Population Genomics: background and tools

26 April 2018 (10:30-12:30)

Practical: "Coalescent simulations and ABC"

Mathias Currat

Dpt 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 the reconstruction of past demography using coalescent simulations combined to Approximate Bayesian Computation (ABC).

Through a series of steps, you will generate two or three datasets under different demographic conditions: 1/ a stationary panmictic population with a small effective size, 2/ a stationary population with a large effective size, 3/ a growing population. You will receive pseudo-observed data which have been generated with one of these scenarios and with known parameters. The goal of the practical is to recover which scenario and parameter values were used to generate these pseudo-empirical data.

PROGRAMS

1- FastSimcoal

Short description: Fast sequential Markov coalescent simulation of genomic data under complex evolutionary models

Download: http://cmpg.unibe.ch/software/fastsimcoal2/

Documentation: http://cmpg.unibe.ch/software/fastsimcoal2/man/fastsimcoal25.pdf

Reference: Excoffier, L., Dupanloup, I., Huerta-Sánchez, E., Sousa, V.C., and M. Foll (2013) Robust demographic inference from genomic and SNP data. PLOS Genetics, 9(10):e1003905.

Input files: *.par

Executable name: fsc26

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: arlsumstat3522_64bit

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: ABCtoolbox

STEP 1: SIMULATION OF DEMOGRAPHIC SCENARIO

First, you are going to simulate a very simple demographic scenario representing a stationary panmictic population of size 1,000 using the program **fastsimcoal2** (executable name = **fsc26**). The program user manual "*fastsimcoal25.pdf*" located in "/home/currat/data/manuals/" may help you if needed.

Exercise:

1. Create a working folder called "*step1*" and copy the file "*FastSimcoal_StationaryPop.txt*" provided in the folder "/home/currat/data/*data_step1*". This file contains a very simple input file for fastsimcoal2.

```
mkdir step1
cd step1
cp /home/currat/data/data_Step1/FastSimcoal_StationaryPop.par .
```

2. Have a look to the input file for **fastsimcoal2** and modify it in order to set the population size to 1,000.

→ *FastSimcoal_StationaryPop1000.par*

```
cat FastSimcoal_StationaryPop.par | sed s/200/1000/> FastSimcoal_StationaryPop1000.par
```

3. Perform one simulation using **fastsimcoal2** using FastSimcoal_StationaryPop1000.txt as input and the option "-g" (diploid data).

→ *FastSimcoal_StationaryPop1000_1_1.arp*

```
fsc26 -i FastSimcoal_StationaryPop1000.par -n1 –g
```

4. Have a look to the output file "*FastSimcoal_StationaryPop1000_1_1.arp*" located in the result folder.

Questions:

- What kind of virtual data have you generated using **fastsimcoal2**?
  Autosomal DNA sequences of a population sample of 25 individuals. Only the variable positions are visible.

Remark:

Note that if you want to output the whole DNA sequence (including monomorphic sites), then you need to use the option "-S".

Optional:

Try to simulate a series of 10 independent SNP instead of DNA sequence.

```
//Number of independent loci [chromosome]
10 0
//Per chromosome: Number of linkage blocks
1
#chromosome 1, //per Block: data type, num loci, rec. rate + optional parameters
SNP 1 0 0
```

STEP 2: COMPUTATION OF SUMMARY STATISTICS

You are going to compute a series of statistics summarizing the genetic diversity of the sample generated during step1.

