ADVANCED APLICATIONS IN BIOLOGY

INDIVIDUAL REPORT

2023

CÁTIA ANTUNES
ID 60494

Ciências ULisboa
Faculdade de Ciências da Universidade de Lisboa

Exercise 1: Permutation Tests

1.1. Methods

The first step involves defining a permutation test function, which takes as input two genotype matrices, representing the two populations, and the number of permutations to be performed. This function computes the observed FST between the two populations, permutes the index of individuals between populations, and computes the FST for each permutation. Finally, it calculates the p-value as the proportion of permuted FST values that are equal or greater than the observed FST.
In the second step, the function is applied to assess the pairwise FST values between all pairwise combinations of populations for the Henn et al. (2015) dataset. The genotype matrix is read from a file, and a separate file provides information about individuals and their corresponding populations. The index of population for each individual is determined using this file. A list is then created with the index of individuals that belong to each population. A matrix is initialized to save the pairwise FST values. A nested loop is used to compute the pairwise FST values between each pair of populations, using the permutation_test function, and these values are saved in the matrix. Finally, the matrix is printed and exported to a file. A plot of pairwise FST values is also generated using the plotFst function and saved as a PNG image. Scripts and further results can be found on the following repository: https://github.com/AntunesCSR/AAB_Project.git.

1.2. Results

Table 1 - Pairwise FST values

	San	Mbuti	Mozabite	Pathan	Cambodian	Yakut	Maya
San	NA	0.004	0	0	0	0	0
Mbuti	0.004	NA	0.019	0	0	0	0
Mozabite	0	0.019	NA	0	0	0	0
Pathan	0	0	0	NA	0	0.001	0
Cambodian	0	0	0	0	NA	0	0
Yakut	0	0	0	0.001	0	NA	0
Maya	0	0	0	0	0	0	NA

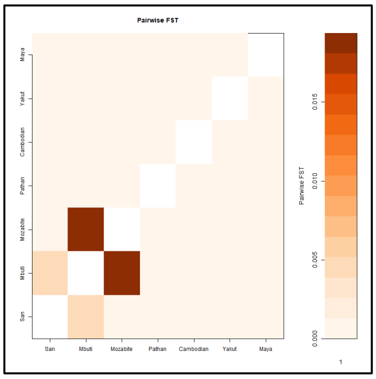

Figure 1 - FST plot of the Henn et al. (2015) dataset

1.3. Discussion

What is the null hypothesis?

The null hypothesis that there is no significant genetic differentiation between the populations.

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

When conducting a permutation test to assess the significance of FST (fixation index) between populations, it is more appropriate to permute individuals between populations rather than alleles between populations. The FST statistic is a measure of population differentiation that compares the genetic variation within populations to the total genetic variation within and between populations.
Permuting individuals between populations allows us to shuffle the distribution of genetic variation among populations and assess whether the observed FST value is significantly different from what we would expect by chance if there was no population differentiation. In contrast, permuting alleles between populations would not be appropriate because it would only shuffle the distribution of genetic variation within populations. Therefore, permuting individuals between populations is a more appropriate approach when using a permutation test to assess the significance of FST between populations.

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

The result shows a pairwise FST matrix and corresponding p-values between 7 populations. The null hypothesis is that the FST values are not significantly different from zero, and the alternative hypothesis is that they are significantly different. Looking at the pairwise p-values, we can see that all the values are less than 0.05, which is the significance level set in the analysis. This means that for all pairs of populations, the p-value is less than 0.05, and the observed FST value is significantly different from zero at a 5% significance level. Therefore, we reject the null hypothesis for all pairs of populations and conclude that there are significant genetic differences between all the pairs of populations considered in this analysis.

Exercise 2: ABC Methods

2.1. Methods

The study aimed to analyse the genetic differences between two subspecies of Chimpanzees by simulating genetic data and comparing it to observed data. The provided dataset of five columns was loaded into the R environment. The prior distribution for the effective population size was defined, and parameter values were randomly sampled from this prior distribution. The sim.tree.mut function was called to simulate data and compute summary statistics, and the number of segregating sites was computed from the simulated data. The absolute difference between the simulated and observed number of segregating sites was computed, and parameter values resulting in large distances compared to the tolerance distance were rejected. Parameter values resulting in small distances were retained, and the closest simulations were identified for both subspecies. The joint distribution of the prior and summary statistics was plotted and saved for both subspecies. Scripts and further results can be found on the following repository: https://github.com/AntunesCSR/AAB_Project.git

