

Exercise 5:

Genetic Data

Aim:

- Learn to work with Genotype Matrix
- Compute genetic diversity metrics
- Study autocorrelation
- Evaluate Population Structure

INTRODUCTION

A crucial step of a landscape genomics analysis is to evaluate how genetic diversity distributes across the landscape. Understanding how genes flow across space and how individuals are related to each other are essential information to describe demographical issues and define conservation priorities. Moreover, this information is a crucial input for adaptation studies. In this section we will start working on genetic data, first of all by operating a quality check on genetic variants and then by analyzing genetic diversity through a landscape. We will continue working on the Moroccan sheep example and introduce an additional case study concerning a fish population from Australia. This case study is based on a publicly available dataset referring to [the 2018 paper from DiBattista et al.](#)

Data Access

Find the genetic data of this exercise on the course website. The environmental data you produced during the previous exercise will serve as input here. If you missed a step, you can find environmental data in the course website as well.

EXERCISE

Create an Rstudio project in the working directory. We will start working with genetic data of sheep from Morocco. Make sure that the working directory contains the following files:

- OA.vcf (genetic data of the sheep samples from Morocco)
- MOOA_ENV_ps.shp (environmental data for the sheep samples)

- The three .R files containing the R code for the exercise.
- plot_map_gradient.Rfun (r custom function for plotting gradients)

The VCF format

The Variant Calling Format is a standard format for representing genetic variations across individuals. It consists in a matrix where rows correspond to genetic variants and columns to individuals. In this part of the exercise, the genetic variants we will work on are SNPs obtained from a Whole Genome Sequencing effort. For computational reason, the dataset was strongly reduced (a whole genome sequencing effort can detect millions of SNPs). Here we retained ~35'000 SNPs distributing across the sheep genome. To open .vcf files in R, we will use the `vcfR` package.

Dataset Size

In our exercises we are working with relatively small datasets (tens of thousands of SNPs). In this case, we can store genotypes in "standard" R matrices. If you happen to work with larger datasets (hundreds of thousands to millions of SNPs), R matrices can become quite complicated to handle. For these situations, you can use the R package *Snprlate* which allows fast computation even on large genotype matrices.

Data Filtering

The first step of the genetic analysis consists in data quality checking, polishing and filtering.

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

Open the *genetic_filtering.R* script and follow the instructions.

Spatial Genetic Variation

There are several indexes of genetic variation that can be employed to study how genetic diversity distributes across a landscape. Here we will show two examples:

- 1) *Classical* Inbreeding Coefficient: a classical meter from population genetics, we will use the environmental data in two populations and then investigate how the mean inbreeding coefficient (calculated SNP-by-SNP) varies between the two groups of samples.
- 2) Observed heterozygosity: we will calculate the amount of heterozygous loci within each individual, and observe how this metric distributes across the map. We will check whether this value can be associated to an environmental condition.
- 3) Pairwise *Fst*: is another classical statistic from population genomics comparing the genetic differentiation between two populations due to genetic structure. The computation of this statistic is more complex and we will see how to use an R package that allows to perform these calculations.

Open the *spatial_genetic_variation.R* script and follow the instructions.

