of genes from one generation to the next is based on the Wright-Fisher model (Fisher 1930; Wright 1931). The genotype of each individual at generation *t* is made of two alleles drawn randomly from the genotypes of two different individuals at generation *t-1*.

Effect of Selection

When selection applies (selection coefficient $s > 0$), the transmission of gametes from one generation to the next in each deme is modified as follows:

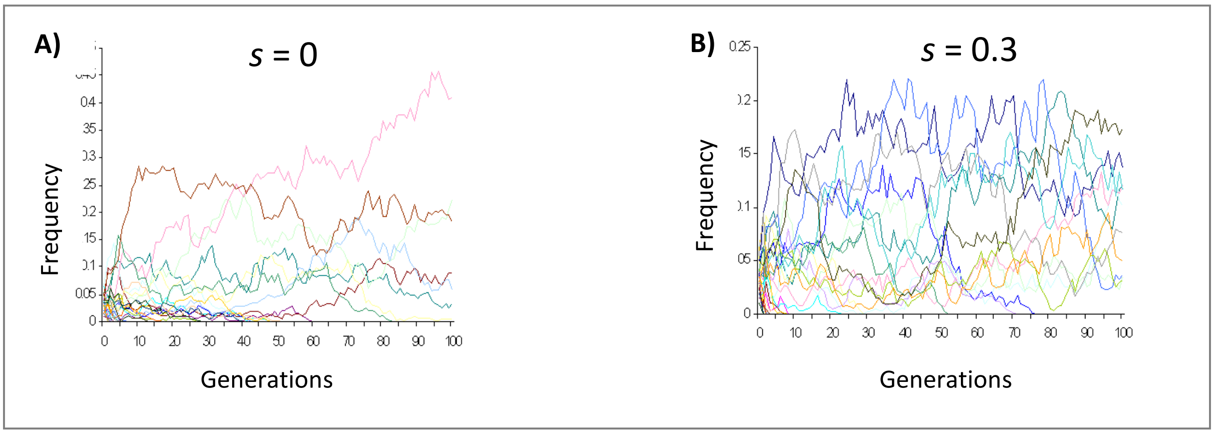


Figure 3. Example of simulated allele frequencies within a single deme. A) Neutral locus: the selection coefficient in favour of heterozygotes *s* is equal to 0. B) Symmetrical overdominant selection model (SOS) where *s* is equal to 0.3. The initial number of alleles within the population is identical in both cases. Note that overdominant selection maintains a high diversity within the population as expected theoretically.

Three different models of selection are implemented in SELECTOR:

1) Symmetrical Overdominant Selection (SOS): overdominant balancing selection with symmetrical heterozygous advantage. In that case, all heterozygotes have the same fitness (Lewontin et al. 1978). This is the basic model which represents the selective advantage for heterozygotes (Doherty and Zinkernagel 1975). Figure 3 shows an example of allele frequencies simulated with SELECTOR under this model.

- If the new genotype is heterozygote, it is accepted with probability $p = 1$
- If the genotype is homozygote, it is accepted with probability $p = 1 - s$

where *s* is the selection coefficient against homozygotes. If a genotype is rejected, another one is drawn.

2) Frequency Dependent Selection (FDS): balancing selection with rare allele advantage. It simulates a selection in favour of alleles that are the less frequent in the population.

- A new allele a_i is accepted with $p = 1 - f(a_i)*s$ where is the current frequency of allele a_i in the deme and *s* is the selection coefficient against frequent alleles. If *f* (ai) tends to 0, then the probability of keeping a_i is close to 1 while if f (a_i) tends to 0, then this probability is close to $1-s$. If the allele is rejected, another one is drawn.

3) Dominant Positive Selection (DPS): positive selection for one specific allele amongst n alleles. Individuals that carry at least one copy of a selected allele are favoured. Under this model, individuals who do not carry the selected allele have a fitness equal to 1-s while carriers have a fitness of 1 (see for example Hedrick 2000).

- If the new genotype has the selected allele (at homozygote or heterozygote state), it is accepted with probability $p = 1$.
- If the new genotype does not have the selected allele, it is accepted with probability $p = 1 - s$. If a genotype is rejected, another one is drawn.

Getting started

SELECTOR is a simple executable file, which takes as several input files (at least two) with generic names located in the same folder. No argument is needed. Two example datasets are provided: BasicExampleFiles and ABCExampleFiles.

How to run SELECTOR

MS windows

Under **Windows**, click on the “SELECTOR_win.exe” file to run SELECTOR if the required input files are present in the same folder (see “Input files” below).

Linux

Under **Linux**, first ensure that the file “SELECTOR_lin64” is executable on your machine by typing the CL “chmod +x SELECTOR_lin64” on a shell. Then type “./SELECTOR_lin64” to run SELECTOR on the current directory if the required input files (chapter Input files, page 8) are present on the same directory.

Input files

selector_param.txt [MANDATORY]

This file contains various parameters read in by SELECTOR. Each line starts with the parameter value and then continues with the parameter definition, after two consecutive slash character “//”. The order of the parameters must be kept as given below:

General settings

1/ [positive integer] //Number of simulations

Each simulation is done independently and generates one output file in format ARLEQUIN (.arp). If the parameter 12 (Population frequency output files) is set to 1 the last simulation outputs the allele trajectory (.deme).

2/ [positive integer] //Number of generations

Total number of generations to iterate the processes through.

Genetics settings

3/ [positive integer] //Number of initial alleles

Number of distinct alleles n_a in the ancestral gene pool.

4/ [0,1] //One ancestral pool of alleles (0) or one for each population group (1)

This parameter allows defining the ancestral gene pools: either all initial population sources are constituted from one single gene pool or each initial population group is derived from independent ancestral gene pools. If this parameter is set to 1 then the parameter n°3 n_a is applied to each population group and the resulting total number of initial allele $n_a max$ is equal to n_a (parameter 3) multiplied by the number of population groups (parameter 15).