2.2. Results

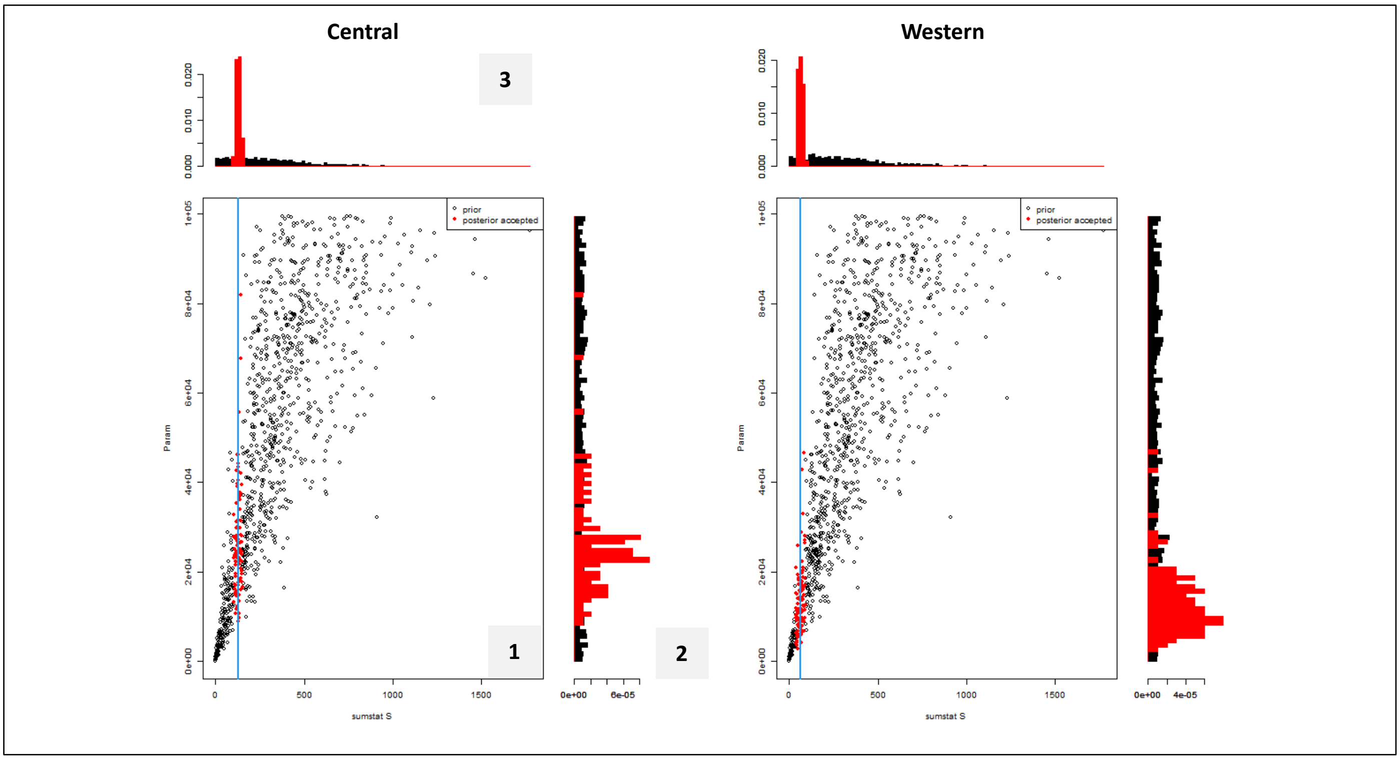

Figure 2 – Graphical representation of the (1) joint distribution of parameters and summary statistics, highlighting the points accepted after the rejection step and (2) the prior and (3) posterior distribution histograms, for both Western and Central Chimpanzee populations.

2.3. Discussion

Given the overlap in the posteriors of the effective size of the two sub-species, do you think there are differences between the two species? Justify.

An overlap in the posterior distributions means that there is a range of Ne values that are plausible for both sub-species. Specifically, the overlap indicates that the data is not strong enough to distinguish between the Ne values of the two sub-species with certainty.

What is the sub-species with the larger effective size? Justify.

Based on the obtained summary statistics, the central sub-species appears to have a larger effective population size than the western sub-species. The median effective population size for the central sub-species (24,280) is higher than the median effective population size for the western sub-species (11,601). Additionally, the 75th percentile of the central sub-species posterior distribution (29,895) is also higher than the 75th percentile of the western sub-species posterior distribution (16,767).

Based on the comparison of the posterior with the prior, do you think we learned about the effective size from the data? Justify.

Based on a visually comparison of the prior and posterior distribution's shape and location, we can argue that because the posterior distribution is different from the prior distribution, some information on the effective size has been learned from the data. This is because, if the histogram of the posterior distribution is peaked around a particular value, it means that the simulations that are closest to the observed data have that value for the parameter. This suggests that this value is a plausible estimate for the effective population size given the observed data. On the other hand, if the histogram of the posterior distribution is flat or spread out, it suggests that the simulations that are closest to the observed data have a wide range of values for the parameter. This means that the data does not provide strong evidence for any particular value of the parameter and that the effective population size cannot be reliably estimated based on the observed data alone.

