Population Genomics:

background, tools and programing

Practical: "Coalescent simulations in time and space"

April 1st 2019 (15:45-17:45)

Mathias Currat (with the help of Jérémy Rio)

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

Email: mathias.currat @unige.ch

Table of Contents

GOAL

The goal of this practical is to learn the main principles of spatially explicit simulations of serial molecular data using the program SPLATCHE3. The simulated data will be used to test for population continuity through time.

Through a series of steps, you will generate data similar to a real dataset of ancient mitochondrial DNA under a null hypothesis of population continuity. You will then test if the null model of population continuity could be rejected with your observed data. The goal of the practical is to decide whether the two serial samples taken from the same location could be considered as coming from one single population evolving through time or if an alternative event, such as the arrival of immigrants, must be invoked between the two sampling times.

PROGRAMS

1- SPLATCHE3

Short description: SPatiaL And Temporal Coalescent in a Heterogeneous Environment, version 3

Download: http://www.splatche.com/splatche3

Documentation: http://www.splatche.com/download/SPLATCHE3_User_Manual.pdf

Reference: Currat, M., Arenas, M., Quilodran C.S., Excoffier L., and Ray N. (submitted) SPLATCHE3: simulation of serial genetic data under spatially explicit evolutionary scenarios including long-distance dispersal.

Input files: *.txt

Executable name: SPLATCHE3

2- Arlsumstat

Short description: An Integrated Software for Population Genetics Data Analysis (efficient, command line version of the software ARLEQUIN)

Download: http://cmpg.unibe.ch/software/arlequin35/Arl35Downloads.html

Documentation: http://cmpg.unibe.ch/software/arlequin35/man/Arlequin35.pdf
http://cmpg.unibe.ch/software/arlequin35/man/arlsumstat_readme.txt

Reference: Excoffier, L. and H.E. L. 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.

Input files: arl_run.ars, ssdefs.txt

Executable name: arlsumstat

3- ABCtoolBox

Short description: A general-purpose program to perform Approximate Bayesian Computation.

Download: https://bitbucket.org/phaentu/abctoolbox-public/downloads/

Documentation: https://bitbucket.org/phaentu/abctoolbox-public/wiki/Home

http://cmpg.unibe.ch/software/ABCtoolbox/ABCtoolbox_manual.pdf

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

Input files: *.obs, *.input, *.est

Executable name: ABCtoolbox2

STEP 1: GRAPHICAL SIMULATION OF POPULATION CONTINUITY

First, you are going to simulate a demographic scenario representing population continuity in Europe during ~40,000 years using the program **SPLATCHE3** (executable name = **SPLATCHE3_GUI_Win.ex**). The program user manual "*SPLATCHE3_User_Manual.pdf*" located in "/data/mathias/manuals/" may help you if needed.

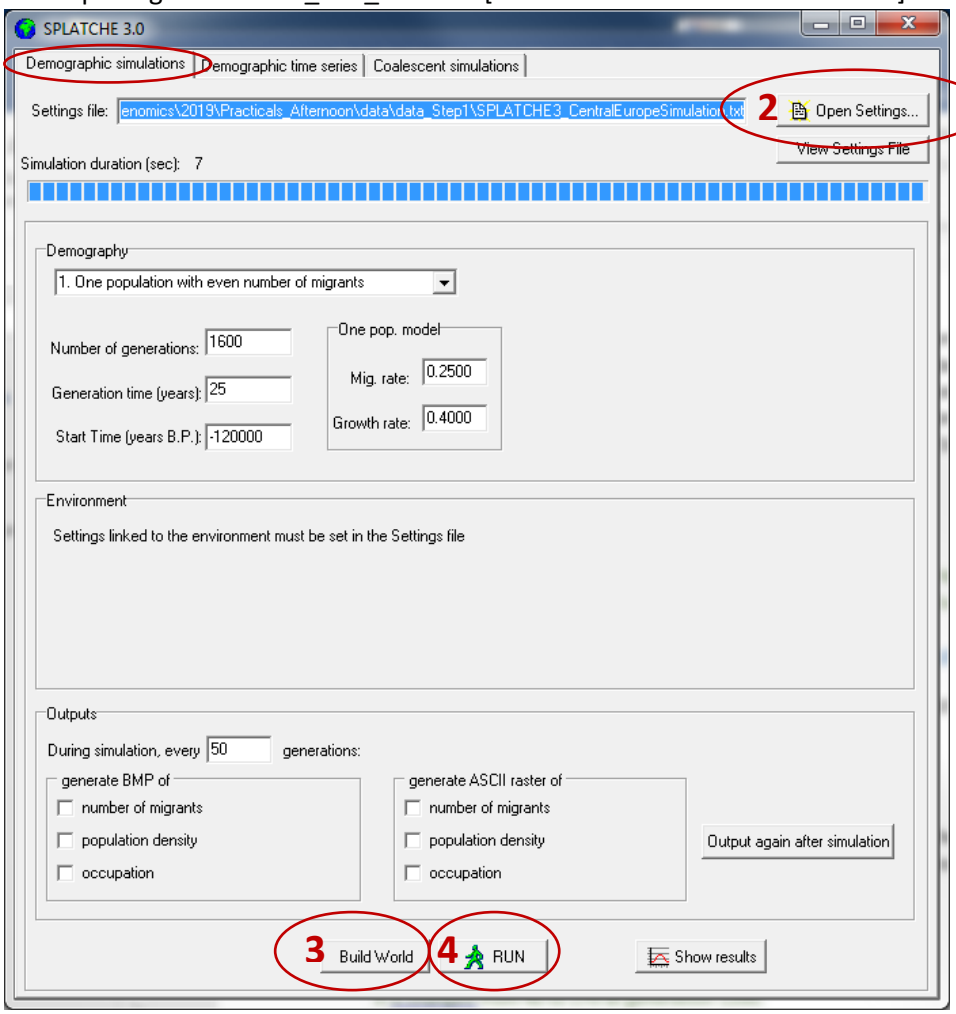
Exercise:

