Exercise 5

Q1) What is the need of this step? What are the major risks to account for? Which methods would you employ?

Low DNA quality or sequencing errors can result in technical noise across the genotype matrix. When working with this data, we want information to be comparable between samples and markers. It is therefore preferable to avoid keeping individuals that missed most of the SNPs (and vice versa). For this reason, we will apply a filter for missingness (% of missing observations across each SNP or individual). Another important point concerns frequencies of alleles and genotypes. Many alleles will be present at extremely low rate (for example, these might be sequencing errors or rare mutations occurring in a couple of individuals). We are not interested in these situations, because they are unlikely to be relevant for our study. For this reason, we usually apply an allele and a genotype frequency filter.

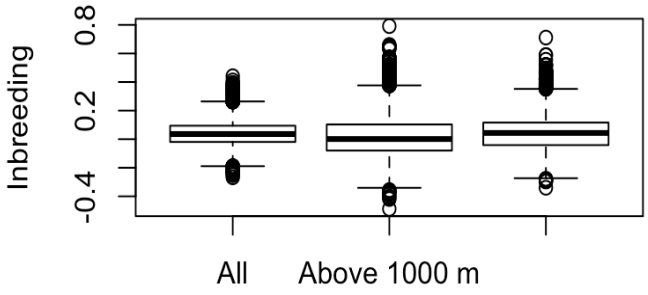
RQ1) MAF: What genotypes would you filter out? Why?

There is no absolute rule on which threshold apply as a MAF, but 5% is often used. As a rule of thumb, try to interpret this frequency considering the sample size. In our case for example, we have 160 individuals which means 320 alleles. With a MAF of 1%, we will have genotypes with 3 minor vs 317 major alleles, which is very low. With 5%, the worst case would provide a ratio of 16:304, i.e. 8 to 16 individuals carry the allele, so this frequency can be more relevant for studying genetic structure and local adaptation.

RQ2) MGF: What genotypes would you filter out? Why?

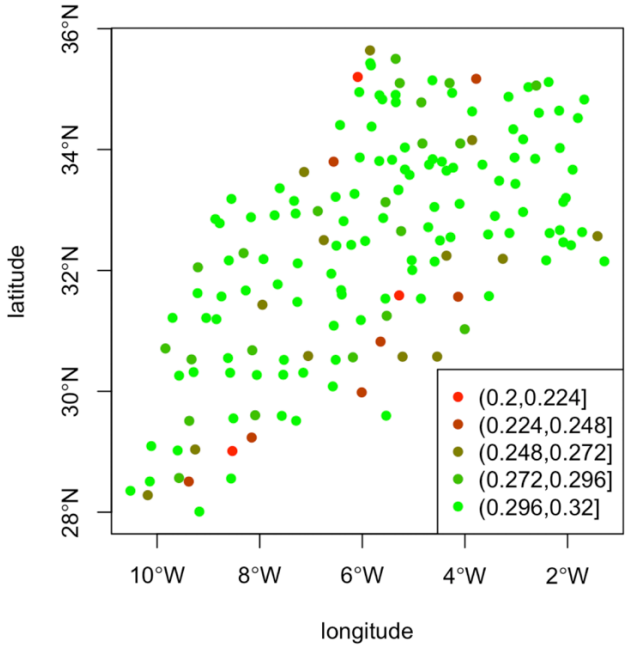
The same as RQ1. Note that minor allele frequency filter usually already filters out most of the dominant genotypes, except those heterozygous which requires a major genotype frequency filter.

RQ3) Compare the three groups? What can we say about the effect of altitude on population isolation?



Not much, since the boxes show that the three groups have roughly the same distribution. An inbreeding coefficient around 0 suggests that there is no particular breeding issue in this population. We can notice though that the range of outliers is larger in samples above 1000 m and that some of these individuals can reach higher value of inbreeding (perhaps due to isolation of living in altitude?).

RQ4) Can you see a spatial structure? Are there sites where heterozygosity drops?



There doesn't seem to be a strong pattern, but it seems as if there are a bit more samples with low heterozygosity in the area behind the mountains facing the desert.