Exercise:

1. Create a new working folder called "*step2*" and copy all the files located in the folder "/home/currat/data/*data_step2*". The file "*arl_run.ars*" contains the settings for the program **arlsumstat** (executable name = **arlsumstat3522_64bit**, linux version of ARLEQUIN 3.5, specifically designed for ABC) and "ssdefs.txt" contains the statistics which will be outputted in a simple format. "LaunchArlSumStatModified.sh" is a script allowing to launch arlsumstat on the ".arp" file generated during step 1 and which must located in a sub-folder. Thus add to your working directory the result folder obtained under step 1 ("step1/FastSimcoal_StationaryPop1000/").

```
mkdir step2
cd step2
cp -r /home/currat/data/data_Step2/* .
cp -r ../step1/FastSimcoal_StationaryPop1000/ .
```

2. Launch **arlsumstat** on the genetic data produced during step1 by calling the script "LaunchArlSumStatModified.sh" with two arguments; 1/ the name of the result subfolder (FastSimcoal_StationaryPop1000), 2/ the name of the outputfile ("output.stats").

```
./LaunchArlSumStatModified.sh FastSimcoal_StationaryPop1000 output.stats
```

→ output.stats

Questions:

- What contains the file output.stats?

The following 4 intra-population statistics computed on the simulated data: mean_K (number of alleles), mean_H (Heterozygosity), mean_S (Segregating sites), mean_Pi (Differences between pairs of sequences).

STEP 3: USE A PARAMETER PRIOR DISTRIBUTION

Modify the demographic scenario by drawing the population size from a uniform prior distribution between 1,000 and 20,000 (instead of fixing it to 1,000 as for step1 and step2) using the program **ABCtoolbox** (executable name = **ABCtoolbox**). Then launch 10 simulations in drawing each time a population size value from the prior distribution.

Exercise:

1. Create a new working folder called "step3" and copy all the files provided in "/home/currat/data/data_Step3". The file "PopGen.input" contains the settings for **ABCtoolbox**, the file "PopGen_st-small-pan.est" contains the prior distribution for the population size parameter and the file "data.obs" the summary statistics of (pseudo) real data.

```
mkdir step3
cd step3
cp /home/currat/data/data_Step3/* .
```

2. Copy all the input files needed to run **arlsumstat** located in "/home/currat/data/data_step2" in your working folder.

```
cp /home/currat/data/data_Step2/* .
```

3. Modify the input file for **fastimcoal** in order to replace the population size value by the keyword "PARAM1" and save it in a new file called "FastSimcoal_StationaryPop.template". Also change the number of simulations from 1 to 10.

→ FastSimcoal_StationaryPop.template

```
cat /home/currat/data/data_Step1/FastSimcoal_StationaryPop.par | sed s/200/PARAM1/ >
FastSimcoal_StationaryPop.template

cat /home/currat/data/data_Step3/PopGen.input | sed s/"numSims 1"/"numSims 10"/ >
PopGen.input
```

4. Then Launch **ABCtoolbox** with the option "simulate".

→ PopGen _output_sampling1.txt

```
ABCtoolbox PopGen.input
```

Questions:

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

This file contains the parameters (second column) and the 4 statistics (remaining columns) for all simulations (one per line).

Sim	POPSIZE	mean_K	mean_H	mean_S	mean_Pi
1	7632	2	0.25621 14	3.58694	
2	2783	2	0.207347 6	1.24408	
3	19974	2	0.203525 22	4.47755	
4	18869	2	0.232212 37	8.59184	
5	12440	2	0.303158 19	5.76	
6	19778	2	0.193003 42	8.10612	
7	7095	2	0.107619 6	0.645714	
8	12141	2	0.250418 21	5.25878	
9	3123	2	0.396463 6	2.37878	
10	14729	2	0.321866 28	9.01224	

STEP 4: GENERATE ABC SIMULATION DATASETS

You are going to generate two datasets of 1,000 simulations obtained with two different demographic scenarios: a small panmictic population and a large panmictic population.

