Population structure

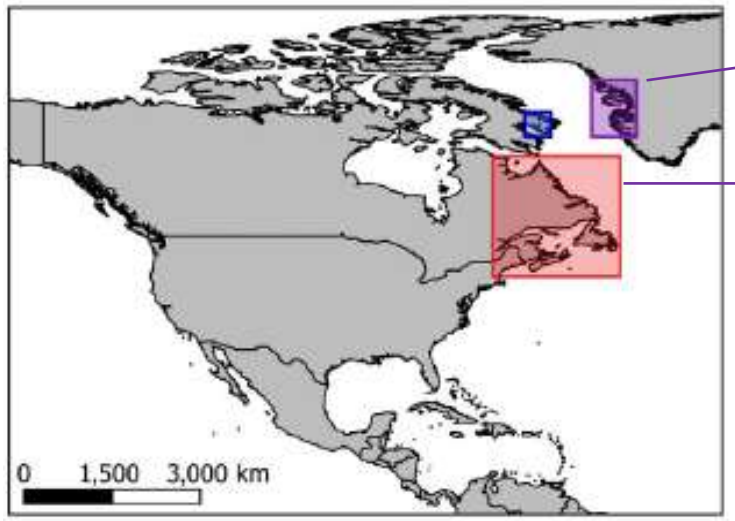
Overview of what we saw yesterday

Dataset 1	Dataset 2	Dataset 3
« 2_lin »	« all »	« canada »
4 populations (2 greenland /2 canadian) => 80 samples	14 populations (2 greenland /12 canadian) => 280 samples	12 populations (12 canadian) => 240 samples
Fst (vcftools) PCA Optional (Fst with Stacks)	Faststructure DAPC	PCA DAPC Optional (Pairwise Fst) -> ALL analyses of day 3-day4-day5

Fst

2 population
N= 40

12 populations
N= 240 (20/pop)