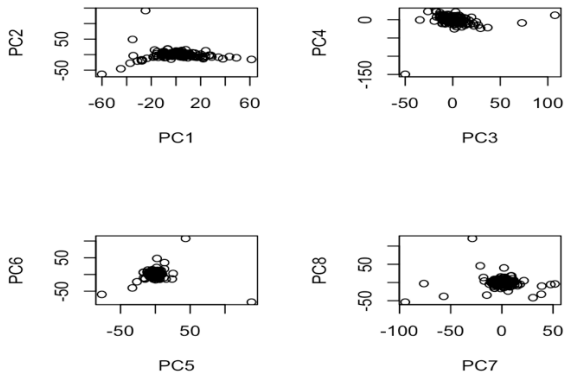
RQ5) Is there a variable that associates with heterozygosity?

Not really, R=0.3 is quite low.

RQ6) What can we say about the Moroccan breeds? Are they isolated between each other?

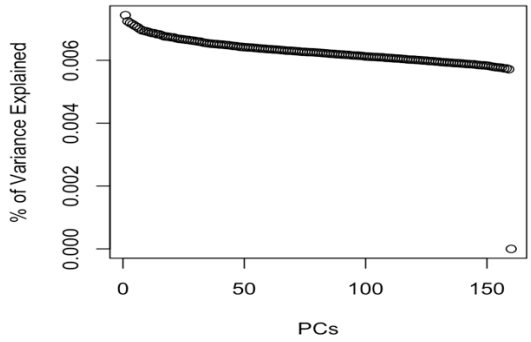
Fst values are very close to 0. This is coherent with what we previously observed: the breeds of Morocco seems to be well interconnected between each other.

RQ7) Comment this graph. Do you think that there is genetic structure?



We can see that there are individuals that appear more related than others, but it is hard to tell how strong this differentiation is from this graph.

RQ8) What can you say about the strength of the population structure? How can this information influence your adaptation study?



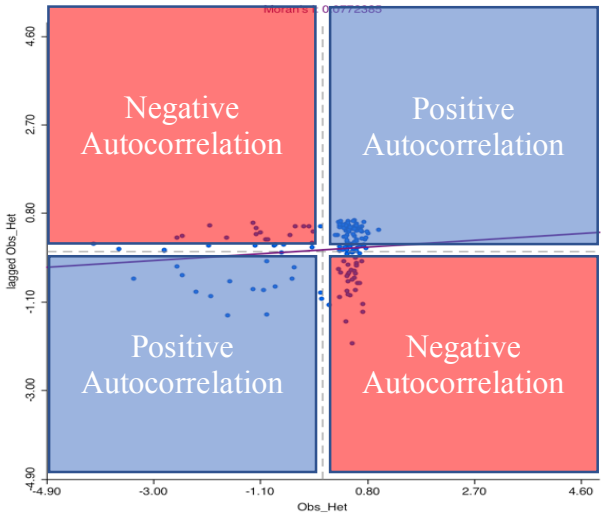
The structure seems very weak, because there does not seem to be a group of markers strongly correlated (if there were, we would have seen the first PCs "standing out" from the distribution). This

is good news for an adaptation study, since the confounding role of neutral structure should be negligible.

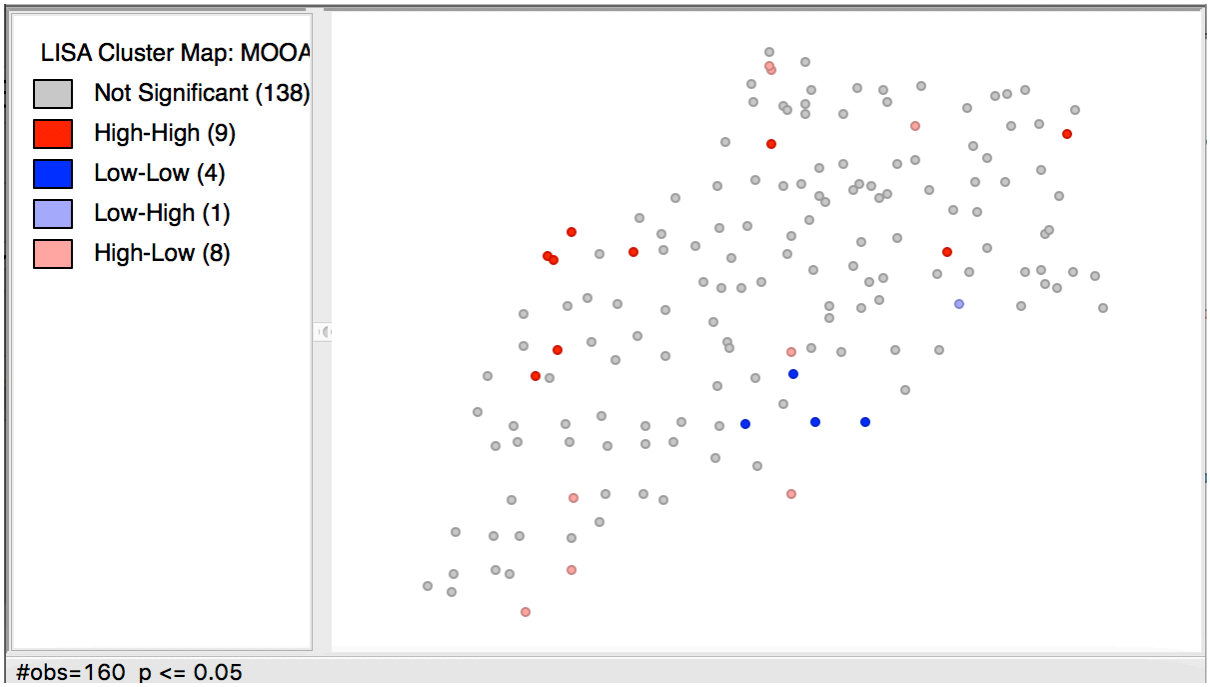
Q2) Would you still say that there are no heterozygosity patterns among the Morocco Sheep population?

