SPLATCHE 3

USER MANUAL

2019

Authors:

Mathias Currat [1,2], Miguel Arenas [3,4], Claudio S. Quilodran [1], Laurent Excoffier [5,6], Nicolas Ray [7,8]

1- Laboratory of anthropology, genetics and peopling history, Department of Genetics and Evolution – Anthropology Unit, University of Geneva, 1205 Geneva, Switzerland.

2- Institute of Genetics and Genomics of Geneva (IGE3), University of Geneva, 1205 Geneva, Switzerland.

3- Department of Biochemistry, Genetics and Immunology, University of Vigo, 36310 Vigo, Spain.

4- Biomedical Research Center (CINBIO), University of Vigo, 36310 Vigo, Spain.

5- Computational and Molecular Population Genetics Lab, Institute of Ecology and Evolution, University of Bern, 3012 Bern, Switzerland.

6- Swiss Institute of Bioinformatics, 1015 Lausanne, Switzerland.

7- Institute of Global Health, GeoHealth group, University of Geneva, 1205 Geneva, Switzerland

8- Institute for Environmental Sciences, University of Geneva, 1205 Geneva, Switzerland.

Acknowledgments

We would like to thank all the people that contributed in a way or the other to the various versions of SPLATCHE.

Thanks to all users who sent us feedbacks, suggestions or who reported bugs.

More specifically and in no particular order, thanks to the following people who contributed to the developments of the options implemented in SPLATCHE3:

Stefan Schneider, Pierre Berthier, Daniel Wegmann, Seraina Klopfstein, Guillaume Laval, Samuel Neuenschwander, Grant Hamilton, Stefano Mona, Anna Sramkova, Matthieu Foll, Nuno Silva, Da Di, Thomas Goeury, Jérémy Rio, Nicolas Broccard, Stephan Peischl, Isabel Alves, Jason Brown and Joe Wakano.

We also would like to acknowledge the following people for their technical help: Jose Manuel Nunes, Stephan Weber and David Roessli.

The development of the various versions of the SPLATCHE program was possible through the following grants:

- Swiss NSF grant No 31-054059.98 to Laurent Excoffier;
- Swiss NSF grant No 3100A0-100800 to Laurent Excoffier;
- EU/FP7 grant No 226852 (SCALES);
- EU Marie Curie initial training network, Grant/Award No BEAN (GA 289966);
- Swiss NSF grant No 31003A-156853 to Mathias Currat.
- Spanish grants RYC-2015-18241 and ED431F 2018/08 to Miguel Arenas.
- Swiss NSF grant No 31003A_182577 to Mathias Currat.

Table of contents

Table of Figures

1 Introduction

The goal of this user manual is to describe the technical aspects of the current version 3.0 of the program SPLATCHE (for SPatiaL And Temporal Coalescences in Heterogeneous Environments). SPLATCHE3 is derived from the second version of SPLATCHE that has been released in 2010 (Ray, Currat et al. 2010), the first version being released in 2004 (Currat, Ray et al. 2004). This coalescent-based program has been designed to model complex demographies, such as range expansions, in a heterogeneous and varying environment, and to simulate the resulting molecular diversity of samples of genes. Its objective is to allow the use of realistic environmental models to study the impact of ecological factors on the genetic structure of populations. This approach has already proved valuable to study the patterns of genetic variation in spatially-explicit contexts (e.g. Ray, Currat et al. 2005, Foll and Gaggiotti 2006) and especially the genetic consequences of range expansion (e.g. Ray, Currat et al. 2003, Hamilton, Currat et al. 2005, Currat, Excoffier et al. 2006, Klopfstein, Currat et al. 2006, Wegmann, Currat et al. 2006). More than 200 articles have cited SPLATCHE (see www.splatche.com).

New functionalities have been added to the previous version of SPLATCHE2 (see next section), allowing the simulation of a wider range of evolutionary scenarios. Most of these new features have already been applied independently and they have now been merged in SPLATCHE3. Minor bugs were corrected. A graphical user interface (GUI) for Windows is provided, allowing to specify the implemented models and to check them graphically. A console version running under Windows, Linux and MacOs are also provided for intensive calculations on computer clusters. All versions (GUI and console) use the same input files.

1.1 Changes from SPLATCHE version 2.0

We list below the main new functionalities of SPLATCHE3:

- Three new demographic models under the one population mode, including **long-distance dispersal** (LDD);
- Three new **models of admixture** between populations under the two populations mode;
- The simulation of **ancient genetic samples**;
- Dynamism and spatial heterogeneity of **migration rate across time and space**;
- New **mutation models of DNA sequence evolution**;
- Reset of the population size when carrying capacity is set to 0 in order to simulate **population contractions**;
- Corrections of minor bugs.

1.2 Citation and download

Mathias Currat, Miguel Arenas, Claudio Quilodran, Laurent Excoffier and Nicolas Ray: SPLATCHE3: simulation of modern and ancient genetic data under spatially explicit evolutionary models including long-distance dispersal. submitted

The last version of SPLATCHE as well as examples datasets can be freely downloaded from the web site http://www.splatche.com/splatche3

2 General approach

SPLATCHE simulations are based on a two-phase process, namely the forward (demography) and the backward (coalescent) steps. The first phase is the simulation, forward in time, of a two dimensional stepping stone (Kimura and Weiss 1964) world, which is an array of equally spaced demes (or sub populations). Each deme exchanges migrants with its four neighbours depending on three parameters: the carrying capacity (the number of individuals that can be sustained by local resources within a deme), the migration rate (the per-generation probability of an individual to move out of a deme) and the friction (the relative ease to move into a deme).

Both carrying capacities and frictions may be kept homogeneous over the whole simulated array, or they may be set to different values for each deme. This latter possibility is done by providing specific maps of corresponding values over the landscape. In this way, it is possible to take into account realistic geographic and environmental features, by varying carrying capacities of the demes according to the quality of the habitat and change friction and migration values in order to model rivers, mountains or roads. Notably, such maps can be held fixed during the whole simulation process or may change after a specified number of generations to take into account historical events, such as recent human induced fragmentation of the habitat (Quemere, Crouau-Roy et al. 2010), or cycle ages with a progressive reduction of the available landscape (Taberlet, Fumagalli et al. 1998, Hewitt 2004), and more.

2.1 Forward demographic simulations

The forward simulation starts from one (or more) user-defined deme(s), which send(s) migrants to neighbouring demes (Figure 1). Migrations to empty demes represent new colonization events. The density of occupied demes is logistically regulated, implying a local demographic expansion of newly colonized demes. Hence, at each generation, a logistic regulation step (regulated by the carrying capacity and the growth rate) is followed by a migration step (regulated by the effective number of individuals, the migration rate and the friction). The whole simulation process, which lasts for a user-defined number of generations, results in a wave of advance of the whole population, with a shape and at a speed determined by the demographic parameters and the friction values.

One important feature of SPLATCHE is that migrations occur from one deme to a neighboring one, but also in the reverse direction and especially that migrations do not stop when a deme is fully colonized. It means that migrations continue even at demographic equilibrium (when all the simulated world has been colonized).

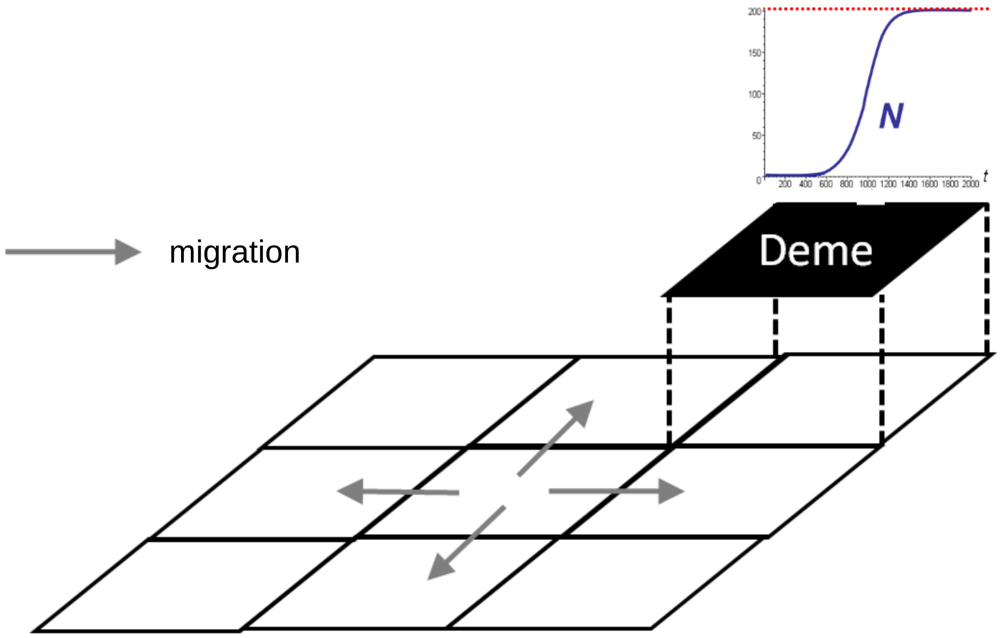


Figure 1. Schematic view of an array of demes where the simulations occur.

2.2 *Backward genetic simulations*

Each generation, the demographic and migration histories of every deme are stored in a database in computer memory (not on disk). This database is then used in the backward coalescent step (Hudson 1990) of the SPLATCHE program. This second phase of the algorithm now starts from the present generation, and proceeds backward in time (Figure 2). The effective number of individuals present in a deme is used to compute the probability of a coalescent event, and the migration rates and friction values determine the probability of each sampled genes to emigrate, backward in time, to the surrounding demes. The coalescent process stops after all genes have coalesced, most often in the deme where the range expansion started. If more than one expansion origins (source populations) have been set, additional parameters defining the migration rates among the origins prior to the expansion need to be specified.

2.3 *Sampling*

Genes can be sampled in any number and in any demes of the simulated world, and therefore at different spatial scales (within deme, within nearby demes -a patch-, or at very remote locations). The genetic markers simulated (DNA, STR and SNPs) can be outputted in text files in ARLEQUIN/ARLSUMSTAT format (Excoffier and Lischer 2010) so that the genetic diversity can be monitored at any of such scales level.

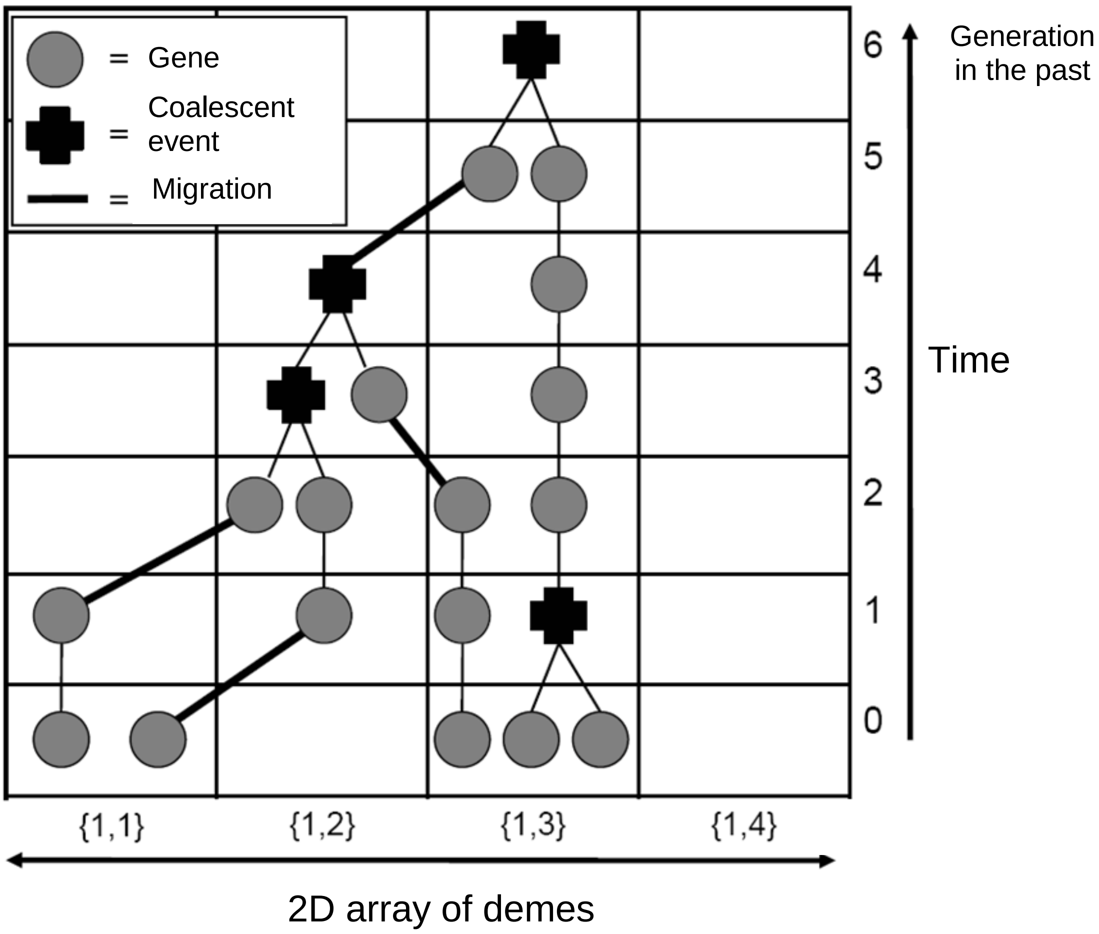


Figure 2. Schematic view of the backward coalescent process.

2.4 Simulation of two coexisting and interacting populations

Two co-existing and interacting populations can be simulated using SPLATCHE3. Technically, two layers of demes representing the virtual world are superimposed (Figure 3). Each layer represents one population. The two populations can coexist independently (without any interaction) in each layer but they can also be in competition for the resources and exchange genes by interbreeding.

Source populations can be located in different places in each layer but there are two main constraints: First, there must be at least one source for the population A at the start of the simulation (generation 0); Second, individuals constituting the source(s) of population B are taken from the corresponding deme in population A (e.g. the deme located at the same place in the layer A). This means that population B is created by a group of individuals belonging initially to population A. It can represent, for example, a speciation event.

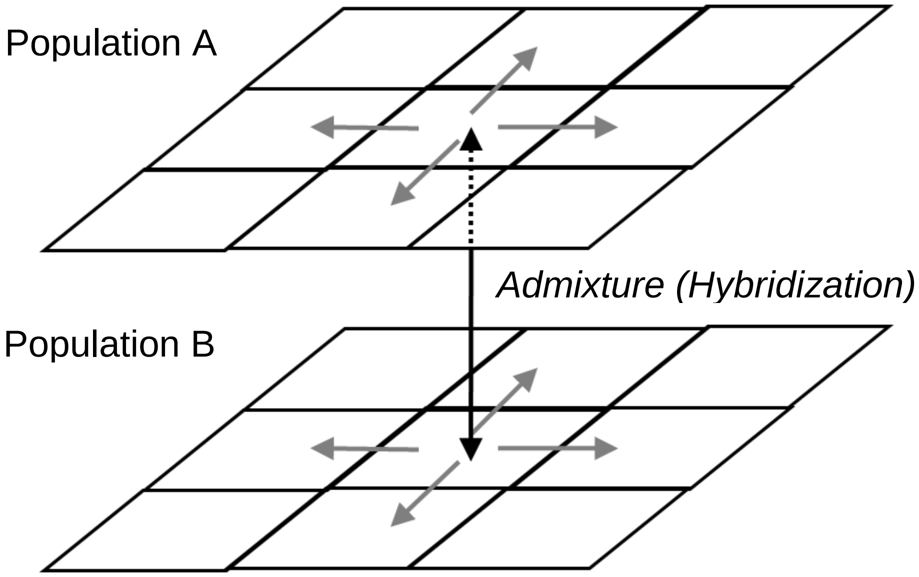


Figure 3. Schematic representation of the two superimposed layers of demes.