Exercise:

1. In a new working folder called "step4", repeat the operation of step 3 but generate 1,000 simulations and save the output file as "SimData_Stationary-Small-Panmictic-Population.txt"

   → SimData_Stationary-Small-Panmictic-Population.txt

```
mkdir step4
cd step4
cp /home/currat/data/data_Step2/* .
cp /home/currat/data/data_Step3/* .
cat /home/currat/data/data_Step1/FastSimcoal_StationaryPop.par | sed s/200/PARAM1/ >
FastSimcoal_StationaryPop.template
cat /home/currat/data/data_Step3/PopGen.input | sed s/"numSims 1"/"numSims 1000"/ >
PopGen.input
ABCtoolbox PopGen.input
cat PopGen_output_sampling1.txt > SimData_Stationary-Small-Panmictic-Population.txt
```

2. Generate a second dataset containing 1,000 simulations of a stationary panmictic population of large size (prior between 10,000 and 200,000)

   → PopGen_st-big-pan.est

   → SimData_Stationary-Big-Panmictic-Population.txt

```
cat PopGen_st-small-pan.est | sed s/"1\tPOPSIZE\tunif\t1000\t20000
output"/"1\tPOPSIZE\tunif\t10000\t200000 output"/ > PopGen_st-big-pan.est
cat PopGen.input | sed s/"estName PopGen_st-small-pan.est"/"estName PopGen_st-big-
pan.est"/ > PopGen-big.input
ABCtoolbox PopGen-big.input
cat PopGen_output_sampling1.txt > SimData_Stationary-Big-Panmictic-Population.txt
```

Questions:

- What does contain those simulation files?

Same answer than under step 3 but for 1,000 simulations.

OPTIONAL STEP 5: GENERATE ANOTHER DATASET WITH TWO PARAMETERS

You are going to generate a third dataset for a new scenario of growing population which involves two parameters (growth rate and population size), instead of a single one (population size) for previous scenarios.

Exercise:

1. In a new working folder called "step5", generate a third dataset for a growing population. You need to add the population growth rate as a second parameter (GROWTHRATE). Use two uniform prior distributions: one from 1,000 to 200,000 for the population size and one from 0.0 to - 0.1 from the growth rate.

```
mkdir step5
cd step5
cp /home/currat/data/data_Step2/* .
cp /home/currat/data/data_Step3/* .

cat /home/currat/data/data_Step1/FastSimcoal_StationaryPop.par | sed -e s/"200"/"PARAM1"/
-e s/"-0"/"PARAM2"/ > FastSimcoal_GrowingPop.template

cat PopGen_st-small-pan.est | sed s/"1\tPOPSIZE\tunif\t1000\t20000
output"/"1\tPOPSIZE\tunif\t1000\t200000 output"/ > PopGen_grow-pan.est
echo -e "0\tGROWTHRATE\tunif\t-0.05\t0.0 output" >> PopGen_grow-pan.est

cat /home/currat/data/data_Step3/PopGen.input | sed -e s/"numSims 1"/"numSims 1000"/ -e
s/"estName PopGen_st-small-pan.est"/"estName PopGen_grow-pan.est"/ -e s/"simArgs
1PopDNA\toutput.stats POPSIZE"/"simArgs 1PopDNA\toutput.stats POPSIZE GROWTHRATE"/ -e
s/"simulates"/"simulates2P"/ > PopGen-grow.input

cat simulates.sh | sed -e s/POPSIZE\=\$3\/POPSIZE\=\$3\\nGROWTHRATE\=\$4/ > temp.1
cat temp.1 | sed s/FastSimcoal_StationaryPop/FastSimcoal_GrowingPop/ | sed s/"sed"/"sed -e
s\/PARAM2\/\$GROWTHRATE\/ -e"/ > simulates2P.sh

chmod +x *.sh
ABCtoolbox PopGen-grow.input

cat PopGen_output_sampling1.txt > SimData_Growing-Panmictic-Population.txt
```

→ SimData_Growing-Panmictic-Population.txt

STEP 6: MODEL CHOICE WITH ABC

Use **ABCtoolbox** to compare the two or three demographic scenarios simulated during steps 4 and 5 (1'000 simulations each) in retaining the 100 best simulations and using the "Standard rejection sampling" procedure.