1. Create a working folder called "*step1*" and copy the files provided in the folder *"/data/mathias/afternoon/data_step1"*. These input files allow simulating with SPLATCHE3 a scenario of colonization of the European continent by hunter-gatherers starting ~40,000 years ago, followed by an increase in population size in all demes at the same time ~10,000 years ago. This scenario is based on Silva et al. BMC Genetics (2017).

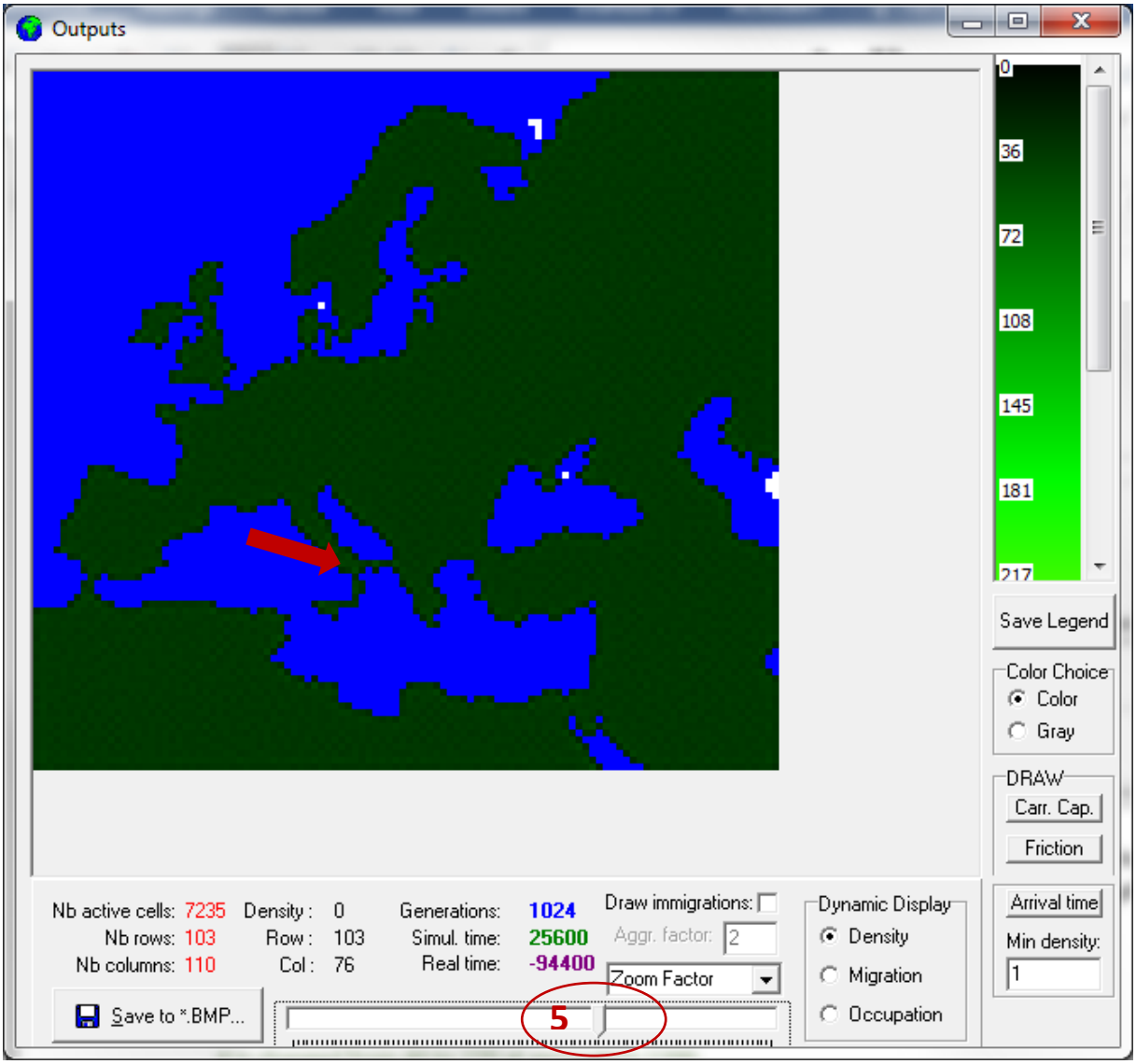
```
mkdir step1
cd step1
cp /data/mathias/afternoon/data_step1/SPLATCHE3_CentralEuropeSimulation.txt .
cp –r /data/mathias/afternoon/data_step1/SPLATCHE3_CentralEuropeSimulation .
```