For instance, Figure 4 shows a schematic representation of two successive spatial and demographic expansions. Each population expansion takes place in one layer and gene flow (admixture) can occur between the two populations. The demographic parameters of each population can be set independently. It is also possible to simulate competition between the two populations, potentially leading to the extinction of one population.

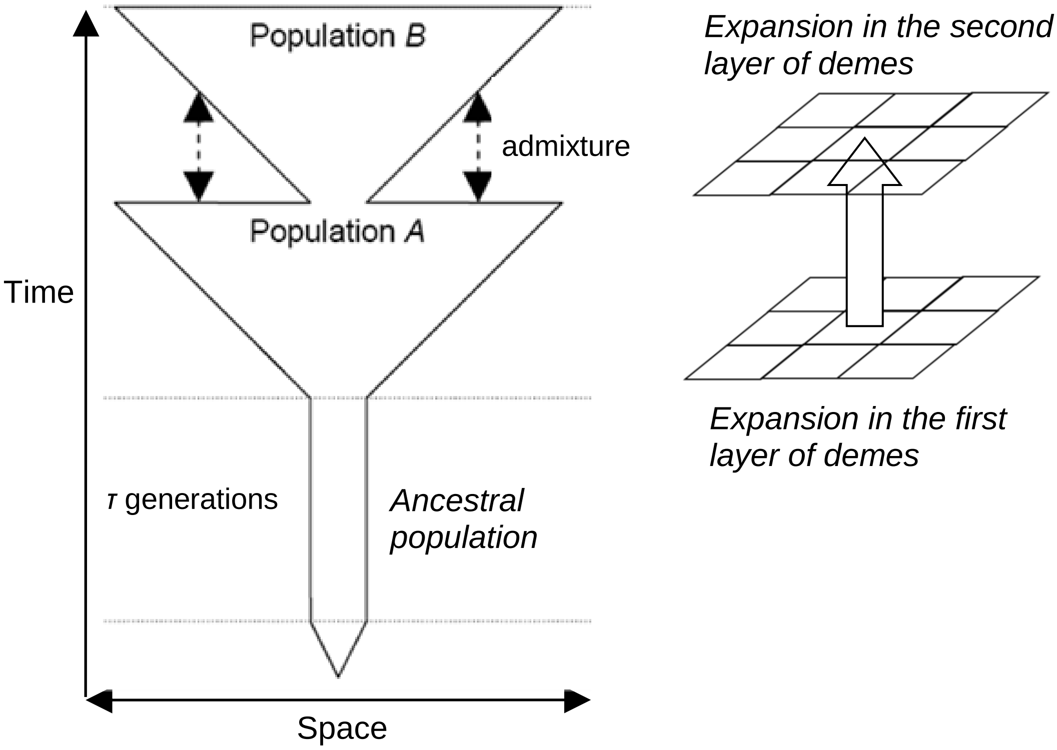

Population B
admixture
Population A
Time
τ generations
Ancestral population
Space
Expansion in the second layer of demes
Expansion in the first layer of demes

Figure 4. Schematic representation of the two successive population expansions with admixture.

2.5 Computer requirement

SPLATCHE3 does not need any minimal computer requirement. However, because demographic information for every deme is stored in computer memory, the complexity of a simulation is limited by the characteristics of the computer. RAM memory requirement is roughly proportional to the product of the number of demes by the number of generations. For example, using the setting file *settings_test_1layer-ver3.txt* distributed with SPLATCHE3, 400 MB of memory are necessary to simulate an expansion in 1,700 active demes during 1,000 generations. Roughly 1 GB of memory is necessary to simulate 5,000 demes during 10,000 generations. One way to decrease memory requirement by almost 50% is to set the parameter *AllowShortIntForNumberOfIndividuals* to 1 (see chapter 5.2). In that case, the storage of the deme densities requires less memory, but these densities must not exceed 16,000 individuals per deme at any generation during the simulation.

In order for the user to have an idea about time requirement when performing simulations with SPLATCHE3, the table below shows the time it takes for various simulations on two different computers and operating systems. We used both datasets given as example in the website (*settings_test_1layer-ver3.txt* and *settings_test_2layer-ver3.txt*, which are constituted by 1,700 active demes). We performed one demographic simulation and two genetic simulations: one with 400 haploid individuals constituted by DNA sequences of 500 bp length and a second one with 100 independent STRs in 200 diploids individuals. Note that those simulations have been computed without recombination. When simulating recombination, the time needed for a genetic simulation increases proportionally with the recombination rate. Also note that the requirement in RAM memory increases with the recombination rate and the proportion of LDD and that the program may ultimately crash if this rate is too high related to the other parameters (mainly sequence length, carrying capacity *K* and migration rate *m*).

File	Model	Generations	Simulations	Win (graphical)*	Win (console)*	Linux (console)*	Mac (console)*
settings_test_1layer-ver3.txt	1	500	Demography	<1 sec	1 sec	<1 sec	<1 sec
			DNA (500bp)	1 sec	<1 sec	<1 sec	<1 sec
			STR (100 STRs)	116 sec	2 sec	1 sec	2 sec
			SNP (10,000)	1,276 sec†	761 sec	227 sec	387 sec
			DNA (10,000x100bp)	3,112 sec†	2,374 sec	480 sec	725 sec
		1000	Demography	1 sec	1 sec	1 sec	<1 sec
			DNA (500bp)	2 sec	<1 sec	<1 sec	<1 sec
			STR (100 STRs)	237 sec	4 sec	3 sec	4 sec
			SNP (10,000)	1,554 sec†	985 sec	365	552 sec
			DNA (10,000 x100bp)	3,394 sec†	2,607 sec	616 sec	915 sec
	4	500	Demography	6 sec	16 sec	3 sec	2 sec
			DNA (500bp)	1 sec	<1 sec	<1 sec	<1 sec
			STR (100 STRs)	119 sec	2 sec	1 sec	2 sec
			SNP (10,000)	1,268 sec†	771 sec	215 sec	361 sec
			DNA (10,000 x100bp)	3,104 sec†	2,374 sec	470 sec	721 sec
		1000	Demography	16 sec	39 sec	7 sec	6 sec
			DNA (500bp)	3 sec	<1 sec	<1 sec	<1 sec
			STR (100 STRs)	242 sec	5 sec	2 sec	4 sec
			SNP (10,000)	1,588 sec†	1,008 sec	360 sec	585 sec
			DNA (10,000 x100bp)	3,419 sec†	2,516 sec	617 sec	956 sec
settings_test_2layer-ver3.txt	7	500	Demography	1 sec	1 sec	<1 sec	<1 sec
			DNA (500bp)	2 sec	<1 sec	<1 sec	<1 sec
			STR (100 STRs)	198 sec	2 sec	2 sec	3 sec
			SNP (10,000)	6,579 sec†	740 sec	208 sec	391 sec
			DNA (10,000 x100bp)	8,359 sec†	2,193 sec	423 sec	660 sec
		1000	Demography	3 sec	3 sec	2 sec	2 sec
			DNA (500bp)	3 sec	<1 sec	<1 sec	<1 sec
			STR (100 STRs)	360 sec	6 sec	3 sec	5 sec
			SNP (10,000)	11,030 sec†	1,040 sec	361 sec	684 sec
			DNA (10,000 x100bp)	12,635 sec†	2,497 sec	582 sec	884 sec
	8	500	Demography	1 sec	1 sec	1 sec	1 sec
			DNA (500bp)	2 sec	<1 sec	<1 sec	<1 sec
			STR (100 STRs)	193 sec	3 sec	1 sec	2 sec
			SNP (10,000)	6,560 sec†	727 sec	213 sec	370 sec
			DNA (10,000 x100bp)	8,274 sec†	2,183 sec	424 sec	618 sec
		1000	Demography	3 sec	2 sec	1 sec	1 sec
			DNA (500bp)	3 sec	<1 sec	<1 sec	<1 sec
			STR (100 STRs)	327 sec	6 sec	3 sec	5 sec
			SNP (10,000)	10,789 sec†	1,009 sec	378 sec	625 sec
			DNA (10,000 x100bp)	12,497 sec†	2,478 sec	587 sec	870 sec

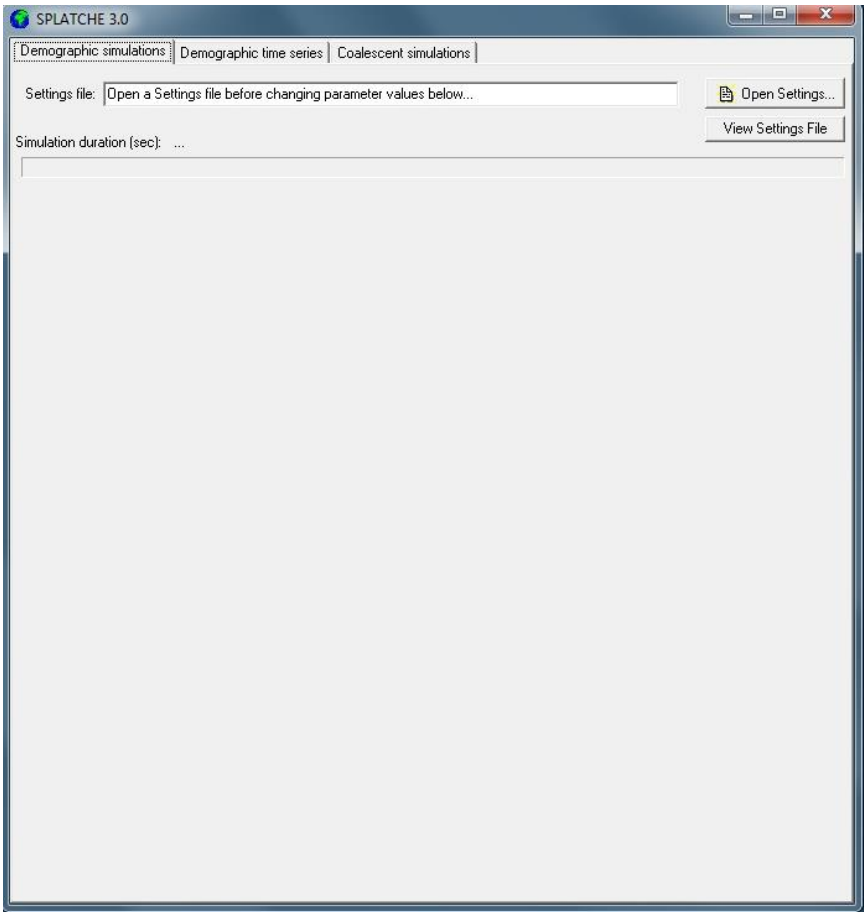
* Win: Windows 7 64 bits (3.6GHz CPU); Linux: Ubuntu 64bits (3.6GHz CPU); Mac: OS X 10.13 64 bits (3.1 GHz CPU). † Refresh rate of the coalescent simulation in the graphical version set to 1,000 in order to accelerate the simulation time (default 10).

3 Overview of the graphical user interface

3.1 Graphical user interface for Windows

The graphical user interface (GUI) for Windows allows the user to interactively modify some of the input data (e.g., carrying capacity, frictions, geographic locations of source populations and sampled populations) and to explore the output results (e.g., range expansion, coalescence process). Over the years, the authors of SPLATCHE3 have found the graphical visualization of the simulations extremely important to (*i*) check the validity of input data, (*ii*) understand the influence of input settings on demographic and genetic outputs, (*iii*) discover patterns and behaviors of expanding populations and their associated coalescent process.

When opened, the GUI looks as follows:

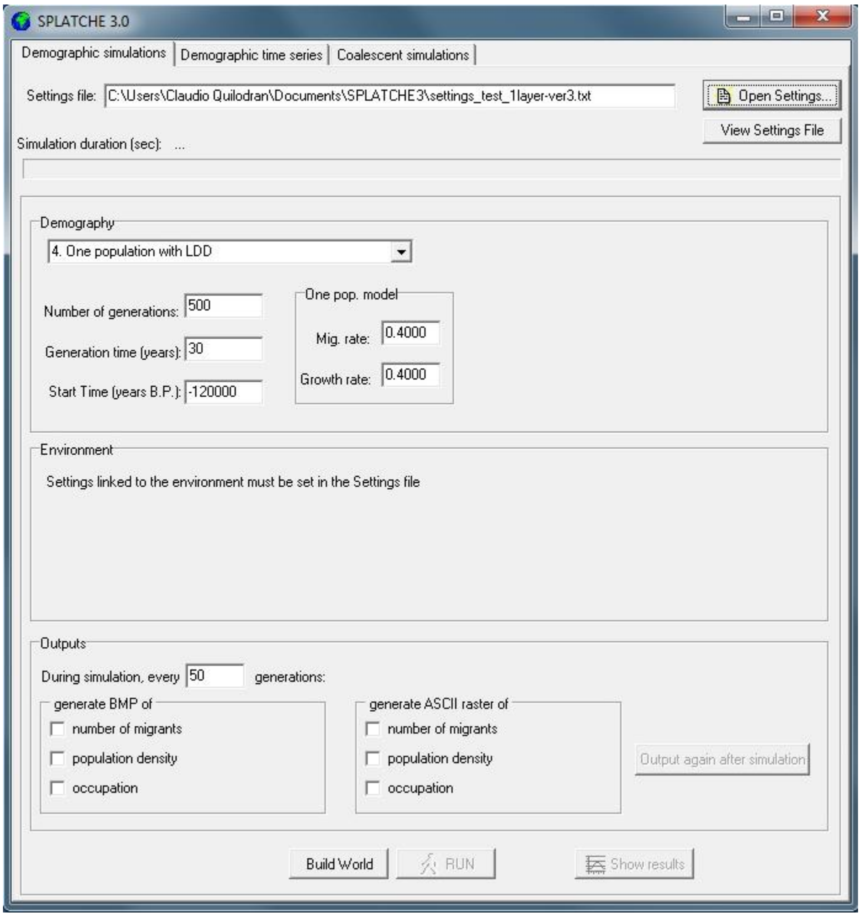


Settings file: Open a Settings file before changing parameter values below...

Simulation duration (sec): ...

Figure 5. *Opening* panel with the GUI.

The first step is to open a settings file, which will fill the corresponding settings fields in the GUI. Note that only a restricted number of parameters may be changed using the GUI, most of them can only be changed in the setting files.



Settings file: C:\Users\Claudio Quilodran\Documents\SPLATCHE3\settings_test_1layer-ver3.txt

Simulation duration (sec): ...

