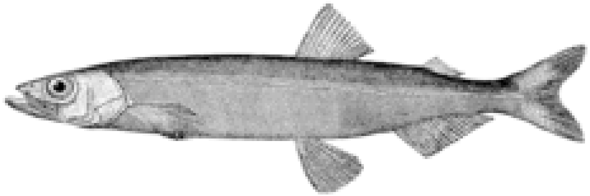
Tutorials overviews and wrap-up

(including additionnal details)

Capelin dataset



DOI: 10.1111/mec.15499

ORIGINAL ARTICLE

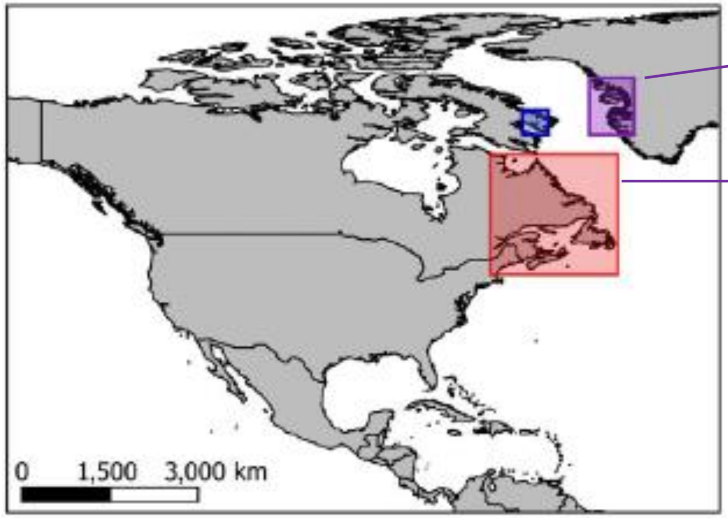
MOLECULAR ECOLOGY WILEY

Shared ancestral polymorphisms and chromosomal rearrangements as potential drivers of local adaptation in a marine fish

Hugo Cayuela[1*] | Quentin Rougemont[1*] | Martin Laporte[1] | Claire Mérot[1] | Eric Normandeau[1] | Yann Dorant[1] | Ole K. Tørresen[2] | Siv Nam Khang Hoff[2] | Sissel Jentoft[2] | Pascal Sirois[3] | Martin Castonguay[4] | Teunis Jansen[5,6] | Kim Praebel[7] | Marie Clément[8,9] | Louis Bernatchez[1]

- *Mallotus villotus*
- Small fish
- Spawn on beaches
- Cold waters of the North Atlantic Ocean