Population Structure

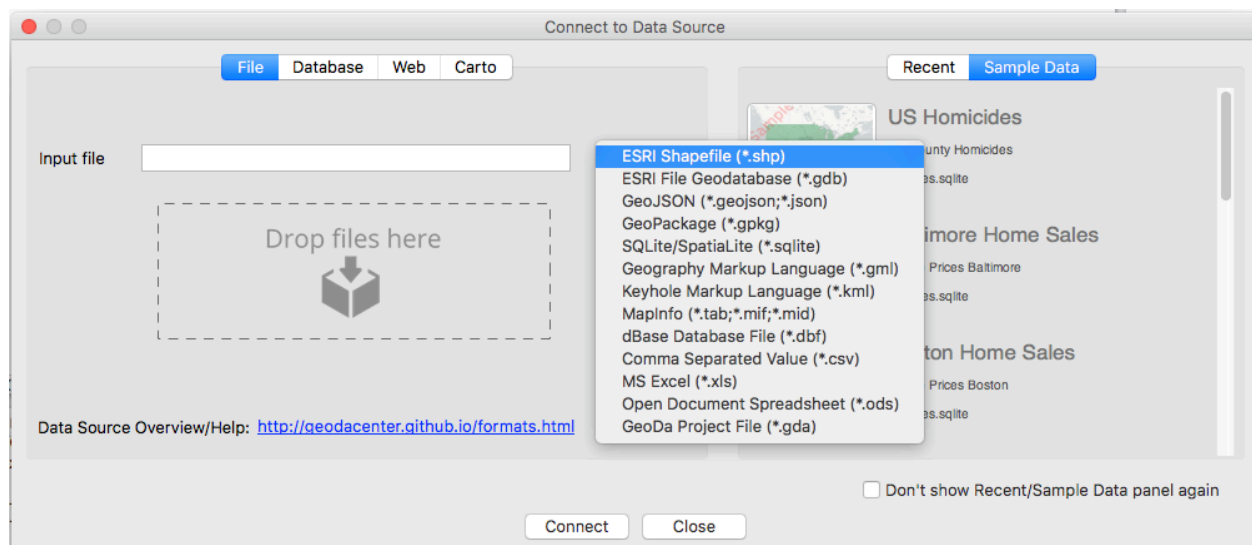
The next step of the analysis of the genotype matrix involves the study of the population structure.

Why is it important to study the population structure? How can this affect the discovery of adaptive traits?


There are several software allowing to study population structure based on genetic information. Here we will show an approach using principal component analysis in R. Open the *population_structure.R* script and follow the instructions.

Spatial Autocorrelation

We now want to better elucidate how heterozygosity distributes across the landscape of Morocco. By looking at heterozygosity associated to each sample, no particular structure emerged. We now want to test whether heterozygous or homozygous individuals tend to be closer to each other or whether their distribution is random. We run a spatial autocorrelation analysis, using the Geoda software. Open geoda. In the main window, click on the folder symbol and select ESRI Shapefile:

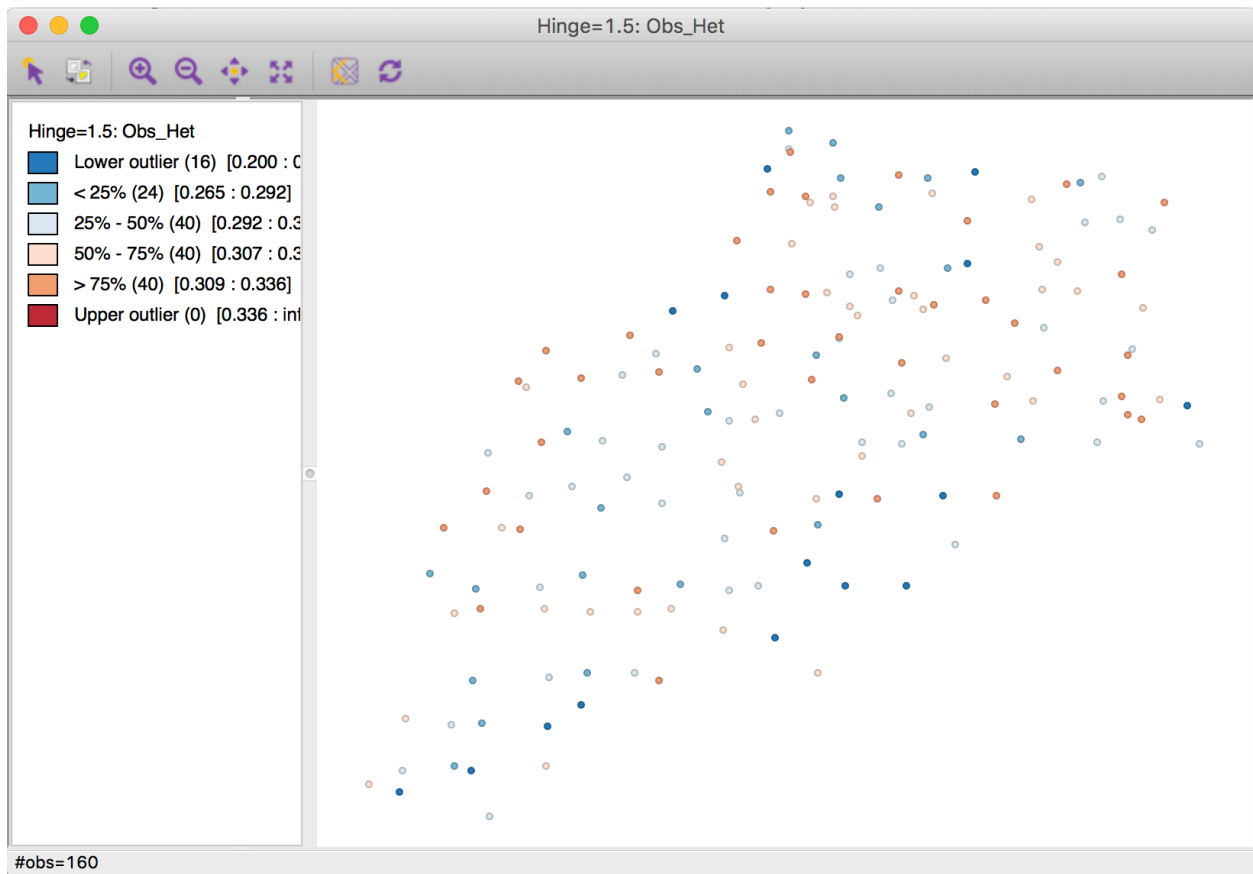


Open the MOOA_env_Gen_Str.shp. Geoda is a software specialized in working with vector data. It allows to explore shapefiles in an interactive way and also provides analysis functions

that allow to perform computations on the spatial objects. The  icon allows to visualize the attribute table. If you select on object on the map, it will be highlighted on the table, and viceversa. You can change the data representation by right clicking on it, then select:

>> Change Current Map Type > Box Map > Hinge 1.5

And select Obs_Het as represented variable. This map should appear:



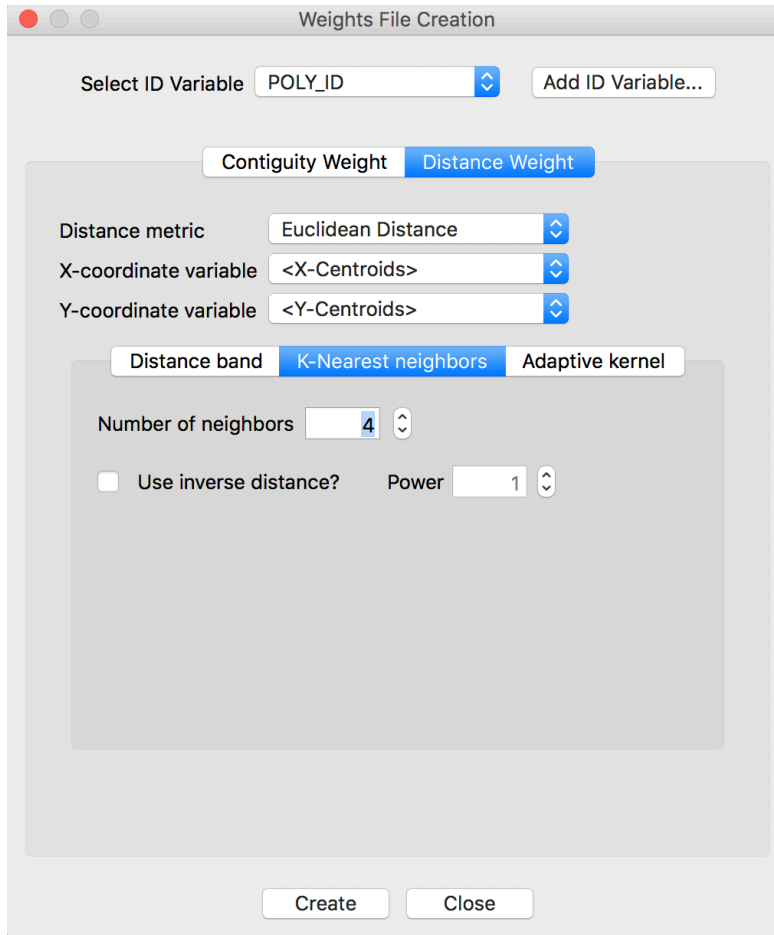
Each point is colored according to the quartile (of the heterozygosity distribution) in which the sample heterozygosity falls.