Demography

4. One population with LDD

Number of generations: 500

Generation time (years): 30

Start Time (years B.P.): -120000

One pop. model

Mig. rate: 0.4000

Growth rate: 0.4000

Environment

Settings linked to the environment must be set in the Settings file

Outputs

During simulation, every 50 generations:

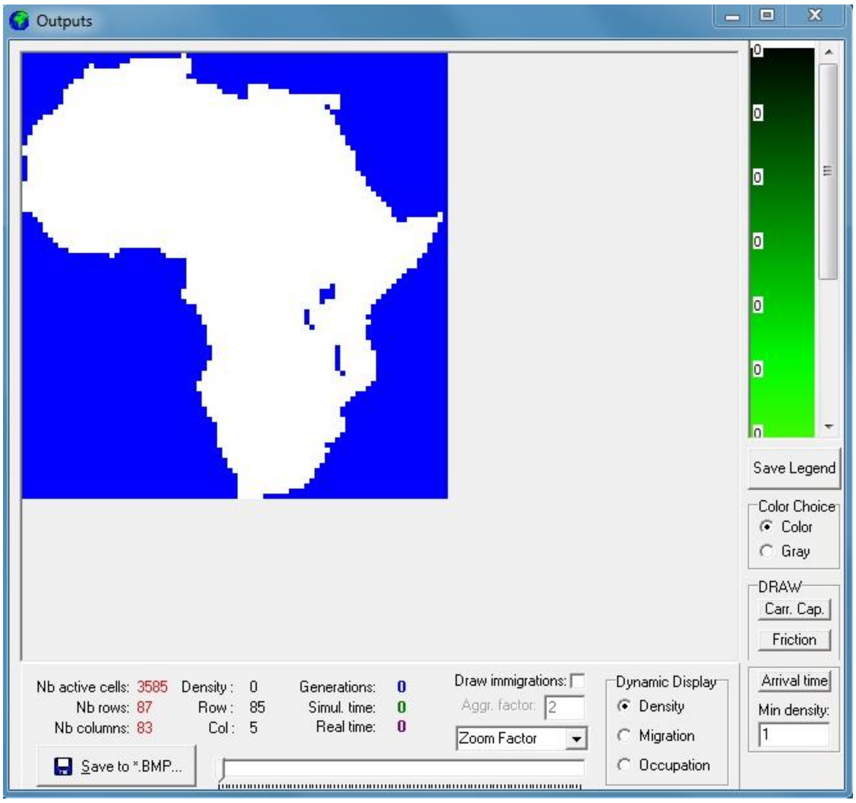
generate BMP of
- □ number of migrants
- □ population density
- □ occupation

generate ASCII raster of
- □ number of migrants
- □ population density
- □ occupation

Figure 6. ***Demographic simulation*** **panel.**

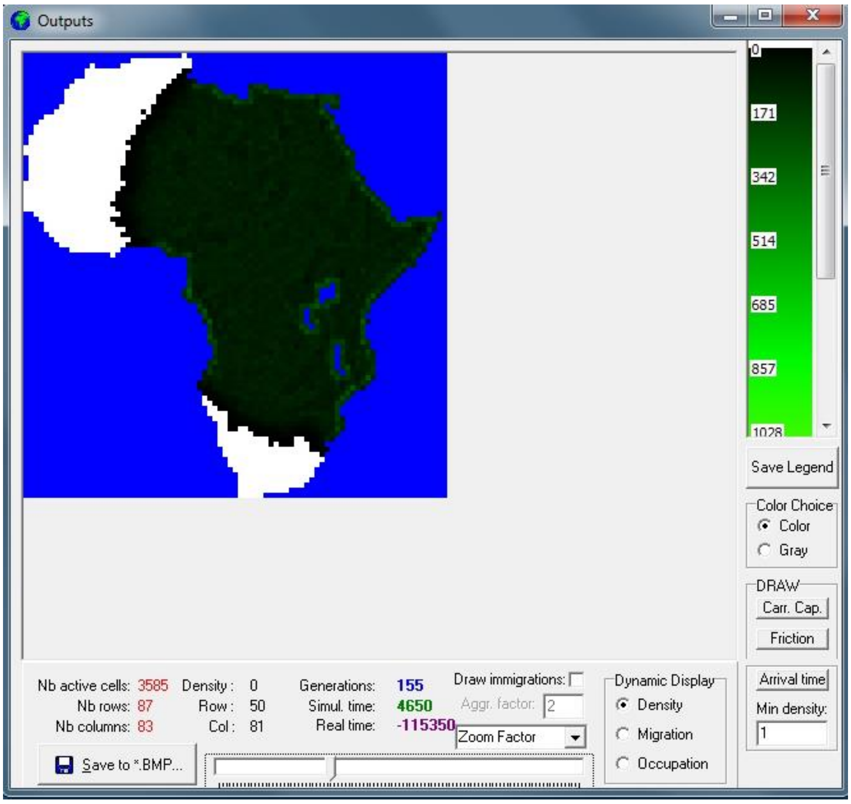
After the successful reading of a settings file, three panels are made available, but only the first one (*Demographic simulations*, Figure 6) is needed at this stage. Some settings can be changed through the interface (e.g. identification of the migration scenario), while others (e.g. carrying capacities, frictions) can only be changed directly in the settings files.

Once all simulation settings values have been chosen, the user must click on the "*Build World*" button. This button opens a new window (called *Outputs*, Figure 7), instantiates the world (i.e. the virtual 2D array of demes), allocates appropriate computer memory for the simulations, and shows the instantiated world (with sea or *NoData* demes in blue, see Figure 7).



Nb active cells: 3585
Nb rows: 87
Nb columns: 83
Density : 0
Row : 85
Col : 5
Generations: 0
Simul. time: 0
Real time: 0
Draw immigrations: □
Aggr. factor: 2
Dynamic Display: ● Density ○ Migration ○ Occupation
Color Choice: ● Color ○ Gray
DRAW
Min density: 1

Figure 7. *Outputs* windows after the "build world" step and before the forward simulation step.



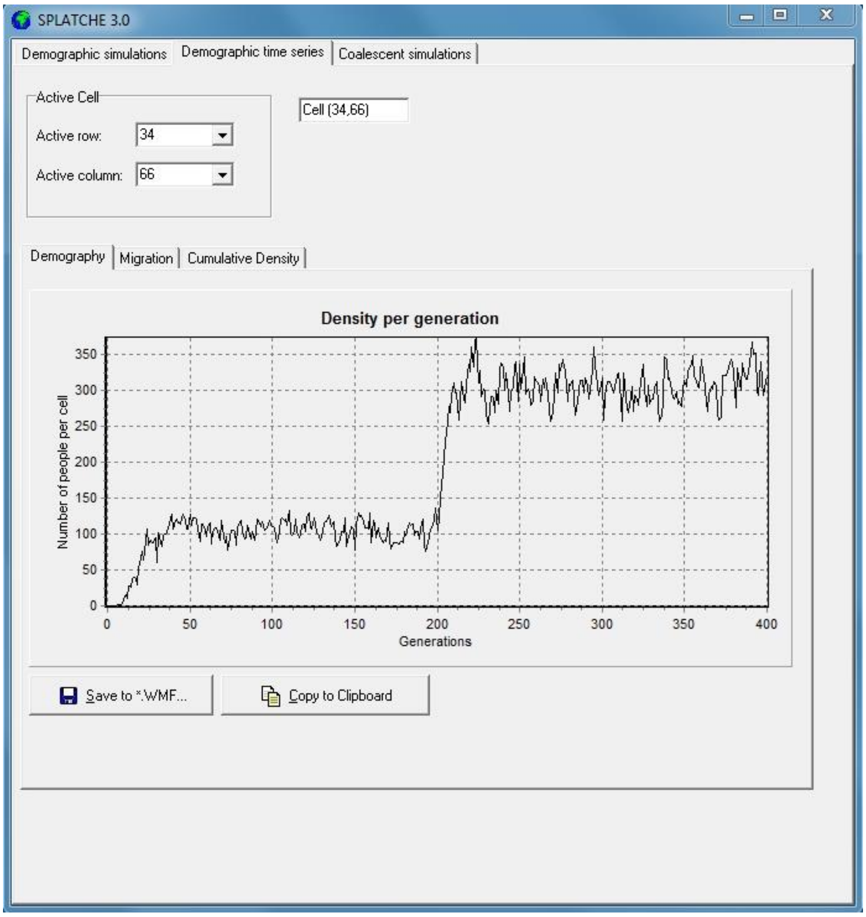
Nb active cells: 3585
Nb rows: 87
Nb columns: 83
Density : 0
Row : 50
Col : 81
Generations: 155
Simul. time: 4650
Real time: -115350
Draw immigrations: □
Aggr. factor: 2
Dynamic Display: ● Density ○ Migration ○ Occupation
Color Choice: ● Color ○ Gray
DRAW
Min density: 1

Figure 8. *Outputs* windows: exploration of the demographic history using the sliding bar.

Once the user is satisfied with input data and the selected settings, the next step is to click on the *Run* button (now available) on the *Demographic simulations* window (Figure 6). At the end of the simulations the output window shows the end state of the demographic simulation. Note that these simulations can take some time depending on the migration model and on the number of generations, but the computation progress is indicated on the *Simulation duration* bar (Figure 6). The demographic history can then be explored using the sliding bar at the bottom of the window (see Figure 8).

By clicking on a deme on the resulting world bitmap, one can access its full demographic history that appears in the second panel (*Demographic time series,* Figure 9) of the main windows. The *Demographic time series* panel shows the demographic history of any simulated cell defined by its geographic coordinates.

The first sub-panel called "*Demography*" shows the evolution of the density within the deme, generation after generation. Two other sub-panels are also available ("*Migration*" and "*Cumulative Density*") that allows one to visualize the migration histories and the cumulative total population size, respectively.

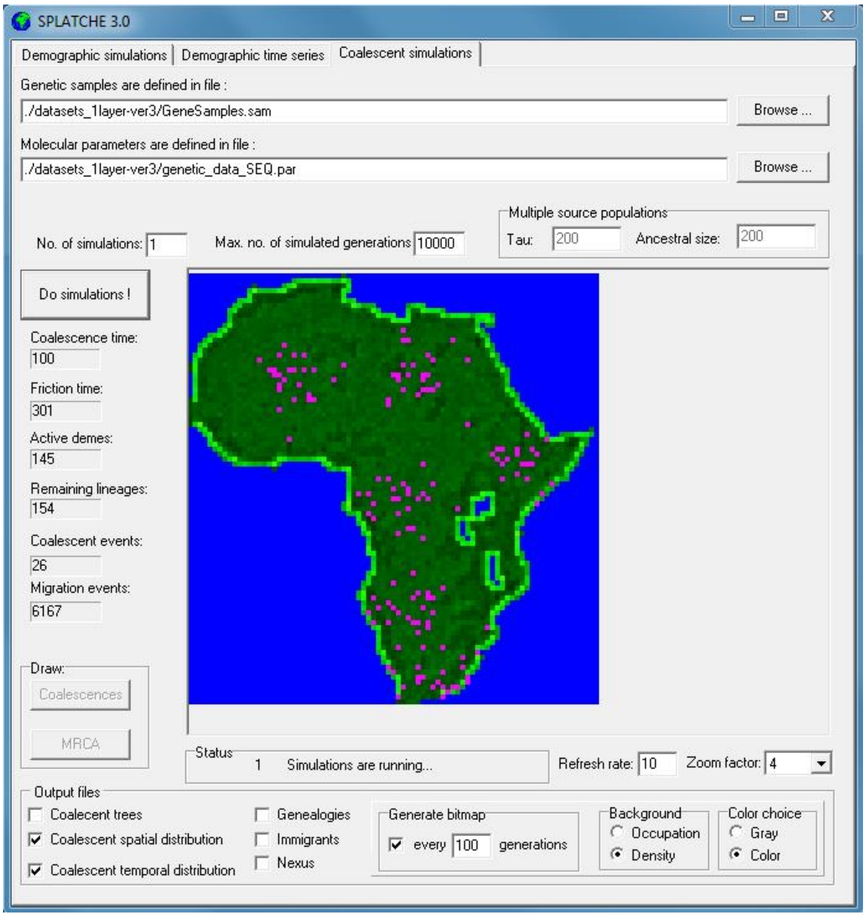


Active Cell
Active row: 34
Active column: 66
Cell (34,66)

Figure 9. *Demographic time series*: exploration of the demographic history of a given deme.

Finally, the third panel (*Coalescent simulations*) on the main window allows one to launch the genetic simulations. This panel is filled with parameters values found in the Settings file, and, again, only a few of them (number of simulations and parameters related to the multiple source populations) can be changed in the graphical interface before launching the simulation, which is done by clicking on the "*Do simulations!*" button. During the coalescent backward simulation, the demes occupied by at least one gene are depicted as violet dots.

Note that one independent coalescent simulation is done per independent marker. At the end of the genetic simulations, bitmaps showing the locations of the coalescent events and the locations of the MRCA can be opened by clicking on the "*Coalescences*" and "*MRCA*" buttons, respectively.



Genetic samples are defined in file : ./datasets_1layer-ver3/GeneSamples.sam

Molecular parameters are defined in file : ./datasets_1layer-ver3/genetic_data_SEQ.par

No. of simulations: 1

Max. no. of simulated generations 10000

Multiple source populations: Tau: 200 Ancestral size: 200

Coalescence time: 100

Friction time: 301

Active demes: 145

Remaining lineages: 154

Coalescent events: 26

Migration events: 6167

Status: 1 Simulations are running...

Refresh rate: 10 Zoom factor: 4

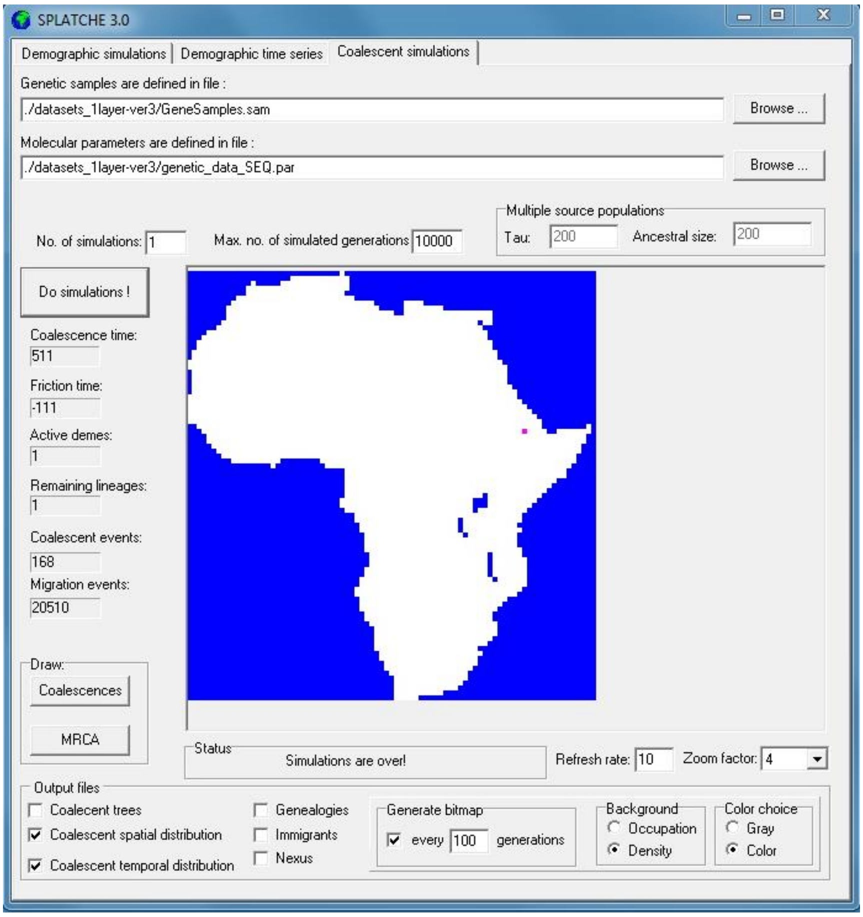
Output files:
- ☐ Coalecent trees
- ☑ Coalescent spatial distribution
- ☑ Coalescent temporal distribution
- ☐ Genealogies
- ☐ Immigrants
- ☐ Nexus

Generate bitmap: ☑ every 100 generations

Background: ○ Occupation ● Density

Color choice: ○ Gray ● Color

Figure 10. *Coalescent simulations* panel showing the moving genes over the simulated world.



SPLATCHE 3.0

Demographic simulations | Demographic time series | Coalescent simulations

Genetic samples are defined in file : ./datasets_1layer-ver3/GeneSamples.sam

Molecular parameters are defined in file : ./datasets_1layer-ver3/genetic_data_SEQ.par

No. of simulations: 1

Max. no. of simulated generations 10000

Multiple source populations: Tau: 200 Ancestral size: 200

Coalescence time: 511

Friction time: -111

Active demes: 1

Remaining lineages: 1

Coalescent events: 168

Migration events: 20510

Status: Simulations are over!