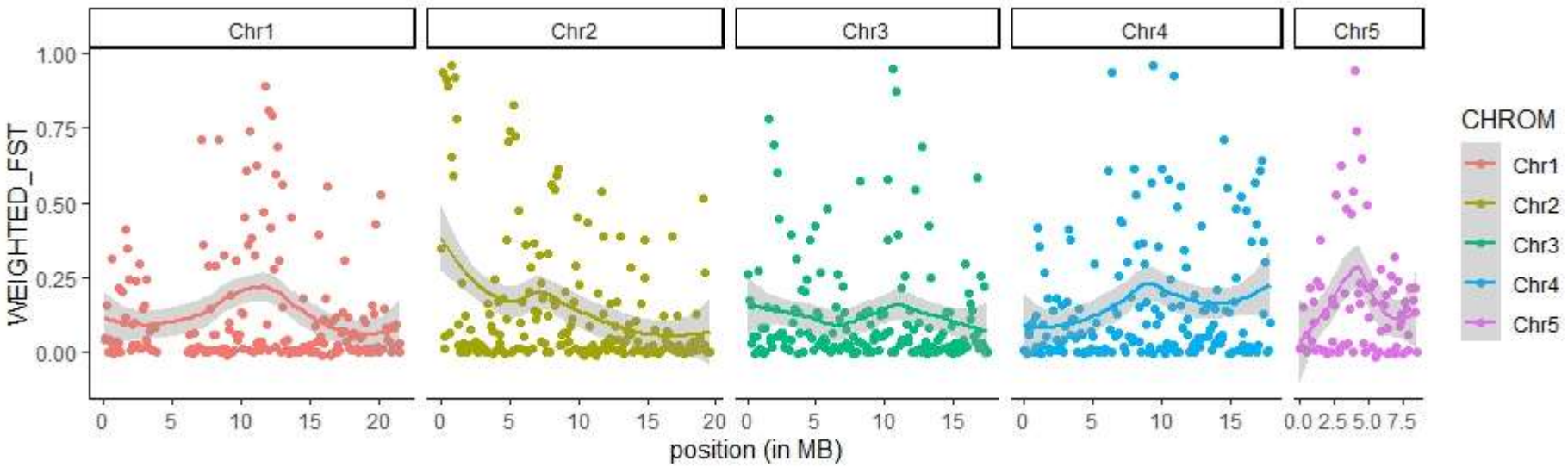


Greenland species

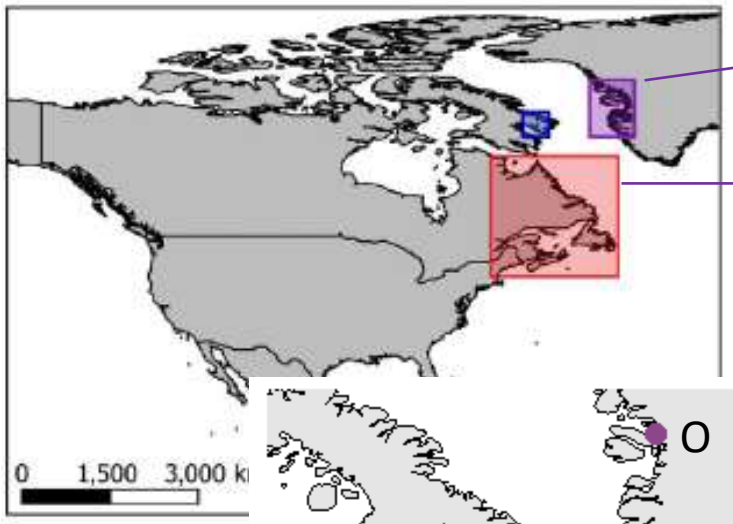
North American species

Fst = 0.23

Mean Fst 0.03 / weighted Fst 0.23

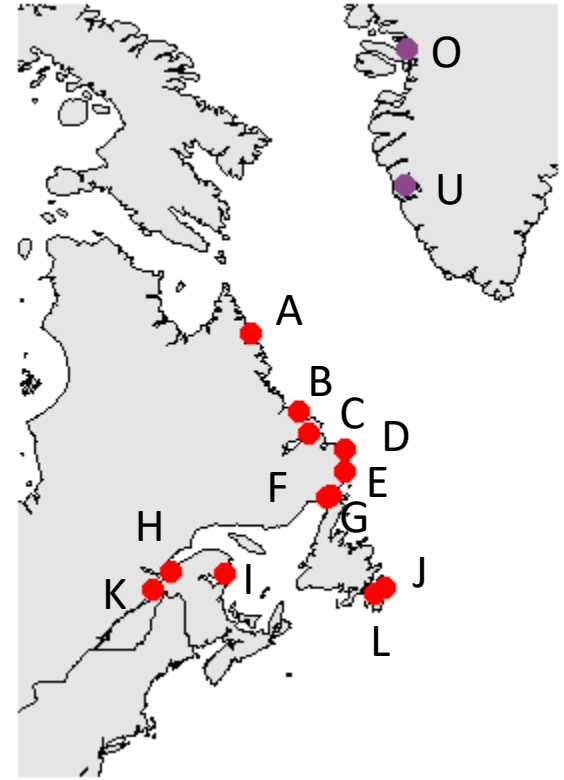


FastStructure



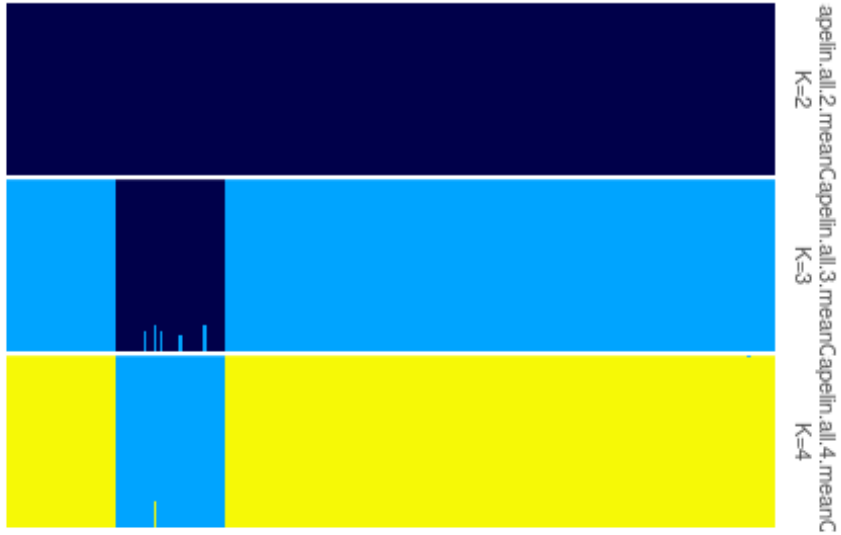
Greenland species

North American species



2 lineages

All

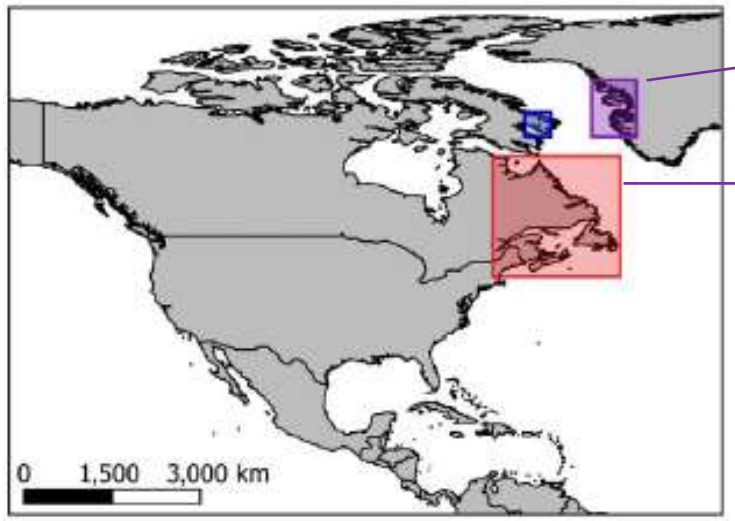


O – U
= Greenland

PCA

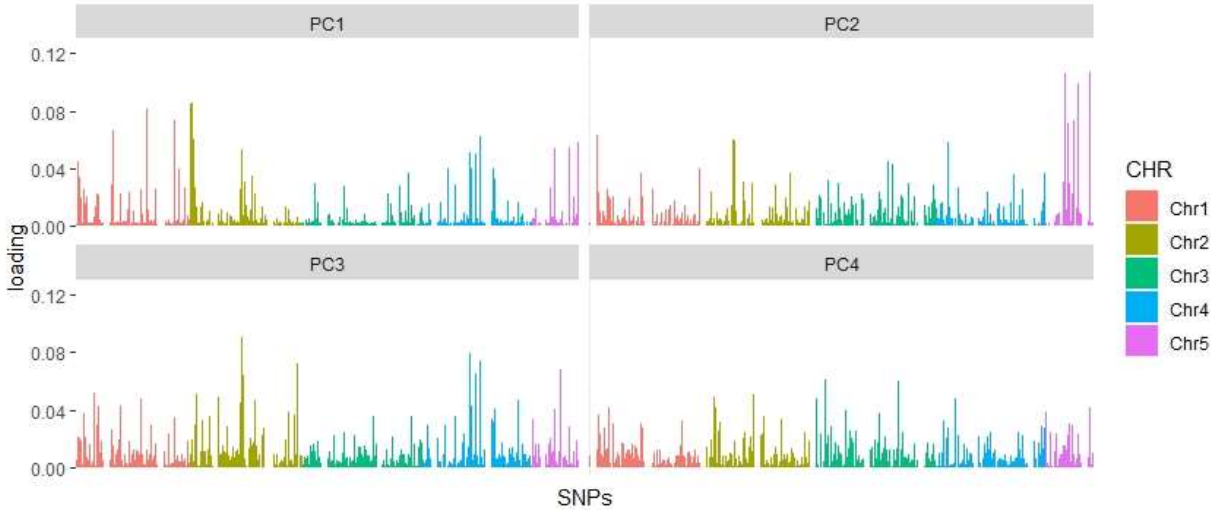
2 population
N= 40

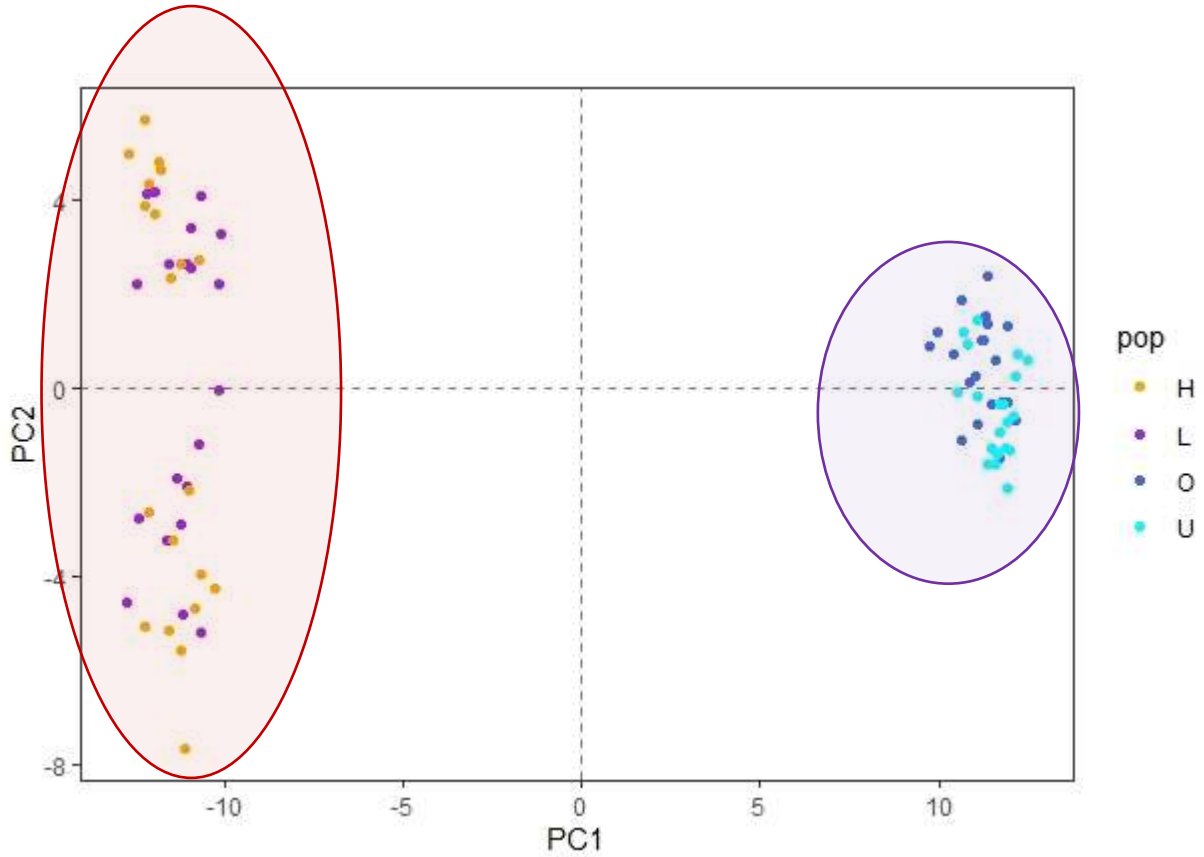
12 populations
N= 240 (20/pop)



Greenland species

North American species

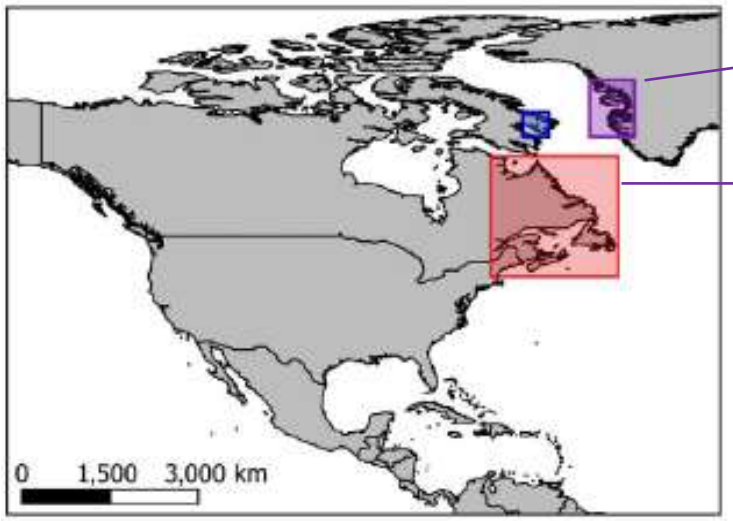




DAPC

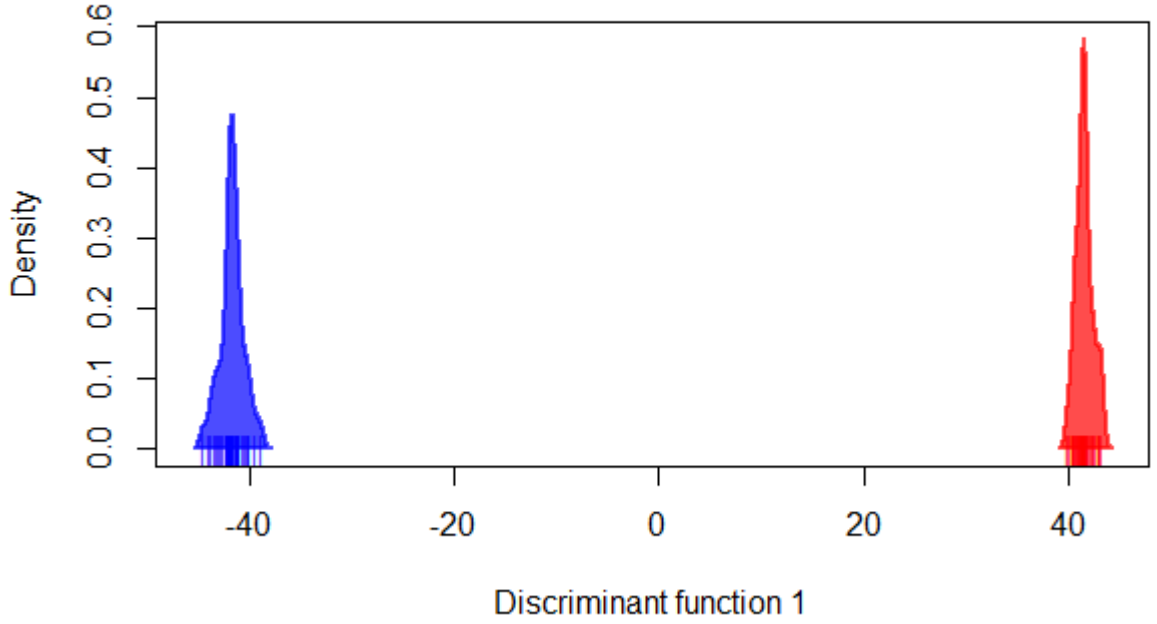
2 population
N= 40

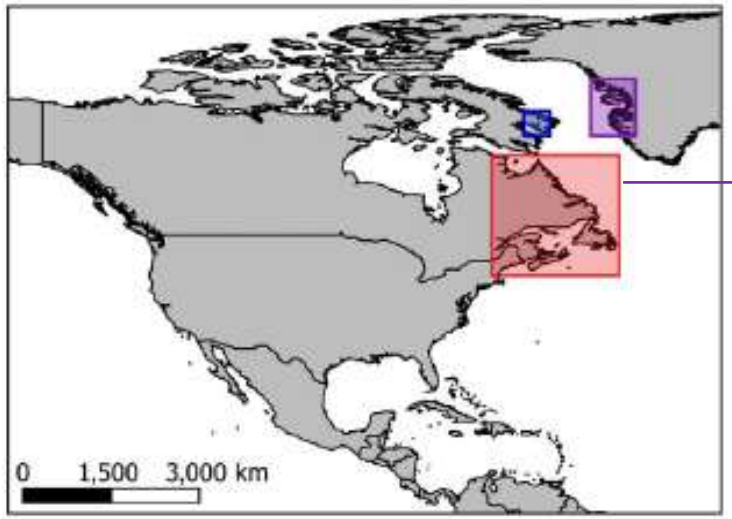
12 populations
N= 240 (20/pop)



Greenland species

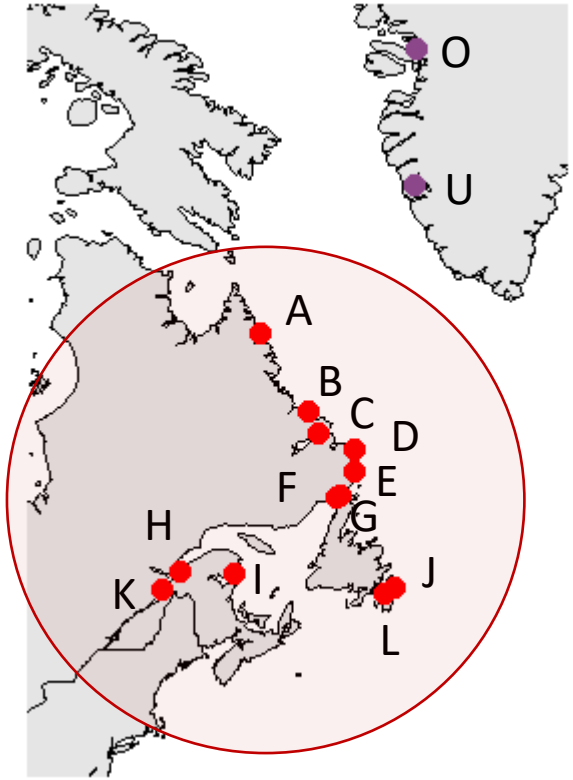
North American species



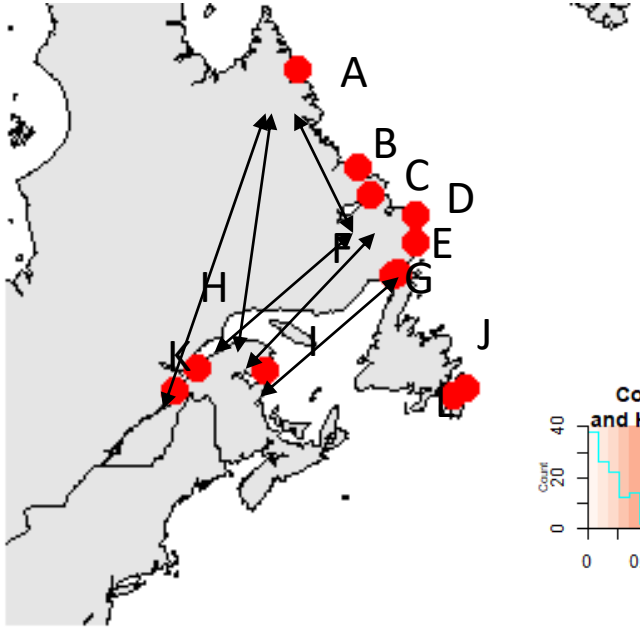


North American species

12 populations
N= 240 (20/pop)

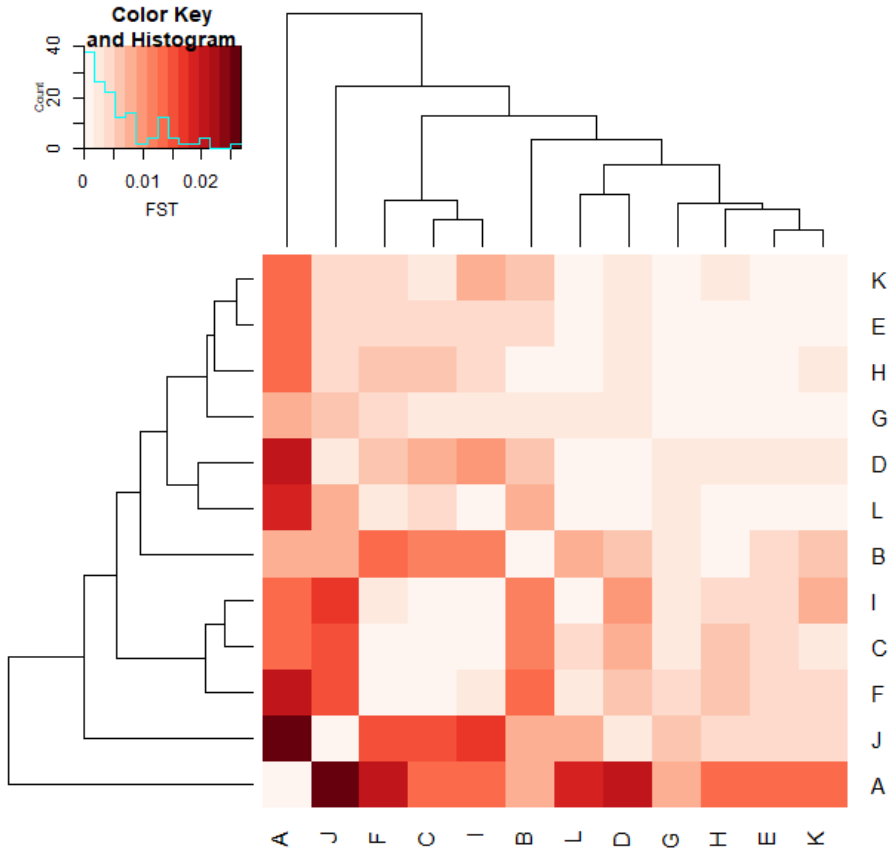


Fst



Do we observe genetic structure ?