2. Download the GUI version of SPLATCHE3 in the web site www.splatche.com/splatche3 and launch it. Then run a demographic simulation by

1. Opening SPLATCHE3_GUI_Win.exe [WORKS ON MS WINDOWS ONLY !]



2. Click on the button “Open settings” and select the file *“SPLATCHE3_CentralEuropeSimulation.txt”*
3. Click on the button “build world”
4. Click on the button “RUN”
5. Once the demographic simulation is finished, then scroll the map to inspect visually the scenario simulated. Click on any deme on the map to inspect its demography through time.



Questions:

- What is the carrying capacity value set at the time of the Neolithic transition?
  K is changed from 40 to 268 at generation 1200.

Remarks:

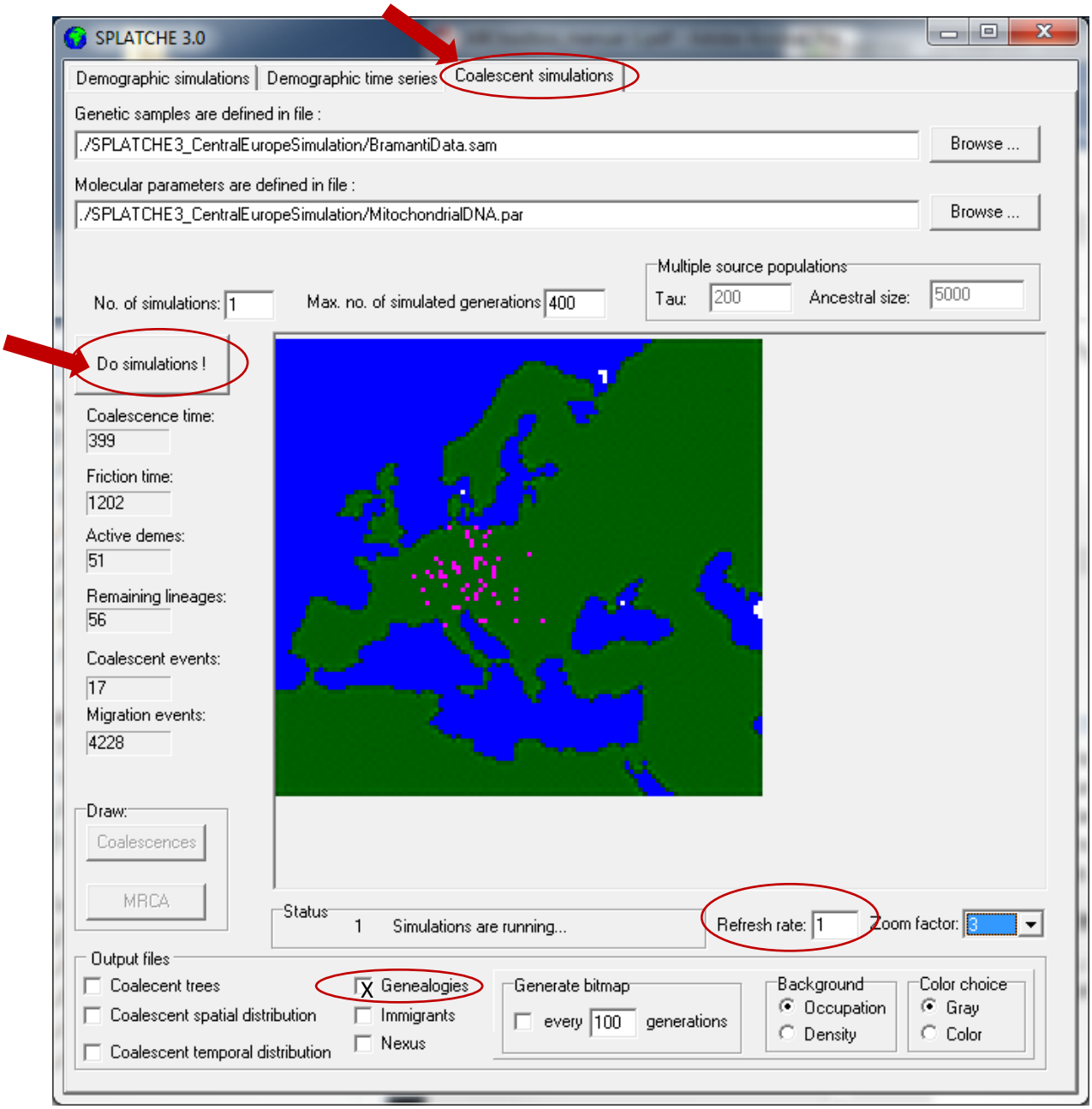
- If you don’t have a MS WINDOWS OS in your computer, then just follow the demonstration.
- SPLATCHE3 produces a “*.log” file where you can find indications about the state of the simulations.