Exercise:

1. Copy all the files located in the folder "/home/currat/data/data_step6" into a new working folder called "step6". The input file for **ABCtoolbox** is called "PopGen_ABC_ModelChoice.input" and is designed for three input matrices. If you have only two matrices, modify it accordingly.

   ```
   mkdir step6
   cd step6
   cp /home/currat/data/data_Step6/* .
   ```

2. Copy the two (or three) simulation datasets generated during steps 4 and 5 into your working folder and check that their names correspond to those in **ABCtoolbox** input file.

   ```
   cp ../step5/SimData_Growing-Panmictic-Population.txt .
   cp ../step4/SimData_Stationary-* .
   ```

3. Launch **ABCtoolbox** and look at the output file "PopGen_modelchoicemodelFit.txt". You can use any tabulator to visualize it better.

   ```
   ABCtoolbox PopGen_ABC_ModelChoice.input
   ```

   → PopGen_modelchoicemodelFit.txt

If you get the following error message: "ERROR: Highly correlated statistics found!", copy the three input files from "/home/currat/data/data_step6" and try again.

```
cp /home/currat/data/data_Step6/*.txt .
ABCtoolbox PopGen_ABC_ModelChoice.input
```

Questions:

- Is there any scenario better supporting the data than the others?
  Yes, the scenario of a stationary population with a small population size is significantly more probable that the two others. It is relative probability compared to both other scenario is > 99.9%.

- If yes, how good is this scenario to reproduce the observed statistics? Could it be considered as plausible?
  The scenario of a stationary population with a small population size is able to reproduced the observed statistics very well as shown by the marginal P value (=0.64) and the Tukey P value (=0.67). A contrary, the two other scenarios are not able to reproduce the observed data, both the scenario of a stationary population of large size (Marginal P value = 0.02 and Tukey P value = 0.0) and the scenario of an expanding population (Marginal and Tukey P values = 0.0).

Step 7: PARAMETER ESTIMATION WITH ABC

Still using **ABCtoolbox**, you are going to estimate the size of the population using the best scenario detected in step 6.

Exercise:

1. Copy all the files located in the folder for step 6 into a new folder called "step7", excepted those with a filename starting with "PopGen_modelchoice…". Also copy in your working folder the files located in "/home/currat/data/data_step6"

```
mkdir step7
cd step7
cp -r ../step6/* .
rm *modelchoice*
cp /home/currat/data/data_Step7/* .
```

2. Use the parameter estimation option of **ABCtoolbox** to estimate the population size value. Use the R-script provided to visualize the results.

```
ABCtoolbox PopGen_ABC_ParamEstimate.input

Rscript DrawPosterior.R
```

→PopGen_Stationary-Small-Panmictic-Population-PopSize-Posterior.pdf
→ PopGen_Stationary-Small-Panmictic-Populationmodel0_MarginalPosteriorCharacteristics.txt

3. In order to evaluate the accuracy of the estimation, look at the output file "PopGen_Stationary-Small-Panmictic-Populationmodel0_RetainedValidation_Obs0.txt".

4. In order to check if the posterior is biased, generate Quantile plot of the true parameter value within the posterior distribution. Using R, perform a Kolmogorov–Smirnov test to check if the distribution reject uniformity and may thus be indicative of bias posteriors. Use the following R commands:

```
Rscript  plotQuantile.R
R
retvalid_genomic<-read.delim("PopGen_Stationary-Small-Panmictic-
Populationmodel0_RetainedValidation_Obs0.txt")
ks.test(retvalid_genomic$POPSIZE_quantile,"punif")
quit()
```

5. Finally, apply the same estimation procedure to the scenario of stationary population with a big size and visualize the posterior distribution of the population size parameter as well as the Quantile plot.