Fst between all poulations



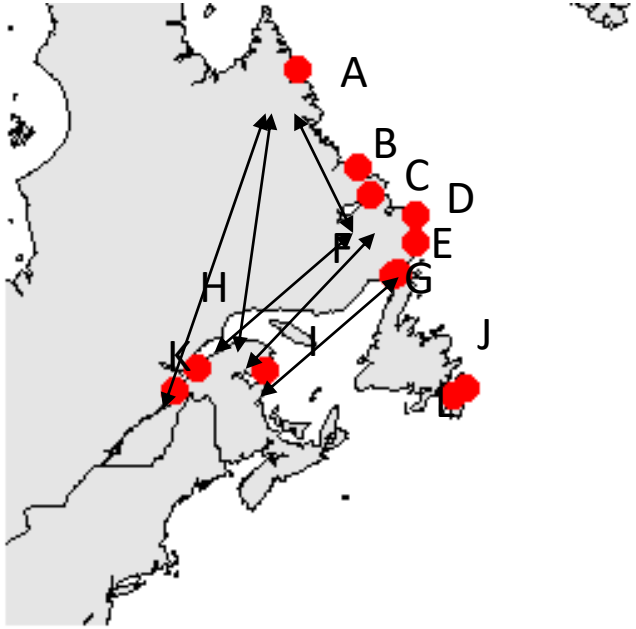
Medium values (Fst = 0.025)?
Lots of heterogeneity…

⇒ pop A: 0 females, 20 males
⇒ pop J: 18 females and 2 males

⇒ Sex-linked markers + unbalance sample size influence differentiation

⇒ Solutions:
- better sampling?
- exclude sex-linked markers (chr 5)!!

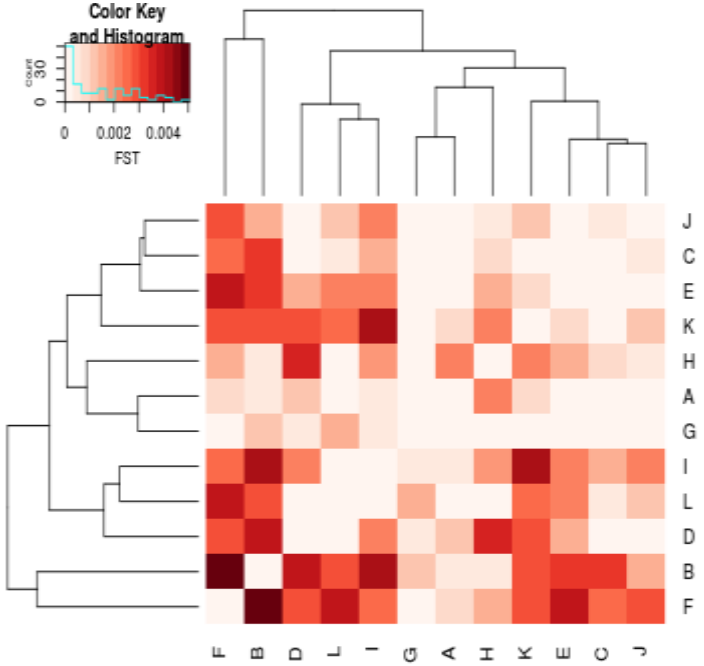
Fst



Do we observe genetic structure ?

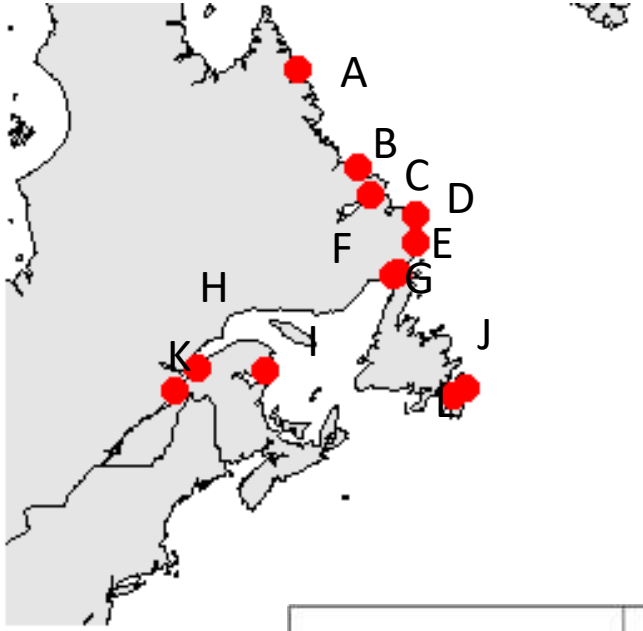
Fst between all poulations
Excluding chromosme 4 (inversion) & chromosome 5 (sex)

Note the highest values : they are now at about 0.005 instead of 0.025



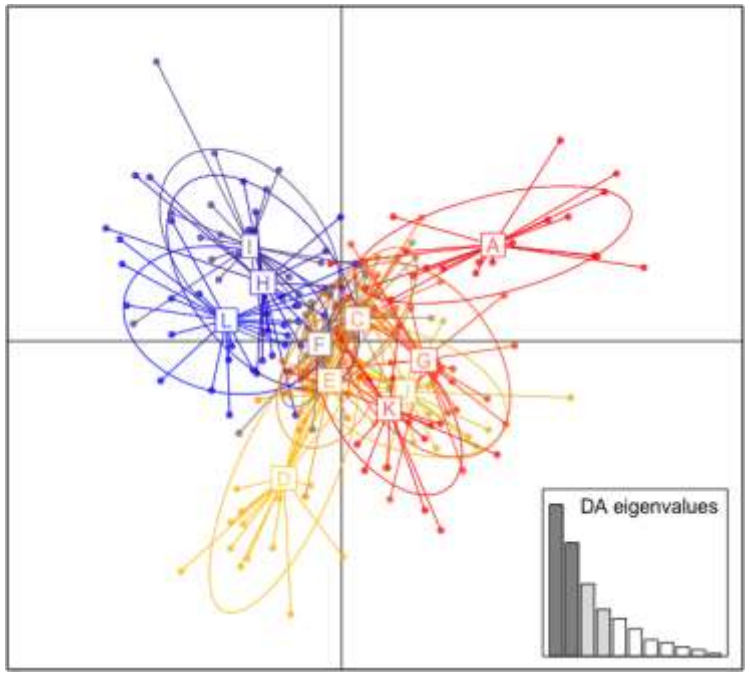
⇒ Almost no geographic structure…

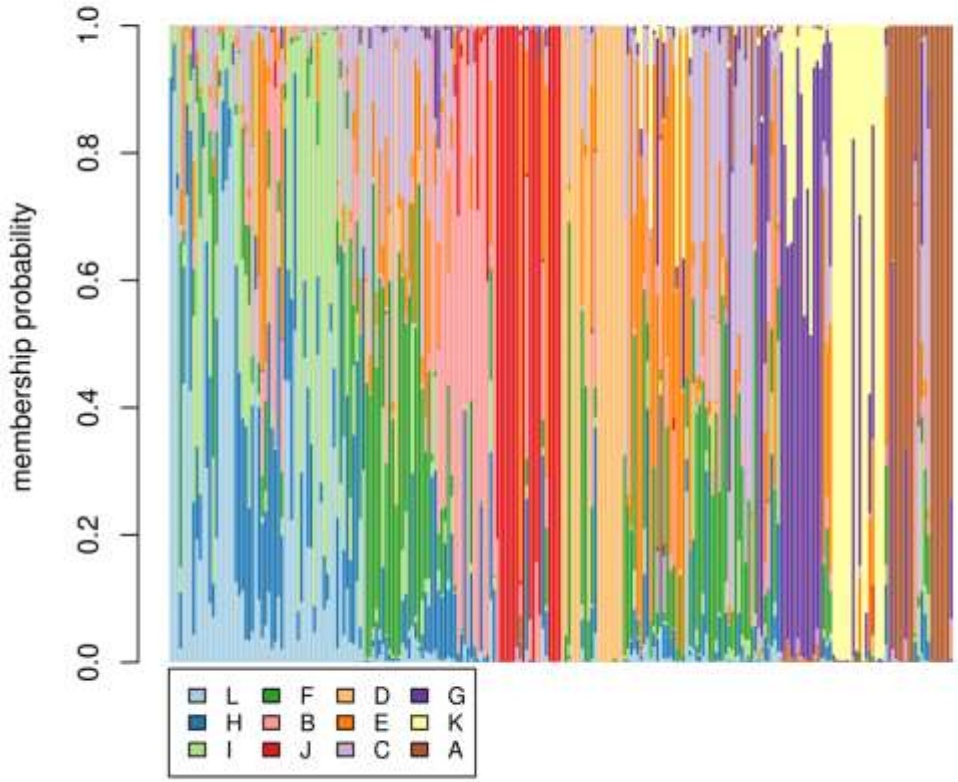
DAPC



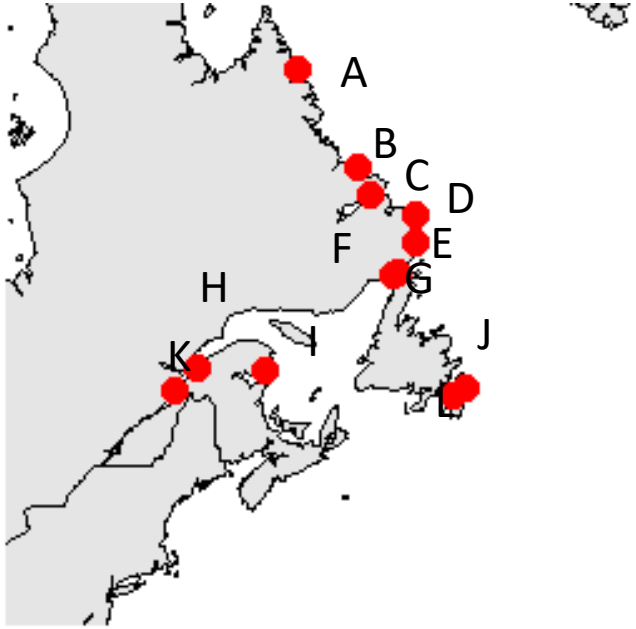
Do we observe genetic structure ?