STEP 2: GRAPHICAL SIMULATION OF GENETIC DIVERSITY

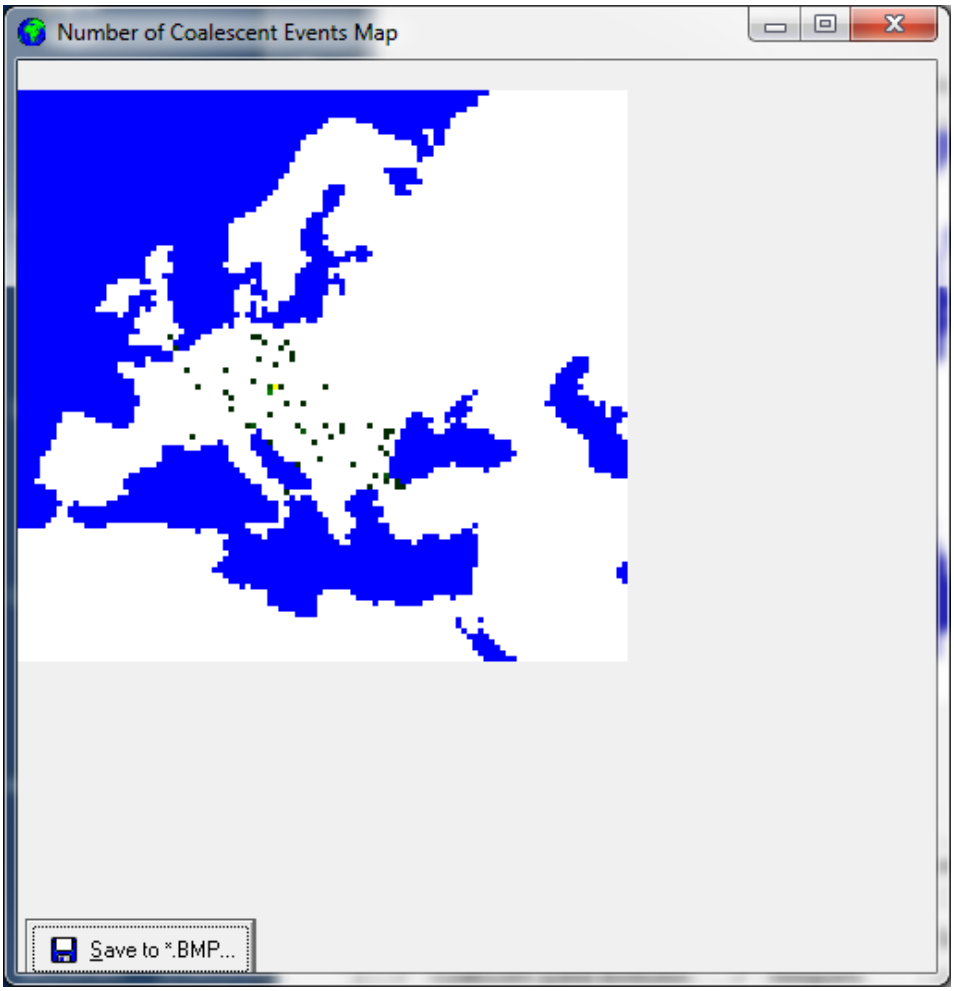
You are going to simulate genetic diversity for a series of samples under the demographic scenario simulated during step1. The samples consist of mitochondrial DNA sequences from 50 modern Germans (Baasner et al. Forensic Sci Int. 1998;98(3):169–78) and 23 mitochondrial DNA sequences from central European early Neolithic farmers dating from around 6,775-7,500 years ago (Bramanti et al Science. 2009;326:137–40.).

