Ciências ULisboa
Faculdade de Ciências da Universidade de Lisboa

Aplicações Avançadas em Biologia

Master's Degree in Bioinformatics and Computational Biology

May 2021

Miguel Casanova Vieira Parente

Nº 24475

Individual Project

2nd Semester – 2020/2021

Material and Methods

For estimating F_{ST}, the *getFst* function provided during the course was used. This function computes F_{ST} according to Hudson's estimator following Bathia[1]. Significance of the observed F_{ST} values was determined by a permutation test with 1000 permutations, using the *PermutFst* function, designed for this project. Briefly, for each permutation, the location of the individuals in the sampled populations was randomly permuted and a simulated F_{ST} recalculated. The p-value was calculated as the proportion of the 1000 simulated F_{ST} that were greater than the observed value. As such, two subpopulations are considered as significantly structured when 50 or less of the 1000 permutations (5%) produce a simulated F_{ST} greater or equal than the observed F_{ST}.

The ABC method works by providing an approximation of a simulation from the posterior distribution, when the distributions associated with both the prior and the likelihood (which is impossible to define analytically) can be simulated. The two studied populations had the following features: 5 sampled individuals (being diploid, this translates to 10 sampled units); 50000 sites considered; mutation rate of 1.2e-8. The number of segregating sites for these sampled populations were 126 and 64, for Central and Western chimpanzees, respectively. This summary statistics was used to predict the effective size (N_e) of the populations. To model the prior distribution of the parameter N_e, an uniform distribution between 10 and 100000 was used. For the simulation of the model, the *sim.tree.mut* function, provided during the course, was used. This function allows simulating the summary statistics: the number of segregating sites for a given parameter value (N_e). For modelling the joint distribution, 10000 simulations were used. Next, the distance between the simulated and observed summary statistics was computed and a rejection step of the simulations was applied. For the rejection, tolerance levels of 1, 10 and 20% for the quantiles of distance distribution were applied. Different plotting functions were designed using *ggplot2*, *cowplot* and *ggextra* to allow visualization of the data, including plots for posterior, prior, joint distribution of parameters and summary statistics, together with marginal plots for prior and posterior distributions and posterior density plots.

F_{ST} and permutation tests

1. Function for permutation tests

a) What is the null hypothesis?

The null hypothesis is that there isn't a genetic differentiation between individuals of two populations: ***H0: F_{ST} = 0 Vs. H1: FST ≠ 0***. In other words, if the null hypothesis was to be true, this would mean that all individuals would belong to the same population (the two sampled populations are identical), and F_{ST} would be equal to 0.

b) Why do you permute individuals between populations and not alleles between populations?

For a permutation test, depending on the question at hand, it is critical to know what and how to resample. In our case, we are testing whether two populations are genetically differentiated, by using the F_{ST} method. This method allows estimation of the proportion of genetic diversity among populations: when FST is zero, there is no genetic differentiation; when FST is equal to 1, the populations are fixed for different alleles. Given we want to determine the significance of the pairwise FST values and the degree of genetic differentiation between two populations, we will need to blindly assign (permute) individuals to each of the subpopulations, and test whether the simulated F_{ST} is comparable to the real F_{ST}. By blindly assigning the individuals, we expect to simulate a sample distribution of the test statistics, under the null hypothesis.

Nevertheless, although in our case we permute individuals, it is important to mention that we could actually use a combination of permutations of both individuals and loci. The biggest issue with making permutations of loci, would be the computational challenge of performing these permutations for all of the variant loci of the genome. In an ideal scenario, the allelic frequencies of each loci, per individual, could be permuted to create simulated individuals belonging to a random population. As this would be computationally very difficult, this strategy would be unrealistic. To tackle this, different tools use the technique of bootstrapping over loci to generate a confidence interval around the observed Fst[2,3].

c) What is the test statistic?

The test statistic we are using is F_{ST}. It is a measure of degree of genetic differentiation between populations, providing the proportion of the total genetic variance contained in a subpopulation (S), relative to the total genetic variance (T).

d) What is the definition of p-value? Explain how to obtain the p-value from the sampling distribution and the observed values of F_{ST}?

The p-value is given by the proportion of permutations where we obtain a value more extreme than the observed sample statistics. As such, it is very simple to obtain the p-value by summing the number of simulated F_{ST} that are higher than the observed F_{ST} and dividing this for the total number of simulations.

2. Statistical significance of the pairwise F_{ST}

a) Based on the p-values and significance level, do you reject the null hypothesis for all the pairs? Justify.

To assess the significance of the pairwise F_{ST} values between all pairwise comparisons for the provided dataset[4], I repurposed a loop used in class 2 (to calculate the pairwise F_{ST} between all populations), to estimate the p-values and significance of the pairwise F_{ST} values. The values obtained are represented in **Fig. 1**.