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

We now want to calculate the spatial autocorrelation of the observed heterozygosity. This is a metric of how the value of each point is related to the one of its neighbors. The first step when calculating spatial autocorrelation is to set a neighborhood weighting system. Click on:

>> Tools > Weights Manager

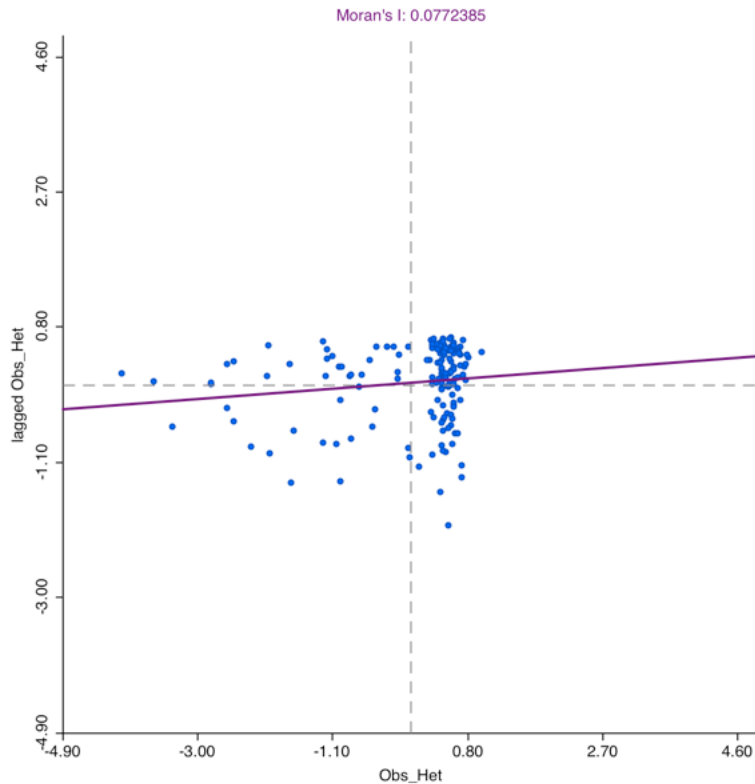
And on the pop-up click on Create. The Weight File Creation window will open, click on Add ID Variable, call it POLY_ID and then click on Add Variable. Set the neighbor weighting methods as shown here below:



Then click on Create. Save the weighting file as MOOA_env_Gen_Str.gwt. Now that the neighborhood weights are computed, we can move on to the Spatial Autocorrelation analysis. Click on:

>> Space > Univariate Local Moran's I

Be careful to select *Univariate **Local** Moran's I* and not *Univariate **Global** Moran's I*. Set Obs_Het as study variable and click on OK (the weighting system should be automatically set to the one you just created). In the pop-up window, select all the three windows to open and click on OK. Three windows appear:



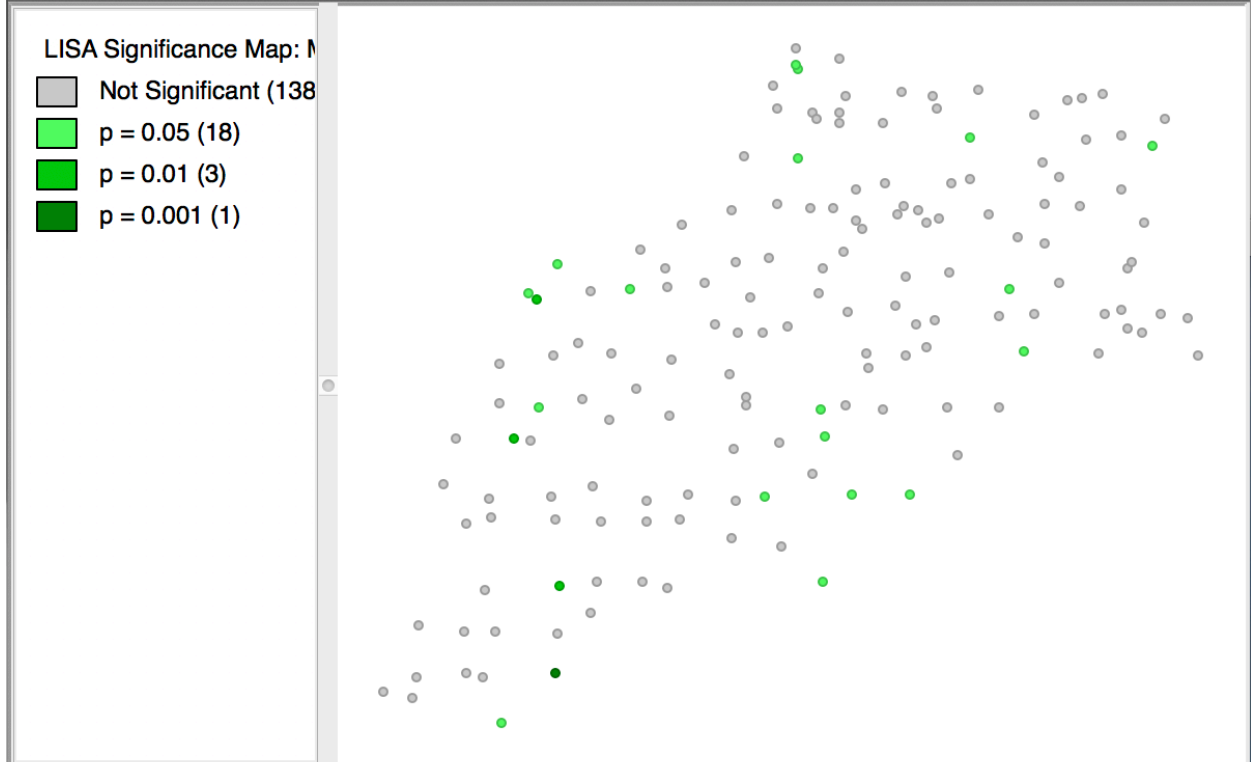
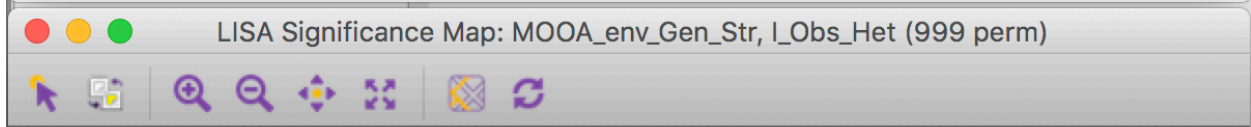
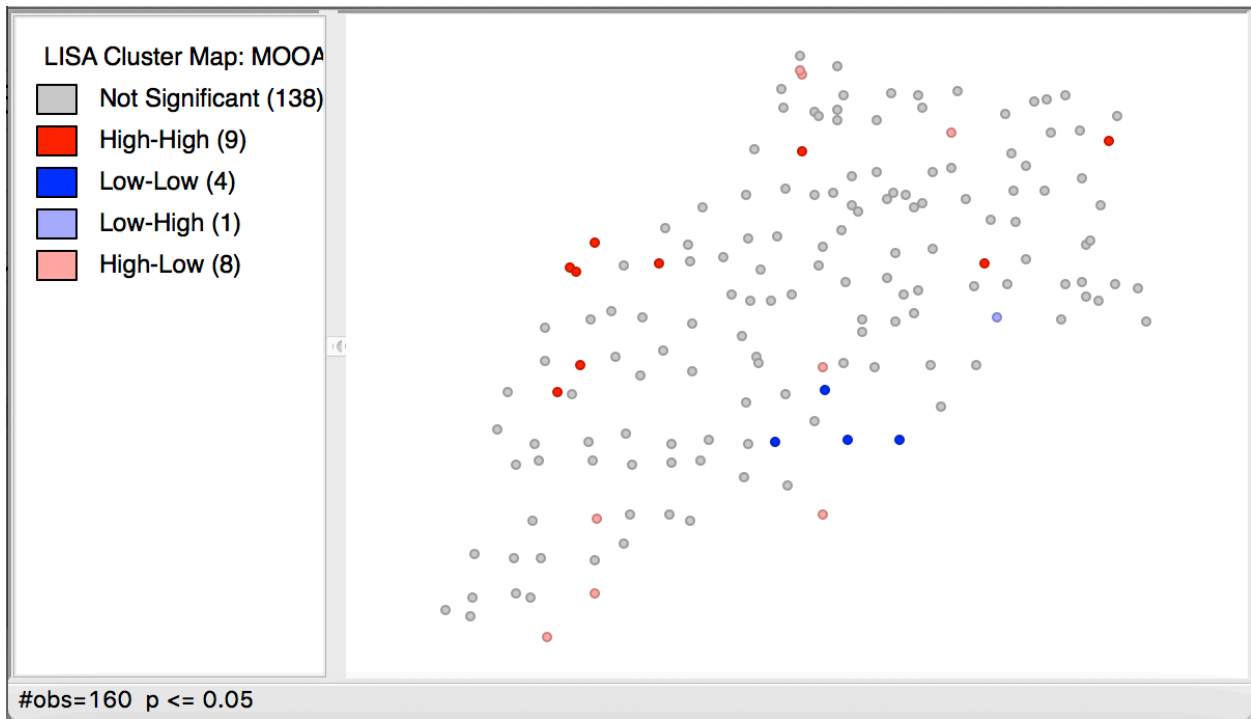
In the first window, it is represented the neighborhood mean heterozygosity (weighted average on the distance, y-axis) as a function of the neighborhood center heterozygosity (x-axis). Note that the values have been scaled. When the two values are similar, it means that neighborhood and center are similar, i.e. it might be a case of positive autocorrelation.