Refresh rate: 10

Zoom factor: 4

Output files

- ☐ Coalecent trees
- ☑ Coalescent spatial distribution
- ☑ Coalescent temporal distribution
- ☐ Genealogies
- ☐ Immigrants
- ☐ Nexus

Generate bitmap: ☑ every 100 generations

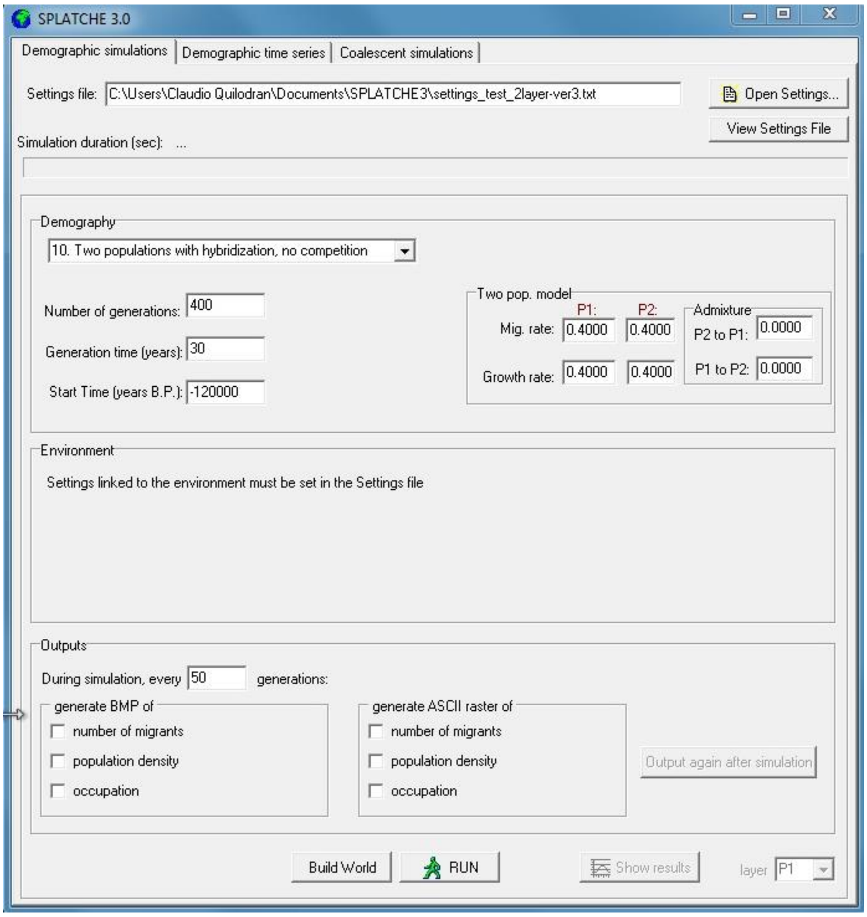
Background: ○ Occupation ● Density

Color choice: ○ Gray ● Color

Figure 11. *Coalescent simulations* panel: end of a genetic simulation.

3.1.1 Graphical user interface for two populations

When opening a setting file defining two populations (DoublePopulationMode=1), a series of new fields appear on the GUI (Figure 12). The acronyms P1 and P2 stand for “Population one” and “Population two”. Admixture rates between P1 and P2 may be asymmetrical and set independently (see chapter 5.2). The “layer” dropdown menu on the bottom right of the Demographic simulations panel indicates which population (P1 or P2) is displayed in the *Demographic time series* panel, *Coalescent simulations* panel, and the Outputs window.



Settings file: C:\Users\Claudio Quilodran\Documents\SPLATCHE3\settings_test_2layer-ver3.txt

Simulation duration (sec): ...

Demography: 10. Two populations with hybridization, no competition

Number of generations: 400

Generation time (years): 30

Start Time (years B.P.): -120000

Mig. rate: P1: 0.4000 P2: 0.4000

Growth rate: P1: 0.4000 P2: 0.4000

P2 to P1: 0.0000

P1 to P2: 0.0000

Settings linked to the environment must be set in the Settings file

During simulation, every 50 generations:

generate BMP of

- □ number of migrants
- □ population density
- □ occupation

generate ASCII raster of

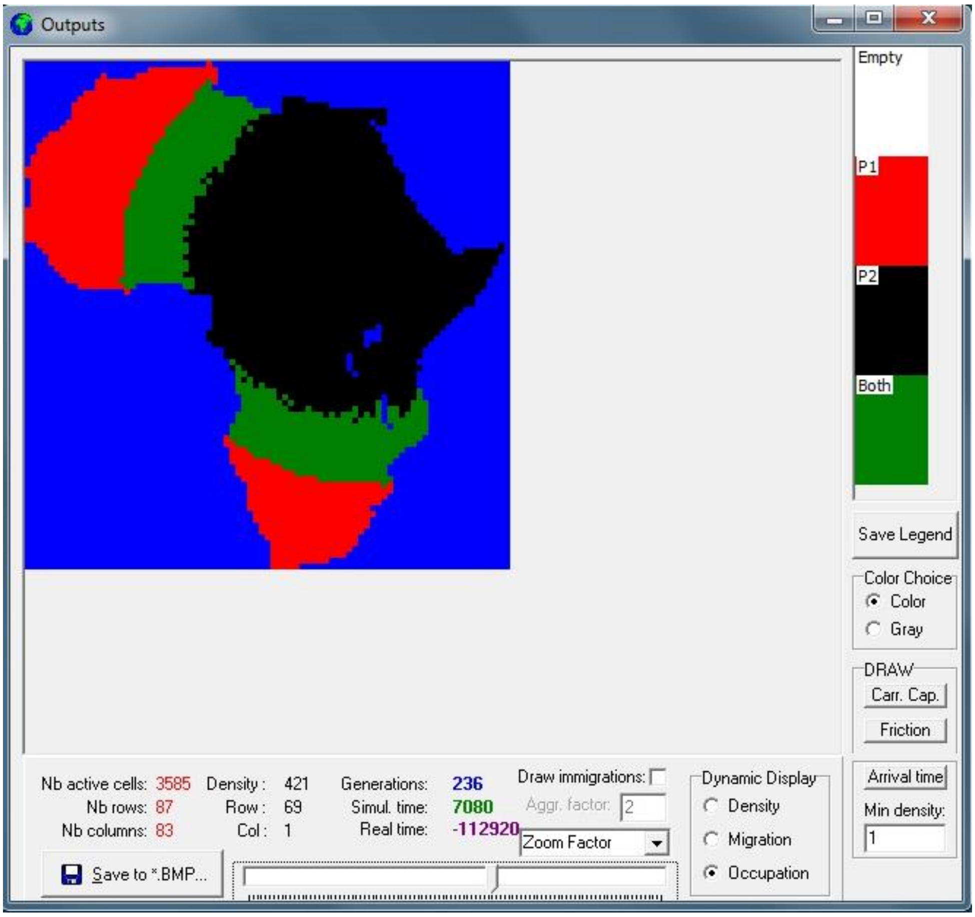
- □ number of migrants
- □ population density
- □ occupation

layer P1

Figure 12. ***Demographic simulations*** **panel for two populations.**

In the outputs window, after a demographic run, there is a new option available called “occupation” (see Figure 13). This option displays in different colors the demes that are occupied either by one or two populations at any time of the simulation.

After a demographic run, there is an additional sub-panel called “Admixture events” available in the *Demographic time series* panel (Figure 14). This panel displays the number of realized admixture events (or introgressions) in both directions (e.g. gene flow from P1 to P2 or the reverse) during the simulation and for any given deme.



Nb active cells: 3585 Density : 421 Generations: 236

Nb rows: 87 Row : 69 Simul. time: 7080

Nb columns: 83 Col : 1 Real time: -112920

Draw immigrations: □

Aggr. factor: 2

Dynamic Display
○ Density
○ Migration
● Occupation

Color Choice
● Color
○ Gray

Min density: 1

Figure 13. *Outputs* window showing the occupation of the world by the two populations.

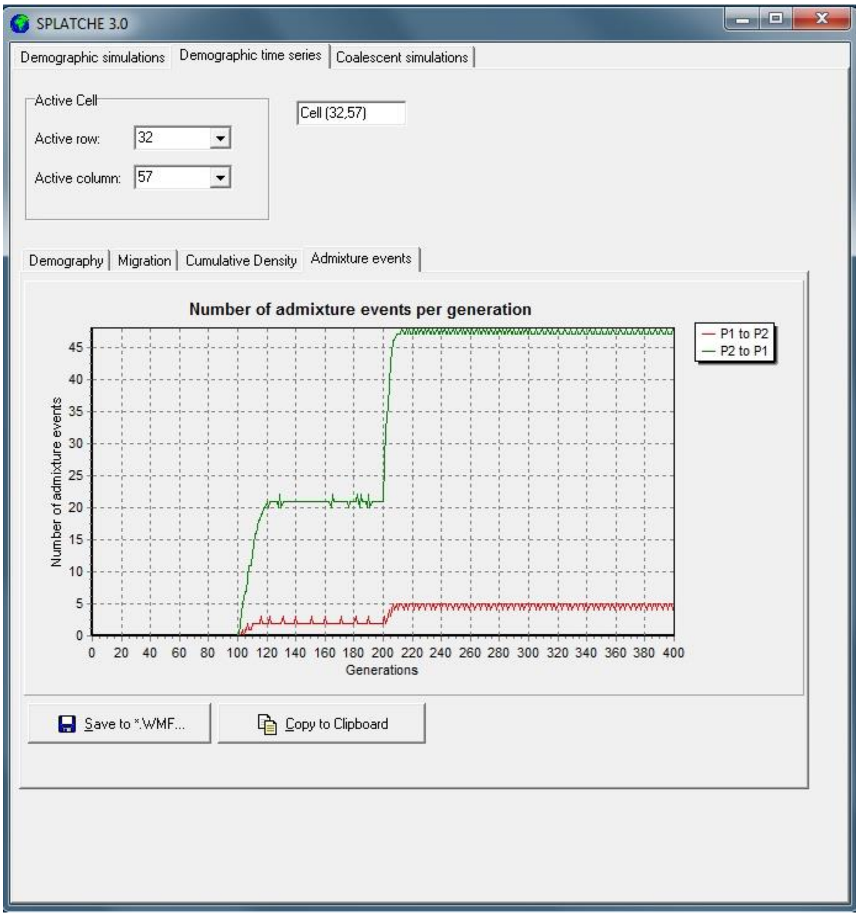


Figure 14. ***Admixture events*** **subpanel.**

3.2 Console interface (Linux)

The console versions (Linux) have been created to make it easier to launch parallel instances of SPLATCHE on computer clusters. With these versions, launching a set of simulations (demographic + genetic) is simple, because it only involves calling the program at the command prompt with the name of the settings file, for example under Linux:

```
$ ./splatche3_lin_64 settings_test_1layer-ver3.txt
```

Output messages will appear during the simulations, for example:

```
$ ./splatche3_lin_64 settings_test_1layer-ver3.txt
Reading inputfile 'settings_test_1layer-ver3.txt' .....
Random generator initialized with seed=-1437263
Building World ... OK
Simulating from 'settings_test_1layer-ver3.txt'
Simulating demography, iteration no. 1...
OK
Simulating genetics, iteration no. 1...
OK
ALL SIMULATIONS AND PROGRAMS TERMINATED SUCCESFULLY!
```

The Linux version is available as a 64-bits version.

4 Demographic models

The demographic simulation regulates local population growth and migration. Six migration models are available when using one layer (one population mode). With two layers (two populations mode), six other models can be chosen based on the level of competition and kind of admixture between the two populations. We describe these models hereafter.

4.1 Local population growth

The local logistic growth is the same for one or two layers. The logistic population growth of each deme follows a standard logistic curve, of the form

$$N_{t+1} = N_t \left(1 + r \frac{K - N_t}{K} \right),$$

where K is the deme-specific carrying capacity, and r is the growth rate identical for all demes of the same population layer. Note that the sequence of the demographic steps at each generation step is as follows: (1) demographic growth, (2) local migration. Because integer values are needed for the coalescent part, N_{t+1} is floored and the remaining part is kept in memory for the next generation, where it is added to N_t before computing the logistic equation. Note that r may result in chaotic or periodic dynamics when it is bigger than 2.

4.2 Friction

In SPLATCHE3, the friction attributed to a focal deme controls the ease/difficulty of immigrating to that focal deme from neighboring demes. When the number of emigrants is computed in a focal deme, the migration model will "look" at the frictions in the neighboring demes and will distribute the emigrants in the neighboring demes according to their relative frictions.

Frictions (f) are expressed as relative numbers and must be set up as $0<f<1$, with friction close to 0 expressing ease of movement and frictions close to 1 expressing difficulty of movements. If $f\leq 0$ or $f\geq 1$, the deme is considered isolated and no migrant will reach it.

When computing the directional percentage of emigrants (p_i), or the directional probability of emigrations, the following formula is applied:

$$p_i = \frac{1}{f_j \sum_{j=1}^{n} \frac{1}{f_j}}$$

where f_j is the friction of the deme in direction j (north, south, east or west) over the n available neighboring demes, and $\sum p_i = 1$.

Note that the total number of emigrants from a focal deme is always spread among the n available neighboring demes, even if only 3, 2, or 1 of such neighboring deme(s) are available (i.e. there is no loss of emigrants, or in other terms there is no absorbing boundaries around demes).

4.3 Migration models with one population

For the migration part of the demography, six models are available in SPLATCHE3. To use any of these models, the ID (1, 2, or 3, 4, 5, 6, as defined below) of the chosen model must be indicated for the settings *ChosenDemographicModel* in the Settings file (see Section 5.2).

4.3.1 Model 1: Migration model with even number of emigrants

The number of emigrants M_i in any of the *n* available directions is computed, for each generation, as $M_i = floor(p_i \cdot m \cdot N_t)$, where m is the migration rate, N_t is the population density of the deme at generation t, p_i is the directional percentage of emigrants, and *floor* means that the fractional part of the number is truncated. The total number of emigrants *M* from a deme is the sum of the M_i, and is therefore always a multiple of the available *n* neighbors.

The fractional parts of M_i at time *t* (prior to applying the *floor* function) are kept in a variable and are accounted for in the next generation by adding them to the deme density N_t.

4.3.2 Model 2. Migration model with absolute number of emigrants

Same as Model 1, but the fractional part of M_i is not truncated. However, the sum of all M_i is truncated as $floor(\sum_1^i M_i)$. Then, a multinomial distribution is used to split M emigrants to the neighboring demes (see Ray 2003), according to the directional probability of emigration (p_i). This ensures that there are always M emigrants that are sent. The drawback of this technique is that it requires the drawing of random numbers, which increases simulation time.

4.3.3 Model 3. Stochastic migration model with absolute number of emigrants

Same as Model 2, but the number of emigrants M (in integer) varies stochastically as a Poisson variable centered around $N_t m$. This model is used in Ray *et al.* (2005).

4.3.4 Model 4: One population with LDD

Same as model 3, but with part of the number of the M emigrants that can be long distance dispersal (LDD) events. The absolute number of these LDD migrants is computed from a binomial distribution $b(M, \alpha)$, where alpha is the probability for an emigrant to migrate over a long distance (more than one deme away from its current location). The distance travelled by each LDD migrant is then drawn from a LDD distribution kernel, following a two-step process.

A random direction is first chosen, and the distance travelled by the migrants is then obtained by drawing it from a Gamma distribution $f(x) = \frac{1}{\Gamma(k)\theta^k} x^{k-1} e^{(\frac{|x|}{\theta})}$, where k is a shape parameter and θ is a scale parameter. In Ray and Excoffier (2010), the shape and scale parameters of the dispersal kernel were set to 0.0419 and 488.5, respectively. These two parameters were initially estimated from human demographic data by Cavalli-Sforza, Kimura et al. (1966) and later used by Novembre, Galvani et al. (2005).