DAPC (yesterday) -> when we avoided over-fitting, no genetic structure related to geography (12 populations)

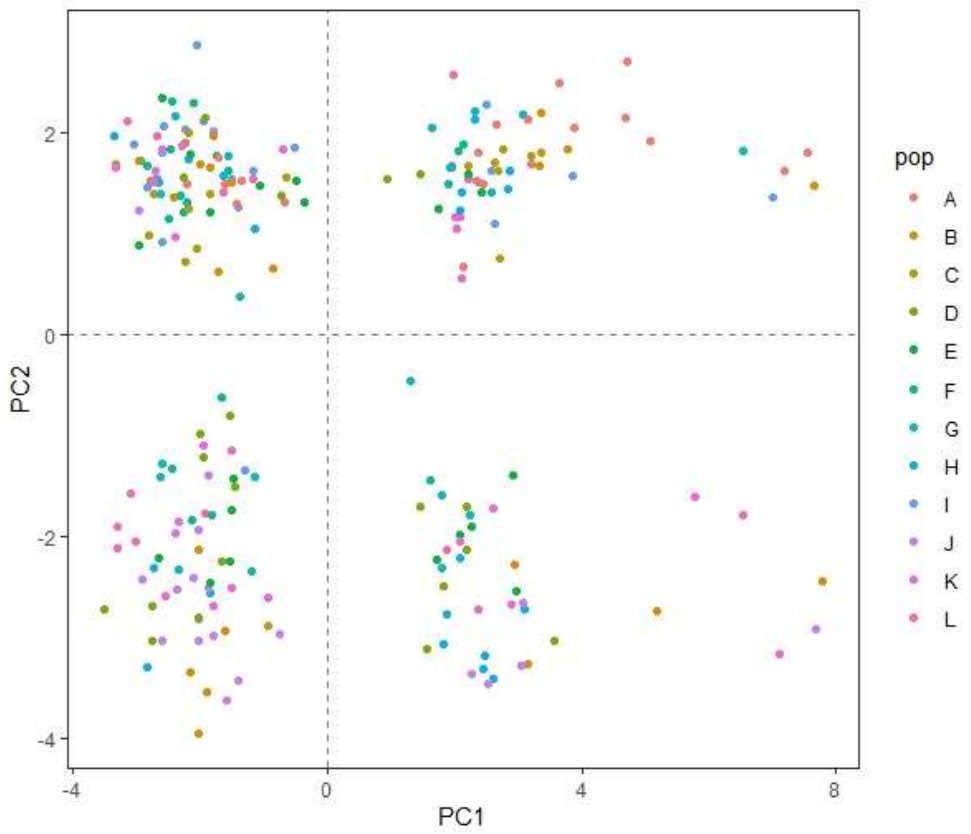




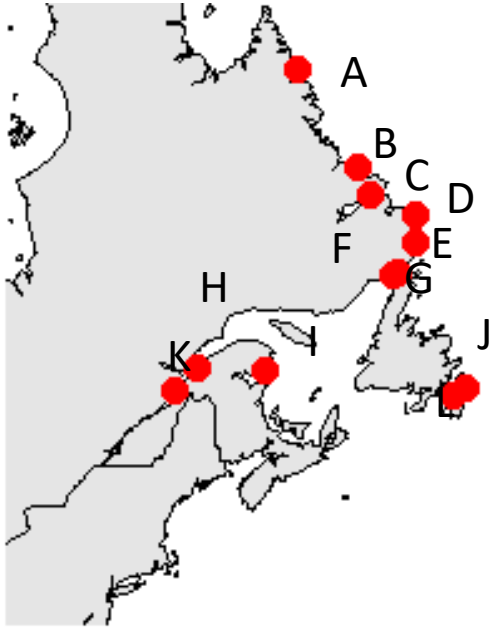
PCA



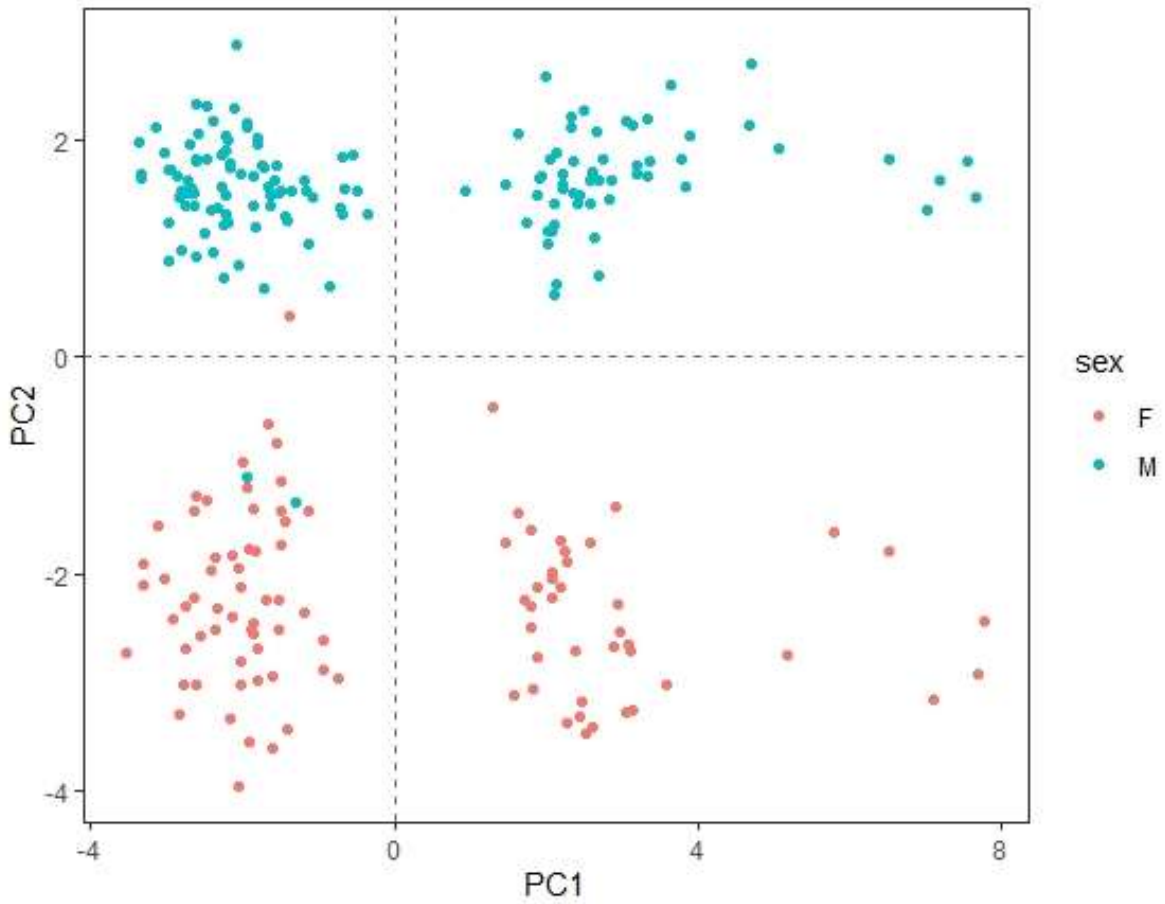
Canada (12 pop)



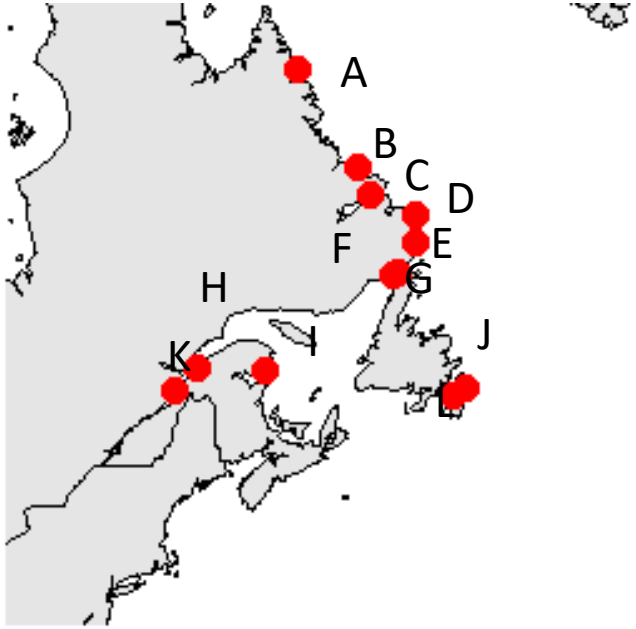
PCA



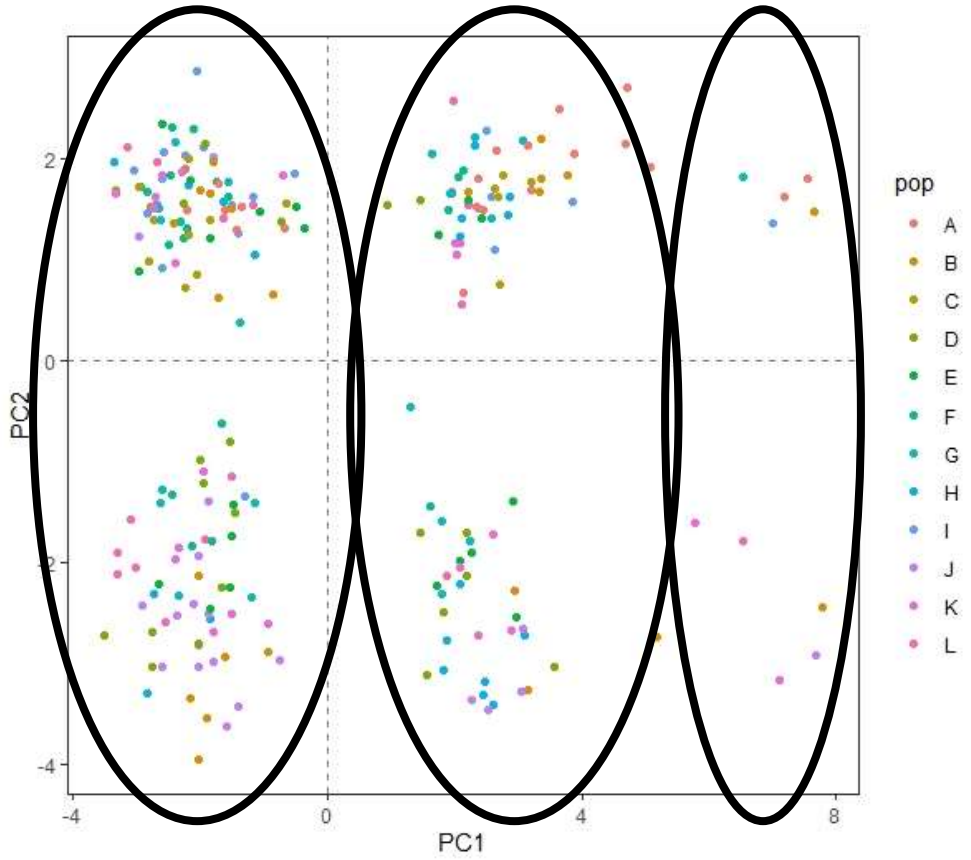
Canada (12 pop)



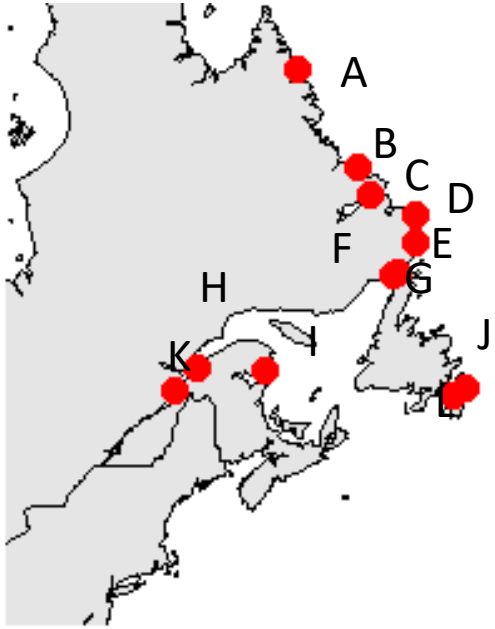
PCA

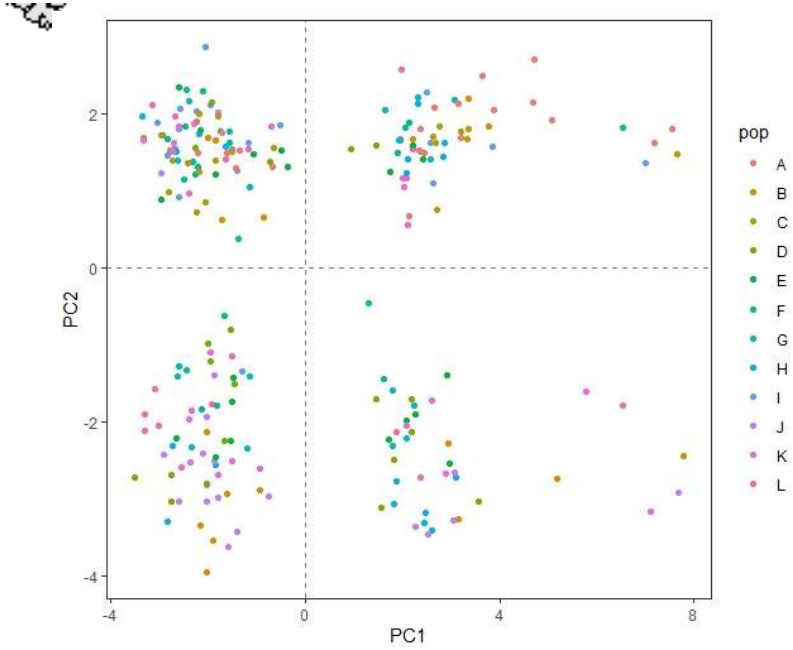


Canada (12 pop)