Assuming a significance level of 5%, we can see that all p-values are below 0.05 and, as such, we reject the null hypothesis for all of the pairs of populations. As such, our data does not provide enough information to accept H_0, and we consider that all subpopulations are genetically different.

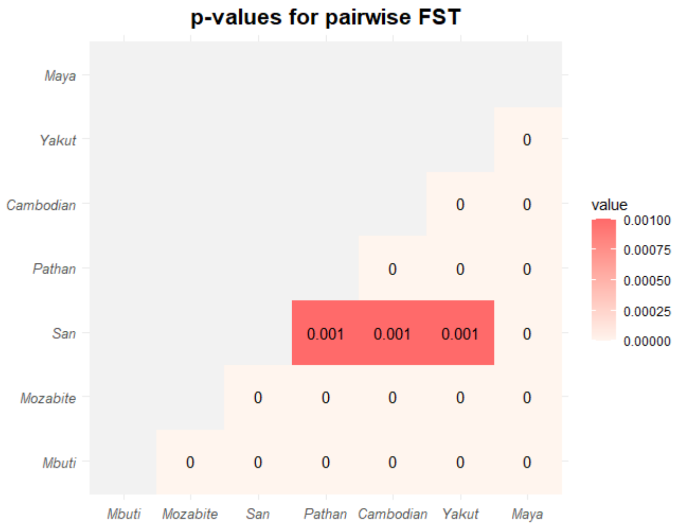

Figure 1- Matrix with the p-values for the permutation tests for the pairwise F_{ST} between all populations.

3. Histogram of the sampling distribution

a) Is the observed value within the expected if the null hypothesis was true? Justify.

By plotting the simulated F_{ST} together with the observed F_{ST}, we have the histogram represented in **Fig. 2**.

As we can clearly see, all of the simulated F_{ST} give values that are lower than the observed F_{ST}. If the observed value was within the expected if the null hypothesis was true, the observed F_{ST} would be located within the distribution curve of the simulated FST, and not at its very extremity (or outside of it). As such, the observed F_{ST} doesn't fall within the values expected for the null hypothesis.

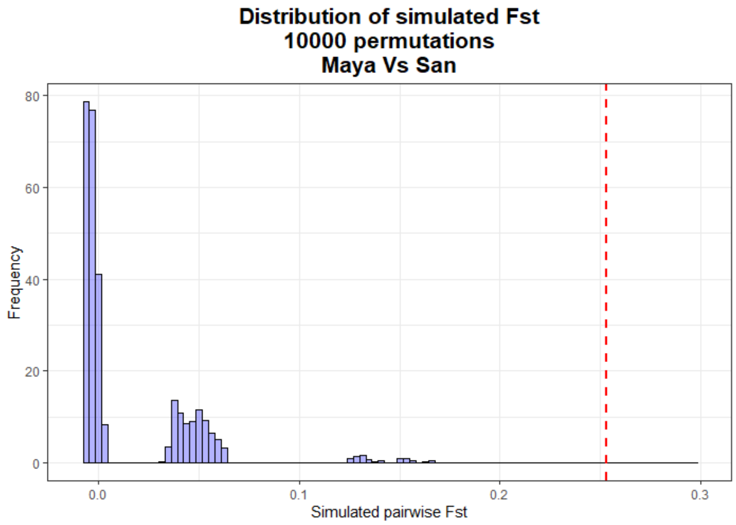

Figure 2 – Simulated F_{ST} values of 10000 SNPs for the Maya and San populations. The red dashed line represents the observed F_{ST}. The histogram represents the distribution of simulated F_{ST} values.

ABC methods

4. Using ABC method

Given that most of the parameters for the populations are similar, we can actually use the same set of simulations, to simulate the summary statistics (number of segregating sites). After a rejection step, taking into account the distance between the observed and simulated statistics, we can have a posterior distribution of accepted parameters (effective population size, in this case). We can represent all these prior and posterior distributions (for both parameters and summary statistics), as well as the simulated data, with the plots in **Fig. 3**.