Once the LDD direction and distance are computed, the corresponding target deme is defined. The LDD event targets a deme if the deme is on land. In this model, The LDD emigrate in the target deme even if a population is already present. However, if the target deme is on water (or outside the world), the LDD event is resampled from the Gamma distribution and a new target deme is found. A maximum of 10 resamples are allowed for each LDD event. If no target deme is found after 10 resamples, the LDD event falls back to a normal nearest-neighbor migration event.

This model is the one used in Ray and Excoffier (2010), but with a small difference: in Ray and Excoffier (2010), when a LDD migrant could not find an appropriate distant deme to "land" after 10 resamples, the migrant was discarded and not used as a nearest-neighbor migrant.

A further parameter (LDD_max_treshold) can be defined in the settings file that can set the maximum distance (in number of demes) for any LDD event. This parameter was set to 6 in Alves, Arenas et al. (2016).

4.3.5 Model 5: One population with LDD only on empty demes

Same as model 4, but LDD migrants can only fall on empty demes (and therefore starts a new population in the arrival deme). A maximum of 10 resamples from the LDD Gamma distribution are allowed for each LDD event that do not satisfy this constraint. This model is used in Ray and Excoffier (2010).

4.3.6 Model 6: One population with LDD only on occupied demes

Same as model 4, but LDD migrants can only fall on demes already occupied (i.e. N>0). A maximum of 10 resamples from the LDD Gamma distribution are allowed for each LDD event that do not satisfy this constraint. This model is used in Ray and Excoffier (2010).

4.4 Migration model with two populations

For all six demographic models (7-12) available with two layers (*DoublePopulationMode* = 1), the migration model is equivalent to the model 1 described above (section 4.3.1). It means that no migration stochasticity is available with two layers, in all cases an even number of emigrants is distributed in the neighboring demes with fractions retained for the next generation. Note that LDD is not available when two interacting populations are simulated.

4.5 *Admixture model with two populations*

There are two kinds of admixture models in SPLATCHE3, one is called “assimilation” and the other one “hybridization”. The “assimilation model” is the former admixture model in SPLATCHE2 while the “hybridization model” is newly implemented in SPLATCHE3.

4.5.1 Assimilation model

The life cycle of a population at a given generation is as follows: admixture (see section 4.5), logistic regulation incorporating competition (see section 4.6), followed by migration (see section 4.4). This life cycle thus assumes that migration is at the adult stage. Gene flow is due to the actual movements of individuals from one population (one layer) to the other population (the other layer), i.e. the assimilation in population B of individuals from population A and the reverse. It thus means that admixture is affecting the demography of both populations. This model has been initially developed to simulate the assimilation of hunter-gatherers in farmer populations (Currat and Excoffier 2005).

The frequency of admixture events is assumed to be density-dependent. Within a given deme, each of the N_i individuals from the i-th population has a probability

$$A_{ij} = \gamma_{ij}(2N_iN_j)/(N_i + N_j)^2$$

to reproduce successfully with one of the N_j members of the j-th population, and γ_{ij} represents the probability that such a mating results in a fertile offspring. Alternatively, γ_{ij} could represent the relative fitness of hybrid individuals or an index of disassortative mating. γ_{ij} and γ_{jj} are defined by the user in setting the parameters *MigrRate_P1_to_P2* and *MigrRate_P2_to_P1* respectively. Following admixture, population densities are then updated as

$$N_i^{t+1} = N_i^t(1 - A_{ij}) + A_{ij}N_j^t$$

4.5.2 Hybridization model

The hybridization model is better adapted to simulate admixture between species than the assimilation model because it does not affect the demography of both populations (species) and only simulate gene flow between them. The life cycle of a population at a given generation is as follows: logistic regulation incorporating competition (see section 4.6), admixture (see section 4.5), followed by migration (see section 4.4).

The hybridization model posits that each of the N_i^{t+1} newborn individuals in the population i must have one parent coming from the same population i. Then assuming local random mating between populations, the probability that the other parent originates from the same population i is

$$\frac{N_i^t}{N_i^t + N_j^t}$$

where N_i^t and N_j^t are the diploid population sizes at the previous generation. Similarly, the probability that the other parent comes from population *j* and therefore that the individual is a hybrid is

$$\frac{N_j^t}{N_i^t + N_j^t}$$

Therefore the expected number of introgression events from population *j* to population *i* per generation is

$$A_{ji} = \frac{\gamma_{ji} N_i^{t+1} N_j^t}{N_i^t + N_j^t}$$

which is similar to the equation used by Zhang (2014) to define the number of admixed individuals except that N_i^{t+1} is used here instead of N_i^t in Zhang (2014). γ_{ij} and γ_{jj} have the same meaning than for the assimilation model (Section 4.5.1) and are defined by the user in setting the parameters *MigrRate_P1_to_P2* and *MigrRate_P2_to_P1* respectively.

4.6 Competition models with two populations

The six following models are only available when the *DoublePopulationMode* parameter has been set to one.

4.6.1 Model 7 & 10: No competition

In that case, there is no competition between the two populations. The demography of both populations is regulated independently using the model 1 described above (section 4.3.1). This model was used in Currat et al. (2008). Model 7 used the assimilation model (see section 4.5.1) while Model 10 used the hybridization model (see section 4.5.2).

4.6.2 Model 8 & 11: Competition density-dependant

These models include competition density-dependant between the two populations. The model of density regulation incorporating competition is based on the Lotka–Volterra interspecific competition model, which is itself an extension of the logistic growth model (Volterra 1926, Lotka 1932). For each population *i*, a new density N_i^{t+1} is calculated from the former density N_i^t as

$$N_i^{t+1} = N_i^t\left(1 + r_i\left(K_i - N_i^t - \alpha_{ij} N_j^t\right)/K_j\right)$$

where r_i is the intrinsic growth rate, K_i the carrying capacity, N_j is the density in the other population from the same cell and α_{ij} is an asymmetric competition coefficient. An α_{ij} value of 1 implies that individuals of the *j*-th population have as much influence on those of population *i* as on their own conspecific, or that competition between populations is as strong as competition within a population. Lower values of α_{ij} indicate lower levels of competition between populations than within populations; a value of zero implies no competition between individuals from different populations. Because integer values are needed for the coalescent part, N_{t+1} is floored and the remaining part kept in memory for the next generation where it is added to N_t before computing the Lotka-Volterra equation.

Contrary to the classical Lotka-Volterra competition model which assume constant values for α, we assumed that competition is density-dependant and α_{ij} is thus calculated as

$$\alpha_{ij} = N_j^t / (N_i^t + N_j^t),$$

reflecting the fact that the influence of the members of a population on the other population grows with its density. This model was first used in Currat & Excoffier (2004, 2005). Model 8 used the assimilation model (see section 4.5.1) while Model 11 used the hybridization model (see section 4.5.2).

4.6.3 Model 9 & 12: Competition density-independent (classical Lotka-Volterra)

These models are the same than model 5 but competition coefficients α_{ij} are set to a constant value of 1 and do not depend on the current densities. This model was used in Currat et al. (2008). Model 9 used the assimilation model (see section 4.5.1) while Model 12 used the hybridization model (see section 4.5.2).

5 Settings, input and output files

5.1 *Description of the Settings file*

Each setting is set through a "*SettingsName=value*" pair, for instance:

```
ChosenDemographicModel=3
```

Note that Settings names should not be changed and that they are case sensitive. No space should be put before and after the "=" sign. Any line in the Settings file with the "#" sign in front of it is commented and therefore ignored. Empty lines are also ignored. The order of the lines in the Settings file does not matter, but we advice users to keep the original order for clarity.

Some of the settings are not mandatory and can therefore be commented if one does not want to use them. For mandatory settings, the program will throw an error if these settings are not defined. For some settings, it is only when defined (i.e. not commented) that the corresponding machinery is available. For example, the taking into account of frictions is only available if the setting "FrictionChoice" is defined.

All settings that are paths to other text files must be set either with absolute paths (e.g. C://documents/file.txt) or with relative path (e.g. ./file.txt) to where the SPLATCHE executable is located.

5.2 List of settings with a short definition

Below is a list of all available settings, presented as they appear in the sample Settings file available on the SPLATCHE website. Although settings can appear in any order in the Settings file, they have been grouped here by way of their use. When necessary, the reader is sent to a detailed description of the settings or its format.

Parameters linked to filenames	
Setting name	**Definition**
PopDensityFile	Path to a text file with the locations, initial densities, etc. of the initial population(s). See section 5.2.1. (MANDATORY).
PresVegetationFile	Path to an Ascii file describing the initial world (array of demes). See section 5.2.2 (MANDATORY).
HydroFile	Path to an Ascii file describing hydrology (rivers). See section 5.3.2 (NOT MANDATORY).
RoughnessTopoFile	Path to an Ascii file with the roughness values. See section 5.3.2 (MANDATORY only if FrictionChoice=1 or =2).
mMapFile	Path to an Ascii file describing the various zones that can hold different migration rates. (NOT MANDATORY) See section 5.2.2. If this parameter is not given, the global migration rate *MigrationRate* is used.
Veg2KFile	Path to a text file holding paths to tables making the correspondence between vegetation categories and carrying capacity values (MANDATORY). See section 5.2.3.
Veg2FFile	Path to a text file holding paths to tables making the correspondence between vegetation categories and friction values (NOT MANDATORY). See section 5.2.3.
Veg2mFile	Path to a text file holding paths to tables making the correspondence between zones from the mMapFile and zone-specific migration rates. (NOT MANDATORY). See section 5.2.3. If this parameter is not given, the global migration rate *MigrationRate* is used.

Parameters linked to demographic simulations	
Setting name	**Definition**
ChosenDemographicModel	Identification (number) for the demographic model (MANDATORY). See Section 2.5.
EndTime	Number of simulated generations (MANDATORY).
GenerationTime	Duration of a generation (in years), used in several functions needing it (MANDATORY).
GrowthRate	Intrinsic growth rate used in the logistic function (MANDATORY).
MigrationRate	Migration rate for neighboring deme migration (MANDATORY). See section 5.2.4. If either mMapFile or Veg2mFile are not defined, MigrationRate is used.
AllowSourcePopulationOverflow	The value indicates whether the density *N* that is attributed to the initial deme is spread over neighboring demes when *N* is greater than the carrying capacity (NOT MANDATORY). See section 5.2.5.
TauValue	Tau value (in generations). Only used with multiple source populations. Backward in time, this is the time between the onset of the expansion and the timing at which all remaining lineages are brought into one small deme (NOT MANDATORY). See section 5.2.6.
AncestralSize	Size of the ancestral deme at time Tau (defined in the setting *TauValue*). (NOT MANDATORY). See section 5.2.6.
ArrivalCellFile	Text file (*.col) with coordinates of demes for which arrival times are needed (MANDATORY). See section 5.2.7.

Parameters linked to long distance dispersal (LDD)	
Setting name	**Definition**
LDDProportion	The proportion of migration events that are LDD events. Same for all demes. (MANDATORY IF ONE OF THE LDD MODELS IS USED: ChosenDemographicModel = 4, 5 or 6).
GammaShapeParamAllCells	Shape parameter for the LDD Gamma distribution. Same for all demes. (MANDATORY IF ONE OF THE LDD MODELS IS USED: ChosenDemographicModel = 4, 5 or 6).
GammaScaleParamAllCells	Scale parameter for the LDD Gamma distribution. Same for all demes. This is the standard scale parameter, which is equal to 1/theta (e.g., 0.002047083=1/488.5). (MANDATORY IF ONE OF THE LDD MODELS IS USED: ChosenDemographicModel = 4, 5 or 6).
LDD_max_treshold	Maximum LDD migration distance (in number of demes from the focal deme). Must be larger than 2 if used. This translates into a truncated gamma for the LDD Gamma distribution. (NOT MANDATORY. Default to no maximum threshold if the parameter is not defined).

Parameters linked to physical environment	
Setting name	**Definition**
FrictionChoice	Choice of friction type (0: vegetation,1: roughness topography,2: both). If friction needs to be taken into account, this parameter must be defined (along with "Veg2FFile" and/or "RoughnessTopoFile"). If not defined, friction is uniform (NOT MANDATORY). See section 5.2.8.
RealBPTime	Real time (in years before present (BP)) of the start of the simulation. The number must be a negative integer. This parameter is linked to the graphical display of the real time in any simulation (MANDATORY).
RiverFrictionChangeFactor	This is a factor [>0] that increases or decreases the friction of the river cells (as defined in the HydroFile input file). The friction of these cells is simply multiplied by this factor. In case of "double populations", the change is applied on both friction maps. (NOT MANDATORY).
RiverCarCapChangeFactor	This is a factor [>0] that increases or decreases the carrying capacity of the river cells (as defined in the HydroFile input file). The carrying capacity of these cells is simply multiplied by this factor. In case of "double populations", the change is applied on both friction maps. (NOT MANDATORY).
CoastFrictionChangeFactor	This is a factor [>0] that increases or decreases the friction of the coast cells (coasts are automatically defined). The friction of these cells is simply multiplied by this factor. In case of "double populations", the change is applied on both friction maps. (NOT MANDATORY).
CoastCarCapChangeFactor	This is a factor [>0] that increases or decreases the carrying capacity of the coast cells (coasts are automatically defined). The carrying capacity of these cells is simply multiplied by this factor. In case of "double populations", the change is applied on both carrying capacity maps. (NOT MANDATORY).

Parameters linked to genetic simulations	
Setting name	**Definition**
SampleFile	Text file containing the coordinates and sizes (+other info) of the genetic samples. It must have a "*.sam" extension (MANDATORY). See section 5.2.9.
GeneticFile	Text file (*.par) containing the definition of markers property. This is the main file for genetic settings, with or without recombination (MANDATORY). See Section 5.2.11.
NumGeneticSimulations	Number of genetic simulations following the demographic simulation. Each genetic simulation independently reconstructs a genealogy of the sampled genes and produces one ARLEQUIN output file (MANDATORY).
GenotypicData	Choice of genotypic or haplotypic data (1:genotypic; 0: haplotypic)(MANDATORY) Note that haploid data are always simulated, only the way output are represented does change.
MaxNumGenerations	Maximum number of total generations for a genetic simulation. This number corresponds to the number of generation of the demographic simulation + the extra generations for the coalescence process prior to time 0. After this time the process stops if the genealogy has not been correctly reconstructed and none output file is produced (MANDATORY).
GenealogiesFile	Generation of genealogy files *.tri (0:no ; 1: yes) (MANDATORY). See section 5.2.14 .
ImmigDistrFile	Generation of immigrants distribution file *.nm (0: no ; 1: yes) (MANDATORY). See section 5.2.15.

Parameters linked to double populations	
Setting name	**Definition**
DoublePopulationMode	Flag to indicate that double populations (two layers) version is used (0: no ; 1: yes) (MANDATORY).
GrowthRate_P2	Intrinsic growth rate for the second population (P2) used in the logistic function. (MANDATORY IF *DoublePopulationMode*=1).
MigrationRate_P2	Migration rate for the second population (P2) to neighboring deme migration. (MANDATORY IF *DoublePopulationMode*=1).
MigrRate_P1_to_P2	Admixture rate from P1 to P2 [0-1], see section 5.2.16. (MANDATORY IF *DoublePopulationMode*=1).
MigrRate_P2_to_P1	Admixture rate from P2 to P1 [0-1], see section 5.3.15. (MANDATORY IF *DoublePopulationMode*=1).
Veg2K_P2_File	Text file for population 2 (P2) holding the carrying capacity values for each vegetation category. (MANDATORY IF *DoublePopulationMode*=1).
Veg2F_P2_File	Text file holding the friction values for each vegetation category. (NOT MANDATORY).
Veg2m_P2_File	Path to a text file holding paths to tables making the correspondence between zones from the mMapFile and zone-specific migration rates for the second layer. (NOT MANDATORY). See section 5.2.3. If this parameter is not given, the global migration rate *MigrationRate* is used.
PropFile	Compute theproportion of sampled genes in the second population (P2) which are descendent from the original source population P2 (0:no ; 1: yes), see section (MANDATORY IF *DoublePopulationMode*=1).

