R practical courses for metagenomic-based population genomics

Amin Madoui and Romuald Laso-Jadart

April 2021

1 Introduction

The massive amount of environmental genomic data generated by recent sequencing projects such as *Tara* Oceans offers new perspectives, but also many challenges. Indeed, plankton is characterized by a wide range of species, for which almost no reference data are available, but is also an interesting choice of model to better understand. Thus, we must find ways to exploit and analyse metagenomic and metatranscriptomic data, without reference, to explore the evolution of the organisms.

One way to understand this evolution is to apply population genetics or genomics concepts and methods. This field relates on the analysis of polymorphic genomic markers across different populations of sampled individuals. Nowadays, population genetics studies are mostly based on single-nucleotide variants or SNVs. To date: an SNV is position in the genome presenting two or more alleles sufficiently represented in the population.

In this workshop, we will present two complete examples (based on previous studies) of reference-free population genomics based on environmental genomics.

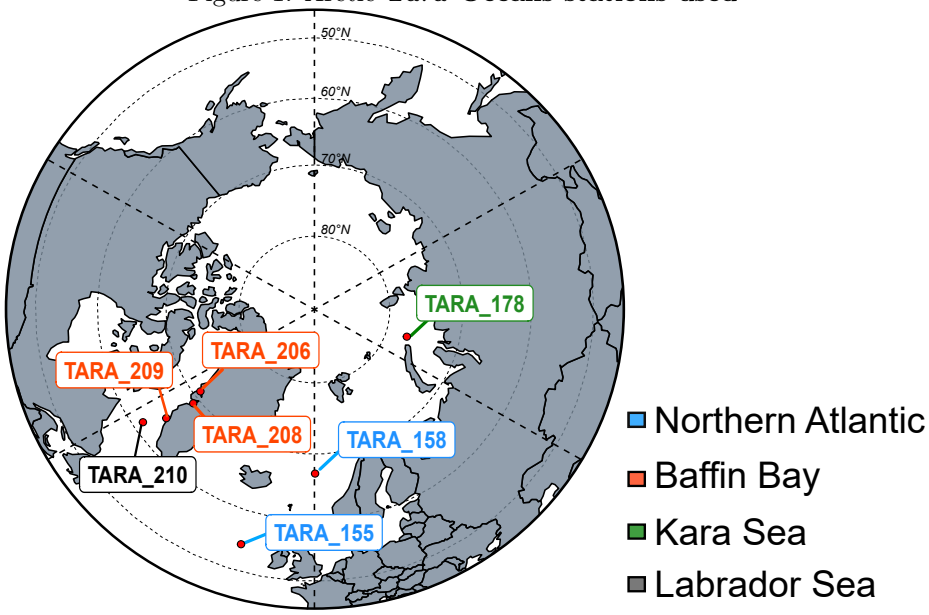
2 Adaptation and gene expression in polar *Oithona similis*

Understanding the adaptation and the acclimation of species to environment is a key to better assess how plankton populations evolve in the oceans. Specifically, this first part of the practical course is based on a previous studies which aimed to find a potential link between genomic adaptation through natural selection and phenotypic plasticity by focusing on gene expression variation.

We focused on a copepod species, *Oithona similis*, present in cold to temperate waters. Preliminary analyses using metabarcoding showed that this species

was present in *Tara* Oceans Polar Circle samples, and was almost the only one found in the size fraction 20-180µm.

Figure 1: Arctic ***Tara* Oceans stations used**



2.1 Extraction of genetic markers and filtering

In this exercise, seven stations of *Tara* Oceans Polar Circle were chosen to conduct the study (figure 1). We will use data generated from *Tara* Oceans expeditions applied on the polar copepod *Oithona similis*. The variants were detected from metagenomic and metatranscriptomic samples of the Polar Circle with a reference-free variant caller called *DiscoSNP++*, and reads were mapped on *de novo* transcriptomes to assess the accuracy of the variant calling. The alternative allele is chosen arbitrary by alphabetic order by *DiscoSNP++*.

This resulted in a dataset of more than 42,000 loci.

```
# First load some useful functions located in the file

source ("popgene_functions.R")

# Load the read count table

data=read.table('data/RC/OsimilisRC.txt',row.names = 1)
head(data)
```

```
poplist = c("155","158","178","206","208","209","210")

# Display depth of coverage for station 206

summary(data$GA206 + data$GB206)

hist(data$GA206 + data$GB206, breaks = 100)
```

Questions 1: A. Describe the shape of the depth of coverage distribution.

