Exercise 6

Q1) What are the main differences between the two dataset? How should such differences be accounted for during the adaptation study? How do you think this would impact the computing performances?

Leaving aside the differences concerning the type of species and environment, the two dataset differs in the fact that Moroccan sheep did not show a genetic structure while Australian Striepey Snapper did. When running SamBada, it is possible to add population structure as co-variable so that the models trying to relate individual genotypes and environment will "automatically" correct their predictions accordingly. In the current version of SamBada, adding population structure as co-variable implies an exponential increase of computation time.

Q2) Compressing environmental information via PCA: What is the main drawback of this approach?

In some cases, Principal Components can be composed by complex combination of environmental variables and therefore the interpretation of results can become less intuitive.

Q3) Building models on genotypes or alleles: What are the pros and cons of each approach?

Natural selection might act on a genotype or on an allele and both approaches generally catch both cases with only one exception: when selection acts on heterozygotes this can only be detected when working on genotypes. For this reason, we usually prefer working on genotypes. Working with alleles has the advantage of being computationally faster, since only two models per marker are built.

Q4) What co-variable can we use to define population structure?

Any continuous variable encoding for genetic structure. Principal components are a perfect example, because they are continuous and not correlated. Alternatively, you can use population membership coefficients calculated by other methods (ex. ADMIXTURE, STRUCTURE, etc.).