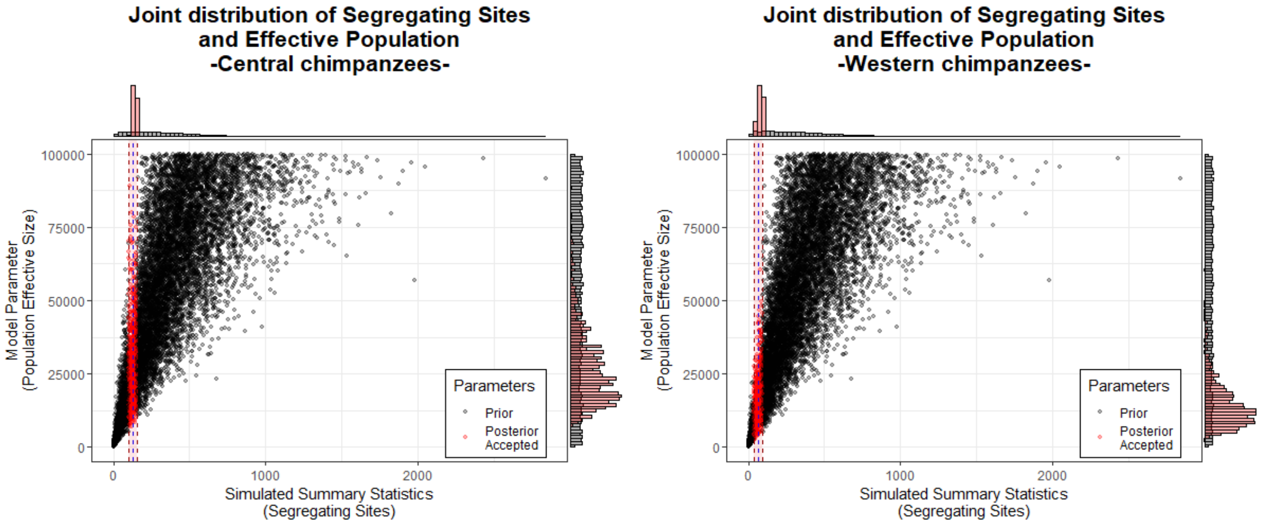

Figure 3 – Representation of the joint distributions of model parameters (N_e) and summary statistics (segregating sites), for Central (left) and Western (right) chimpanzee populations. Points accepted after the rejection step, are represented in red. Marginal histograms, with the prior (black) and posterior (red) distributions of parameters (left) and summary statistics (top), are also represented.

a) What is the population with the larger effective size? Justify.

Our sampled Central and Western chimpanzee populations[5] have similar sample size, number of sites and mutation rates. The only differentiating factor, is the number of segregating sites: 126 for Central and 64, for Western populations. This is used as the summary statistics for our ABC exercise. Assuming the above, it is not surprising to observe that the population with the highest number of segregating sites, corresponds to the population with the larger effective size. In this case, the population of Central chimpanzees. This can be observed in the density curves of the posterior effective sites (**Fig. 4**).

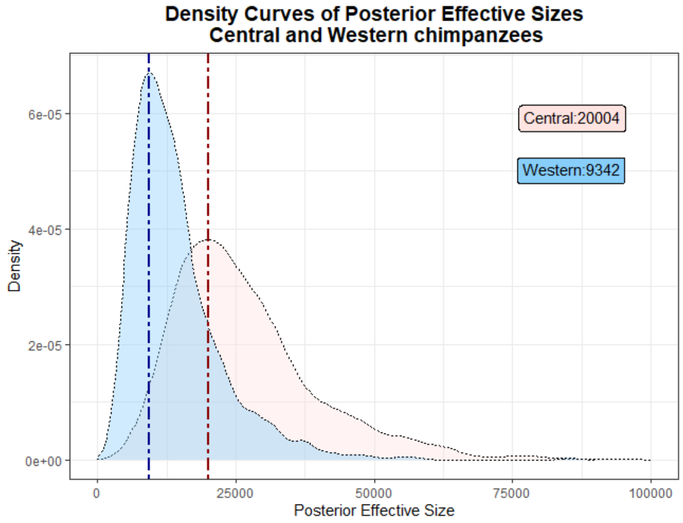

Figure 4 – Density curve for the posterior distribution of effective sizes for the Central (blue) and Western (rose) chimpanzee populations. The blue line represents the most likely value for effective size for the Western population; the red line represents the most likely value for the effective size for the Central population. These values are represented in textboxes of the matching color.

b) Is the posterior different from the prior? Is there enough information in the data? Justify.

ABC methods provide an approximation of the posterior probability, when a likelihood function can’t be obtained. In these methods, we replace data by a summary statistics and assume that both summary statistics and model parameters are random variables. By doing so, and using a given model (a genetic coalescence model, in this case), we can calculate the joint distribution of both summary statistics and parameters. This is done, by simulating summary statistics from parameter values taken from the prior distribution. If the summary statistics are not related to the parameters, we consider that the data does not have enough information. As a consequence, the posterior and prior distributions will be similar, as the data (our observed summary statistics) will not allow changing our prior knowledge about the population parameters.

As we can clearly see in **Fig. 3**, the posterior and prior distributions for our model are different (marginal histograms on the right). As such, we can objectively say that the information we have about our data (the summary statistics) is sufficient to change our prior knowledge about the parameters of the populations to a posterior distribution, calculated from the available data and prior distribution. In other words, the available data is sufficient to learn about the population parameters.

c) Select two tolerance levels and repeat the inference with the two values. Are your conclusions affected by the tolerance level? Justify.