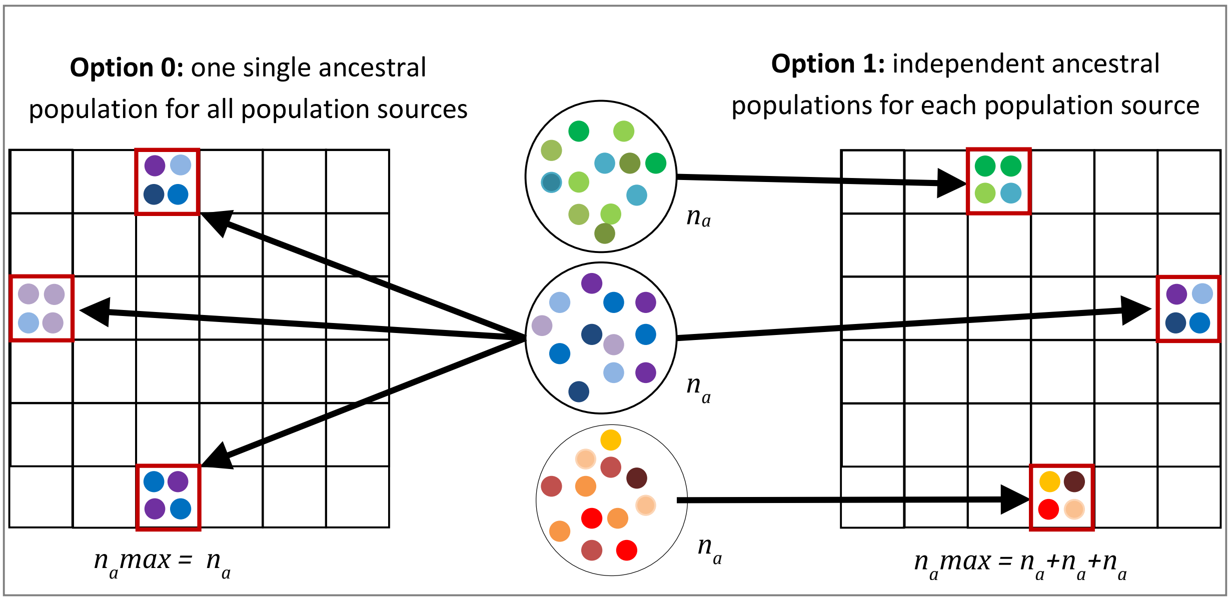

Option 0: one single ancestral population for all population sources
Option 1: independent ancestral populations for each population source
n_a
n_a
n_a
$n_a max = n_a$
$n_a max = n_a + n_a + n_a$

Figure 4. Two available options to define initial population sources. Black cells represent demes, red cells represent population sources; small coloured circles represent distinct alleles for a single locus and large black circles represent ancestral population(s).

5/ [float 0.0-1.0] //Initial frequency for allele "a1"

This parameter defines the expected initial frequency of allele "a1". If this value is equal to 0, then this parameter is not taken into account, in which case a random initial allele frequency is drawn as for all other alleles.

6/ [float] //Mutation rate per generation and per individual

This parameter defines the probability that a new allele appears (mutation rate per individual per generation), following the infinite allele model (Kimura and Crow 1964). When a new allele appears, it either takes the name of an old allele that has already disappeared (accross all demes), or it takes a new name "a*x*" (where *x* is a new number).

Selection settings

7/ [1,2,3] //Type of selection model

Type of selection model to be applied:

1. Symmetric overdominant selection (SOS, see definition above)
2. Frequency dependent selection (FDS, see definition above)
3. Dominant positive selection (DPS, see definition above)

8/ [0,1,2] //Selection heterogeneity (0: uniform, 1: latitudinal, 2: longitudinal)

This parameter defines how the intensity of selection is varied over space. 0 means uniform while 1 corresponds to a latitudinal gradient and 2 a longitudinal gradient. The gradient is set up linearly between the Northern and Southern (i.e. option 1), or Eastern and Western (option 2), with bound values defined in parameters 10 and 11 below.

9/ [float 0.0-1.0] //Selection coefficient (if uniform)

This parameter defines the value of the selection coefficient *s*. This is used when selection is uniform over space (parameter 8 above set to 0), but ignored otherwise.

10/ [float 0.0-1.0] // Northern or Eastern selection coefficient (if gradient)

This parameter is used when selection heterogeneity parameter (parameter 8) is not uniform. It defines the value of the selection coefficient *s* in the northernmost demes if the selection heterogeneity model (parameter n°8) is set to 1 or in the easternmost demes if it is set to 2.

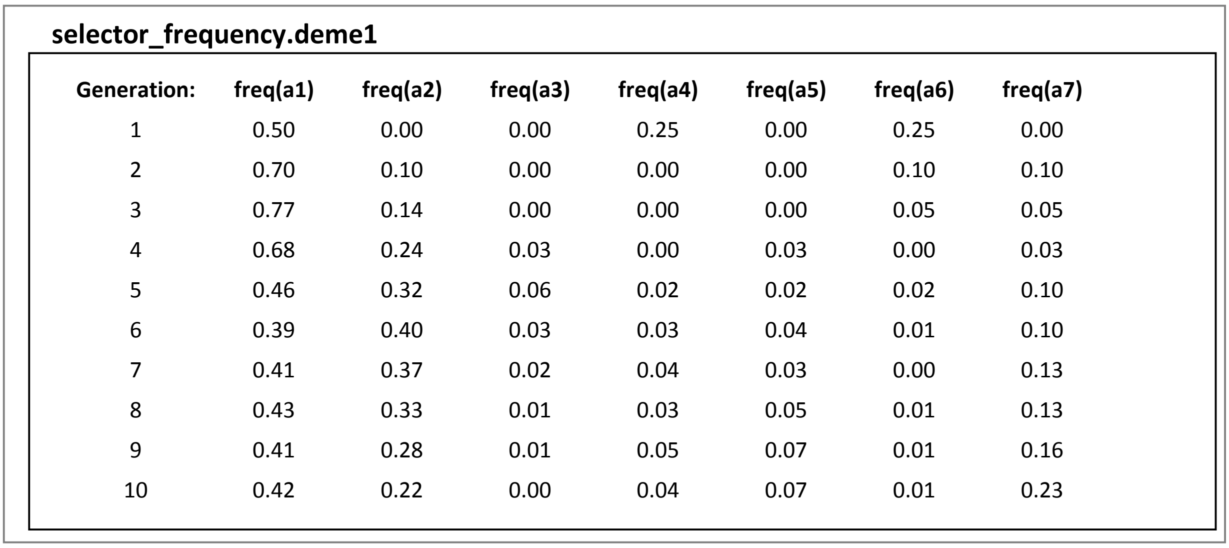
11/ [float 0.0-1.0] // Southern or Western selection coefficient (if gradient)

This parameter defines the value of the coefficient of selection *s* in the southernmost demes if the selection heterogeneity model (parameter n°8) is set to 1 or in the westernmost demes if it is set to 2.