Capelin dataset

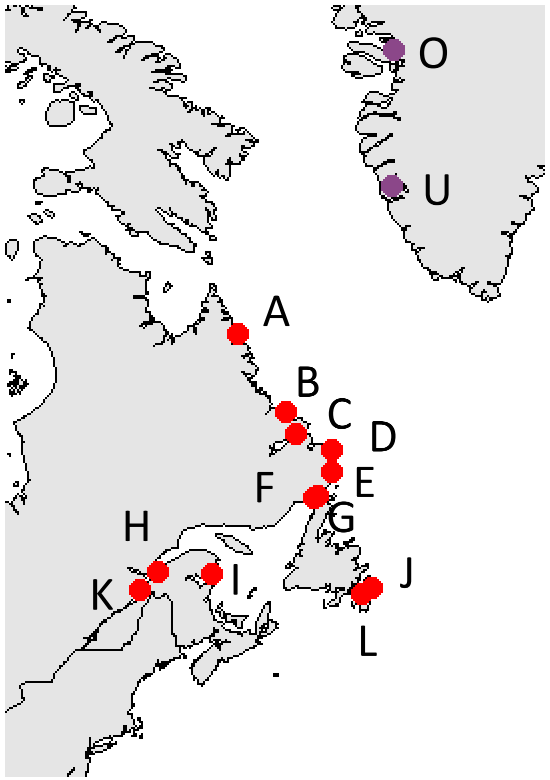


Greenland species

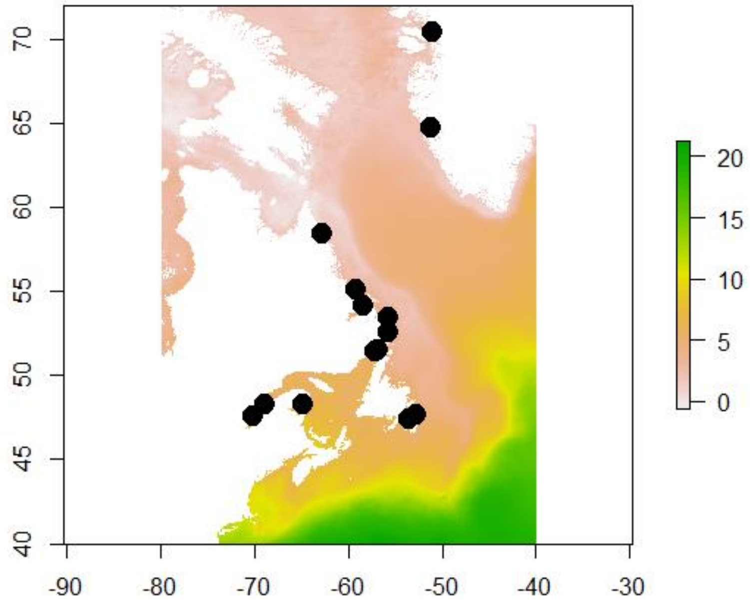
2 population
N= 40

North American species

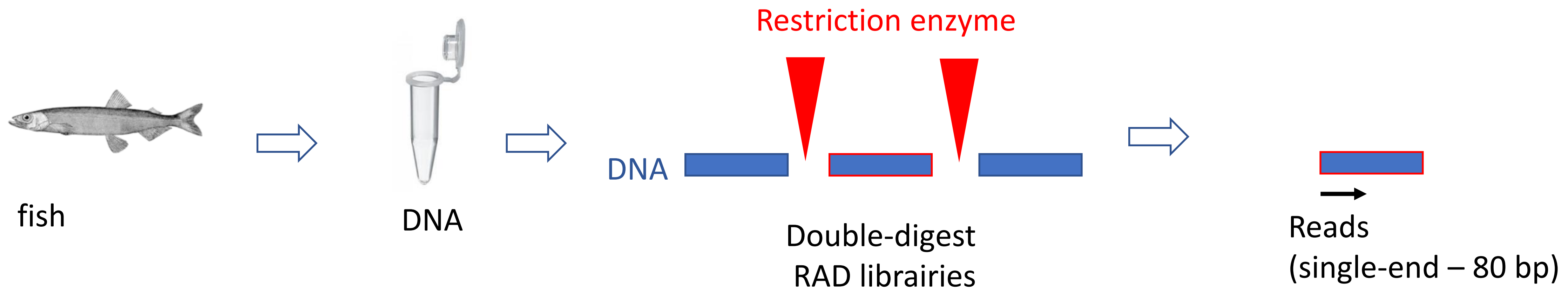
12 populations
N= 240 (20/pop)

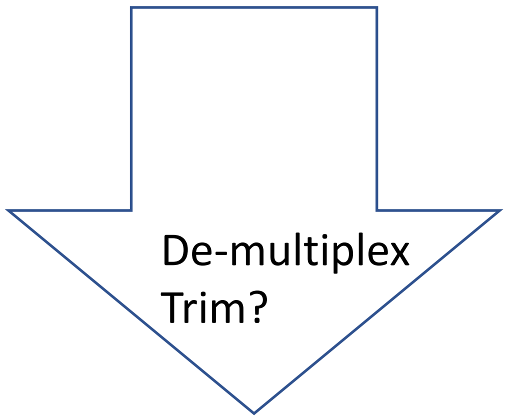


Sea temperature
(from MARSPEC)



Capelin dataset





N fastQ files

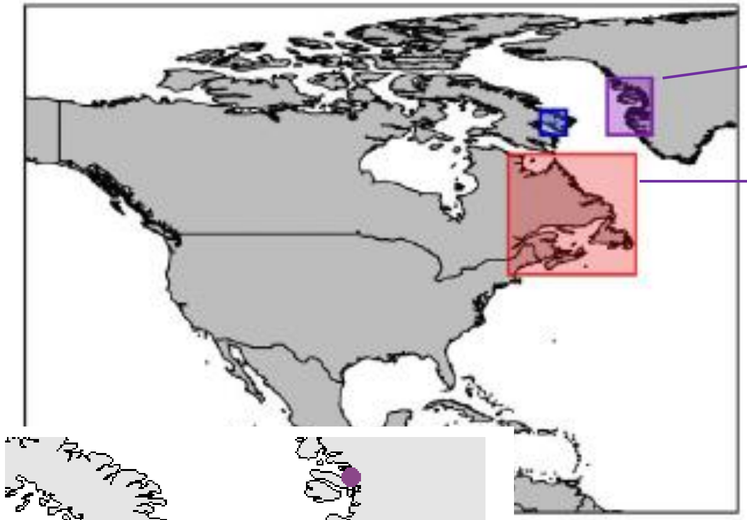
```
@70ZFD:01332:11598/1
TGCATCAACTTTAAGATACGCTATTGGAGCTGGAATTACCGCGGCTGCTGGCACCCAGACTTGCCCTCCAATGGATCCTC
+
7<<=<;<4676*115345::=;<=6;5<;<;;7<1918<199<6<<::9:5:556+38469166=3;<6<655-477-4/
@70ZFD:01334:11636/1
TGCATCCTGTGGAAGTAGCTGCACACCTGCTCATGCTGTGCCAGGAAGGGAGGGTGGGATCAGCCAATCGGGGAACAGAG
+
5;?;;;5855;4:4<A<;<<;;<<<B9B=<<<><;;;=<<:69:58-55)533)/893<:;:9:496888<:1;599;;B
@70ZFD:01335:11615/1
TGCATGGCAGAGTGGAGAGGAGCGCCCTCTACTGGAACTTCCTGGAACAGGTCCTCCGAATGTCCAAGGTACAACGGTTC
```