PCA

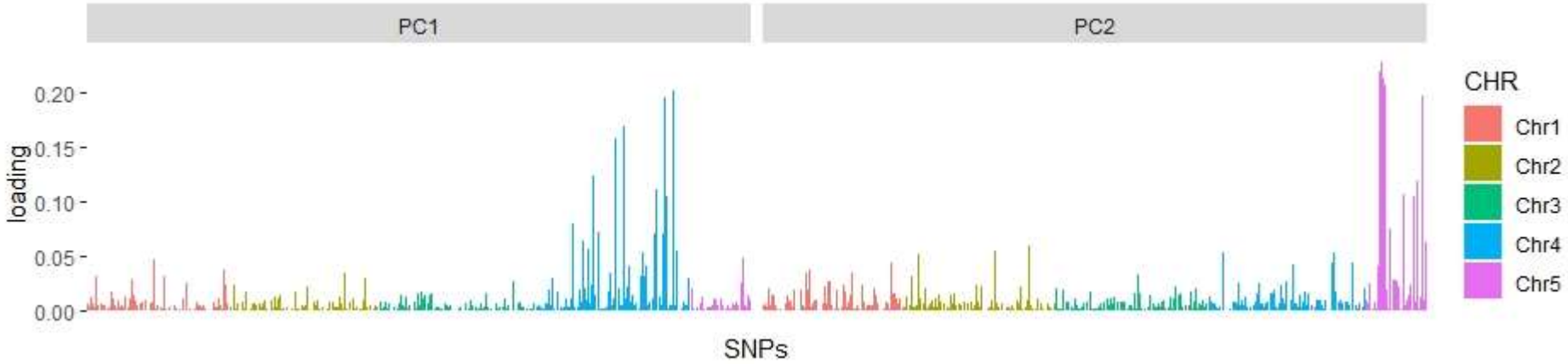




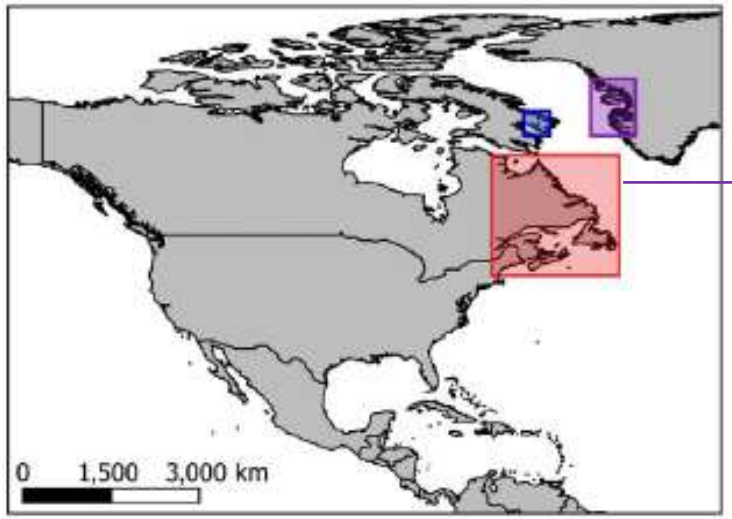
12 pop)

Rearrangement on Chr 4

Sex-determining locus on Chr 5

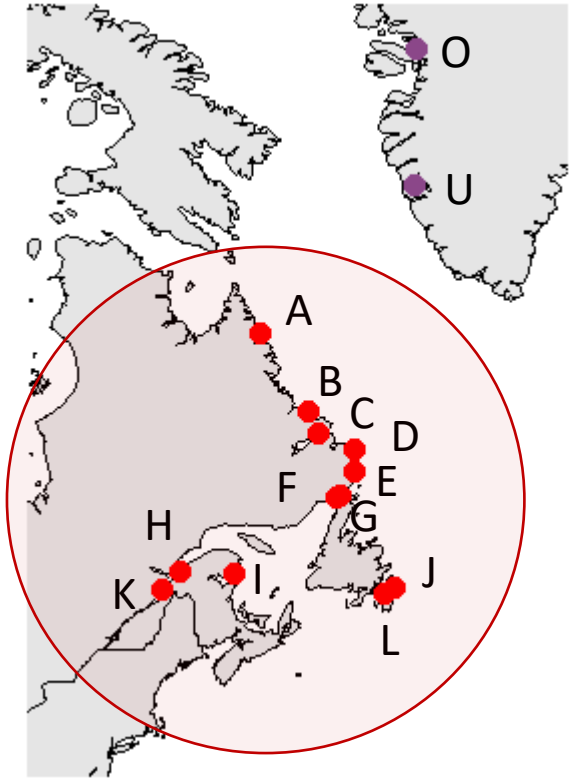


Today’s practical

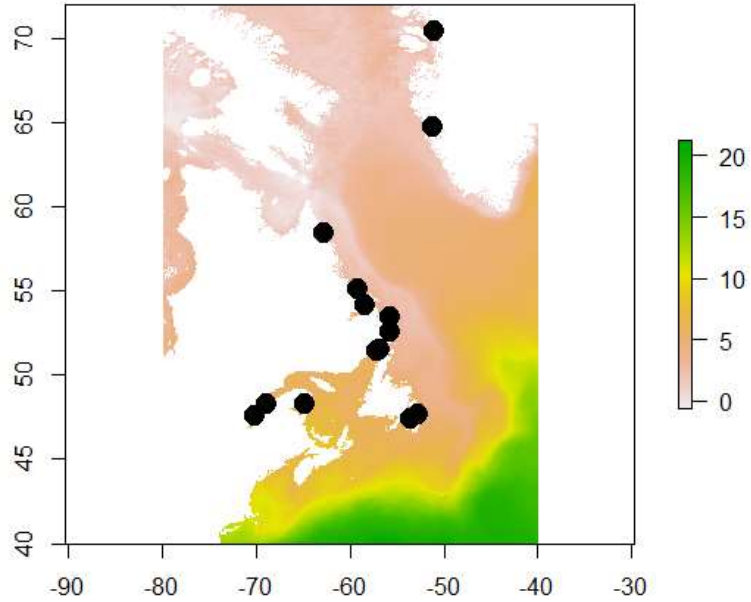


North American species

12 populations
N= 240 (20/pop)



Sea temperature (from MARSPEC)



Day 3: Local adaptation

Disentangle population structure & putative signature of adaptation…

3-1 Fst statistics & geography

→ We did so yesterday! (short manipulation to do LD-pruning today)

3-2 Outliers of differentiation

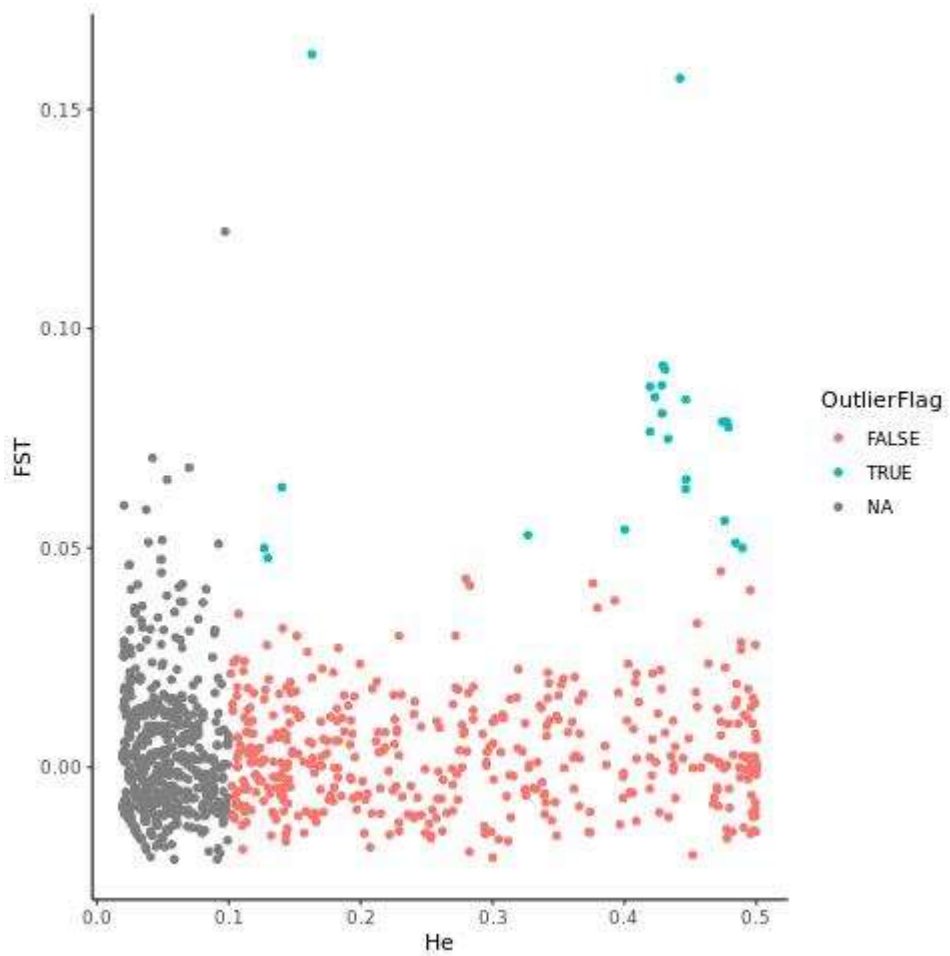
3-3 Genotype-environnement assocations

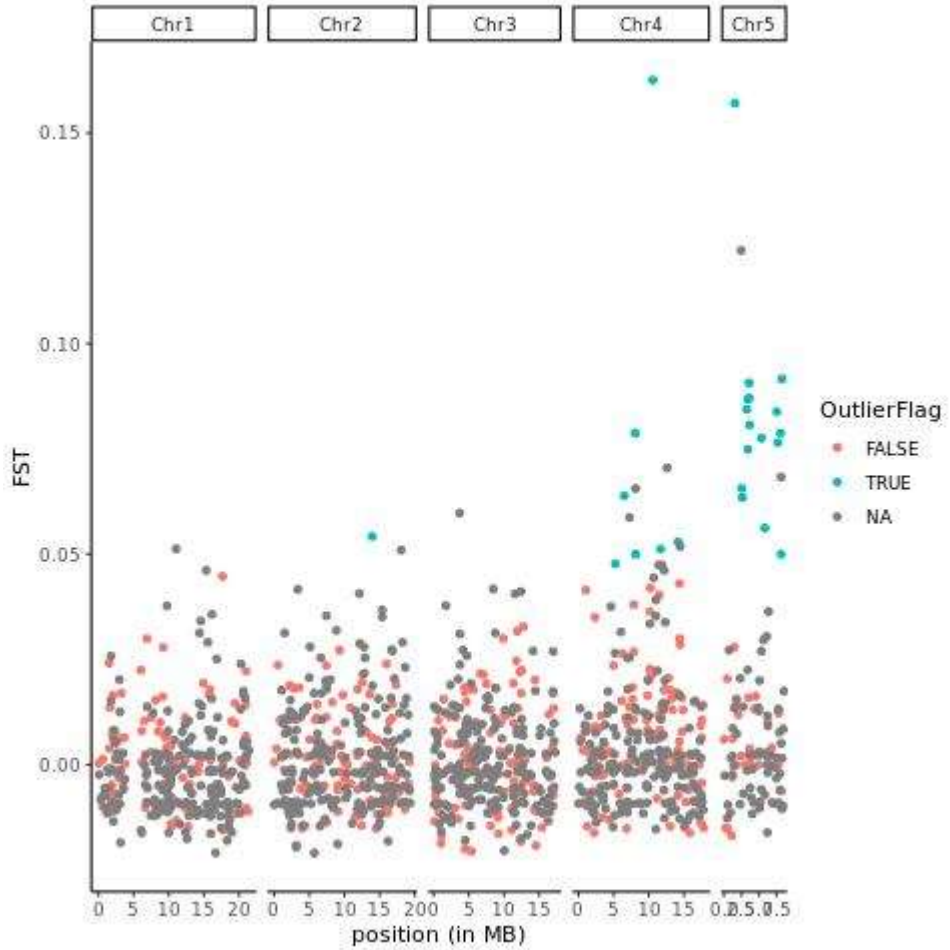
3-1 Create a subset of LD-pruned SNPs -> we will use plink