Output settings

12/ [0,1] //Population frequency output files (1=yes, 0=no)

This setting selects if files *".deme"* containing the evolution of allele frequencies in every demes during the whole simulation have to be generated. There is one output file per deme (.deme1, .deme2, .deme3, etc…) and only for the last simulation run.



selector_frequency.deme1

```
Generation:  freq(a1)  freq(a2)  freq(a3)  freq(a4)  freq(a5)  freq(a6)  freq(a7)
     1         0.50      0.00      0.00      0.25      0.00      0.25      0.00
     2         0.70      0.10      0.00      0.00      0.00      0.10      0.10
     3         0.77      0.14      0.00      0.00      0.00      0.05      0.05
     4         0.68      0.24      0.03      0.00      0.03      0.00      0.03
     5         0.46      0.32      0.06      0.02      0.02      0.02      0.10
     6         0.39      0.40      0.03      0.03      0.04      0.01      0.10
     7         0.41      0.37      0.02      0.04      0.03      0.00      0.13
     8         0.43      0.33      0.01      0.03      0.05      0.01      0.13
     9         0.41      0.28      0.01      0.05      0.07      0.01      0.16
    10         0.42      0.22      0.00      0.04      0.07      0.01      0.23
```

Figure 5. Example of output file showing the evolution of allele frequencies in deme n°1 (.deme1) during 10 generations (7 different alleles, with alleles 2, 3, 5 and 7 which were not present in the deme 1 at the beginning but entered the deme later due to migration).

Virtual World Structure

13/ [integer] // Number of demes along the X axis

Number of demes along the x axis of the modelled area (longitude).

14/ [integer] // Number of demes along the Y axis

Number of demes along the y axis of the modelled area (latitude).

15/ [integer] //Number of population groups (at least 1)

This permits to define groups of demes (groups or populations) which can then be used in two ways: (1) sampled demes belonging to the various groups are defined in the "population structure" of the ARLEQUIN output file and can thus be used to compute group statistics (e.g. AMOVA, or other GROUP_LEVEL measures). (2) When parameter 16 (Demes from the same group with identical demography) is set to 1, identical demographic characteristics (*K*, *m* and *r*) can be attributed to all demes belonging to the same group.

16/ [0,1] //Demes from the same group with identical demography (0=no, 1=yes)

When this parameter is set to 1, then the carrying capacity *K*, the growth rate *r* and the migration rate *m* in all demes are identical to the values defined in a file called *“selector_alldemes_1.txt”* and they replace the individual values set in “*selector_structure_1.txt”*. Note that the initial size, the group number and the sample size for each deme are always defined in “*selector_structure_1.txt”*, even if this parameter is set to 1. This option is particularly useful to use an ABC approach (see below) and draw one value from a prior distribution, which applies to all demes of one given group. When this parameter is set to 0, then all carrying capacities, growth rates and migration rates are set independently for each deme in “*selector_structure_1.txt”*.

17/ [integer] //Number of routes to modify

This refers to the number of pairs of demes (routes) between which the migration rate has to be modified compared to the general rules defined in the population structure file (“*selector_structure_*1.txt”). It is especially useful to set barrier to migrations between demes. There must be as many number of routes to modify as there are lines defining number of routes defined in parameter 18 (Migration routes to modify).

18/ [integer] //Migration routes to modify (source, target, new relative rate)

This parameter defines the routes (pairs of demes) for which the migration rate is modified compared to the general rule defined in the population structure file (“*selector_structure_*1.txt”). Each route is defined by parenthesis containing three numbers separated by commas, as follow: (*demefrom*, *demeto*, *m*) where *demefrom* is the index number of the source deme, *demeto* is the index number of the target deme and *m* is migration rate modifier (e.g. 0.0 means no migration, 0.5 half the migration rate that has been defined in the population structure file and 1.0 means no modification). Each bracket represent a "route" and they must all be on one line. Here the two routes are both directions between the same pair of demes (from deme 5 to deme 1 and from deme 5 to deme 1).

```
2               //Number of routes to modified (positive integer number)
(5,1,0.5) (1,5,0.5)      //List of routes to be modified (deme1,deme2,relativemigrationrate)
```

Figure 6. This example shows that the migration rate between deme 1 and 5 is cut by half, in both direction (from deme 1 to deme 5 and from deme 5 to deme 1).

19/ [integer] //Number of different population structure to load

It is possible to define one (mandatory) or several population structures to be loaded at different times (the first structure is mandatory and must be loaded at generation 0). There must be one line per population structure, with two columns each, the first one specifying the index number of the population structure events (1,2,3,.. to z if there are z structure events to occur) and the second one specifying the generation time at which the population structure has to be loaded.

```
1           //Number of different population structure to load
1           0
```

Figure 7. This example shows the minimum population structure which must be defined, which means that the population structure "selector_structure_1".txt must be loaded at generation 0 (at the beginning of the simulation).

```
3           //Number of different population structure to load
1           0
2           100
3           400
```

Figure 8. In this example, three population structures are loaded, the first one at generation 0, the second at generation 100 and the third at generation 400.