Dummy genome

Smaller genome :
5 chromosomes

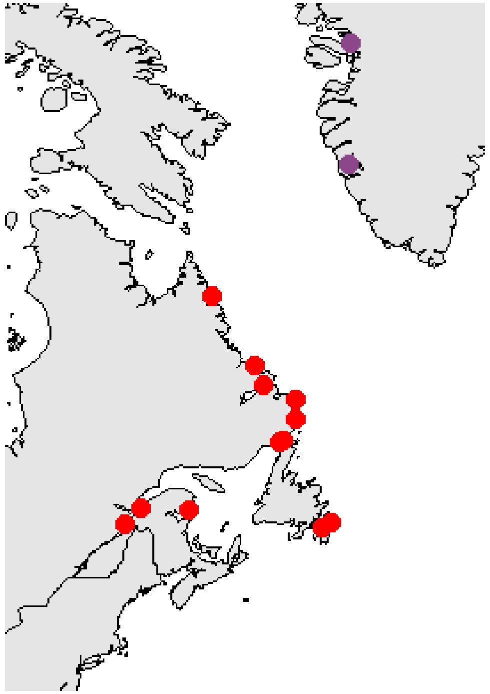
We aligned fastq files = the raw reads on that dummy reference genome

$\Rightarrow$ BAM files that you will play with in STACKS.



2 population
N= 40

12 populations
N= 240 (20/pop)



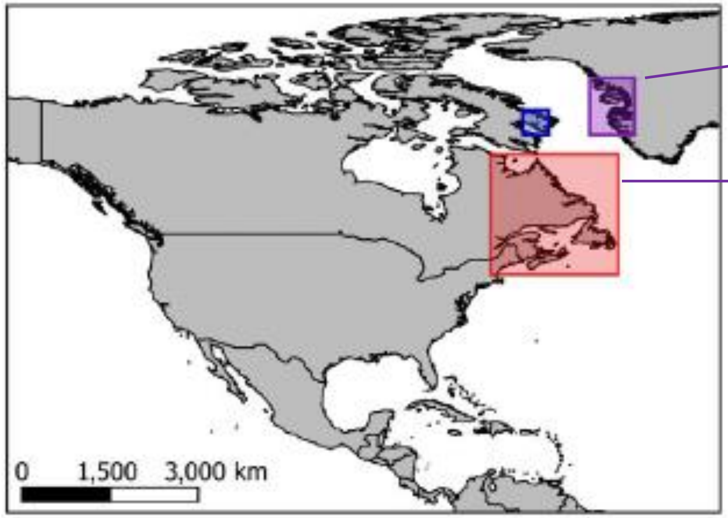
Dataset 1	Dataset 2	Dataset 3
« 2_lin »	« all »	« canada »
4 populations (2 greenland /2 canadian) => 80 samples	14 populations (2 greenland /12 canadian) => 280 samples	12 populations (12 canadian) => 240 samples
Fst (vcftools) PCA Optional (Fst with Stacks)	Faststructure DAPC	PCA DAPC Optional (Pairwise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