B. Why a locus can be over represented in read counts? Which issues can arise from it?

C. Why a locus can be under represented in read counts? Which issues can arise from it?

We thus need to chose a window of depth of coverage to retrieve loci with a correct depth of coverage, among all the samples i.e loci from single copy regions of the genome.

```
# Filter on the depth of coverage

data = FilteringCov(data, poplist, dev=2, minCov=5, maxCov = 150)

hist(data$GA206 + data$GB206, breaks = 100)
```

A second filtering step has to be performed on frequencies (cf Question 1).

Question 2: How are computed B-allele frequencies?

```
# Create an allele frequency matrix

freq = CreateFreqMatrix ( data, poplist )

# Example of two loci with unwanted frequencies (fixed to 1 or under 0.05

barplot(freq["13564462"])

barplot(freq["10789206"])
```

Questions 3: A. Describe the frequencies of these variants among the populations.
B. What does it mean to have a B-allele frequency too close to 1 in each sample?
C. What does it mean to have a B-allele frequency too close to 0 in each sample?
D. Which issues does it create for further statistical analysis?

```
# Create a clean data set after filtering on frequency and expression
 data = FilteringFreqExpression(data, poplist ,GAF=0.1, TAF=0.05)
```

2.2 Estimation of the genomic differentiation between zooplankton populations

2.2.1 Allele frequency spectrum

```
# Calculate the B-allele frequencies

 Gfreq = data$GB206/(data$GA206+data$GB206)

# Observe the distribution of the B-allele frequencies

 hist(Gfreq, breaks = 10, main="B allele frequency distribution")
```

Questions 4: A. What is a typical distribution of allele frequency?
B. How is the distribution of your allele frequencies?
C. Why?

2.2.2 Pairwise-F_{ST}

We will use the Wright formulation of the F_{ST} as an estimator of the genomic differentiation with:
$F_{ST} = \frac{var(p)}{mean(p)(1-mean(p))}$.

This estimator is a biaised estimator, but is here used for the sack of clarity.

```
# The load the B-allele frequencies in a data frame called freq
freq = CreateFreqMatrix(data, poplist)

# Rename the columns of freq

colnames(freq) = c("TARA_155","TARA_158","TARA_178",
"TARA_206","TARA_208","TARA_209","TARA_210")

# compute the pairwise-FST matrix

pairwiseFST = pwFst(freq)

print(pairwiseFST)

# plot the pairwise-FST matrix
```

```
plotFST(pairwiseFST)
```

Questions 5: A. How does the genomic structure of *Oithona similis* in the Arctic look like?
B. Do you see any structure?
C. How can you explain this result?

2.3 Isolation by currents

We can compare the genetic structure with geographic distances and marine currents.

```
# test the isolation by distance using euclidean distances

euclideans = read.table("data/Euclidean_Arctic.txt",
row.names=1,header = TRUE)

MantelPlot(euclideans,pairwiseFST)

# Perform a Mantel plot

mantel.euclideans = mantel(vegdist(pairwiseFST,na.rm=TRUE),
vegdist(euclideans,na.rm=TRUE))

print(mantel.euclideans)

# test the isolation by distance using lagrangian distances

lagrangians = read.table("data/Lagrangian_Arctic.txt",
row.names=1,header = TRUE)

# Perform a Mantel plot

MantelPlot(lagrangians,pairwiseFST)

mantel.lagrangians = mantel(vegdist(pairwiseFST,na.rm=TRUE),
vegdist(lagrangians,na.rm=TRUE))

print(mantel.lagrangians)
```

Question 5: Is there a link between genetic distances and geographic distances? Marine currents?

3 Detection of loci under natural selection

After analysing the genetic structure of *Oithona similis*, we can scan the loci to detect potential loci under natural selection. We will use two different approaches

3.0.1 LK outliers method

We will use the LK, as $LK = \frac{(n-1)FST}{(FST)}$, n being the number of populations which is supposed to follow a chi square distribution under neutral evolution. LK outliers can be considered under selection.

```
freq = CreateFreqMatrix(data, poplist)

colnames(freq) = c("TARA_155","TARA_158",
"TARA_178","TARA_206","TARA_208","TARA_209","TARA_210")

# Compute LK and test outliers, significance is store in LK$q_value
LK = LK(freq)

head(LK)

# Plot the observed LK distribution

hist ( LK$LK, freq = F, xlab = "LK" , breaks = 50 ,
main = "LK distribution")

# add the neutral evolution

n_pop = ncol(freq)-1

curve ( dchisq( x, n_pop ) , lwd=3, col="orange" , add=T )

# select loci with higher LK value than expected

selection_LK = LK[LK$q_value<0.1,]

dim(selection_LK)
```

Questions 6 : A. Which property should be met by variants to allow to detect selection?
B. How do we call the historical theory behind it?