selector_param.txt

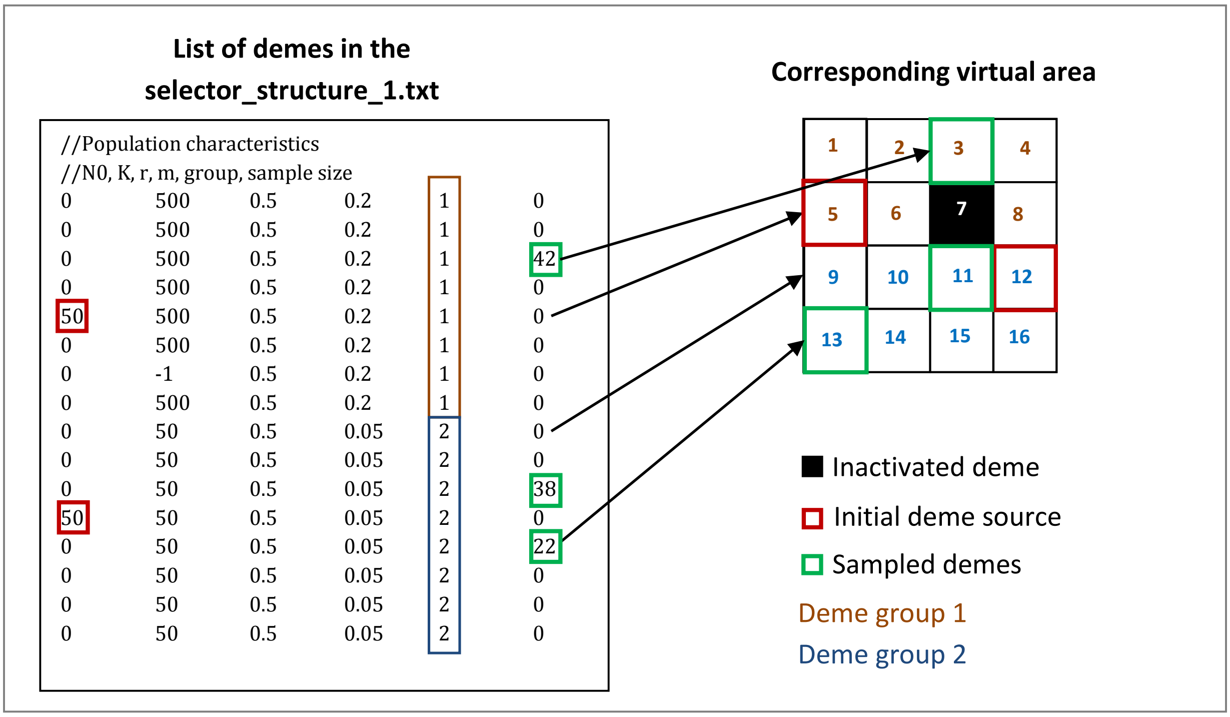
```
//GENERAL SETTINGS
1           //Number of simulations (positive integer number)
400         //Number of generations (positive integer number)
//GENETIC SETTINGS
20          //Number of initial alleles (positive integer number)
0           //One unique pool of initial alleles (0) or one independent pool for each group of population (1)
0.0         //Initial frequency for allele no1 ([0.0-1.0]; if 0 not taken into account)
0.0         //Mutation rate per generation and per individual (positive floating number)
//SELECTION SETTINGS
1           //Type of selection model (1=SOS, 2=PS)
0           //Selection heterogeneity model (0 uniform, 1 latitudinal gradient, 2 longitudinal gradient)
0.01        //Selection coefficient, used only with homogenous selection (positive floating number [0.0-1.0])
0.0         //Northern or Eastern selection coefficient, used only with gradient (positive floating number [0.0-1.0])
0.0         // Southern or Western selection coefficient, used only with gradient (positive floating number [0.0-1.0])
//OUTPUT SETTINGS
0           //Population Allele frequency output files (1=yes, 0=no)
//WORLD STRUCTURE AND DEMOGRAPHIC SETTINGS
4           //Number of demes along the X axis (positive integer number)
4           //Number of demes along the Y axis (positive integer number)
2           //Number of population groups (at least 1, positive integer number)
0           // Demes from the same group have identical demography (0=no, 1=yes)
2           //Number of routes to modify (positive integer number)
(5,1,0.5) (1,5,0.5)     //List of routes to be modified (deme1,deme2,relativemigrationrate)
3           //Number of different population structure to load (at least 1 0, 2 positive integer numbers)
1           0 //selector_structure_1.txt
2           100 //selector_structure_2.txt
3           300 //selector_structure_3.txt
```

Figure 9. Example of main input file "*selector_param.txt*" for SELECTOR. Together with the file "selector_structure_1.txt" presented in Figure 10, they define the simulation of two different populations with different demographic characteristics (group 1 and group 2) during 400 generations, starting with 20 different alleles distributed randomly in the two source populations and under the effect of symmetrical overdominant balancing selection with a selection coefficient of 1%.

selector_structure_1.txt [MANDATORY]

The "*selector_ structure_1.txt"* file defines the demographic characteristics of all demes as well as the sample size. The first two lines constitute the header, then each subsequent line defines one deme starting from the most northwestern one and ending with the most southeastern one (see figure 3).

For each deme, the first column represents the initial population size, the second column the carrying capacity *K*, the third column the growth rate *r*, the fourth column the migration rate *m* and the fifth column the sample size (number of individuals *n* sampled in that deme at the end of the simulation, it thus corresponds to *2n* genes in the ARLEQUIN output file). The population size, the group number and the sample size are fixed in "*selector_structure_1.txt*" once for all and cannot be modified during the simulation while *K*, *r* and *m* can be modified by subsequent *"selector_structure"* or "*selector_alldemes"* files (see below). Note that if *K* is equal to -1, the deme is inactive during the whole simulation (useful to simulate water or mountain demes for example).



List of demes in the selector_structure_1.txt

```
//Population characteristics
//N0, K, r, m, group, sample size
0     500   0.5   0.2    1    0
0     500   0.5   0.2    1    0
0     500   0.5   0.2    1    42
0     500   0.5   0.2    1    0
50    500   0.5   0.2    1    0
0     500   0.5   0.2    1    0
0     -1    0.5   0.2    1    0
0     500   0.5   0.2    1    0
0     50    0.5   0.05   2    0
0     50    0.5   0.05   2    0
0     50    0.5   0.05   2    38
50    50    0.5   0.05   2    0
0     50    0.5   0.05   2    22
0     50    0.5   0.05   2    0
0     50    0.5   0.05   2    0
0     50    0.5   0.05   2    0
```

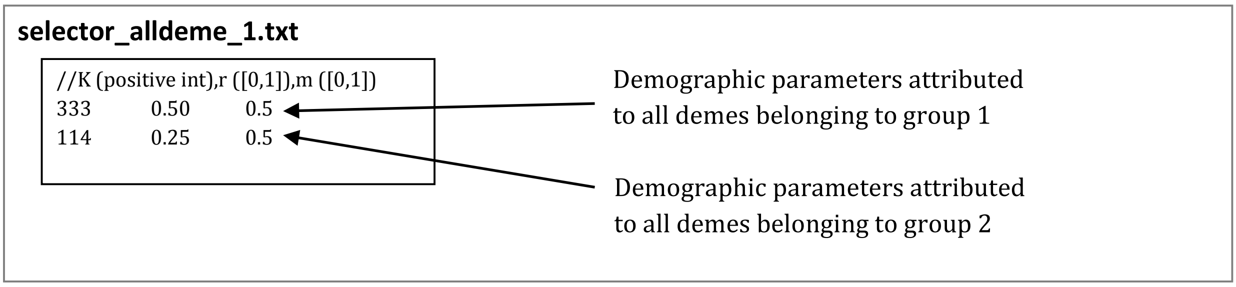
Corresponding virtual area

Figure 10. Example of a "selector_structure_x.txt" input file and how it relates to the virtual area where the simulation occurs. In that example, there are 16 demes arranged in a square area of 4 by 4 demes and divided into two groups (group 1 = demes 1 to 8; group 2, demes 9 to 16). The simulation starts with 50 individuals in deme 5 and 50 individuals in deme 12. All demes have the same growth rate (= 0.5) but demes belonging to group 2 are smaller in size and more isolated (carrying capacity=50 and migration rate=0.05) than demes in group 1 (carrying capacity=500 and migration rate=0.2). Deme 7 has been inactivated (e.g mountain). At the end of the simulation, three samples of various sizes are taken in demes 3 (*n*=41), 11 (*n*=38) and 13 (*n*=22).

selector_alldemes_1.txt [OPTIONAL]

The files "*selector_alldemes_1.txt*" is used to replace the demographic parameters defined in "*selector_structure_x.txt*" when parameter 17 (Demes from the same group have identical demography) is set to 1. In this case, all the demographic parameters *K*, *r* and *m* are identical for all the demes belonging to the same population group (defined in "*selector_structure_1.txt*"). In "*selector_alldemes_1.txt*", the first line contains the legend, then there is one line per population group with three columns each. The first column is the new value for the carrying capacity *K*, the second one the growth rate *r* and the third one, the migration rate *m*. If there are as many groups as there are demes, then there is one line per deme.