Day1 SNP calling with STACKS

Day2 Population structure

Fst

2 population
N= 40



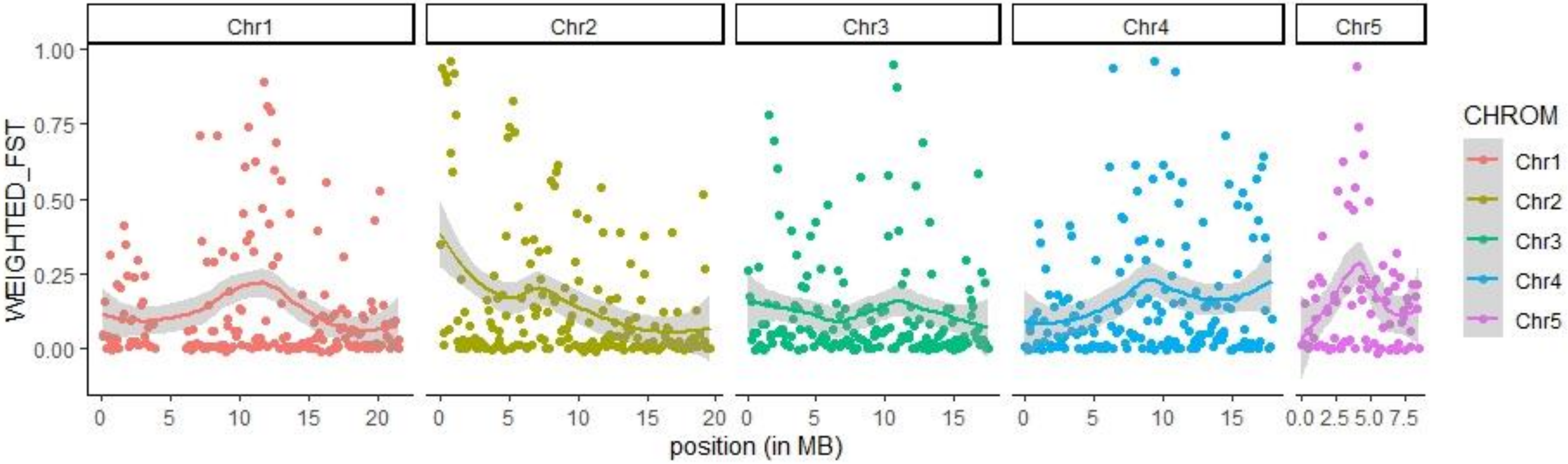
Greenland species

North American species

12 populations
N= 240 (20/pop)