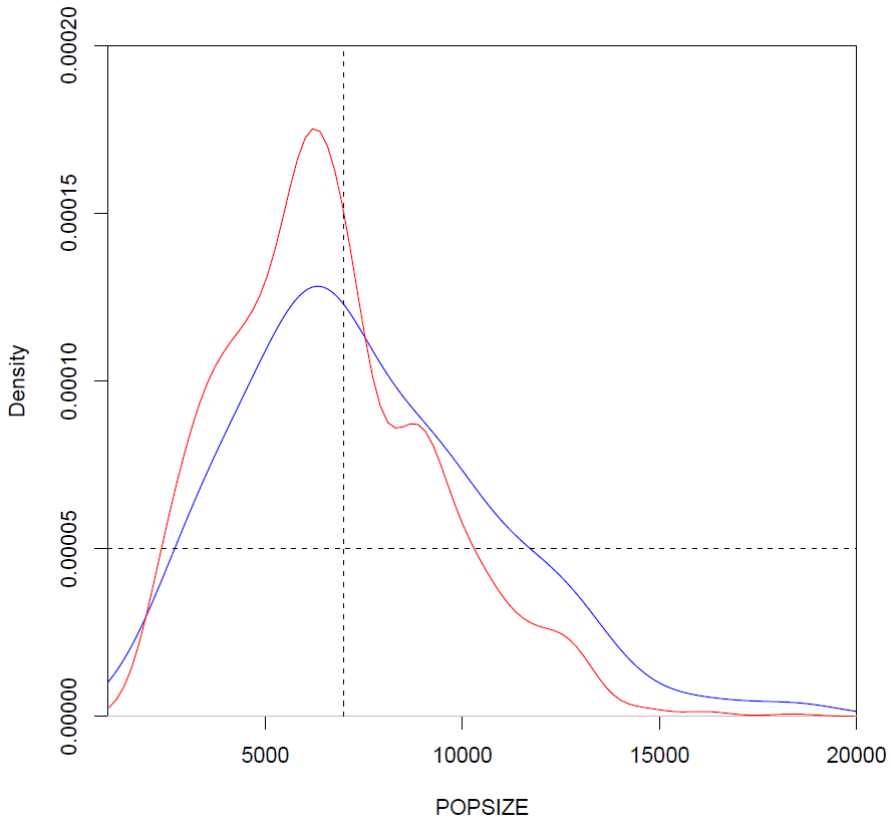
```
cat PopGen_ABC_ParamEstimate.input | sed s/"Stationary-Small-Panmictic"/"Stationary-Big-
Panmictic"/ > PopGen_ABC_ParamEstimate-Big.input

ABCtoolbox PopGen_ABC_ParamEstimate-Big.input
```

```
cat DrawPosterior.R | sed s/PopGen_Stationary-Small-Panmictic-Population/PopGen_Stationary-Big-Panmictic-Population/ | sed s/",from=1000,to=20000"/",from=10000,to=200000"/ | sed s/h=0.00005/h=0.000005/  > DrawPosterior-big.R

cat plotQuantile.R | sed s/PopGen_Stationary-Small-Panmictic-Population/PopGen_Stationary-Big-Panmictic-Population/ > plotQuantile-big.R

Rscript DrawPosterior-big.R
Rscript plotQuantile-big.R
R
retvalid_genomic<-read.delim("PopGen_Stationary-Big-Panmictic-Populationmodel0_RetainedValidation_Obs0.txt")
ks.test(retvalid_genomic$POPSIZE_quantile,"punif")
quit()
```

Questions:

- What value of population size is estimated and what is its 95 HDI?

The mean of the posterior is 6,660 with a 95 HDI [1,900-12,088]. This is a relatively large interval, but remember that the number of simulations is small.