For ABC, a too high tolerance will lead to accepting all simulations for the summary statistics and, consequently, the posterior will be the same as the prior distribution. On the other hand, decreasing the tolerance normally increases the quality of the approximation to the correct posterior (assuming that the prior covers all the values that the posterior can assume). However, if the tolerance is too low, the number of accepted simulations will be too low to make good approximations. As such, the appropriate level of tolerance has to be finely tuned, in coordination with other factors, like the number of simulations.

For this exercise, three tolerance levels were tested: 1, 10 and 20%. This is represented in **Fig. 5**.

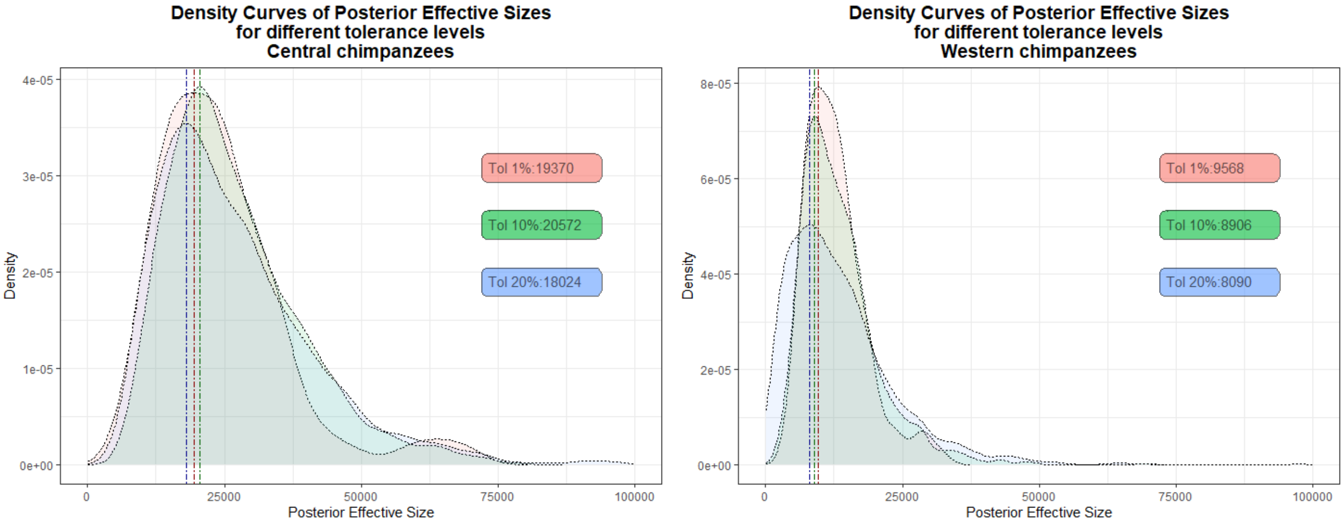

Figure 5 – Effect of the tolerance on the density curves for the posterior distribution of effective sizes for the Central (left) and Western (right) chimpanzee populations. Density curves for 1% (red), 10% (green) and 20% (blue) are overlaid in each plot. Lines intersecting the modes for the different density functions are represented.

As we can observe, for our study case, the effect of changing the tolerance is not dramatic. For lower tolerances, we have a sharper density curve for the posterior, with tighter confidence intervals and lower dispersion. Nevertheless, the most likely effective size does not show radical changes.

Importantly, independently of the tolerance level, the conclusion that the population of Central chimpanzees has a larger effective size remains unchanged.

References

1. Bhatia, G., Patterson, N., Sankararaman, S. & Price, A. L. Estimating and interpreting FST: The impact of rare variants. *Genome Res.* **23**, 1514–1521 (2013).
2. Pembleton, L. W., Cogan, N. O. I. & Forster, J. W. StAMPP: an R package for calculation of genetic differentiation and structure of mixed-ploidy level populations. *Molecular Ecology Resources* **13**, 946–952 (2013).
3. Keenan, K., McGinnity, P., Cross, T. F., Crozier, W. W. & Prodöhl, P. A. diveRsity: An R package for the estimation and exploration of population genetics parameters and their associated errors. *Methods in Ecology and Evolution* **4**, 782–788 (2013).
4. Henn, B. M. *et al.* Distance from sub-Saharan Africa predicts mutational load in diverse human genomes. *PNAS* **113**, E440–E449 (2016).
5. Manuel, M. de *et al.* Chimpanzee genomic diversity reveals ancient admixture with bonobos. *Science* **354**, 477–481 (2016).