Fst = 0.23

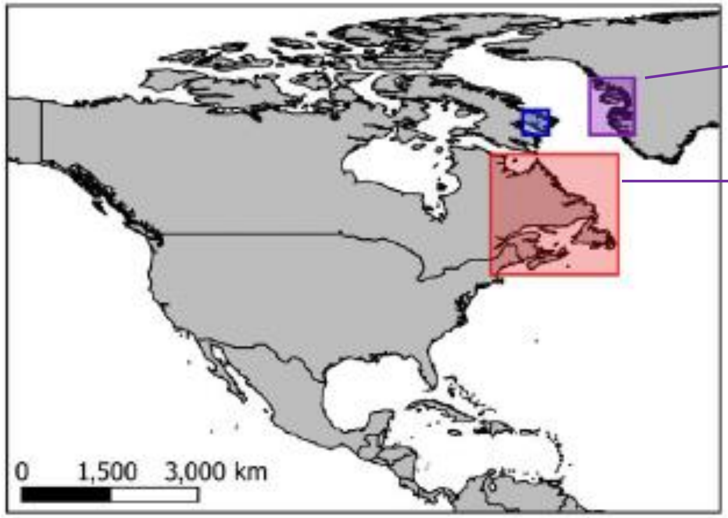
Mean Fst 0.03 / weighted Fst 0.23



PCA

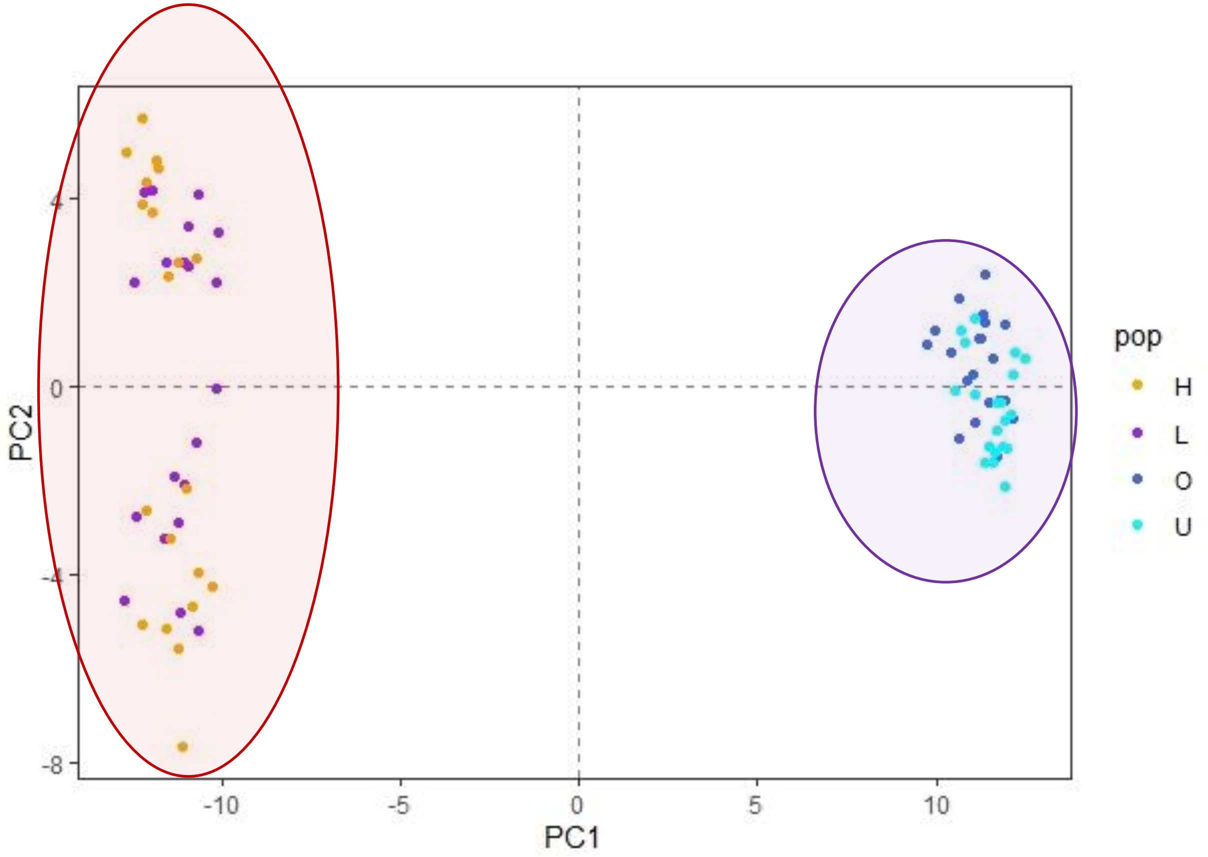
2 population
N= 40

12 populations
N= 240 (20/pop)