Exercise:

1. On the graphical instance of SPLATCHE3 opened during step1, select the panel "Coalescent simulations", set the "Refresh rate" to 1, check the box entitled "Genealogies" and click on the button "Do simulations!"



2. Once the coalescent simulation is finished, you can click on the button “Coalescences” to visualize were the coalescent events took place. If you perform more than one simulation, then the map will cumulate the coalescent events for all the simulations and thus be more instructive about the spatial behavior of the lineages.



If you look to the results folder called “GeneticsOutput” located in the folder “*SPLATCHE3_CentralEuropeSimulation”*, you will find a file with the extension “*.tri*” that contains all spatial and temporal information about the coalescent events.

More importantly, you will also find a file with the extension “*.arp*”, which contains the mitochondrial sequences simulated in ARLEQUIN format.

Have a look at those output files

Questions:

- What represent the pink dots moving on the map when the genetic simulation is running?
  Each pink dot represents a deme where at least one sampled individual or one of their ancestors (i.e. coalescent lineages) is located at the current time. It allows following in “real” time the movements of lineages.

STEP 3: SIMULATION USING THE CONSOLE LINUX VERSION

Now you will make the same simulation but on Linux environment using the corresponding console version (executable name = **SPLATCHE3**).

Exercise:

1. Create a working folder called "*step3*" and copy the files provided in the folder *"/data/mathias/afternoon/data_step1"*.

```
mkdir step3
cd step3
cp /data/mathias/afternoon/data_step1/SPLATCHE3_CentralEuropeSimulation.txt .
cp -r /data/mathias/afternoon/data_step1/SPLATCHE3_CentralEuropeSimulation .
```

2. Have a look to the settings file "*SPLATCHE3_CentralEuropeSimulation.txt*" and to the files in the folder *"SPLATCHE3_CentralEuropeSimulation"*. All the parameters for the simulations are listed in those input files.

3. Launch SPLATCHE3 console version using the command line

```
SPLATCHE3 SPLATCHE3_CentralEuropeSimulation.txt
```

4. Once the simulation is done, have a look on the results folder "*GeneticsOutput*" which has been created.

5. For this practical, we use mitochondrial data but SPLATCHE3 allows one to simulate nuclear data as well. To simulate nuclear loci, you must do the following changes in the main setting file and save it as *"SPLATCHE3_CentralEuropeSimulation_Nuc.txt"*:

- set the parameter `GenotypicData` to 1
- change the `GeneticFile` parameter to call the file *"NuclearSNP.par"* to simulate nuclear SNP (for example)
- change the `SampleFile` parameter to call the file "*BramantiDataNuc.sam*" to sample two gene copies instead of 1 per individual
- called different files for setting the carrying capacity: "*vegK_L1_time_1_Nuc.txt*" instead of "*vegK_L1_time_1.txt*" and "*vegK_L1_time_2_Nuc.txt*" instead of "*vegK_L1_time_2.txt*".

You need to increase the carrying capacity values by four to reflect nuclear effective size instead of mitochondrial effective size, in the corresponding files "*vegK_L1_time_1_Nuc.txt*" and "*vegK_L1_time_2_Nuc.txt*".

```
cat SPLATCHE3_CentralEuropeSimulation/vegK_L1_time_1.txt | sed 's/40/160/g' >
SPLATCHE3_CentralEuropeSimulation/vegK_L1_time_1_Nuc.txt

cat SPLATCHE3_CentralEuropeSimulation/vegK_L1_time_2.txt | sed 's/268/1072/g' >
SPLATCHE3_CentralEuropeSimulation/vegK_L1_time_2_Nuc.txt
```

Finally, launch SPLATCHE3 on this new setting file and look at the changes in the results.

```
SPLATCHE3 SPLATCHE3_CentralEuropeSimulation_Nuc.txt
```

Questions:

- Which version is the most rapid, the GUI or the console?
  The console version is much quicker; this is the reason why the GUI is useful to set up the scenarios while the console version is useful for extensive simulations
- Is there any difference in the input and output files produced by the two version of SPLATCHE3 (GUI and console)?
  No, there is no differences, both versions use the exact same input files and produce the same output files. The only difference is that some graphical outputs (such as .bmp files) are not produced by the console version to increase its speed.
- Look at the file NuclearSNP.par, how many SNP are simulated here?
  20 independent SNP are simulated

STEP 4: MERGE ANCIENT SAMPLES

As you have seen in the ARLEQUIN file (".arp") produced by SPLATCHE3, ancient lineages constitute different groups depending on their location and age. We would like to merge them into one single "ancient" population sample before computing statistics of gene diversity (statistics H) and of genetic differentiation with the modern population sample (statistics *Fst*).

Exercise:

1. Create a working folder called "*step4*" and copy the files provided in the folder *"/data/mathias/afternoon/data_step1"*.

```
mkdir step4
cd step4
cp /data/mathias/afternoon/data_step1/SPLATCHE3_CentralEuropeSimulation.txt .
cp -r /data/mathias/afternoon/data_step1/SPLATCHE3_CentralEuropeSimulation .
```

2. Launch SPLATCHE3 console version using the command line

```
SPLATCHE3 SPLATCHE3_CentralEuropeSimulation.txt
```

3. Look at the ".arp" file produced by the program.

4. Copy in your working folder the file "*MergeAncientSample.sh*" from *"/data/mathias/afternoon/data_step4"* and execute it using the simulation filename "*SPLATCHE3_CentralEuropeSimulation*" as argument.

```
cp /data/mathias/afternoon/data_step4/MergeAncientSample.sh .
./MergeAncientSample.sh SPLATCHE3_CentralEuropeSimulation
```

Questions:

- What is the differences between the .arp file produced by SPLATCHE and the .arp file after the execution of the bash script?
  The arp file outputted by SPLATCHE3 contains one modern population samples+ 10 neolithic population samples while the arp file after the execution of the script contains only 2 population sample: one modern and one ancient grouping all the Neolithic sequences.
- Why do we need to group all ancient DNA samples together in a single ancient population sample.
  When simulating the data, SPLATCHE3 considers both the age and the location of each lineages, which is one of the interest of the approach. However, in many studies, real ancient lineages are grouped together for the analyses to increase the sample size (e.g. Bramanti et al Science. 2009;326:137–40) so we reproduced this strategy here. Note that an alternative strategy could have been to keep the Neolithic samples separated, but the same should have been done with the real data when computing statistics.

Remarks:

- If the Bash script is not running, please type the command "chmod +x MergeAncientSample.sh" before calling it again. This should be done for all bash scripts provided during the practical.

STEP 5: DRAW NEOLITHIC POPULATION SIZE FROM PRIOR DISTRIBUTION

The accuracy of the population continuity test depends on the amount of gene flow (measured by *Nm*) among various sub-populations (demes). You are going to draw the carrying capacity (parameter *K*) since the Neolithic era from a prior distribution going from 50 to 1000. It allows to explore *Nm* values from ~10 to 250, representing various amount of gene flow and thus taking into account this uncertainty. *Nm* is used by convention to describe gene flow among demes but in SPLATCHE3 it would rather be *Km* which corresponds to Nm at demographic equilibrium. Here we will modify K (called POPSIZE) to explore *Nm* while keeping m constant to 0.25.

Exercise:

1. Create a working folder called "*step5*" and copy the files provided in the folder *"/data/mathias/afternoon/data_step1"*.

```
mkdir step5
cd step5
cp /data/mathias/afternoon/data_step1/SPLATCHE3_CentralEuropeSimulation.txt .
cp -r /data/mathias/afternoon/data_step1/SPLATCHE3_CentralEuropeSimulation .
```

2. Create a template file "vegK_L1_time_2.template" in the settings folder *"SPLATCHE3_CentralEuropeSimulation"* with PARAM1 instead of the current numeric value 268. This word PARAM1 will then be replaced by another numerical values drawn from the prior distribution for each simulation.

→ vegK_L1_time_2.template

```
cat SPLATCHE3_CentralEuropeSimulation/vegK_L1_time_2.txt | sed 's/268/PARAM1/g' >
SPLATCHE3_CentralEuropeSimulation/vegK_L1_time_2.template
```

3. Copy all the files from /data/mathias/afternoon/data_step5/ into your working folder. Look at the different files to understand what are their roles.

```
cp /data/mathias/afternoon/data_step5/* .
```

4. Launch ABCtoolbox with the task "simulate" (already set in the input file) and look at the result file called "*PopGen_output_sampling1.txt*".

```
ABCtoolbox2 PopGen.input
```

Questions:

- What does contain the file "*PopGen_output_sampling1.txt*"?