Parameters linked to various other outputs	
Setting name	**Definition**
GenerateOutputMigrationBMP	Generate output BMP of migrations (0:no ; 1: yes) (MANDATORY).
GenerateOutputMDensityBMP	Generate output BMP of densities (0:no ; 1: yes) (MANDATORY).
GenerateOutputOccupationBMP	Generate output BMP of occupations (0:no ; 1: yes) (MANDATORY).
GenerateOutputMigrationASCII	Generate output ASCII of migrations (0:no ; 1: yes) (MANDATORY).
GenerateOutputMDensityASCII	Generate output ASCII of densities (0:no ; 1: yes) (MANDATORY).
GenerateOutputOccupationASCII	Generate output ASCII of occupations (0:no ; 1: yes) (MANDATORY).

Miscelaneous Parameters	
AllowShortIntForNumberOfIndividuals	Allow "short int" (instead of int) to be used for the database. This reduces by half the amount of RAM necessary. Only do that if you are sure that your numbers of individuals (i.e. population densities and number of migrants) never exceed 16,000 in any single deme! (NOT MANDATORY).

Detailed description of settings and outputs

5.2.1 PopDensityFile

A file, called "dens_init_europe_small.txt" in the example Settings file, is used to specify the place(s) of origin of the simulated population. This file contains a first line with the number of source populations to be considered, and then one line per source population. Each of these lines consists of 10 fields separated by "tab" or "space" character.

Example:

```
3
Paleo         100    12     60     5      10     100    0.4    0      0
Iberic        100    12     10     0      50     50     0.1    0      0
Central       50     20     35     0      150    150    0.2    0      0
```

These 10 fields (parameters) correspond to:

1. The name of the source populations (without space).

2. The size of the source population, in number of effective haploid individuals.

3. & **4**. The geographic coordinates of the source population (latitude and longitude). SPLATCHE will determine in which particular deme the coordinates of the population corresponds. Coordinates do not need to be in a particular unit (e.g. decimal degrees), but they needs to be in the same units that the coordinates defined in the header of the environmental files (see Section 5.2.2). Each coordinate must also fall within the extent of the world defined by the environmental files and they should not fall in "sea" (*NoData*) demes.

5. A resize parameter: it is the size of the source population before the beginning of the expansion. This parameter is used only for genetic simulations. If this parameter is set to 0, then the size of the source population before the onset of the expansion is regarded as being equal to the initial size (set in the second column).

6 & 7. The time (6) in number of generations before the beginning of the expansion when a second resize (7) occurs. It could be seen as the duration of the bottleneck and the population size before the bottleneck.

8. The emigration rate from the current source to the other sources before the time of the expansion. This rate is only used in case of multiple source populations (e.g. multiple sources for the first population P1 at generation 0).

9. Index of the layer where the source appears: 0 means the first layer (P1) while 1 indicates the second layer (P2). Note that: *i)* If there is only one layer (population) simulated (*DoublePopulationMode*=0), all sources must start in P1; *ii)* if two layers are simulated (*DoublePopulationMode*=1), one source must start in P2, otherwise the second population will not be created. *iii)* If this parameter is set to 1 (starting in population P2) then a number of individual equal to the size of the source population (column 2) will be taken from P1 at the same location in the array of demes, in order to found the new population P2. It results that if the number of individuals currently belonging to the deme in P1 is smaller or equal to the size of the population source in P2, then all individuals then all individuals in P1 will be taken. It also means that if P1 has not been colonized, then P2 source cannot be created, and the second layer is not colonized.

10. The expansion time. If two populations are simulated (*DoublePopulationMode*=1), the founding of the second population (P2) should start after the beginning of the expansion of the first population (P1). This parameters indicates the time (in number of generation) at which the source must start. Note the following constraints: *ii)* it must be at least one source starting in P1 at generation 0; *ii)* P1 sources can only start at generation 0.

Figure 15 shows a schematic example of two parallel expansions in the first layer. Two source populations start in the first layer (P1), both at generation 0. Both source populations passed through a bottleneck prior to their expansions and share a common ancestral population *tau* generations before the onset of their expansions. Compare this scheme to the one represented in Figure 4 that shows another example of two source populations but in two different layers. One source for the first population (P1) at generation 0 and one source for the second population (P2) a given number of generations later.

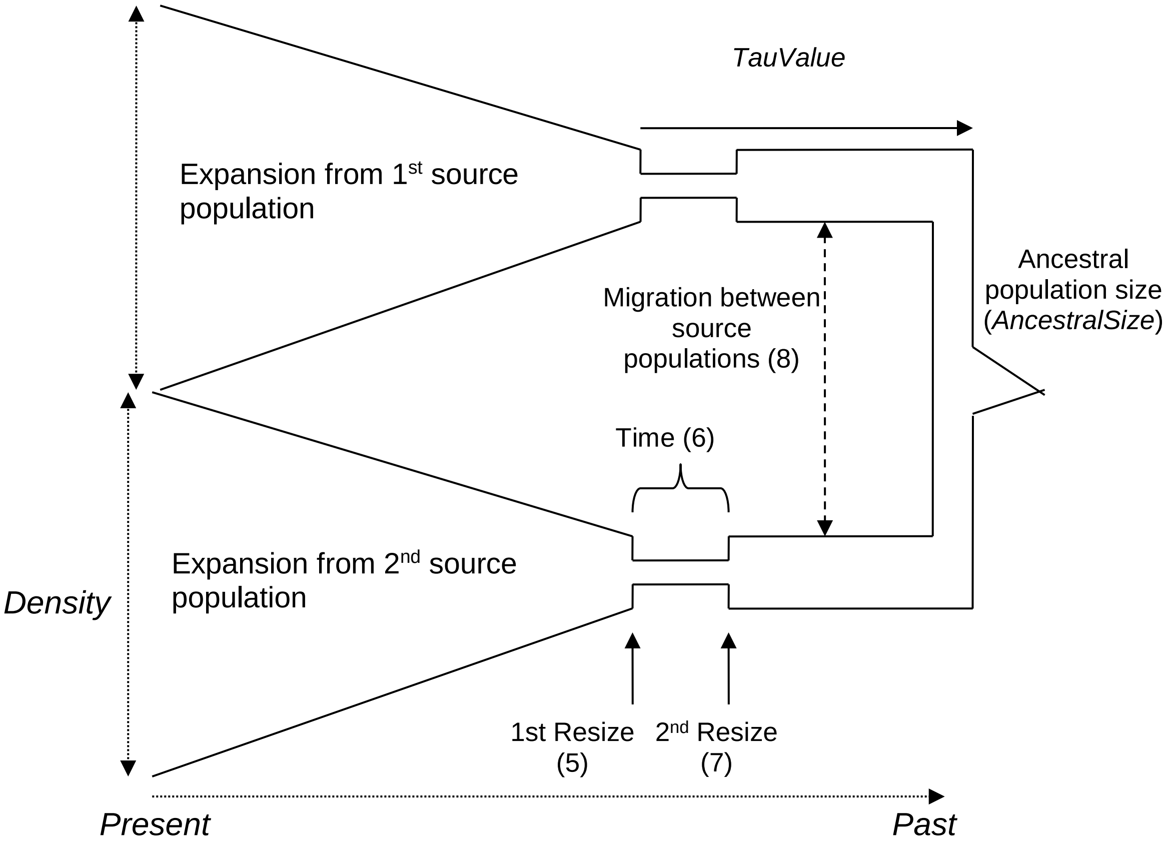

TauValue
Expansion from 1st source population
Ancestral population size (AncestralSize)
Migration between source populations (8)
Time (6)
Expansion from 2nd source population
Density
1st Resize (5)
2nd Resize (7)
Present
Past

Figure 15. Schematic representation of two parallel expansions in one layer (P1).

5.2.2 PresVegetationFile, HydroFile, RoughnessTopoFile, mMapFile

The environmental data sets define the "world" (size, continent contours, and geographic coordinates) in which the simulations take place, but also some of the environmental or demographic values of deme-specific variables. The values of these settings are the paths to the data sets that are text files in ASCII raster format. An ASCII raster file begins with a header (first six lines) containing information on the file, followed by a matrix of values in rows and columns.

The header information is as follow:

```
ncols          : number of columns
nrows          : number of rows
xllcorner      : longitude coordinate of the lower-left deme
yllcorner      : latitude coordinate of the lower-left deme
cellsize       : width of a deme (cell size), in same units than the
                coordinates
NODATA_value   : value indicating that a deme must not be considered (e.g.
                because it is in the sea)
```

Example of an environmental data set,

```
ncols         88
nrows         91
xllcorner     -19.845388
yllcorner     -36.897187
cellsize      0.83
NODATA_value  -9999
-9999 -9999 -9999 -9999 -9999 -9999 -9999 -9999 7 7 7 ...
-9999 -9999 -9999 -9999 -9999 -9999 -9999 7 7 7 7 7 7
-9999 -9999 -9999 -9999 -9999 -9999 7 7 7 7 7 7 7 7 7
-9999 -9999 -9999 -9999 -9999 -9999 7 7 7 7 7 7 7 7 7
-9999 -9999 -9999 -9999 -9999 7 7 7 7 7 7 7 7 7 7 7 7
-9999 -9999 -9999 -9999 -9999 7 7 7 7 7 7 7 7 7 7 7 7
...
```

This format is quite standard in many GIS package, and can be exported from many of them (e.g. ArcGIS).

The “PresVegetationFile” data set defines vegetation categories (integer numbers) for the world (or for the first population in case of two-population setting). For each vegetation category, a carrying capacities and a friction can be defined through the Veg2K and Veg2F files (see next Section). The “HydroFile” defines rivers, i.e. demes with any value (e.g. "0", "1", "2"…) other than “NoData” values. Non-river demes must have the "NoData" value. The “RoughnessTopoFile” defines relative continuous friction values, such as friction derived from topography, which can be done outside of SPLATCHE in any GIS software.

The "mMapFile" data set defines vegetation (or other) categories (integer numbers) for the world (or for the first population in case of two-population setting) that can be attributed a different migration rate. This data set can typically be equal to the "PresVegetationFile" if one wants to set different migration rates for the different vegetation types. But it can also represent another segmentation of the landscape, for example a North to South gradient, when one wants to model a directional range contraction. This was initially implemented by Arenas, Ray et al. (2012) and further developed for the LGM range contraction of European human populations in Arenas, Francois et al. (2013) and Alves, Arenas et al. (2016)).

Note that the RAM necessary to run SPLATCHE is proportional to the number of demes set through the *PresVegetationFile* data set and to the number of simulated generations. If the program crashes due to RAM limitation, either the resolution of the *PresVegetationFile* data set or the number of generations must be reduced.

5.2.3 Veg2K, Veg2F and Veg2m files

Carrying capacities, frictions and migrations can stay unchanged (static) during a demographic simulation, or they can vary through it. The way this variation is implemented is by loading other sets of carrying capacity, friction and migration values at user-defined times during the simulation. If this happens, the simulation is called "dynamic". In order to set at what time(s) the changes occur, different files are needed.

The two first files, which are set in the Settings file through the settings *Veg2KFile* and *Veg2KFile*, are typically called "Dynamic_K.txt" and "Dynamic_F.txt". On the first line of each of these files appears the number of changes during a simulation. Then each line (one per change) includes the time of change (in generations), the filename of the corresponding table (see below), and an arbitrary description (spaces are allowed here). The three components of each line must be separated by a white space. For a non-dynamic simulation (0 or 1 on the first line), only the first filename is considered.

Example of "Dynamic_K.txt" file:

```
2
0 ./dataSets_africa/veg2K.txt vegetation at time 0
500 ./dataSets_africa/veg2K_500.txt doubling of vegetation at time 500
```

Each file name must target a valid "association table" that makes the link between a particular vegetation category and a carrying capacity (or friction) value.

With several tables for the carrying capacity and/or the friction values, it is possible to simulate complex changes in the environment through time.

An association table lists a series of vegetation category numbers, followed by a carrying capacity value (integer values greater than zero), and by a description. The vegetation category numbers must correspond to the numbers found in the input "vegetation" dataset (see Section 5.2.2). Note that the description of each category in the third column is mandatory (the program crashes if it is absent).

Example of "veg2K.txt" file:

```
1      200    Tropical_rainforest
2      200    Monsoon_or_dry_forest
3      500    Tropical_woodland
4      500    Tropical_scrub
5      100    Tropical_semi_desert
6      1000   Tropical_grassland
7      50     Tropical_extreme_desert (50)
8      1000   Savanna
```

Similar files (Dynamic_F.txt) are required for setting friction values. However, this file is not mandatory. If not defined, friction values are uniform over the world. Friction values (*f*) are relative values between 0 (low friction, strong migration) and 1 (extreme friction, no migration possible). If $f \leq 0$ or $f \geq 1$, the deme is considered isolated and no migrant will reach it.

Example of "veg2F.txt" file:

```
1      0.8    Tropical rainforest
2      0.8    Monsoon or dry forest
3      0.5    Tropical woodland
4      0.5    Tropical thorn scrub and scrub woodland
5      0.2    Tropical semi-desert
```

Based on the same principles than the two files above, the settings ***Veg2mFile*** is typically called "Dynamic_m.txt" and holds filenames of the corresponding tables of migration rates (see below).

Example of "Dynamic_m.txt" file:

```
2
0 ./datasets_1layer-ver3/veg2m_pop1_time_1.txt
150 ./datasets_1layer-ver3/veg2m_pop1_time_2.txt
```

Each filename must target a valid "association table" that makes the link between a particular vegetation category and migration rates. The association table lists a series of vegetation category numbers, followed by four migration rates (one for each direction of migration: North, East, South, West), and by a description. The vegetation category numbers must correspond to the numbers found in the input "mMapFile" dataset (see Section 5.2.2). The number of migrants from each deme of a given vegetation category is therefore: N×(m_north+m_east+m_south+m_west), and the sum of the four directional migration rates should not be larger than 1.

Example of "veg2m.txt" file with 15% of the whole population *N* is going south, and 5% of the whole population *N* is going in each of the three other directions:

```
1      0.05   0.05   0.15   0.05   Tropical rainforest
2      0.05   0.05   0.15   0.05   Monsoon or dry forest
3      0.05   0.05   0.15   0.05   Tropical woodland
4      0.05   0.05   0.15   0.05   Tropical thorn scrub and scrub woodland
5      0.05   0.05   0.15   0.05   Tropical semi-desert
```

In that example, the total m value for the focal deme is (0.15+0.05+0.05+0.05) = 0.3, meaning 30% of the density in the deme emigrating at the current generation.

5.2.4 MigrationRate

This is the percentage/fraction (*m*) of the population emigrating at each generation, and it used in the demographic models. In other words, when used in stochastic demographic models, it is the per generation probability for any gene to emigrate. For a deme population of size *N* the number of emigrants is then equal to $N \times m$ at each generation. Note that numbers of emigrants are always integer numbers, so the effective number of emigrants can vary slightly from the exact $N \times m$ value.

5.2.5 AllowSourcePopulationOverflow

If this parameter is set to 0, and if the initial population size (N) is greater than the carrying capacity (K) of the source population, the remaining N-K individuals are spread around the neighboring demes (without ever exceeding the carrying capacities of these demes) until N individuals are placed. The "overflow" of individuals allows one to start with a "patch" of source populations, rather than a single source deme. Also note that the “Resize parameter” (number 5 in section 5.3.1) should be adjusted accordingly (if not, the population size before expansion is still assumed to be equal to the carrying capacity of the source population). If this parameter is set to 1, all individuals are found in the original deme (even if N>>K), but in this case the population of the source deme can be quickly downward regulated by the logistic growth function. Note that overflow can occur only in the first layer P1, not in the second layer P2.