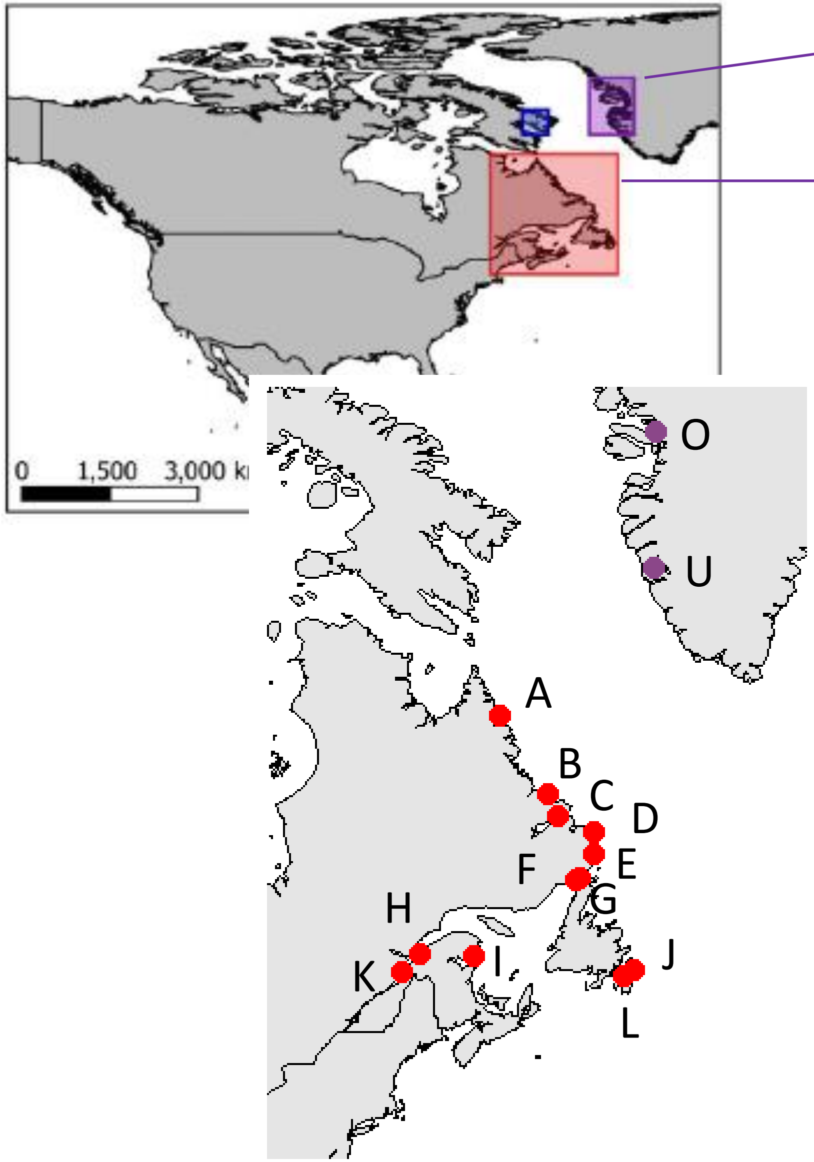


Greenland species

North American species



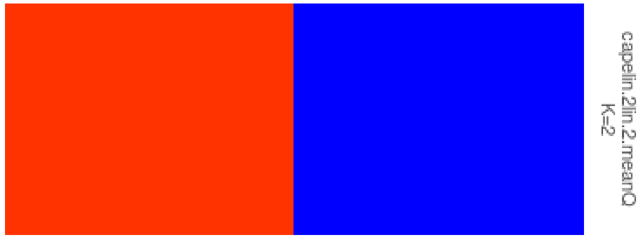
LEA – Clustering methods



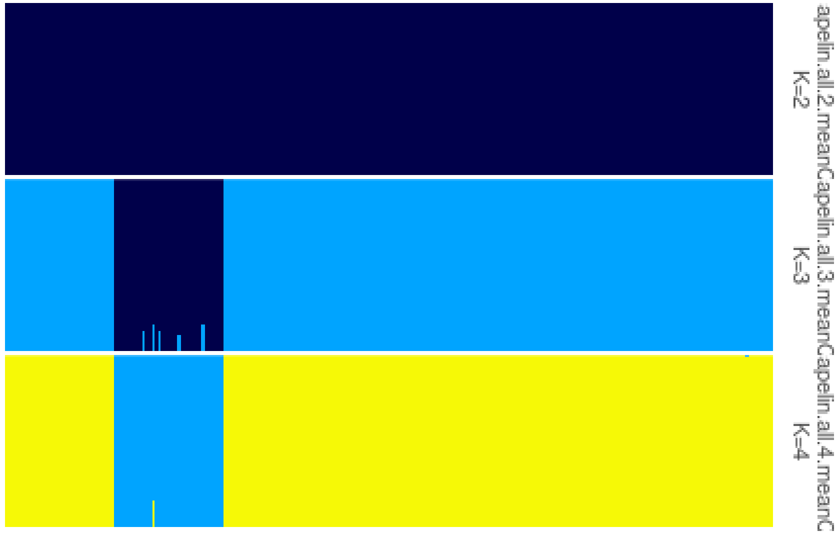
Greenland species

North American species

2 lineages