It contains for each simulation (each line) the simulation number (first column), the parameter value *K* drawn from the prior distribution (second column) and many statistics of diversity within and between samples.

- What are the input files used by ARLEQUIN ?

"arl_run.ars" (settings) and "ssdefs.txt" (statistics to output)

- What are the input files used by ABCtoolbox ?

“PopGen.input” (input file), “PopGen_EuropeanNeol.est” (parameter prior), “CentralEuropeRealData.obs” (observed data) and “simulates.sh” which itself calls the script “MergeAncientSample.sh” (merging ancient sample in arp files), SPLATCHE3 (simulations) and the script “LaunchArlSumStatModifiedSPLATCHE.sh” (computation of statistics)

STEP 6: SELECTION OF THE BEST SET OF SIMULATIONS USING ABC

At this stage, you are going to estimate the simulations that fits at best genetic diversity within the ancient sample. The assumption is that those simulations are obtained with the most probable amount of gene flow (*Nm*) and they are the most accurate to perform the continuity test. We provide a file PopGen_output_sampling1.txt with a larger number of simulations (10,000) than what you did in the previous step to avoid long computational time.

Exercise:

1. Create a working folder called "*step6*" and copy the files provided in the folder *"/data/mathias/afternoon/data_step6"*.

```
mkdir step6
cd step6
cp /data/mathias/afternoon/data_step6/* .
```

2. Then you will select the 500 "best" simulations based on the number of alleles (statistics K2), the gene diversity (H2) and the mean number of pairwise differences (statistics Pi2), all in the ancient sample through an ABC estimation on this sole statistics.

For that, you need to create a new file called Nm.sim that contains the simulated dataset with K2, H2 and Pi2 as statistics instead of all statistics outputted by arlsumstat. So column 1 is the simulation number, column 2 is the value of the parameter K and column 3, 4 and 5 are K2, H2 and Pi2 respectively.

→ *Nm.sim*

```
cat PopGen_output_sampling1.txt | cut  -f 1,2,4,6,10 > Nm.sim
```

You also need to create another file called H2.obs with the observed statistics H2 as sole column.

→ *H2.obs*

```
cat CentralEuropeRealData.obs | cut  -f 2,4,8 > diversity.obs
```

Finally, launch ABCtoolbox on the input file "*PopGen_ABC_ParamEstimate.input*" that has been provided to you.

```
ABCtoolbox2 PopGen_ABC_ParamEstimate.input
```

3. Have a look on the output files of the previous steps. The file "*PopGen_ContinuityTestmodelFit.txt*" does tell you if the model is able to reproduce the observed statistics of diversity adequately. The file "*PopGen_ContinuityTestmodel0_BestSimsParamStats_Obs0.txt*" display the best 500 retained simulations, that will be used in the next step to test for population continuity.