5.2.6 TauValue and AncestralSize

TauValue is the time before the expansion (in generations) at which the multiple source populations are pooled together into one initial population of size *AncestralSize*. These two settings are therefore only used with multiple source population simulations. Note that *TauValue* was expressed in years in previous versions of SPLATCHE. See the illustration under Section 5.2.1.

5.2.7 ArrivalCellFile

This file allows the user to specify the coordinates of the demes for which colonization times are required. An output files called “Arrival_cell_output.txt” lists the arrival time of the demes listed in the colonization file *.col.

This file follows a similar format than the initial population density file (section 5.2.1). The number of demes for which one wants information about their colonization time is given on the first line. If this number is set to 0, then nothing is done. Then, for each line the following information is given:

1. Name of the deme.

2. At which population layer the deme belongs. 0 if only one layer is simulated.

3. & 4. Latitude and longitude of the deme

Example of "colonization.col" file in case of a two-population model:

```
6
Israel        0      4      60
MiddleEast    0      12     60
Iberia        0      12     10
Turkey        1      4      60
MiddleEast    1      12     60
Iberia        1      12     10
#PopName     #Layer #Lat    # Long
```

5.2.8 FrictionChoice

Parameter allowing to choose how the friction values are computed, (0: vegetation, 1: roughness topography, 2: both). When "vegetation" or "roughness" is chosen, friction values are only computed from the corresponding input data set (see Section 5.2.2). If "both" are chosen, friction values are computed by taking, for each deme, the mean value between the friction value from the vegetation data set and the friction value from the roughness data set.

5.2.9 SampleFile

Path to a file with the extension “*.sam” allowing one to specify the localization of the sampled sub-populations, as well as the number of genes sampled in each of these sub-population.

On the first line of this file, the user must specify the number (integer) of sampled sub-populations. The second line is reserved for the legends for each column. Then, each line defines a sample with 5 fields separated by “tab” or “space” character.

1. Name of the sampled sub-population.

2. Number of genes belonging to that sample.

3. Identification of the population layer to which the sampled deme belongs (0 if only one population is simulated: 1 to set it to the second population layer).

4 & **5.** Geographic location of the sampled deme (latitude and longitude). As for the source deme coordinates (see Section 5.2.1), the sampled demes must fall within the extent of the world defined by the environmental files and they should not fall in "sea" (NoData) demes. If one or several sampled-subpopulation fall in the "sea", an error message will appear in the log file (and as a message box in the graphical version), but the simulation will still take place with the reminding sampled sub-populations.

6. Date of the sample given in generations after the onset of the population expansion.

Example of a genetic input file (.sam) for 6 samples in Africa,

```
6
#Name         #Size  #PopLayer     #Lat   #Long         #Date
sample1       30     0             20     20            400
sample2       25     0             20     0             400
sample3       28     0             0      20            400
sample4       32     0             -20    20            350
sample5       30     0             -30    25            300
sample6       40     0             5      40            100
```

5.2.10 Ancient DNA

In order to simulate ancient samples with SPLATCHE3, it is necessary to specify in the input file defined by the parameter "SampleFile" the time at which each sample must be taken. It corresponds to the 6th column.

Generation 0 means the beginning of the simulations and it can go till the last simulated generation. In the example file above, the duration of the simulation is 400 generations, so if the user wants to sample at present for this example, the user has to put 400 at the end of the corresponding sample line. If one wants to sample 100 generations before the present (300 generations after the beginning of the simulation in this example), one has to put 300 at the end of the line, etc.

Note that if this last column is missing or the number given is larger than the number of simulated generations, then the last generation is set by default (400 in the example below).

5.2.11 GeneticFile

Path to a text file (*.par) containing the definition of markers property and recombination parameters. The format of this file is identical to the genetic marker section of SIMCOAL2 (Laval and Excoffier 2004) input files, but including some additional parameters for the simulation of DNA sequences under a variety of mutation models of evolution.

Each line corresponds to one of the following elements:

1. Number of independent chromosome segments: The first line refers to the number of independent chromosome segments that needs to be simulated. It is possible to simulate several chromosomes with different kind and numbers of markers, different patterns of recombination rates and different mutation rates. Note however that ARLEQUIN/ARLSUMSTAT (Excoffier and Lischer 2010) is not able to analyze various kinds of markers in the same file.

2. Number of blocks: For each chromosome, the number of blocks must be defined. This number refers to the number of sets of partially linked loci (that we shall call blocks, hereafter) of the same data type, having the same mutation and recombination rates. For instance, to simulate *L* partially linked microsatellites with varying recombination rates and with different mutation rates, we need to specify *L* blocks, each one containing one microsatellite. To simulate a chromosome segment with *L* partially linked microsatellites uniformly spaced (with identical recombination rates between adjacent loci) and having identical mutation rates, one would only need to define a single block.

3. Block specificities: For every block, we need to specify 3 mandatory parameters, as follows:

3.1. Data type.
3.2. Number of loci to simulate.
3.3. Recombination rate immediately to the right of each locus.

In addition to these 3 required parameters, several additional parameters need to be added depending on the data type**.**

Additional parameters for RFLP:
3.4. Mutation rate per locus.

Additional parameters for STR (microsatellite):
3.4. Mutation rate per locus. A pure stepwise mutation model is used with possible range constraints.

3.5. Geometric parameter for the GSM model. Ranges between 0 and 1, a value of 0 implies no insertion/deletion of more than one repeat (strict SMM model).

3.6. Range constraint. Number of allowed different alleles. 0 means no range constraint.

Additional parameters for SNP:
3.4. Minimum frequency for the derived allele (MAF).

Additional parameters for DNA:
3.4. Mutation rate per locus.

The following parameters are required to define the mutation model of DNA evolution. They include the 4 base frequencies and the 6 relative rates of change, which are required to obtain the exchangeability matrix (Yang 2006).

3.5. Frequency of nucleotide A. It should be a value between 0 and 1.

3.6. Frequency of nucleotide C. It should be a value between 0 and 1.

3.7. Frequency of nucleotide G. It should be a value between 0 and 1.

3.8. Frequency of nucleotide T. It should be a value between 0 and 1.

3.9. Relative mutation rate A↔C. It should be a positive value.

3.10. Relative mutation rate A↔G. It should be a positive value.

3.11. Relative mutation rate A↔T. It should be a positive value.

3.12. Relative mutation rate C↔G. It should be a positive value.

3.13. Relative mutation rate C↔T. It should be a positive value.

3.14. Relative mutation rate G↔T. It should be a positive value.

These parameters allow the consideration of any reversible mutation model of DNA evolution. Next, we describe some examples of commonly used models:

JC (Jukes and Cantor 1969): This is the most basic model where all nucleotide frequencies are equal (0.25) and all relative mutation rates are equal (1.0): fA=fC=fG=fT=0.25; rAC=rAG=rAT=rCG=rCT=rGT=1.0. Note that this model does not distinguish between transition and transversion mutations (transitions and transversions are equally likely). It was implemented in the previous version of SPLATCHE.

K80 (Kimura 1980): The JC model can be complicated by considering a distinction between transition and transversion mutations through a transition/transversion ratio *titv*: fA=fC=fG=fT=0.25; rAG=rCT=1.0 (transition) and rAC=rAT=rCG=rGT=titv (transversion). This model was implemented in the previous version of Splatche.
For example (titv=3.0), fA=fC=fG=fT=0.25; rAG=rCT=1.0;
rAC=rAT=rCG=rGT=3.0.

F81 (Felsenstein 1981): The JC model can be complicated by considering a distinction between base frequencies instead of the relative rates of change. For example, fA=0.35, fC=0.15, fG=0.20, fT=0.30; rAC=rAG=rAT=rCG=rCT=rGT=1.0.

HKY (Hasegawa, Kishino et al. 1985): K80 and F81 models can be combined to account for different base frequencies and transition and transversion mutations (*titv).* For example (titv=3.0), fA=0.35, fC=0.15, fG=0.20, fT=0.30; rAG=rCT=1.0; rAC=rAT=rCG=rGT=3.0.

SYM (Zharkikh 1994): This model considers equal base frequencies and different relative rates of change. For example, fA=fC=fG=fT=0.25; rAG=2.1, rCT=0.7; rAC=0.9, rAT=3.2, rCG=2.0, rGT=1.0.

GTR (Tavaré 1986): This is the most complex model of DNA evolution and it considers different base frequencies and different relative rates of change. For example, fA=0.35, fC=0.15, fG=0.20, fT=0.30; rAG=2.1, rCT=0.7; rAC=0.9, rAT=3.2, rCG=2.0, rGT=1.0.

In addition to the cited models, many other models can be specified by specifying different combinations of these parameters (Lio and Goldman 1998). Note that the best fitting model for a given multiple sequence alignment can be obtained with evolutionary frameworks such as jModelTest2 (Darriba, Taboada et al. 2012). SPLATCHE3 allows simulating sequence segments with different models of DNA evolution, a capability convenient to mimic real evolutionary processes (Arbiza, Patricio et al. 2011).

In summary, the input parameters for every marker are the following:

STR *[number of loci] [recombination rate] [mutation rate] [Geometric parameter] [range constraint]*

SNP *[number of loci] [recombination rate] [minimum frequency]*

DNA *[number of loci] [recombination rate] [mutation rate] [fA] [fC] [fG] [fT] [rAC] [rAG] [rAT] [rCG] [rCT] [rGT]*

RFLP *[number of loci] [recombination rate] [mutation rate]*

Next we describe the meaning of these parameters:

[number of loci] : Number of partially linked loci in a block

[recombination rate] : recombination rate between adjacent markers in a block (Global recombination rate of the block = [recombination rate] *([number of loci] -1)).

[mutation rate] : mutation rate for every marker in a block. (Global mutation rate of the block = [mutation rate] *[number of loci]).

[Geometric parameter]: parameter *p* for Generalized Stepwise Mutation *GSM(model . The number of repeats inserted (deleted) in one mutation event is treated as a random variable, and is randomly drawn from a Geometric distribution with parameter *p*. It may vary between 0 and 1 (0 means that the number of repeats inserted or deleted for **every** mutation event is 1 while 0.5 means that the **expected** number of repeats is 2). See http://cmpg.unibe.ch/software/simcoal2/#GSM for more details.

[range constraint] : 0 means no constraints, higher numbers give the maximum possible number of repeats.

[minimum frequency] : minimum frequency of a SNP in the whole set of population (ascertainment bias).

[fA] [fC] [fG] [fT] : base frequencies. Each base frequency should vary between 0 and 1. The sum of all base frequencies should be equal to 1 (*fA+fC+fG+fT=1*).

[rAC] [rAG] [rAT] [rCG] [rCT] [rGT] : relative mutation rates. They should be positive values. The reference value is 1.0 (Lio and Goldman 1998).

Here is an example of a genetic file for two chromosomes (independent loci) constituted by DNA sequences of 100 base pairs, no recombination, various mutation rates and two different evolutionary models (K80 and GTR, respectively):

```
2 //Number of independent chromosomes
#chromosome 1
1
DNA 100 0.0 0.001 0.25 0.25 0.25 0.25 3.0 1.0 3.0 3.0 1.0 3.0
#chromosome 2
1
DNA 100 0.0 0.0000001 0.15 0.35 0.40 0.10 0.60 1.30 0.47 2.75 4.10 1.00
```

Additional examples of genetic input files for the other genetic markers can be found in the web site of the program SIMCOAL2: ***http://cmpg.unibe.ch/software/simcoal2***

5.2.12 Coalescent distribution

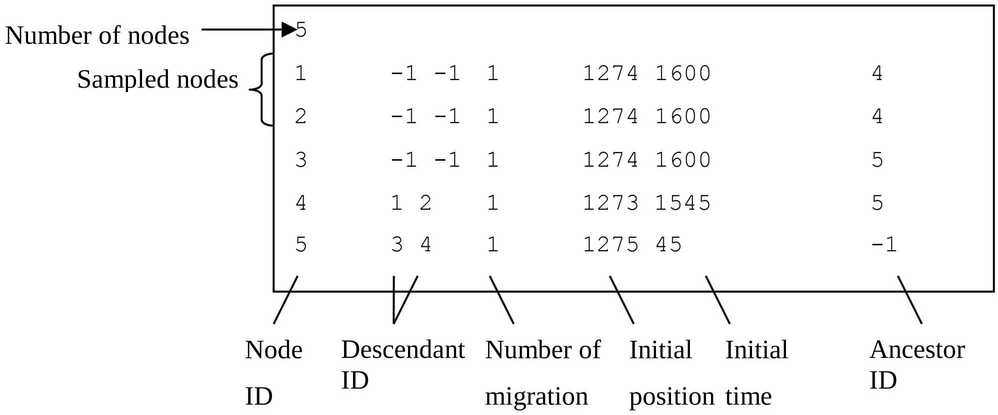
In settings the checkbox “Generate bitmap” on the graphical interface, it is possible to generate (only for the last chromosome simulated) bitmaps representing the location of genes either over densities of layer1 or layer2 (choice made in the “General Output” panel) or alternatively over occupation. Bitmaps can be generated every *n* generations in color or grayscale. Those file names start with*”.genetic”* and end with *”.bmp”*.

5.2.13 Coalescent tree files

In settings the checkbox “coalescent trees” on the graphical interface, it is possible to generate for each simulation a bitmap representing the branches of the coalescence tree for the last chromosome simulated, laid out spatially. These files are terminated with *“*_CoalTree_*.bmp”*. This feature is only available with one population layer and without recombination (the file is not generated when recombination is used). If more than one chromosome is simulated, the file is generated only for the last chromosome without recombination.