3.0.2 pcadapt method

Pcadapt is a method enabling to detect variants under selection. It is based on a principal component analysis (PCA) computed using a matrix of allele frequency. Compared to LK, it takes into account the structure of the population. It computes a corrected mahalanobis distance which is expected to follow a chi square distribution under neutral evolution. Similarly to LK, outliers are candidate to natural selection.

```
# Get the matrix ready for pcadapt by transposing it

freq_forPCA = read.pcadapt(t(freq), type = "pool")

# Run pcadapt

pcadapt_result <- pcadapt(freq_forPCA)

# Plot results

plot(pcadapt_result, option = "scores", pop = colnames(freq))

plot(pcadapt_result, option = "screeplot")

plot(pcadapt_result, option = "qqplot")

# Correction of p-values with Benjamini-Hochberg

q_values = p.adjust(pcadapt_result$pvalues, method = "BH")

freq$pcadapt_qval = q_values

freq[is.na(freq)] <- 1

dim(freq[freq$pcadapt_qval<0.05,])

# Compare LK and pcadapt

freq$LK_qval = LK$q_value

dim(freq[freq$pcadapt_qval<0.05 & freq$LK_qval<0.1,])
```

Questions 7 : A. What do the PCA reflect? Is the trend similar to the pairwise-Fst matrix?
B. Is the intersection between pcadapt and LK high?
C. Which approach seems more accurate?

3.1 Population-scale allele expression

At the cellular, tissue, or individual levels, the expression of one allele can vary compared to the other one (for a biallelic locus). To observe acclimation in *Oithona similis*, we extrapolate variation of expression at the population-scale, using environmental genomics. The goal is to find loci with alleles having an expression that varying from their genomic abundance. In other words, alleles which are more (or less) present in metagenomic data than in metatranscriptomic data. We suppose that this mechanism plays a role in plankton rapid acclimation.

3.1.1 Comparison of allele expression to genomic abundance

The first step is to find this observation in our data.

```
# Compute genomic allele frequecis
 Gfreq = data$GB206/(data$GA206+data$GB206)


# Compute relative allele expression

 Tfreq = data$TB206/(data$TA206+data$TB206)

 # Observe between allele expression and genomic abundance

 plot(Gfreq,Tfreq)

 # Test their correlation
summary(lm(Tfreq~Gfreq-1))
```

Question 8 : A. What is the trend observed between genomic abundance and relative expression?
B. Is is expected?

3.1.2 Detection of allele-specific expression

```
# We will focus on population from TARA_206

ASE_206 = data[,c("GA206","GB206","TA206","TB206")]

# Use only heterozygous sites in TARA_206

ASE_206 = ASE_206[ASE_206$GA206 >= 1 & ASE_206$GB206 >= 1,]

# test for specific allele expression
```

```
Fisher_pvalue = apply(ASE_206,1,
function (x) fisher.test (matrix(x, nrow = 2))$p.value)

# Apply FDR for multiple tests

Fisher_qvalue = p.adjust(Fisher_pvalue, method="BH")

ASE_206$Fisher_qvalue = Fisher_qvalue
```

Question 9: A. How many alleles are under specific expression? Which proportion of the dataset?

3.1.3 Link between natural selection and allele specific expression

Now that loci under selection and under population-scale differential allele expression were found, we can hypothesis that some loci are subjected to adaptation and to acclimation in *Oithona similis* populations.

```
# Finding loci under selection and under allele specific
# expression in TARA_206

# Find loci under selection

selection_ASE_S206 = intersect(rownames(freq[freq$pcadapt_qval<0.05,]),
                               rownames(ASE_206[ASE_206$Fisher_qvalue < 0.05,]))

# Test the size of the intersection by a Hypergeometric test

phyper(length(selection_ASE_S206),
#white balls drawn

       nrow(freq[freq$pcadapt_qval<0.05,]),
       #total white balls in the urn

       nrow(freq)-nrow(freq[freq$pcadapt_qval<0.05,]),
       # total black ball

       length(Fisher_qvalue[Fisher_qvalue<0.05]),
       # black ball drawed in the urn

       lower.tail = F)
```

Questions 10 : A. Why using a hypergeometric test? B. Is the intersection significant, and what would this mean biologically speaking?