Using "*selector_alldeme_1.txt*" instead of "*selector_structure_1.txt*" is particularly useful when performing simulation within an ABC estimation framework (see below) because it is then possible to draw one value of any demographic parameters (e.g. *K*) which applies to all demes of the same group, instead of drawing one different value for each deme.



selector_alldeme_1.txt

```
//K (positive int),r ([0,1]),m ([0,1])
333       0.50      0.5
114       0.25      0.5
```

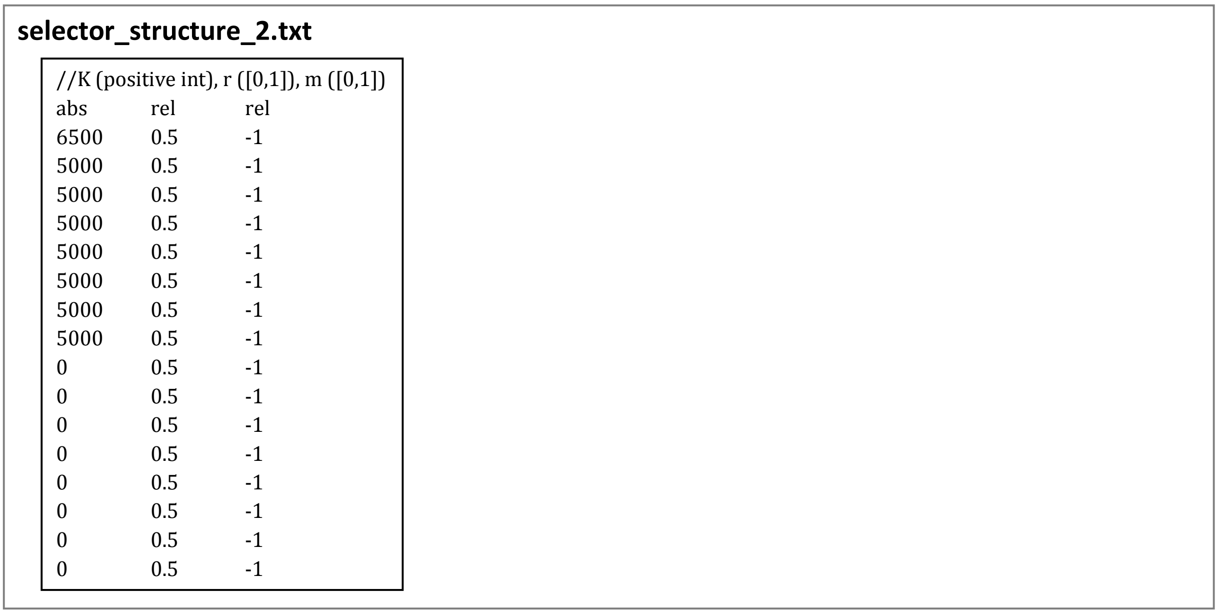
Demographic parameters attributed to all demes belonging to group 1

Demographic parameters attributed to all demes belonging to group 2

Figure 11. Example of selector_alldeme_1.txt file which replaces the demographic values of all demes specified in selector_structure_1.txt. Each line corresponds to demes of one population group, in numerical order.

selector_structure_x.txt [OPTIONAL]

In SELECTOR, it is possible to change dynamically the demographic parameters of all demes during the course of the simulation. It is similar to the "historical events" in the program SIMCOAL (Excoffier and Foll 2011). If several "*selector_structure_n.txt"* files are defined and parameter 17 is set to 0 (Demes from the same group have identical demography), the demographic parameters of demes are replaced at the generation specified under parameter 19 (Number of different population structure to load) by the new ones defined in the corresponding *selector_structure_n.txt* file. In the example of Figure 9, the file "selector*_structure_2.txt"* is loaded at generation 100 and the file "*selector_structure_2.txt"* is loaded at generation 300.



selector_structure_2.txt

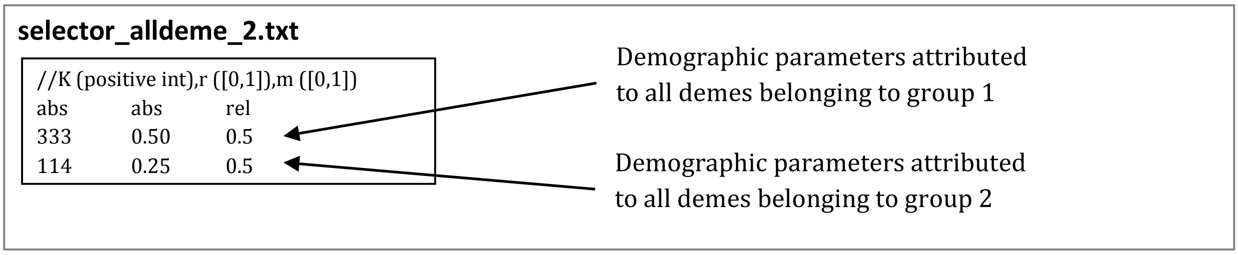
```
//K (positive int), r ([0,1]), m ([0,1])
abs       rel       rel
6500      0.5       -1
5000      0.5       -1
5000      0.5       -1
5000      0.5       -1
5000      0.5       -1
5000      0.5       -1
5000      0.5       -1
5000      0.5       -1
0         0.5       -1
0         0.5       -1
0         0.5       -1
0         0.5       -1
0         0.5       -1
0         0.5       -1
0         0.5       -1
0         0.5       -1
```

Figure 12. Example of "*selector_structure_2.txt*" input file, which modify the demographic values. Here carrying capacities of all demes are changed to a new value, while the growth rate is cut by half and the migration rate remains unchanged. When parameter 17 (all demes have the same characteristics than the first one) is set to 1, then all the values of this file are automatically replaced in SELECTOR by the values of the first deme.