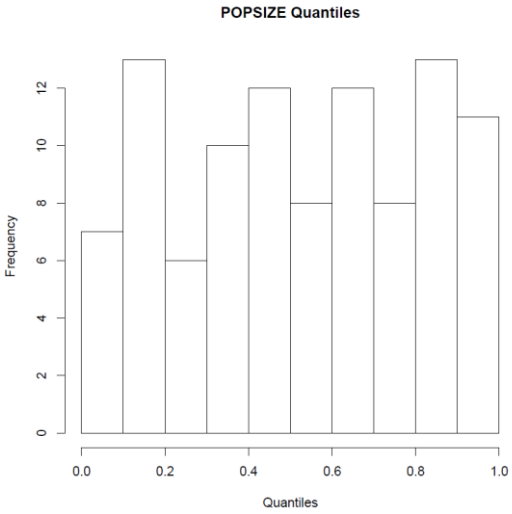


In this figure the blue and red lines correspond to the rejection and GLM-adjusted estimated posteriors, respectively. The horizontal and vertical dashed lines represent the prior and the observed data, respectively.

- What is the best point estimator, the mean or the mode of the posterior distribution?
  The mean is a slightly better point estimate than the mode of the posterior distribution, average difference between true and estimated value is 2,337 and 2,618 respectively.

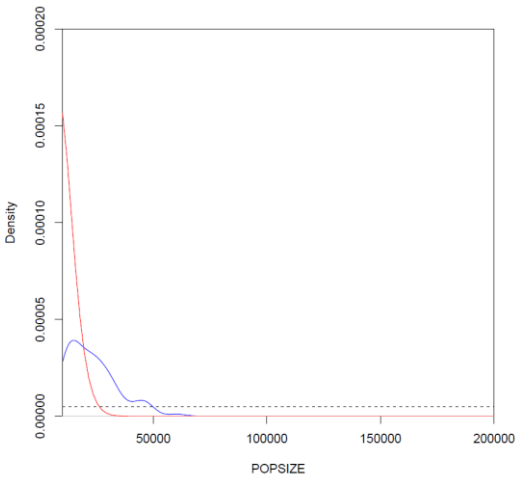
- Is the posterior distribution biased ? If yes in which direction ?

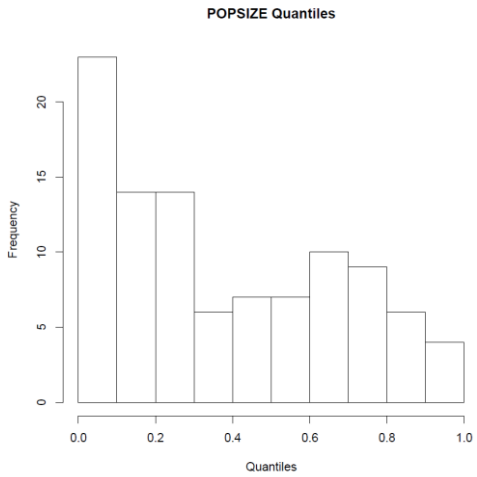
  The Kolmogorov–Smirnov test shows that a uniform distribution cannot be rejected for the Quantile distributions (D = 0.0565, p-value = 0.9072). Consequently, the estimation can be considered as not biased.



- What kind of posterior distribution is given when estimating the population size with the scenario of a stationary population of big size?

  The posterior distribution tends to be distributed toward lower values.





D = 0.224, p-value = 8.784e-05 biased toward too large values.

OPTIONAL STEP 8: EXPLORE AN ADDITIONAL SCENARIO

Exercise:

If you have time, you can then modify your files to simulate with **fastsimcoal** an additional scenario with a stationary population structured in 4 demes exchanging migrants at a rate *m* (parameter MIGRATE).

```
mkdir step8
cd step8
cp /home/currat/data/data_Step8/* .
./LaunchMigrScenario.sh
Cat PopGen_output_sampling1.txt > SimData_Stationary-Structured-Population.txt
cp ../step7/SimData* .

ABCtoolbox PopGen_ABC_ModelChoice_4scenarios.input
```

Questions:

- Does this scenario better explain the observed data?
  No, the structured population has a relative probability = 0%.