This representation shows that the pattern might be a bit stronger than what observed in R.

Q3) In which quadrant of the plot you expect positive autocorrelated values to fall?
Q4) In which quadrant of the plot you expect negative autocorrelated values to fall?



Q5) What does the High-High, High-Low, Low-High and Low-Low notation stands for? Are there heterozygosity clusters in Morocco?



High-high indicates that individuals with high value are surrounded by other individuals with high values. Low-low indicates the opposite situation, when low values are surrounded by low values. High-low and low-high indicates the case when the values of one individual contrast with those of its neighborhood. We can see a low-low cluster of heterozygosity in a region facing the desert, which might indicate the presence of an inbreeding situation.

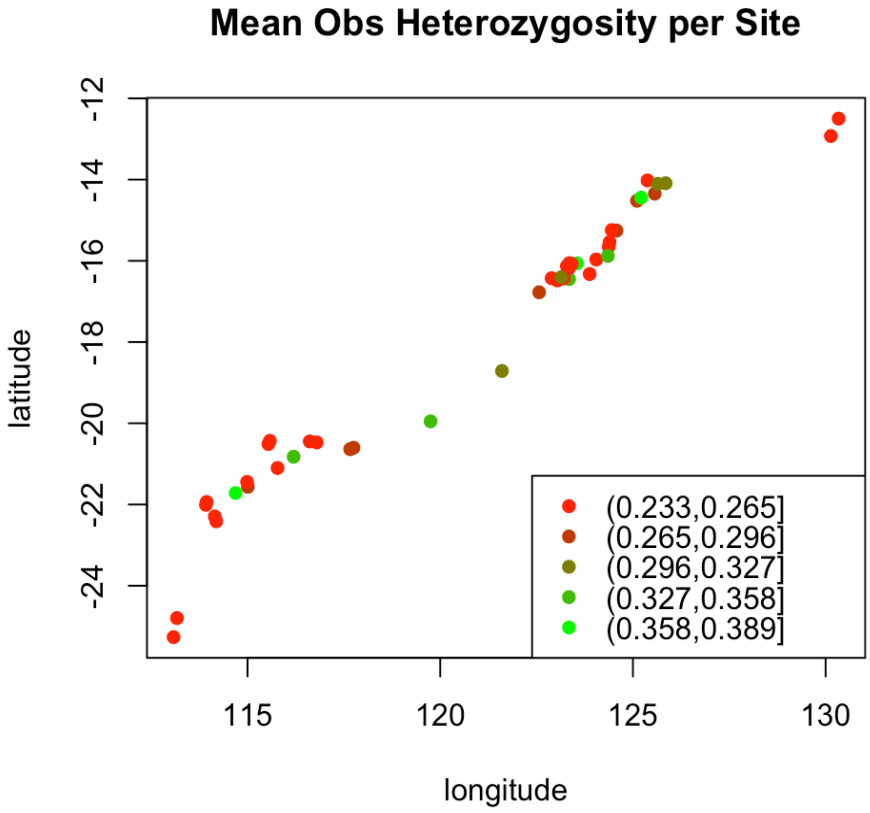
Q6) What differences do you expect between the population structure of land and water species?