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

When, on the other hand, neighborhood and center have a contrasting value, we talk about negative autocorrelation.

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

The Moran I is the slope of the regression line and represent an index of global autocorrelation, the closer to 1 the more the set is positively autocorrelated, the closer to -1 the more negatively autocorrelated the set is. If the value is close to zero we say that the space is neutral. The global Moran I is the average of the local Moran I (for each point vs. his neighborhood). In the other two graphs we observe the significant maps of local I and their directions.



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

Try it by yourself!

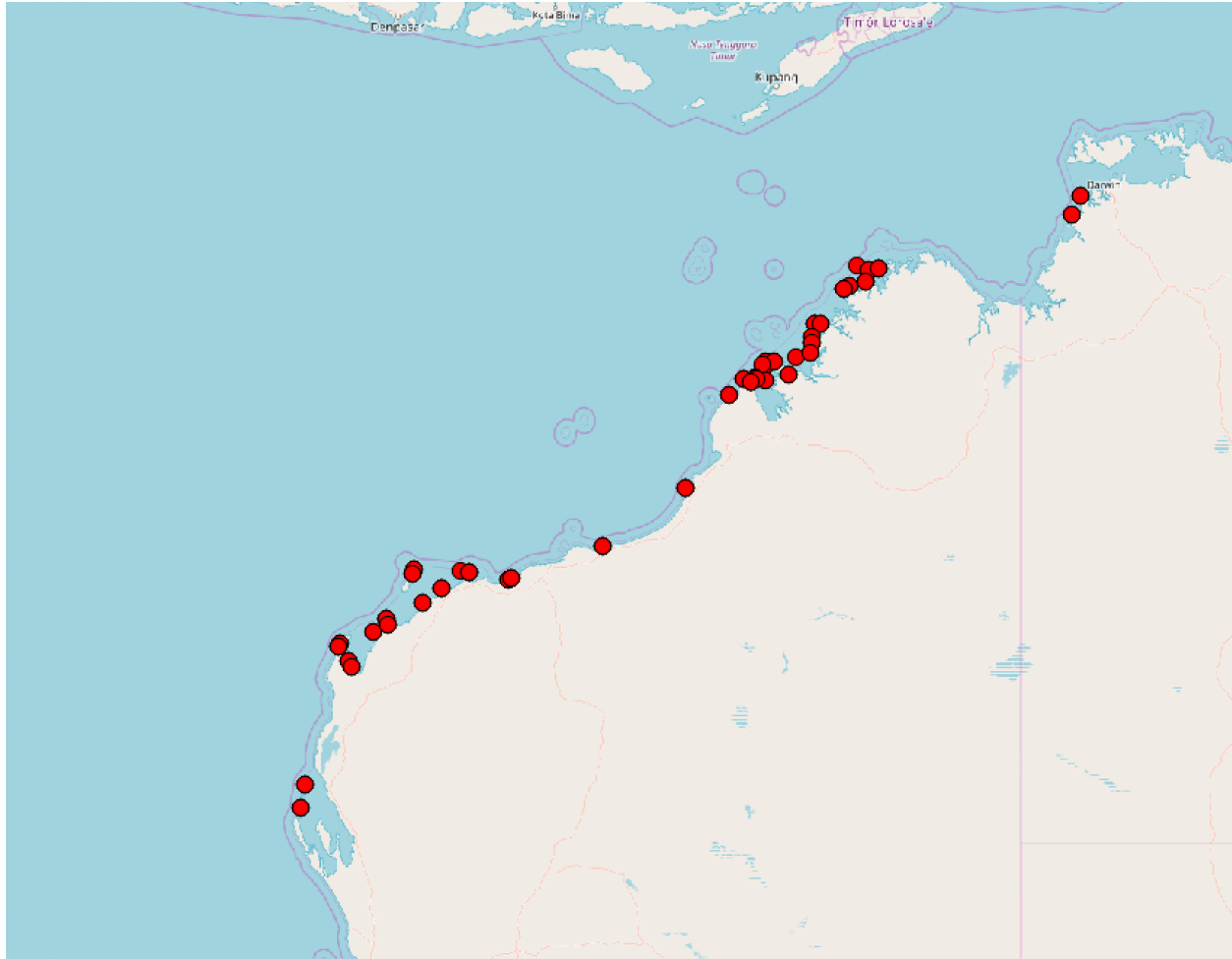
In the last part of the exercise we will apply the tools provided in the exercise to study the genetic structure of a fish population from western Australia. This dataset contains 300 individuals of Stripey Snapper (*Lutjanus carponatus*) genotyped for 17'007 SNPs using a Rad-seq genotyping strategy. This sequencing approach is becoming more and more common for the study of non-model species.

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

For this exercise, we are providing two files that you will have to load on R:

- StSn_SNPS.robj: is a non-filtered genotype matrix.
- StSN_env.shp: is the shapefile containing all the environmental information.

Here below it is shown the geographic distribution of the samples across the western coast of Australia.



HINT1: You can load the SNP matrix in your R session using the command:

```
load("StSn_SNPS.robj")
```

HINT2: You will notice that several samples shares almost identical coordinates. This is a frequent case in landscape genomics studies, since often researchers collect more samples at the same site. It is important to be careful when plotting gradients on a map (for instance, observed heterozygosity) since different values might superpose. To cope with it, the `plot_map_gradient` function can take additional arguments. You have two options:

- 1) `plot_map_gradient(shp=..., gradient=..., superposing='expand')` this flag allows to scatter superposing points and requires that the "plotrix" library is loaded in your R session.
- 2) `plot_map_gradient(shp=..., gradient=..., superposing='bysite', sites=...)` this flag calculates the mean value of the gradient variable at each site. It requires that the name of sites is specified using the `sites=` option (for instance, you can use the Reefs column in the shapefile).

Try to write your own code to:

- Filter the SNPS matrix for missingness, minor allele frequency and major genotype frequency.

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?

- Calculate Observed Heterozygosity and plot it on R.

Q8) Can you observe heterozygosity patterns?

- Evaluate the population structure using a Principal Component Analysis.

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