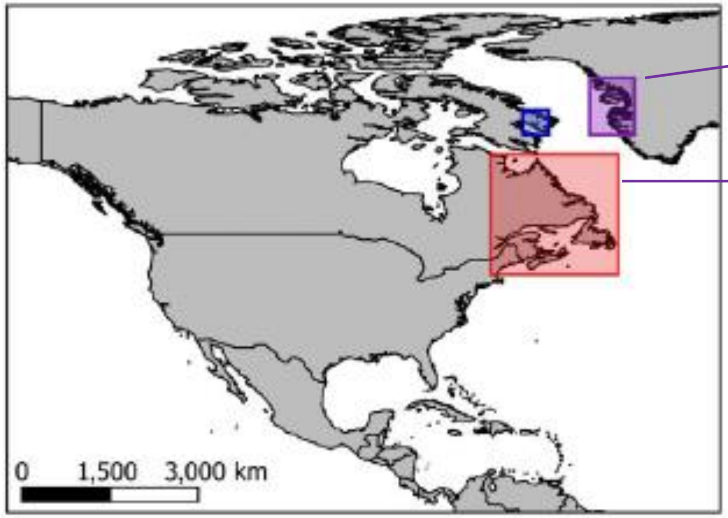
All



O – U
= Greenland

DAPC

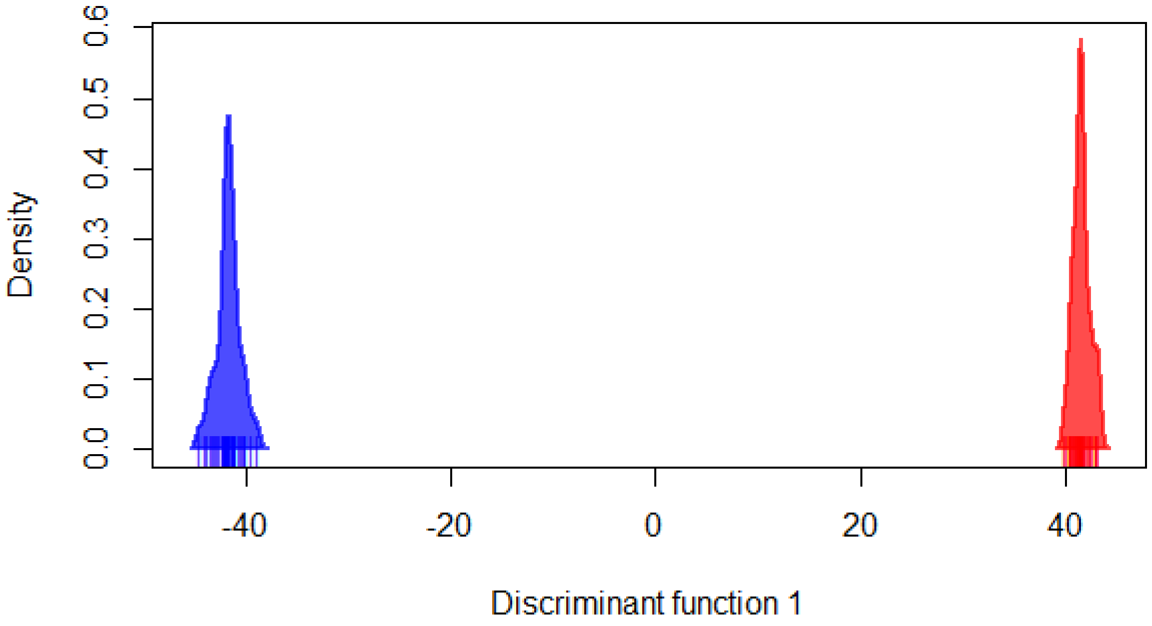
2 population
N= 40

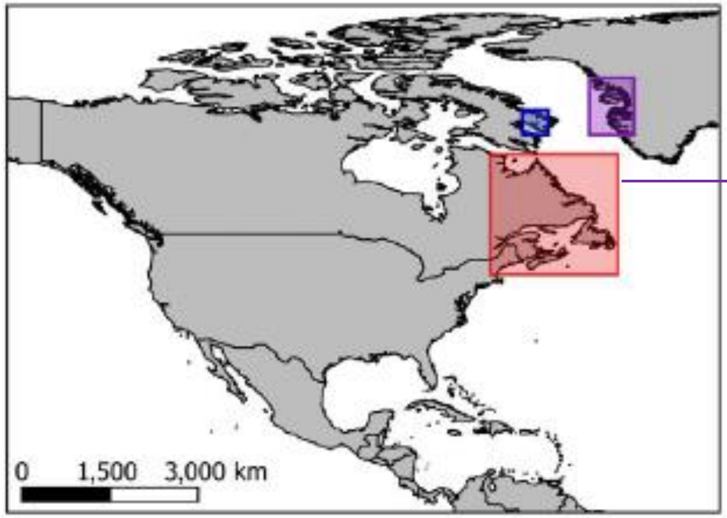


Greenland species

North American species

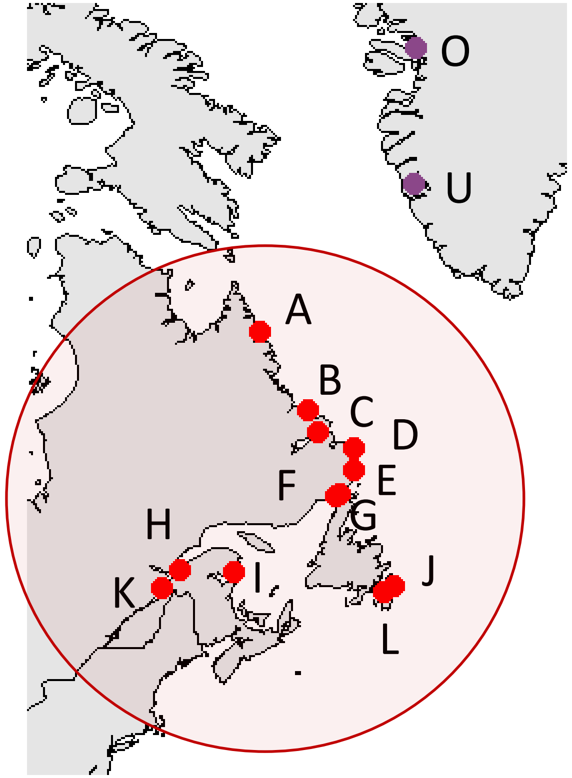
12 populations
N= 240 (20/pop)