5.2.14 GenealogiesFile

This setting controls whether a “*.tri” file(s) is generated to store the geographic position of the coalescent events having occurred in the coalescent process, as well as the migration of the nodes between demes. These files (one per chromosome) are only generated when no recombination is used. With one layer, the *tri files looks as follow:

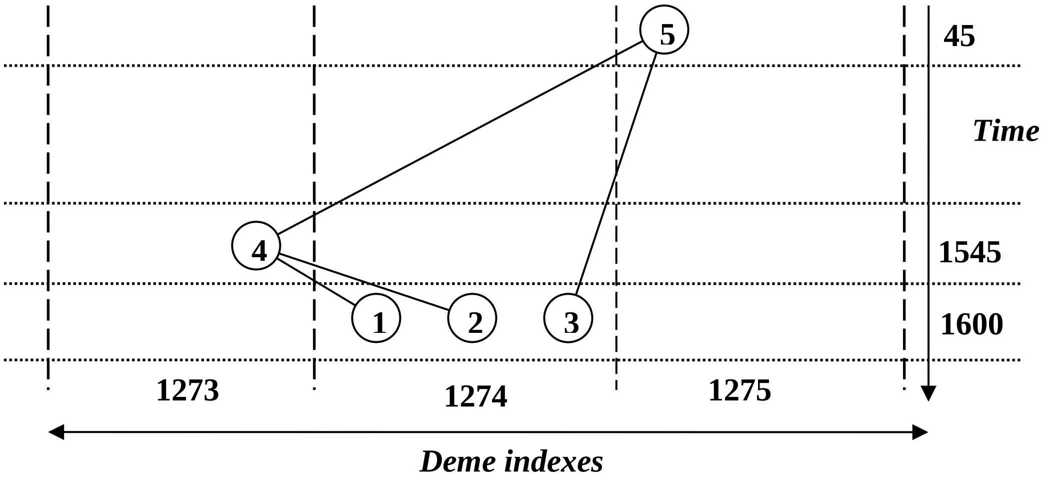


```
5
1      -1 -1  1      1274 1600          4
2      -1 -1  1      1274 1600          4
3      -1 -1  1      1274 1600          5
4      1  2   1      1273 1545          5
5      3  4   1      1275 45            -1
```

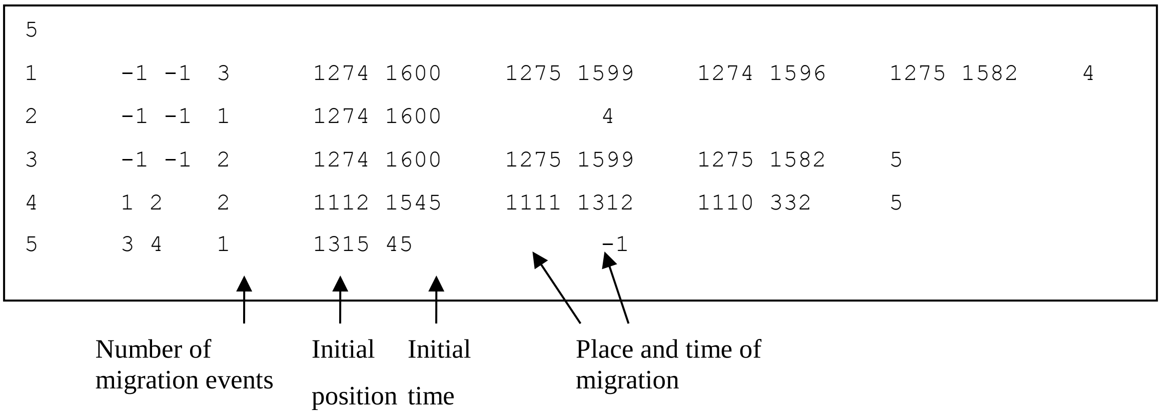
The first column of the first line corresponds to the number *n* of nodes (5 in this case, the first three nodes are sampled genes). For each node, with IDs between 1 and *n*, corresponds a line where the numbers are:

1. Node ID
2. Descendent nodes ID (-1 if the node is sampled)
3. Number of migration events for this node. This number is always equal to 1, and thus uninformative except when the parameters *DoublePopulationMode* and *PropFile* are both set to 1 (see below).
4. Index of the cell in which the node appears.
5. Time (in generations) at which the node appears. In the case of our example, the simulation ends after 1,600 generations, which is the current sampling time.
6. Ancestor node ID (-1 for the MRCA).

For the “*.tri” file example mentioned above, see below a schematic representation of the corresponding coalescent tree displayed over space and time.



In the particular case of two populations (*DoublePopulationMode*=1) and if the setting *PropFile* is set to 1, then every node migrations are stored into the file as follows:



```
5
1      -1 -1  3      1274 1600     1275 1599     1274 1596     1275 1582     4
2      -1 -1  1      1274 1600          4
3      -1 -1  2      1274 1600     1275 1599     1275 1582     5
4      1 2    2      1112 1545     1111 1312     1110 332      5
5      3 4    1      1315 45             -1
```

Number of migration events
Initial position
Initial time
Place and time of migration

This example shows that node 1 migrated 3 times, from deme 1274 to deme 1275 at time 1599, from 1275 to deme 1274 at time 1596 and finally from deme 1274 to deme 1275 at time 1582, etc.

5.2.15 ImmigDistrFile

This setting controls whether a file is created with the number of immigrants in sampled demes for every generations. One file ”*.immigrants_DemeIndex.nm” is created per sampled deme if this setting is equal to 1.

The first column corresponds to the generations and the second one to the number of individuals that arrived in the current deme generation after generation.

```
Generation Immigrants
1      0
2      0
3      0
4      2
5      3
6      5
7      8
8      8
9      8
```

5.2.16 Admixture rate from P1 to P2 and Admixture rate from P2 to P1

This is the rate of gene flow from demes belonging to the first population layer (P1) into demes belonging to the second population layer (P2) or the reverse. It represents admixture or hybridization and could only occur between demes belonging to an identical geographical cell (same location in both arrays of demes). This parameter corresponds to γ_{ij} in the equations in section 4.5. This admixture rate can be set between 0 (no admixture) and 1 (full interbreeding). If the value of *Admixture rate from P1 to P2* is different from the value of *Admixture rate from P2 to P1* then the admixture between both layers is asymmetrical.

5.2.17 PropFile (*.prop)

When simulating two populations (*DoublePopulationMode*=1), it is possible to generate a file that gives the proportion of genes sampled in layer two (P2) that comes originally from the source deme of that layer. In other words, this is the proportion of genes in the invasive population P2 which have not been introgressed by local P1 genes. See (Currat, Ruedi et al. 2008) for more details.

One file is generated for each simulation. In each file, there is one line per chromosome in which the proportions of genes sampled in layer two (P2) that comes originally from the source deme of that layer, are given for all sampled genes taken together, and also for each sample separately.

```
Proportion of lineages sampled in P2 which derived from the P2 source
population
#Chrom Total_prop    1      2      3      4      5
Chrom_1      0.415          0.5    0.3    0.45   0.65   0.4
Chrom_2 Not available with recombination!
```

5.3 *Other generated outputs*

5.3.1 Log files

When running SPLATCHE3, information and potential errors are written in a log file called "splatche3_win.log" for the GUI, and "splatche3.log" for Linux console version.

5.3.2 ARLEQUIN output file

The genetic data generated by one simulation are directly output in an ARLEQUIN project file, with the extension "*.*arp*". This file format allows one to analyse the data with the ARLEQUIN program in order to obtain different statistics, see ARLEQUIN manual (Excoffier and Lischer 2010) and website (http://popgen.unibe.ch/software/arlequin35) for more details. If more than one simulation is performed using one demographic simulation (which is usually the case) then an ARLEQUIN batch file (with extension "*.*arb*") is also generated, listing all simulated files, and allowing one to compute statistics on the whole set of simulated files. Note also that the ARLEQUIN software has a file conversion utility for exporting input data files into several other format like BIOSYS, PHYLIP, or GENEPOP, so that files produced by SPLATCHE3 could be also analyzed by these software after file conversion. Alternatively, PGDSpider (http://www.cmpg.unibe.ch/software/PGDSpider/), a program developed by the CMPG lab, can also be used to convert population genetic data between many formats.

Note that different types of markers can be simulated at the same time in SPLATCHE3 (see how to do it in Section 5.2.11). These simulated markers will be output in the ARLEQUIN files, but the current version of ARLEQUIN (ver. 3.5) is not able to read and further analyze different set of markers. The user of such an output file would therefore need to develop his/her own routines to treat separately the different sets of markers.

The output ARLEQUIN files will be located in a folder called "GeneticsOutput" located in the folder holding the "*.sam" and "*par" files.

5.3.3 Coalescence distribution

With the graphical version, a bitmap representing the spatial distribution of the coalescent events over all simulations (and chromosomes) is automatically created with the "*_TotNumCoal.bmp" suffix. This bitmap can also be visualized by means of the button "Draw Coalescence" on the graphical interface. The bitmap is only created in absence of recombination events.

By checking the "coalescent spatial distribution" checkbox on the graphical interface, similar bitmaps of the spatial distribution of coalescent events are generated for every simulation (with the "*_NumCoal_simi_chromj.bmp" termination).

By checking the "coalescent temporal distribution" checkbox on the graphical interface, times for each coalescent event and each simulation are listed on a file with "*.coal" extension. Those times are indicated in generation units, with the largest number corresponding to the end time of the simulation.

5.3.4 MRCA files

With the graphical version, SPLATCHE also gives information on the location and the timing of the Most Recent Common Ancestors (MRCAs) of the sampled genes,. A bitmap file with the termination "*_TotMRCADensity.bmp" is automatically generated and shows the spatial distribution of MRCA for all simulations. These maps can also be visualized by checking the button "Draw MRCA" on the graphical interface. These maps are only created when no recombination is used.

5.3.5 Arrival cell files

A file called "Arrival_cell_output.txt" is generated, listing the arrival time (in generation) in the cells specified in the "*.col" input file.

5.3.6 Nexus

Two other types of file produced by SPLATCHE are compatible with the NEXUS (Newick) tree file format: for each simulation, a file with "*.trees" extension could be generated. This file lists all the simulated genes together with their true genealogical structure (trees are provided in Newick format). This file can be analyzed by many programs (e.g. with David Swofford's PAUP* software http://paup.csit.fsu.edu or with FigTree http://tree.bio.ed.ac.uk/software/figtree/). A batch file, with extension "*.bat" is also generated.

6 References

- Alves, I., M. Arenas, M. Currat, A. Sramkova Hanulova, V. C. Sousa, N. Ray and L. Excoffier (2016). "Long-Distance Dispersal Shaped Patterns of Human Genetic Diversity in Eurasia." Mol Biol Evol **33**(4): 946-958.
- Arbiza, L., M. Patricio, H. Dopazo and D. Posada (2011). "Genome-Wide Heterogeneity of Nucleotide Substitution Model Fit." Genome Biology and Evolution **3**: 896-908.
- Arenas, M., O. Francois, M. Currat, N. Ray and L. Excoffier (2013). "Influence of admixture and paleolithic range contractions on current European diversity gradients." Mol Biol Evol **30**(1): 57-61.
- Arenas, M., N. Ray, M. Currat and L. Excoffier (2012). "Consequences of range contractions and range shifts on molecular diversity." Mol Biol Evol **29**(1): 207-218.
- Cavalli-Sforza, L. L., M. Kimura and I. Barrai (1966). "The probability of consanguineous marriages." Genetics **54**(1): 37-60.
- Currat, M. and L. Excoffier (2004). "Modern humans did not admix with Neanderthals during their range expansion into Europe." PLoS Biol **2**(12): 2264-2274.
- Currat, M. and L. Excoffier (2005). "The effect of the Neolithic expansion on European molecular diversity." Proc Biol Sci **272**(1564): 679-688.
- Currat, M., L. Excoffier, W. Maddison, S. P. Otto, N. Ray, M. C. Whitlock and S. Yeaman (2006). "Comment on "Ongoing adaptive evolution of ASPM, a brain size determinant in Homo sapiens" and "Microcephalin, a gene regulating brain size, continues to evolve adaptively in humans"." Science **313**(5784): 172; author reply 172.
- Currat, M., N. Ray and L. Excoffier (2004). "SPLATCHE: a program to simulate genetic diversity taking into account environmental heterogeneity." Molecular Ecology Notes **4**(1): 139-142.
- Currat, M., M. Ruedi, R. J. Petit and L. Excoffier (2008). "The hidden side of invasions: massive introgression by local genes." Evolution **62**(8): 1908-1920.
- Darriba, D., G. L. Taboada, R. Doallo and D. Posada (2012). "jModelTest 2: more models, new heuristics and parallel computing." Nature Methods **9**(8): 772-772.
- Excoffier, L. and H. Lischer (2010). "Arlequin suite ver 3.5: a new series of programs to perform population genetics analyses under Linux and Windows." Molecular Ecology Resources **10**: 564-567.
- Felsenstein, J. (1981). "Evolutionary Trees from DNA-Sequences - a Maximum-Likelihood Approach." Journal of Molecular Evolution **17**(6): 368-376.
- Foll, M. and O. Gaggiotti (2006). "Identifying the Environmental Factors That Determine the Genetic Structure of Populations." Genetics **174**(2): 875-891.
- Hamilton, G., M. Currat, N. Ray, G. Heckel, M. Beaumont and L. Excoffier (2005). "Bayesian estimation of recent migration rates after a spatial expansion." Genetics **170**(1): 409-417.
- Hasegawa, M., H. Kishino and T. A. Yano (1985). "Dating of the Human Ape Splitting by a Molecular Clock of Mitochondrial-DNA." Journal of Molecular Evolution **22**(2): 160-174.
- Hewitt, G. M. (2004). "Genetic consequences of climatic oscillations in the Quaternary." Philos Trans R Soc Lond B Biol Sci **359**(1442): 183-195; discussion 195.

- Hudson, R. (1990). Gene genealogies and the coalescent process. Oxford, Oxford University Press.
- Jukes, T. and C. Cantor (1969). Evolution of protein molecules. Mamalian Protein Metabolism. H. N. Munro. New York, Academic press**:** 21-132.
- Kimura, M. (1980). "A simple method for estimating evolutionary rate of base substitution through comparative studies of nucleotide sequences." J. Mol. Evol. **16**: 111-120.
- Kimura, M. and W. H. Weiss (1964). "The stepping stone model of genetic structure and the decrease of genetic correlation with distance." Genetics **49**: 561-576.
- Klopfstein, S., M. Currat and L. Excoffier (2006). "The fate of mutations surfing on the wave of a range expansion." Mol Biol Evol **23**(3): 482-490.
- Laval, G. and L. Excoffier (2004). "SIMCOAL 2.0: a program to simulate genomic diversity over large recombining regions in a subdivided population with a complex history." Bioinformatics **20**(15): 1485-2487.
- Lio, P. and N. Goldman (1998). "Models of molecular evolution and phylogeny." Genome Research **8**(12): 1233-1244.
- Lotka, A. J. (1932). "The growth of mixed populations: two species competing for a common food supply." Journal of the Washington academy of Sciences **22**: 461-469.
- Novembre, J., A. P. Galvani and M. Slatkin (2005). "The geographic spread of the CCR5 Delta32 HIV-resistance allele." PLoS Biol **3**(11): e339.
- Quemere, E., B. Crouau-Roy, C. Rabarivola, E. E. Louis, Jr. and L. Chikhi (2010). "Landscape genetics of an endangered lemur (Propithecus tattersalli) within its entire fragmented range." Mol Ecol **19**(8): 1606-1621.
- Ray, N. (2003). Modélisation de la démographie des populations humaines préhistoriques à l'aide de données environnementales et génétiques Thèse, Université de Genève, Switzerland.
- Ray, N., M. Currat, P. Berthier and L. Excoffier (2005). "Recovering the geographic origin of early modern humans by realistic and spatially explicit simulations." Genome Research **15**(8): 1161-1167.
- Ray, N., M. Currat and L. Excoffier (2003). "Intra-deme molecular diversity in spatially expanding populations." Molecular Biology and Evolution **20**(1): 76-86.
- Ray, N., M. Currat, M. Foll and L. Excoffier (2010). "SPLATCHE2: a spatially-explicit simulation framework for complex demography, genetic admixture and recombination." Bioinformatics.
- Ray, N. and L. Excoffier (2010). "A first step towards inferring levels of long-distance dispersal during past expansions." Mol Ecol Resour **10**(5): 902-914.
- Taberlet, P., L. Fumagalli, A. G. Wust-Saucy and J. F. Cosson (1998). "Comparative phylogeography and postglacial colonization routes in Europe." Molecular Ecology **7**(4): 453-464.
- Tavaré, S. (1986). Some probabilistic and statistical problems in the analysis of DNA sequences. Some mathematical questions in biology - DNA sequence analysis. R. M. Miura, Amer Math Soc**:** 57-86.
- Volterra, V. (1926). Variations and fluctuations of the numbers of individuals in animal species living together (Reprinted in 1931). Animal Ecology. R. N. Chapman. New York, Mc Graw Hill**:** 409-448.
- Wegmann, D., M. Currat and L. Excoffier (2006). "Molecular diversity after a range expansion in heterogeneous environments." Genetics **174**: 2009-2020.

- Yang, Z. (2006). Computational Molecular Evolution, Oxford University Press.
- Zhang, D. Y. (2014). "Demographic model of admixture predicts symmetric introgression when a species expands into the range of another: A comment on Currat et al. (2008)." Journal of Systematics and Evolution **52**(1): 35-39.
- Zharkikh, A. (1994). "Estimation of Evolutionary Distances between Nucleotide-Sequences." Journal of Molecular Evolution **39**(3): 315-329.