It depends on the species and particularly on the reproductive strategy. For instance, many sessile marine species reproduce by spawning gametes or larvae that can then travel considerable distance following sea current. This results that population can be very large and asymmetrical. This is less common on land, but still can apply to plant species dispersing their propagules by wind.

Q7) What are the main differences in comparison to the Moroccan sheep case study? How is the quality of data different and how does this impact the filtering?

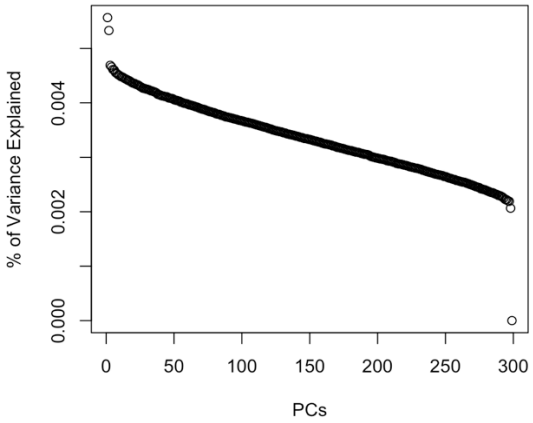
This dataset was obtained using a RAD-seq alike sequencing strategy. This resulted in a genotype matrix with less genetic markers and with more missing values. For this reason, after filtering, the genotype matrix is remarkably smaller, compared to the one of the Moroccan Sheep.

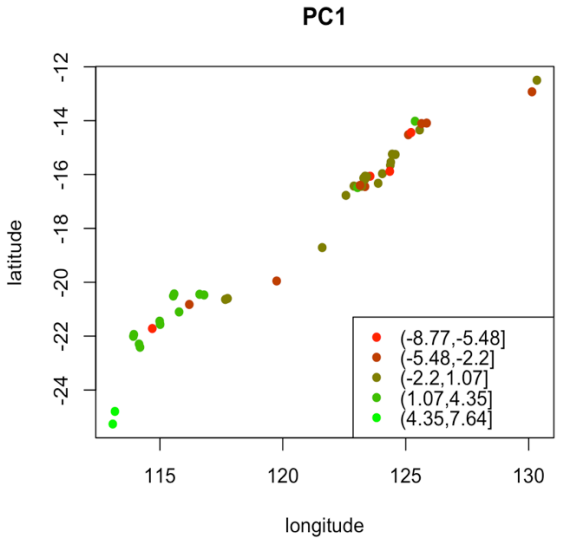
Q8) Can you observe heterozygosity patterns?

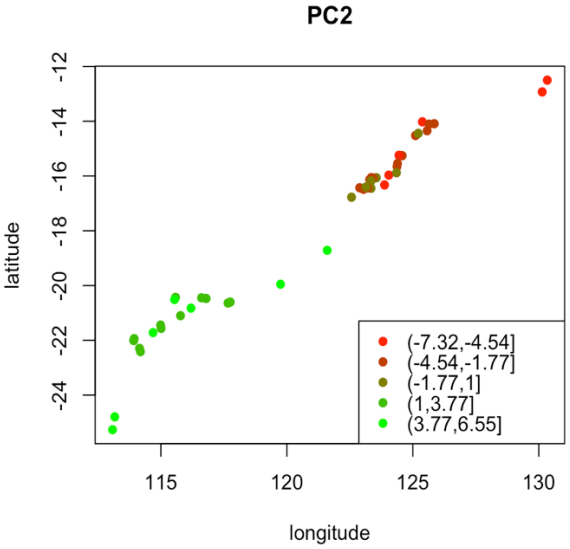


Not particularly, there are some reefs where heterozygosity tend to be higher but they are no following a particular structure.

Q9) Is the population of this species structured or not? How could this impact an adaptation study?







We can see that the first two principal components clearly "stand out" from the others, in the graph of the % of explained variance. This means that there are cluster of highly correlated markers and this is something we would expect if there is a genetic structure in the population. In fact, as we visualize the two PCs on the map, we see that there is a geographical structure in the distribution of these values. We will need to account for this during the adaptation study, in order to avoid the confounding effect of genetic structure.