As shown on Figure 12, the first line of the file is the legend, the second line specifies if each parameter has to be modified in a relative (*rel*) or absolute way (*abs*). If relative, then the previous value of the parameter is multiplied by the value indicated in the file, if absolute, the previous value is replaced by the new one. A "-1" means that the previous value remains unchanged, independently of *abs* or *rel*.

selector_alldemes_x.txt [OPTIONAL]

The files "*selector_alldemes_x.txt*" are used instead of the files "*selector_structure_x.txt*" when parameter 17 is set to 1 (Demes from the same group have identical demography). In this case, all the demographic parameters *K*, *r* and *m* are identical for all the demes belonging to the same population group (as defined in "*selector_structure_1.txt*"). In "*selector_alldemes_x.txt*", the first line contains the legend and then, every following lines contains three columns for the three demographic parameters: the carrying capacity *K*, the growth rate *r* and the migration rate *m*. The second line contains keywords "rel" or "abs" in order to state if the new values are absolute or relative to the previous ones.



selector_alldeme_2.txt

```
//K (positive int),r ([0,1]),m ([0,1])
abs      abs      rel
333      0.50     0.5
114      0.25     0.5
```

Demographic parameters attributed to all demes belonging to group 1
Demographic parameters attributed to all demes belonging to group 2

Figure 13. Example of "*selector_alldemes_2.txt*" input file, which modify the demographic values of all demes belonging to identical groups (one line per group). Here carrying capacities of all demes are changed.

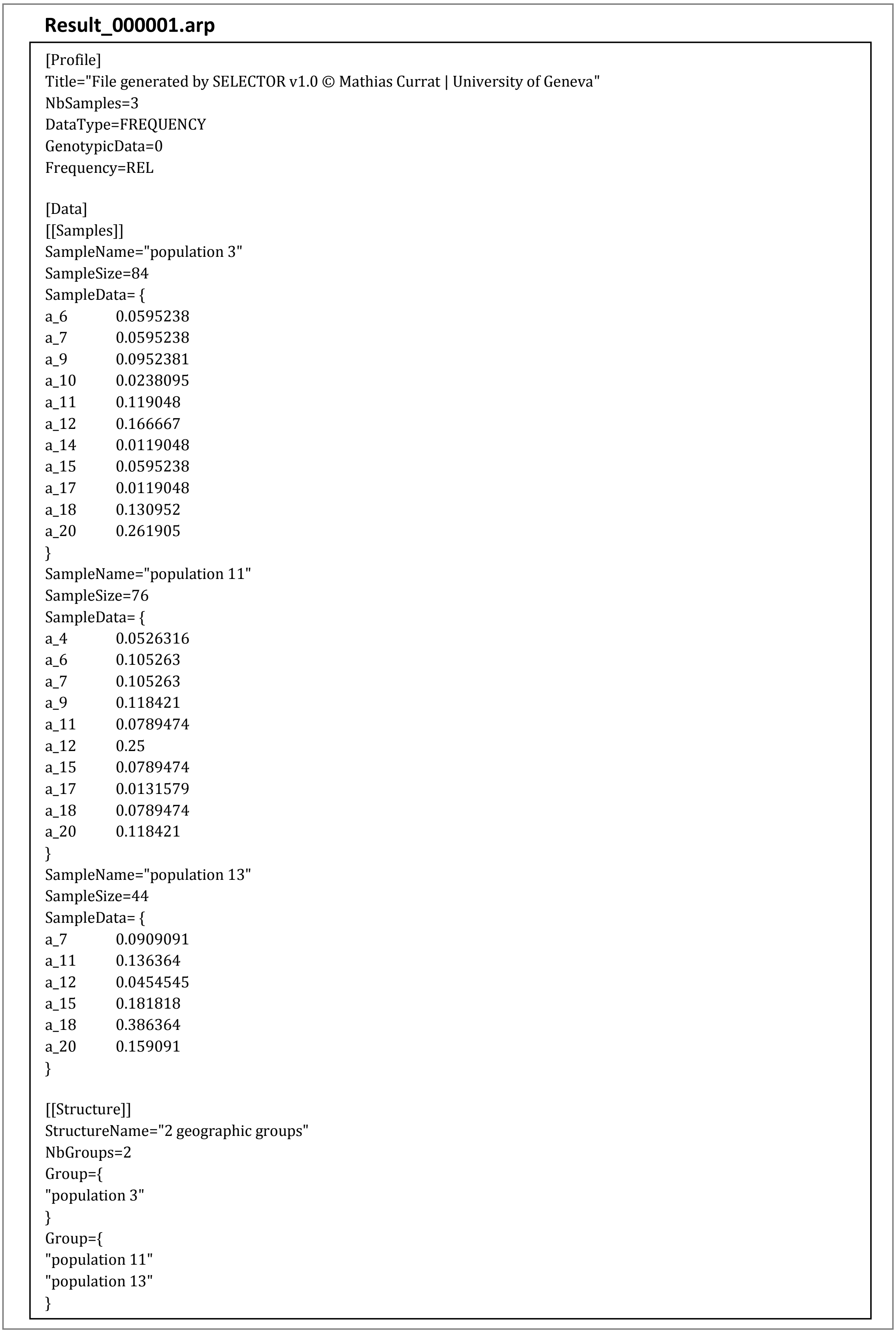
Output files

Selector log file

A file called "*selector.log*" is generated after each run of SELECTOR. It contains information about the program execution and is very helpful to identify problems with the input parameters or to check that the scenario modelled is the one the user wishes to simulate.

Arlequin files

Final allele frequencies in every sampled demes are outputted in a file called "*Result_n.arp*" where *n* is the number of simulations. This file is in a format directly compatible with the program ARLEQUIN (Excoffier and Lischer, 2010). Each sample is listed with its name, its size (in number of gene copy) and the list of present alleles and their respective frequency within the deme (see Figure 14). Allele names are simply given as "*ax*" with *x* being numbers, from 1 to the total number of alleles (*na*, parameter 3). In addition, an ARLEQUIN batch file "*arlequin.arb*" is also created and can be used with ARLEQUIN (see ARLEQUIN user manual).