Useful to have a genetic structure less biased by LD
Will be use to correct for population structure in Outflank, Baypass, etc

3-2 Outlier detection
-> with OutFlank

Based on Fst outliers across all pairs of populations



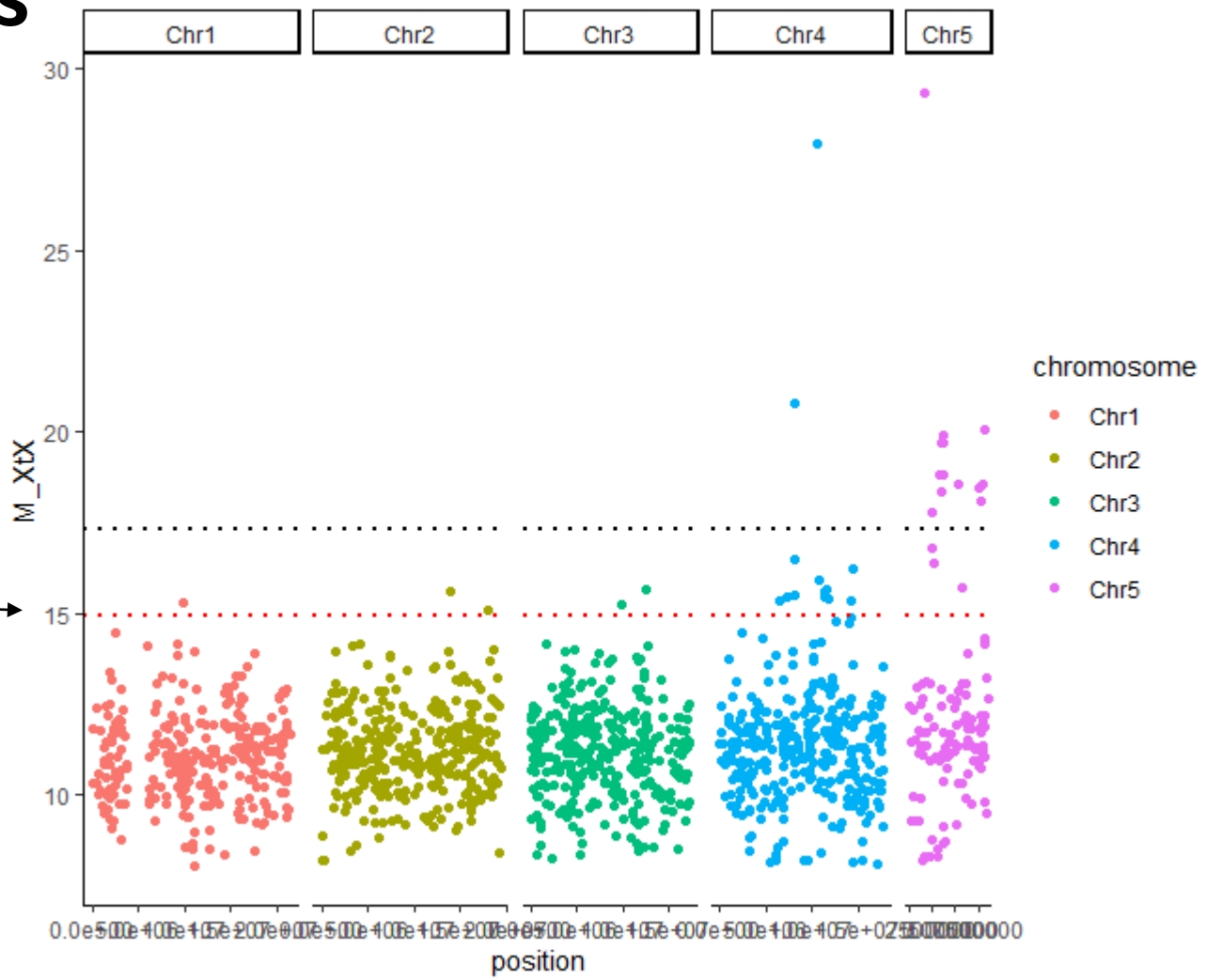


3-2 Outlier detection
-> with Baypass

Get a covariance matrix on Ld-pruned SNPs
Use it to correct the run on all SNPs

⇒ XtX is a measure of differentiation

Run Baypass on simulated SNPs to get thresholds of significance



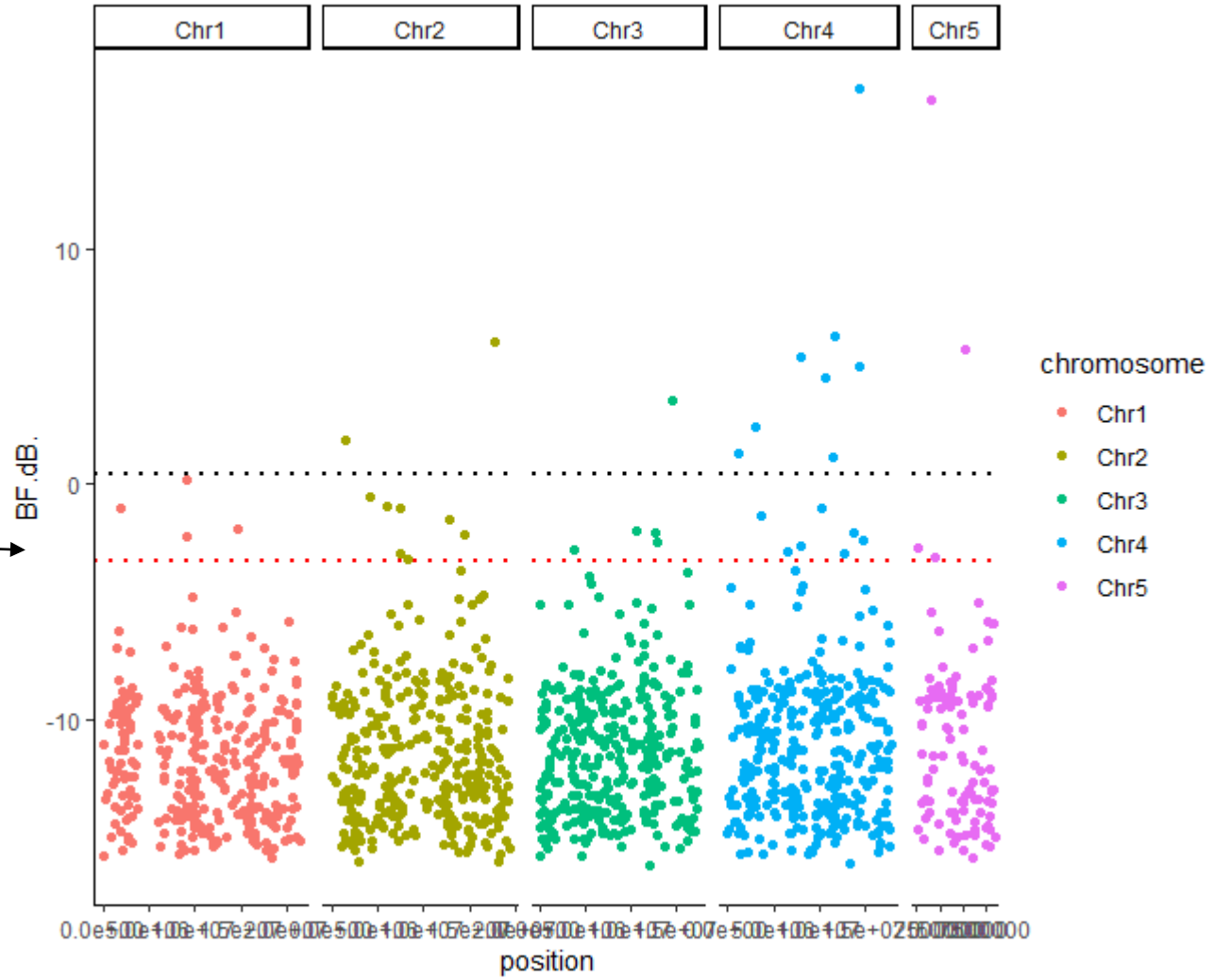
3-3 Environmental associations -> with Baypass

Get a covariance matrix on Ld-pruned SNPs
Use it to correct the run on all SNPs

⇒ XtX is a measure of differentiation

Run Baypass on simulated SNPs to get thresholds of significance

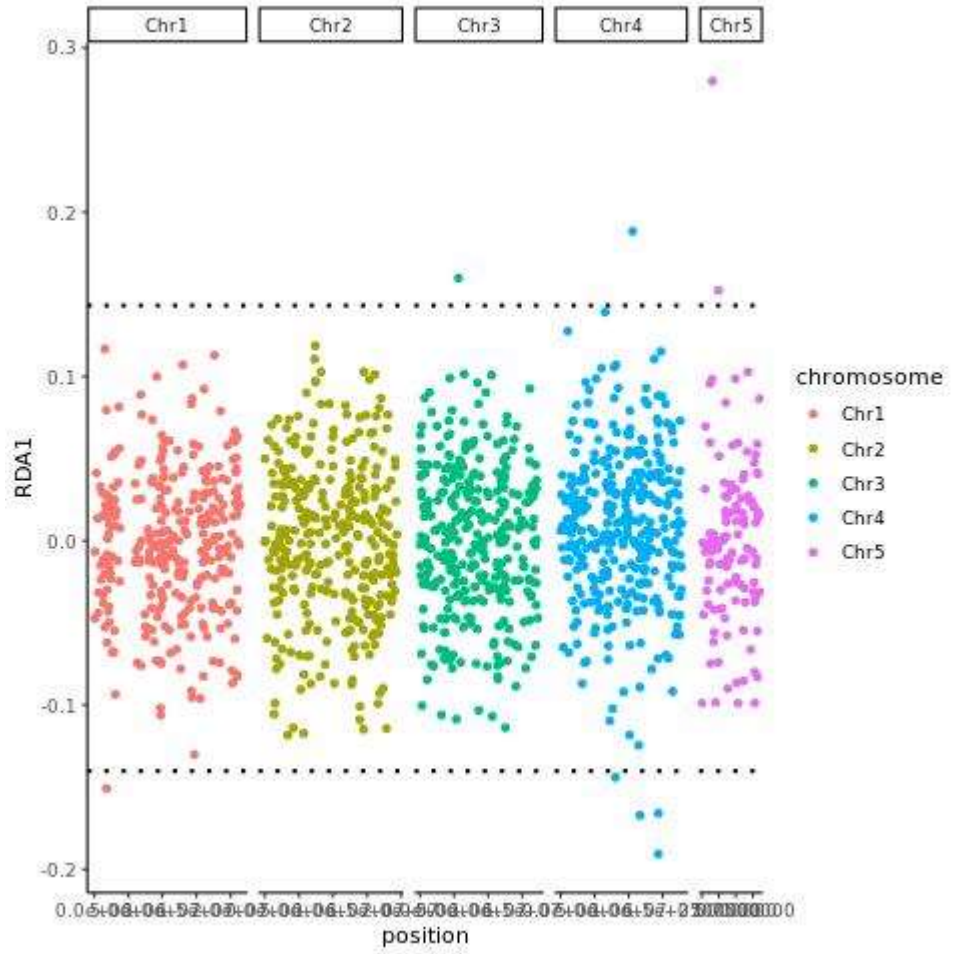
Simply add a co-variable matrix describing environmental variations between pop



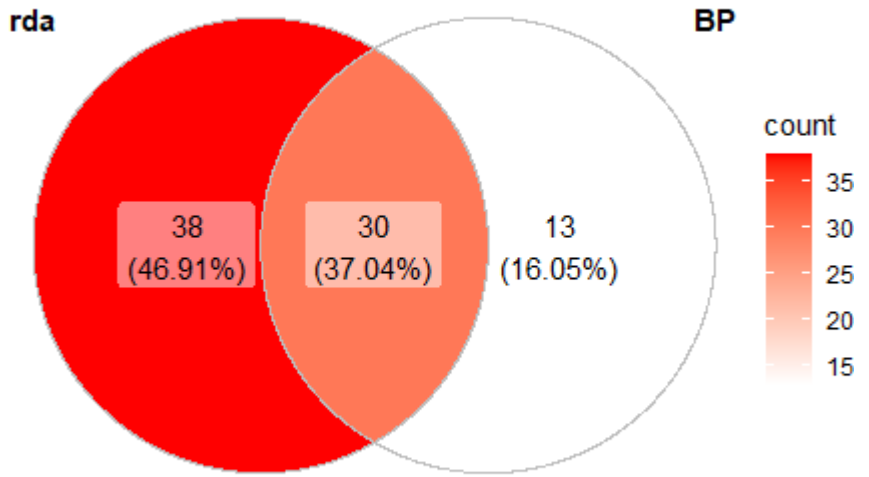
3-3 Environmental associations -> with RDA