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

The tolerance is defined as the proportion of the closest simulated data points to the observed data that are retained for analysis. A low tolerance value would result in a smaller set of accepted simulations, which would provide a more stringent filtering of implausible parameter values, leading to a higher accuracy in posterior estimates. However, a low tolerance value would also result in a larger rejection of true parameter values, as some simulations that are very similar to the observed data may be excluded. On the other hand, a high tolerance value would result in a larger set of accepted simulations, which would provide less filtering of implausible parameter values and increase the risk of accepting false positives. However, a high tolerance value would also result in a lower rejection of true parameter values, as some simulations that are not very similar to the observed data may be included.

In this case, tolerances of 5%, 10% and 40% were compared and overall the conclusions didn't change much, with the exception that at tolerance 5% the posteriors of the subspecies did not overlap.

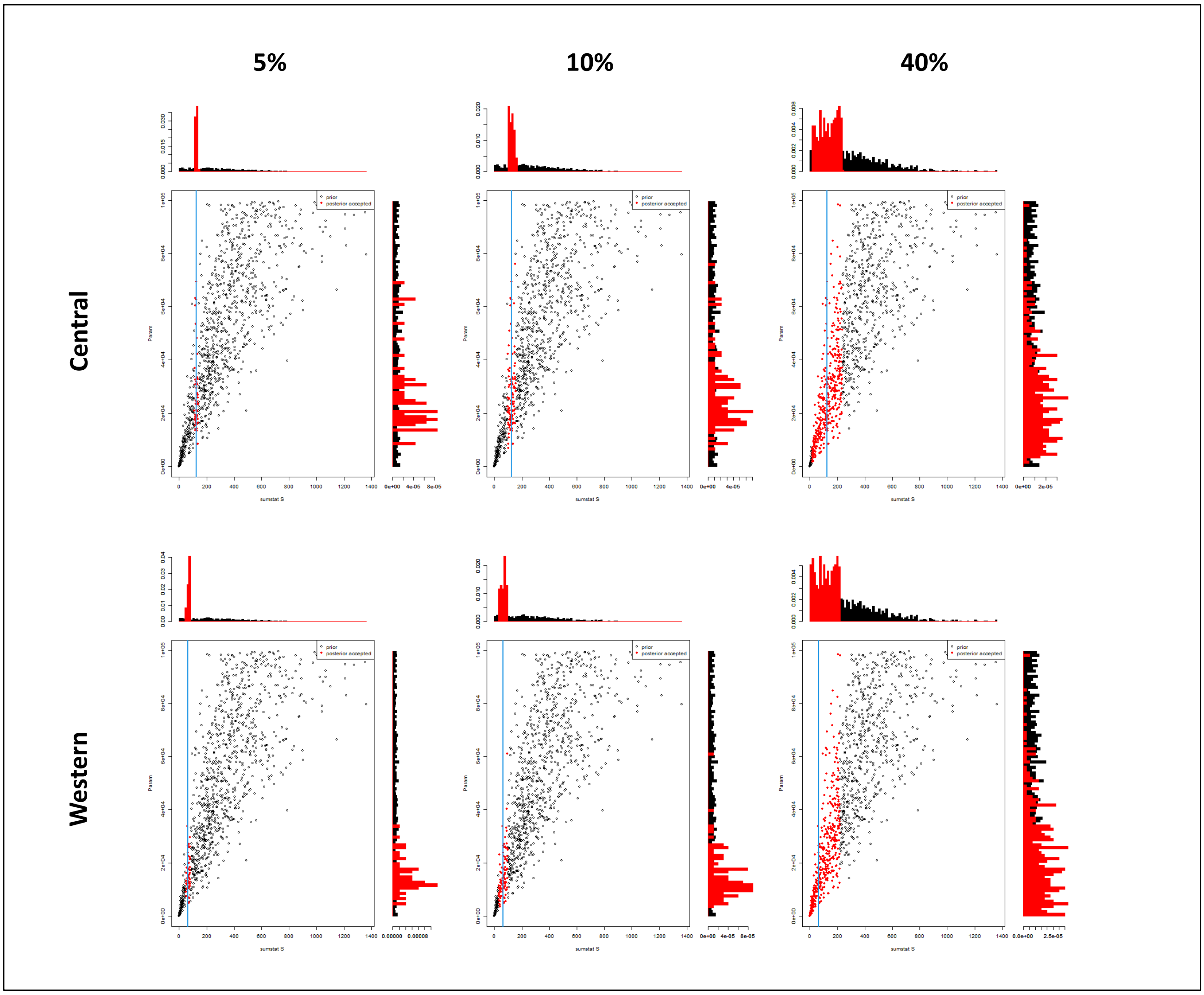

Figure 3 — Graphical representations of the joint distribution of parameters and summary statistics, highlighting the points accepted after the rejection step and the prior and posterior distribution histograms, for both Western and Central Chimpanzee populations, for different tolerance levels of 5%, 10% and 40%.