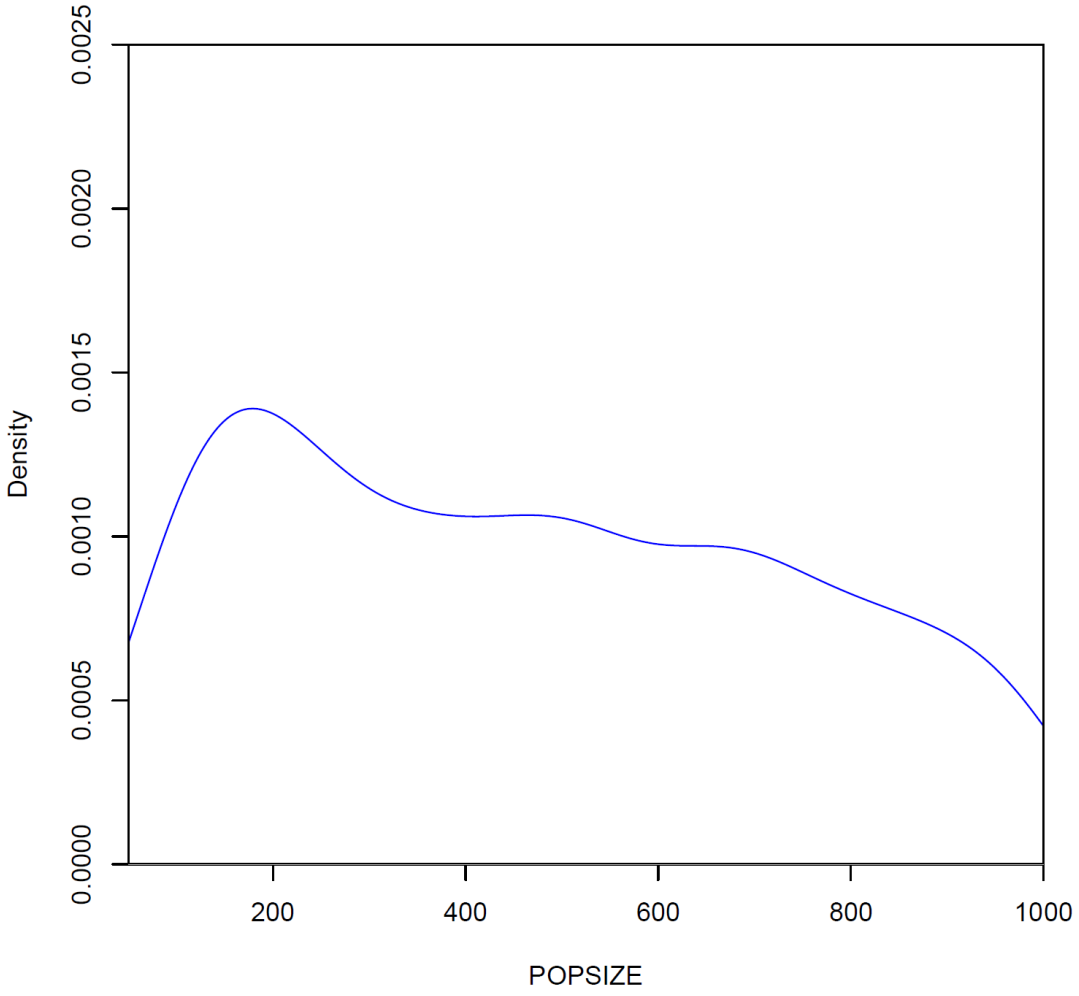
Using the R script "*DrawPosterior.R"* generate a pdf with the posterior distribution of the parameter *K*.

→ *PopGen_ContinuityTestmodel0_BestSimsParamStats_Obs0.pdf*

```
Rscript DrawPosterior.R
```

Questions:

- Is the model compatible with the genetic diversity in the ancient population sample?
  Yes it is, as the *P value* outputted by ABCtoolbox is > 0.05 (Tukey depth P value =0.66 and Marginal density P value = 0.51).
- Is there any signal toward a specific value of *K* ?
  The signal is weak but slightly goes toward smaller values of *K* with a posterior distribution mode around 200. A larger number of simulations could possibly confirm whether there is any signal.



STEP 7: TESTING POPULATION CONTINUITY

You will now test for population continuity using the 150 "best" simulations from the previous step.

1. Create a working folder called "step7" and copy the files provided in the folder "/data/mathias/afternoon/data_step7" as well as the files "*PopGen_ContinuityTestmodel0_BestSimsParamStats_Obs0.txt*" and "*PopGen_output_sampling1.txt*" from step6.

```
mkdir step7
cd step7
cp /data/mathias/afternoon/data_step7/* .
cp ../step6/PopGen_ContinuityTestmodel0_BestSimsParamStats_Obs0.txt .
cp ../step6/ PopGen_output_sampling1.txt .
```

2. Now sort both files "*PopGen_ContinuityTestmodel0_BestSimsParamStats_Obs0.txt*" (containing the 500 best simulations based on ancient genetic diversity) and "*PopGen_output_sampling1.txt"* (containing all statistics for all simulations) using the command "sort -k 1,1". This step is needed to merge them correctly afterward.

→ *best.sorted*
→ *all.sorted*

```
sort -k 1,1 PopGen_ContinuityTestmodel0_BestSimsParamStats_Obs0.txt > best.sorted
sort -k 1,1 PopGen_output_sampling1.txt > all.sorted
```

3. Now join both sorted files and output the results in a file called "*Fst.sim"* that must contain only the *Fst* values for the 150 best simulations (one value per line).

→ *Fst.sim*

```
join best.sorted all.sorted | sed -r 's/ /\t/g' | cut -f 16 > Fst.sim
```

3. Extract the observed *Fst* value into a single file called "*Fst.obs*".

```
cat CentralEuropeRealData.obs | cut -f 9 | tail -n 1 > Fst.obs
```

4. Finally, run the R script called "*testContinuity.R*" and look at the results

```
Rscript testContinuity.R
```

Questions:

- What conclusion can you make from this test? Would you reject population continuity in Central Europe from the Neolithic until today based on this mitochondrial dataset?

  Yes, the *P* value < 0.05 (0.008), thus rejecting the null hypothesis of population continuity during this period of time.