Polygenic multivariate model

-> Can be much more complexified (test several variables, control for geography, etc)
See the optional tutorial



3-3 Environmental associations
-> Overlap



More details

Baypass
about making independant runs

What we did

- Run baypass once
- Use 1 CPU!
- Take the value of xtx (or BF) from this run
- Keep as outliers SNPs with xtx (or BF) above the 99% of Xtx from simulated values
- Look at outliers SNPs that were shared with RDA *(but remember that RDA and Baypass works differently)*

Recommended Practices for your dataset

- Run baypass 3 to 5 times with a different seed
- Use 5 to 10 CPU (nthreads) if available
- Take median value of xtx (or BF) for each SNP
- Keep as outliers SNPs with xtx (or BF) above the 99,99…% of Xtx (or BF) from simulated values – Avoid considering BF below 3 (look at Jeffrey's rule)
- Look at outliers SNPs that were shared with any other method of genotype-environment association

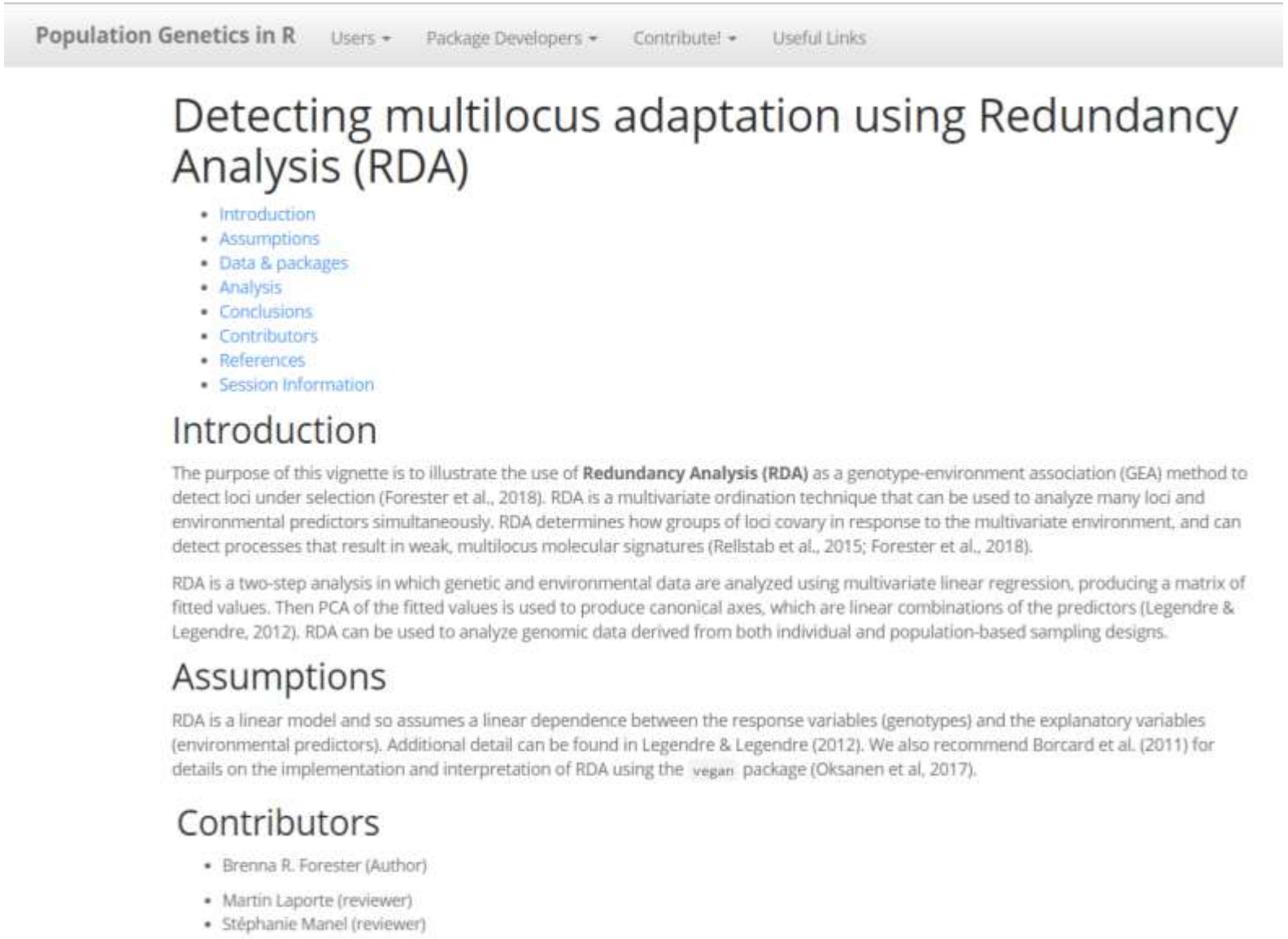
RDA

how to interpret the triplot?
advanced options for geographic variables?

(prepared with the help of
Dr. Martin Laporte)

A super good vignette to understand and do Rda analysis:

https://popgen.nescent.org/2018-03-27_RDA_GEA.html



Population Genetics in R Users Package Developers Contribute! Useful Links

Detecting multilocus adaptation using Redundancy Analysis (RDA)

Introduction

The purpose of this vignette is to illustrate the use of **Redundancy Analysis (RDA)** as a genotype-environment association (GEA) method to detect loci under selection (Forester et al., 2018). RDA is a multivariate ordination technique that can be used to analyze many loci and environmental predictors simultaneously. RDA determines how groups of loci covary in response to the multivariate environment, and can detect processes that result in weak, multilocus molecular signatures (Rellstab et al., 2015; Forester et al., 2018).

RDA is a two-step analysis in which genetic and environmental data are analyzed using multivariate linear regression, producing a matrix of fitted values. Then PCA of the fitted values is used to produce canonical axes, which are linear combinations of the predictors (Legendre & Legendre, 2012). RDA can be used to analyze genomic data derived from both individual and population-based sampling designs.

Assumptions

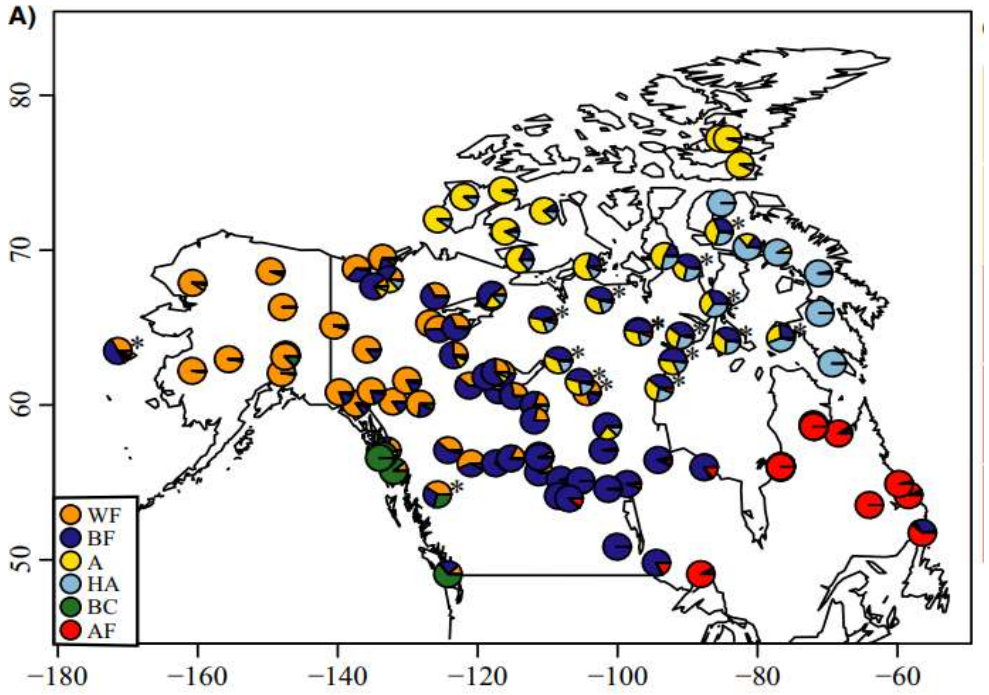
RDA is a linear model and so assumes a linear dependence between the response variables (genotypes) and the explanatory variables (environmental predictors). Additional detail can be found in Legendre & Legendre (2012). We also recommend Borcard et al. (2011) for details on the implementation and interpretation of RDA using the `vegan` package (Oksanen et al, 2017).

Contributors

- Brenna R. Forester (Author)
- Martin Laporte (reviewer)
- Stéphanie Manel (reviewer)

RDA

Multivariate associations:

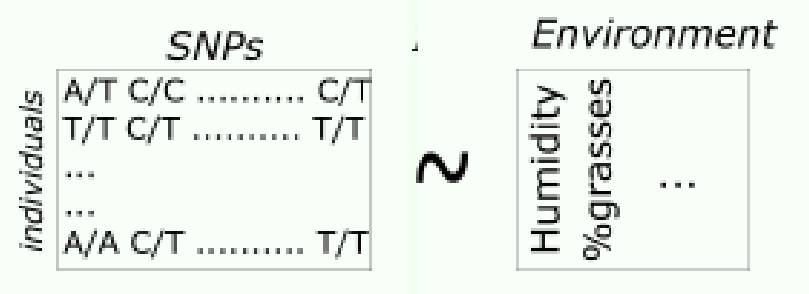


94 wolves
42 597 SNPs

species

sites

In community ecology (package vegan!)

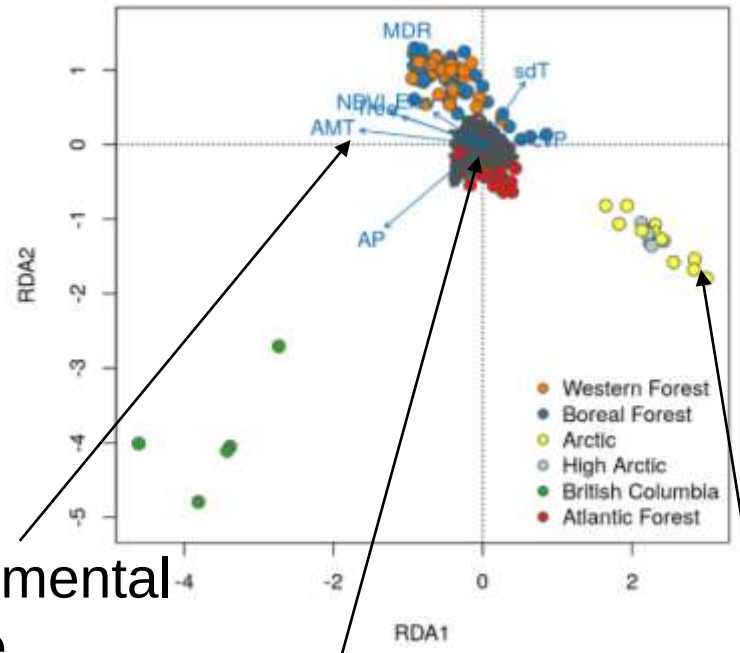


Forester et al 2018 Mol Ecol

RDA

```
points(X.rda, display="sites")
```

Triplot RDA (individual centered)