Result_000001.arp

```
[Profile]
Title="File generated by SELECTOR v1.0 © Mathias Currat | University of Geneva"
NbSamples=3
DataType=FREQUENCY
GenotypicData=0
Frequency=REL

[Data]
[[Samples]]
SampleName="population 3"
SampleSize=84
SampleData= {
a_6       0.0595238
a_7       0.0595238
a_9       0.0952381
a_10      0.0238095
a_11      0.119048
a_12      0.166667
a_14      0.0119048
a_15      0.0595238
a_17      0.0119048
a_18      0.130952
a_20      0.261905
}
SampleName="population 11"
SampleSize=76
SampleData= {
a_4       0.0526316
a_6       0.105263
a_7       0.105263
a_9       0.118421
a_11      0.0789474
a_12      0.25
a_15      0.0789474
a_17      0.0131579
a_18      0.0789474
a_20      0.118421
}
SampleName="population 13"
SampleSize=44
SampleData= {
a_7       0.0909091
a_11      0.136364
a_12      0.0454545
a_15      0.181818
a_18      0.386364
a_20      0.159091
}

[[Structure]]
StructureName="2 geographic groups"
NbGroups=2
Group={
"population 3"
}
Group={
"population 11"
"population 13"
}
```

Figure 14. Example of result files "Result_000001.arp" for the first simulation.

ABC mode

Selector has been designed to be easily incorporated into an ABC estimation framework such as the ABCtoolbox (Wegmann et al., 2010). Many parameter values can be drawn from prior distributions using the following definition (*DISTRIBUTION/PARAMETER1/PARAMETER2) where DISTRIBUTION is the shape of the distribution, and PARAMETER1 and PARAMETER2 the parameters associated to it, as listed in Table 1.

Distribution Keyword	**Distribution parameters**
UNIFORM	2 parameters: minimum and maximum bounds
LOG_UNIFORM	2 parameters: minimum and maximumbounds
UNIFORM_DISCRETE	2 parameters: minimum and maximumbounds
POISSON	1 parameter: λ (positive real number)
BINOMIAL	2 parameters: probability *p* [0.0-1.0], *n* experiments (positive integer)

Table 1 List of distributions and parameters.

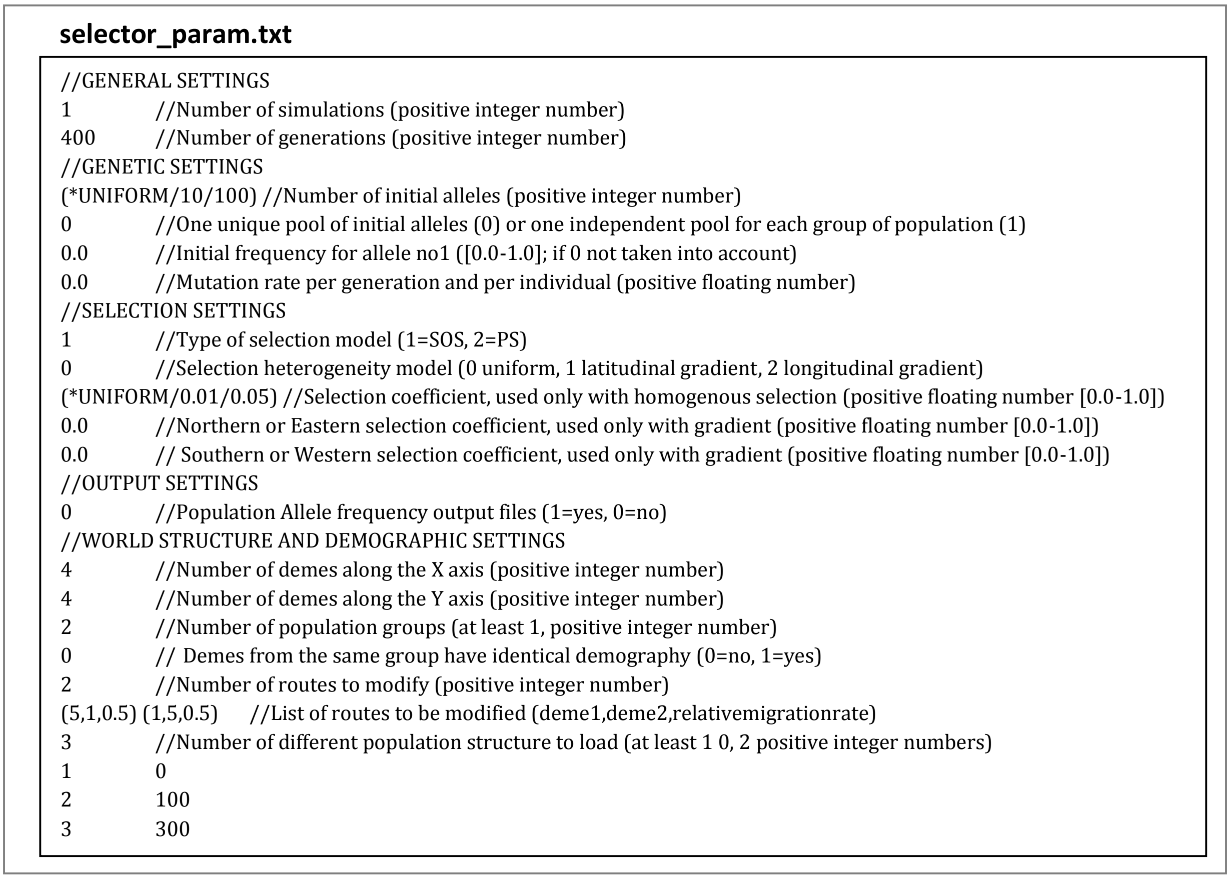
Note that only the three main input files may contain prior distributions: *“selector_param.txt”*, *“selector_structure_1.txt”* and “*selector_alldemes_1.txt*”. All supplementary input files such as *“selector_structure_2.txt”*, *“selector_structure_3.txt”*, etc… and *“selector_alldemes_2.txt”*, *“selector_alldemes_3.txt”*, etc… cannot contain prior distributions.

Input File	**Parameter name**	**SELECTOR ABC code**
param	Number of generations	NUM_GENERATIONS
param	Number of initial alleles	NUM_ALLELES
param	Initial frequency for allele a1	INIT_FREQ
param	Mutation rate	MUTATION_RATE
param	Uniform selection coefficient	SELECTION_RATE
param	North/East selection coefficient	SELECTION_RATE_NORTH_EAST
param	South/West selection coefficient	SELECTION_RATE_SOUTH_WEST
structure_1	Deme initial size	INIT_K_DEME_*n*
structure_1	Deme Carrying capacity K	FINAL_K_ DEME_*n*
structure_1	Deme growth rate r	GROWTH_RATE_DEME_*n*
structure_1	Deme migration rate m	MIG_RATE_DEME_*n*
alldemes_1	Deme Carrying capacity K	FINAL_K_ GROUP_*n*
alldemes_1	Deme growth rate r	GROWTH_ GROUP_DEME_*n*
alldemes_1	Deme migration rate m	MIG_RATE_ GROUP_*n*
-	Sampled Deme Colonization time	COL_TIME_DEME_*n*

Table 2 List of parameters that can be drawn from a prior distribution. *n* corresponds to the deme index number.