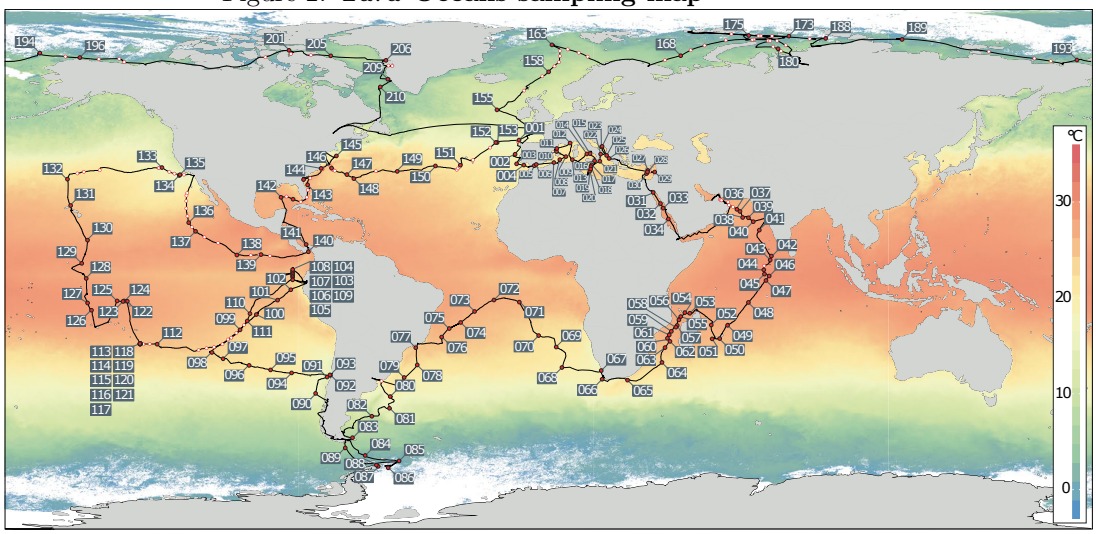
Question 11 : A. Run these analyses for few other stations.
B. Do we find the same trend in the other populations?

4 Reference-free approach and metavariant species

Most of plankton species lack a genome reference. This problem can be bypassed by introducing the concept of metavariant species (MVS), meaning that species are represented only by their polymorphic loci. This can be constructed by detecting variants using (*DiscoSNP++*) and by clustering co-abundant loci expected to belong to the same species using the R package *metaVaR*.

This approach has been applied on *Tara* Oceans metagenomic data from 35 stations from Mediterranean Sea and Atlantic Ocean (figure 2). In this section you will manipulate MVSs and try to find which environmental factors drive their genomic differentiation. MVSs are already produced.

Figure 2: ***Tara* Oceans sampling map**



4.1 Genomic differentiation of metavariant species

We will take the example of a MVS identified as a Ciliophora.

```
# Reading metavariant species

MVS.pwFst = read.table("data/MVS/Ciliophora/pwFst.txt",row.names = 1,
header=T)

# Fst matrix construction

plotFST(pairwiseFST = MVS.pwFst)
```

Question 12: Do you observe any structure?

4.2 Role of the environmental factors

In order to identify the environmental drivers of genomic differentiation, we downloaded environmental parameters from a public database called World Ocean Atlas, covering the 35 *Tara* stations and the dates of the expedition. In addition, using a data-driven method, based on a database containing the positions and trajectories of drifters launched in the oceans since 1979, we obtained the estimation of the transport time (or Lagrangian time) between each *Tara* station.

```
# Reading environmental data and extract data

EnvTable = read.table("data/WorldOceanAtlas_data.txt",header = T)

envlist = list(Temperature = EnvTable[,c(1,7)],
               Salinity = EnvTable[,c(1,8)],
               Silicate = EnvTable[,c(1,4)],
               Po4 = EnvTable[,c(1,6)],
               No3 = EnvTable[,c(1,5)])

# Reading Lagrangian matrix

LagrangianMatrix = read.table("data/Lagrangian_Matrix.txt",header = T)

# Plot Lagrangian matrix

ggplot(data = melt(as.matrix(LagrangianMatrix)),
       aes(Var1,Var2, fill = value)) +
       geom_tile(color = "white",names ="Travel_Time_(days)") +
       scale_fill_viridis() +
       theme(panel.border = element_blank(),
       panel.grid.major = element_blank(),
       panel.grid.minor = element_blank(),
       axis.line = element_line(colour = "black"),
       axis.text.x = element_text(angle=90,vjust = 0.8 ),
       axis.text.y = element_text(vjust = 0.8 )) +
       xlab("") + ylab("")
```

Question 13: Which trends do you observe in the heatmap?

4.3 Variance partitioning by environmental factors

We will compare the Fst matrix to environmental matrices, by applying a linear mix model to decompose the influence of the different environmental factors.

```
# Apply a linear mixed model on pairwise-Fst matrix

VariancePartition.Res = LinearMixedModelMVS(MVS.pwFst,
                                            LagrangianMatrix, envlist)

# Plot the result

PlotLMM(VariancePartition.Res)
```

Question 14: Which parameter drive the genomic differentiation of this Ciliophora?
Question 15: Perform similar analysis on other MVS present in the data directory and compare your results between the different species.
Question 16: Do you a specific trend?