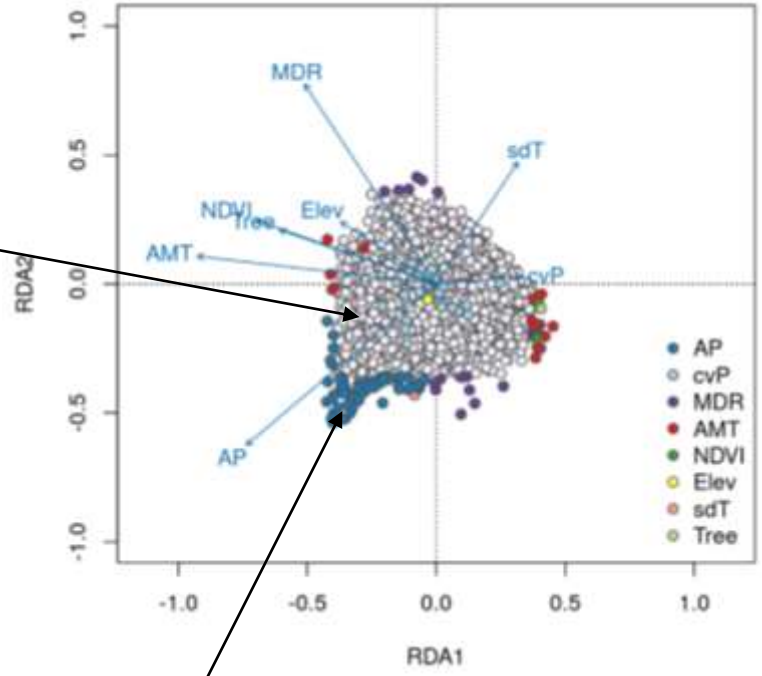
Environmental variable

Genetic markers

Individual

```
points(X.rda, display="species")
```

Triplot RDA (SNPs centered)



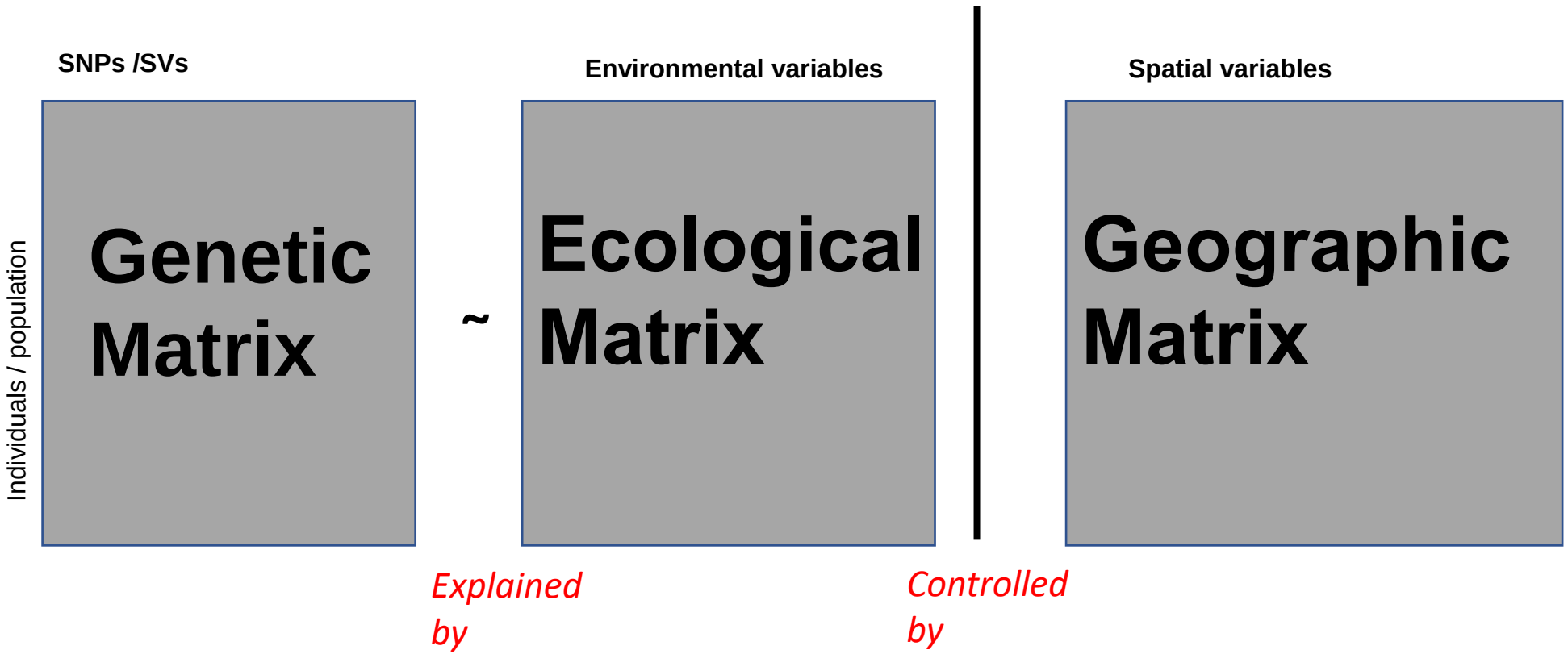
Neutral marker

Outlier marker putatively associated to *AP* variation

Forester et al 2018 Mol Ecol

Use the contribution of genetic markers along the different axis to detect putatively-selected loci

RDA



https://doi.org/10.1016/B978-0-444-53868-0.50014-9

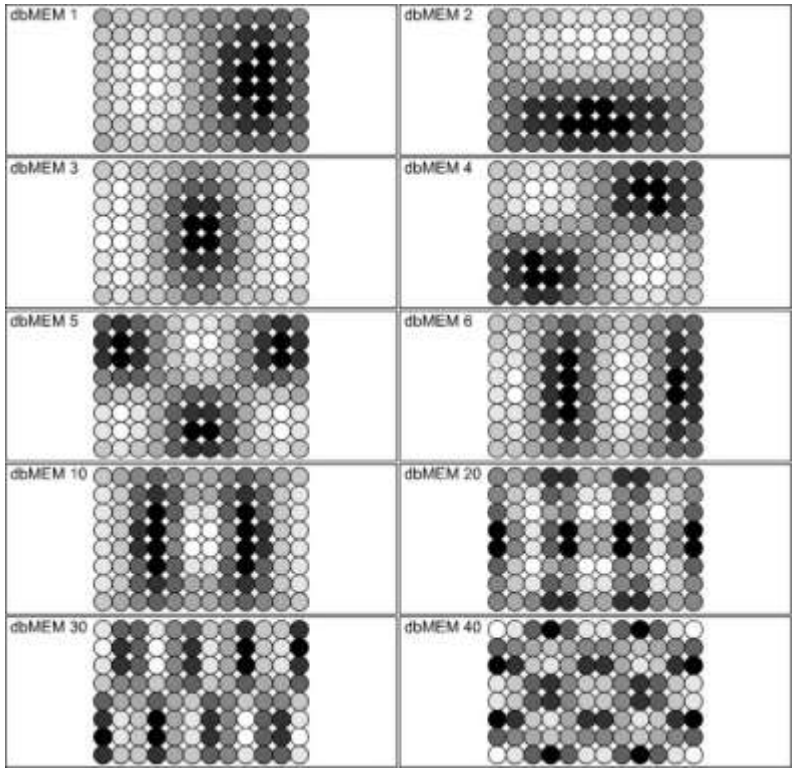
RDA

Spatial-eigen vectors are a way to reduce a distance matrix between samples/populations
-> not necessarily neutral
-> describe different possible spatial combination

G ~ E | S

Latitude + Longitude
or
Spatial eigenvectors

= db-MEM



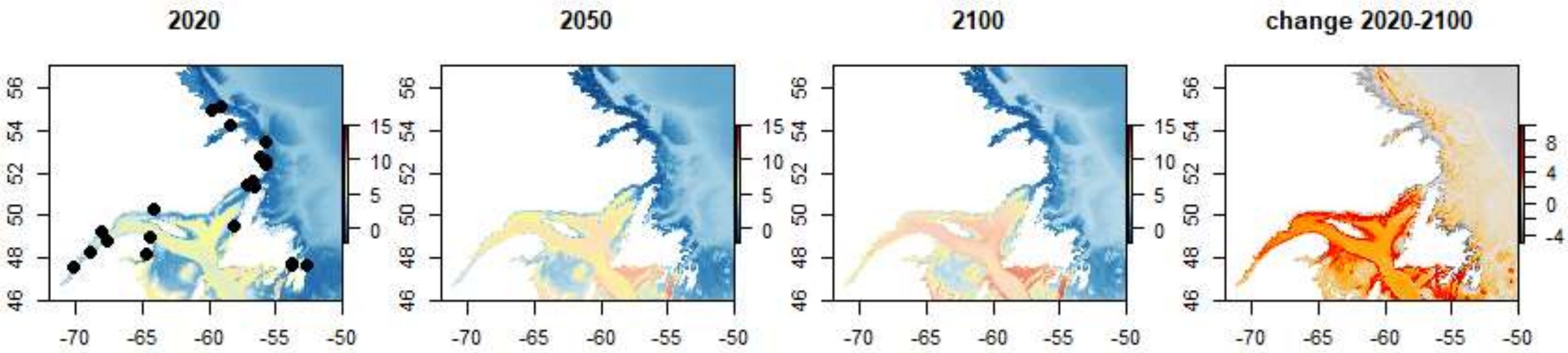
More information:
Legendre & Legendre
https://doi.org/10.1016/B978-0-444-53868-0.50014-9

Climatic Variables
how to extract them from databases?

https://www.worldclim.org/

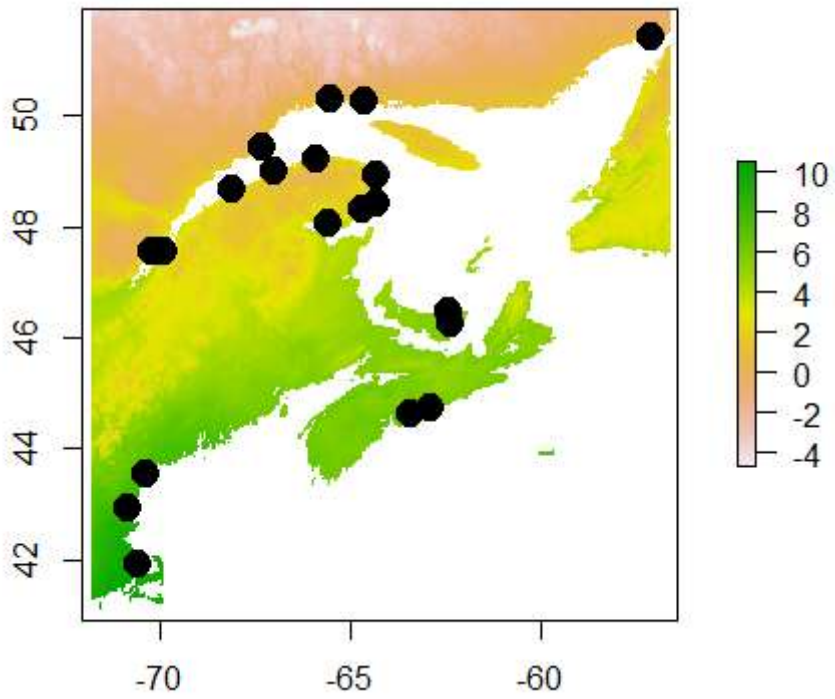
http://www.marspec.org/
(with useful tutorials)

https://www.bio-oracle.org/
(with prediction under GIEC scenarios)



WORLDCLIM: R will gather the data itself

```
location_GPS<- read.delim("location_GPS.txt")
r <- getData("worldclim",var="bio",res=2.5)
div=10 #precision of the data

#1 is mean temp, 12 is annual precipitations, et...
Annual_mean_temp<-r[[1]]
variable<-paste0("bio1")

#make a plot of the area
aoi_area <- extent(min (location_GPS$GPS_EW)-1,max (location_GPS$GPS_EW)+0.5,min (location_GPS$GPS_NS)-1,max (loca
plot((crop(Annual_mean_temp, aoi_area)/div))
points(location_GPS$GPS_EW,location_GPS$GPS_NS, pch=19, col=1, cex=2)

# to get data round a point of your choice like pop 1
i=1
#determine the coordinates around your point
long_min<-floor(location_GPS$GPS_EW[i]*10)/10
long_max<-ceiling(location_GPS$GPS_EW[i]*10)/10
lat_min<-floor(location_GPS$GPS_NS[i]*10)/10
lat_max<-ceiling(location_GPS$GPS_NS[i]*10)/10

#prepare the area
aoi <- extent(long_min, long_max, lat_min, lat_max)

#get the value of the layer in the area
Annual_mean_temp.crop <- crop(Annual_mean_temp,aoi)
mean_value_i<-mean(Annual_mean_temp.crop@data@values, na.rm=T)/div
range_value_i<-(range(Annual_mean_temp.crop@data@values, na.rm=T)[2]-range(Annual_mean_temp.crop@data@values, na.rm=T)[1])/div

#print value
location_GPS[i,]
mean_value_i
range_value_i
```



MARSPEC Download data

I'll drop a tutorial on the github page of the course within day3