Table 2 provides a list of the parameters that can be drawn from prior distributions and the input file to which they belong. In addition, when the ABC mode is activated (which is the case if at least one parameter is drawn from a prior distribution), the colonization time of all demes where genetic sampled are taken are given in the output file "*selector_abc.txt*" with the legend COL_TIME_DEME_*n*, where *n* is the deme number. If a deme has not been colonized, then a "-1" is returned.

For example, in order to use the same files than in the example above (Figure **9** and Figure **10**) but with a number of initial alleles ranging from 10 to 100 and a selection coefficient ranging from 1% to 5%, the file "selector_param.txt" should be as follows:

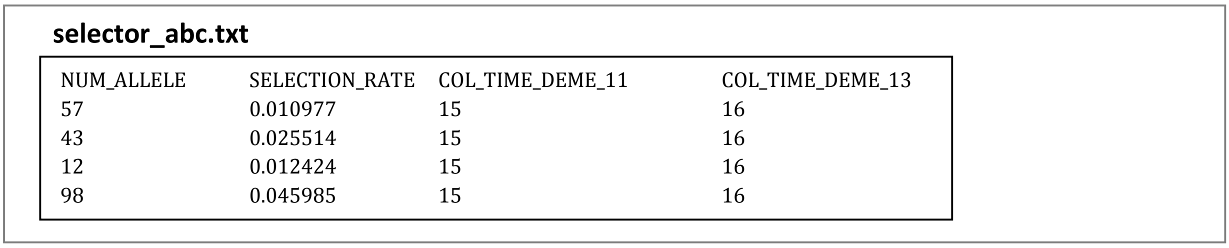


selector_param.txt

```
//GENERAL SETTINGS
1           //Number of simulations (positive integer number)
400         //Number of generations (positive integer number)
//GENETIC SETTINGS
(*UNIFORM/10/100) //Number of initial alleles (positive integer number)
0           //One unique pool of initial alleles (0) or one independent pool for each group of population (1)
0.0         //Initial frequency for allele no1 ([0.0-1.0]; if 0 not taken into account)
0.0         //Mutation rate per generation and per individual (positive floating number)
//SELECTION SETTINGS
1           //Type of selection model (1=SOS, 2=PS)
0           //Selection heterogeneity model (0 uniform, 1 latitudinal gradient, 2 longitudinal gradient)
(*UNIFORM/0.01/0.05) //Selection coefficient, used only with homogenous selection (positive floating number [0.0-1.0])
0.0         //Northern or Eastern selection coefficient, used only with gradient (positive floating number [0.0-1.0])
0.0         // Southern or Western selection coefficient, used only with gradient (positive floating number [0.0-1.0])
//OUTPUT SETTINGS
0           //Population Allele frequency output files (1=yes, 0=no)
//WORLD STRUCTURE AND DEMOGRAPHIC SETTINGS
4           //Number of demes along the X axis (positive integer number)
4           //Number of demes along the Y axis (positive integer number)
2           //Number of population groups (at least 1, positive integer number)
0           // Demes from the same group have identical demography (0=no, 1=yes)
2           //Number of routes to modify (positive integer number)
(5,1,0.5) (1,5,0.5)      //List of routes to be modified (deme1,deme2,relativemigrationrate)
3           //Number of different population structure to load (at least 1 0, 2 positive integer numbers)
1           0
2           100
3           300
```

Figure 15. Example of main input file "*selector_param.txt*" for SELECTOR with two parameters drawn from a uniform prior distribution (Number of initial alleles and selection coefficient).

If some parameters are defined as prior distributions, then a new output file is produced with the name "*selector_abc.txt*". This file contains a header line and then one line per simulation with the values of the parameters drawn from the distributions. Here is an example of such a file obtained with the parameter file above.



selector_abc.txt

```
NUM_ALLELE      SELECTION_RATE   COL_TIME_DEME_11       COL_TIME_DEME_13
57              0.010977         15                     16
43              0.025514         15                     16
12              0.012424         15                     16
98              0.045985         15                     16
```

Figure 16. Example of "selector_abc.txt" output file when two parameters have been drawn from a uniform prior distribution and four simulations (see Figure 15). In this example, deme 11 and deme 13 are always colonized at generations 15 and 16, respectively.

Acknowledgments

Alicia Sanchez-Mazas, Pascale Gerbault and José Manuel Nunes contributed to the development of SELECTOR in various ways: design, writing parts of the code, algorithms, suggestions, etc... I also would like to warmly thank Da Di who contributed to SELECTOR by extensively testing it, reporting bugs and proposing new developments or improvements.

The development of SELECTOR has been financially supported by the Swiss National Science Foundation (SNSF, Switzerland) grants no. 31003A-144180 and 31003A_156853.

References

Beaumont MA, Zhang W, and Balding DJ. 2002. Approximate Bayesian Computation in Population Genetics. Genetics 162(4):2025-2035.

Doherty PC, and Zinkernagel RM. 1975. A biological role for the major histocompatibility antigens. Lancet 1(7922):1406-1409.

Excoffier L, and Foll M. 2011. fastsimcoal: a continuous-time coalescent simulator of genomic diversity under arbitrarily complex evolutionary scenarios. Bioinformatics 27(9):1332-1334.

Excoffier L, and Lischer HE. 2010. Arlequin suite ver 3.5: a new series of programs to perform population genetics analyses under Linux and Windows. Mol Ecol Resour 10(3):564-567.

Fisher RA. 1930. The Genetical Theory Of Natural Selection: At The Clarendon Press.

Hedrick PW. 2000. Genetics of Populations. Sudbury, Massachussets: Jones and Bartlett. 553 p.

Kimura M. 1953. "Stepping-stone" model of population. Annual Report of National Institute of Genetics 3:62-63.

Kimura M, and Crow JF. 1964. The Number of Alleles That Can Be Maintained in a Finite Population. Genetics 49:725-738.

Lewontin R, Ginzburg LR, and Tuljapurkar SD. 1978. Heterozis as an explanation for large amounts of genic polymorphism. Genetics:149-170.

Ray N, Currat M, Foll M, and Excoffier L. 2010. SPLATCHE2: a spatially-explicit simulation framework for complex demography, genetic admixture and recombination. Bioinformatics.

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

Wright S. 1931. Evolution in Mendelian Populations. Genetics 16(2):97-159.