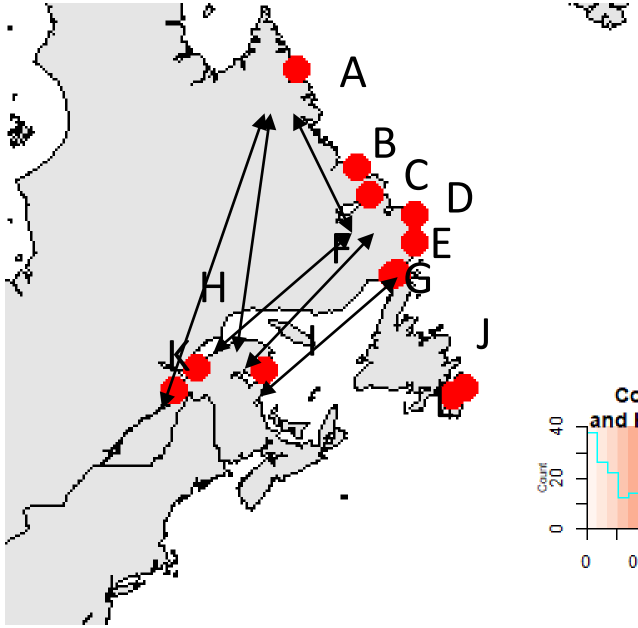


North American species

12 populations
N= 240 (20/pop)

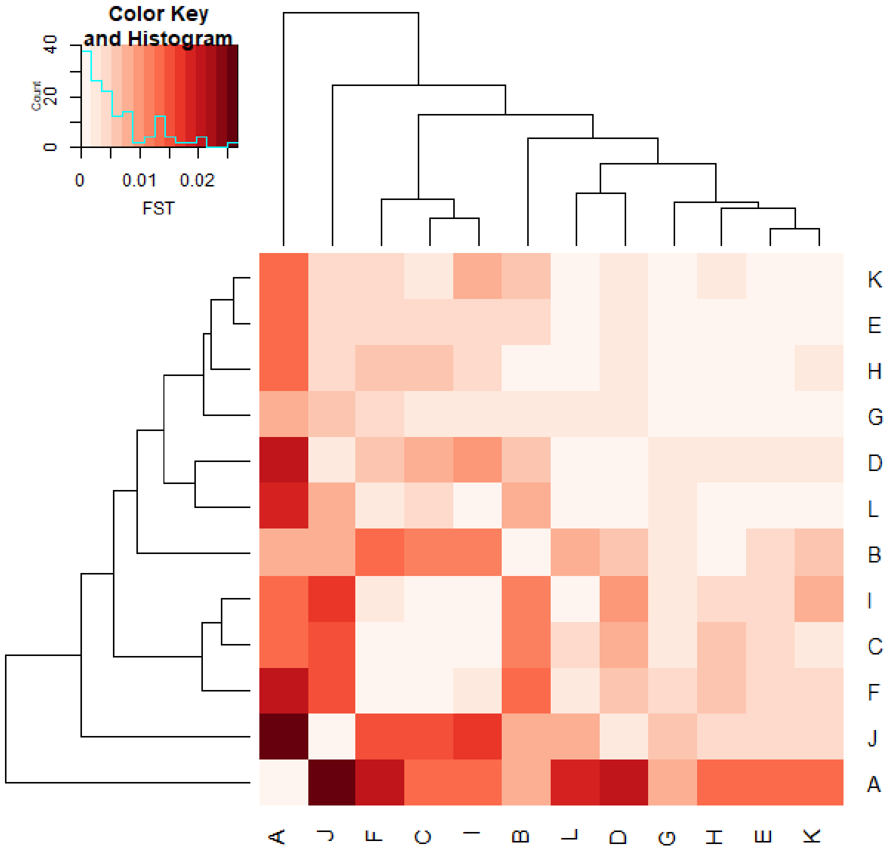


Fst



Do we observe genetic structure ?

Fst between all poulations



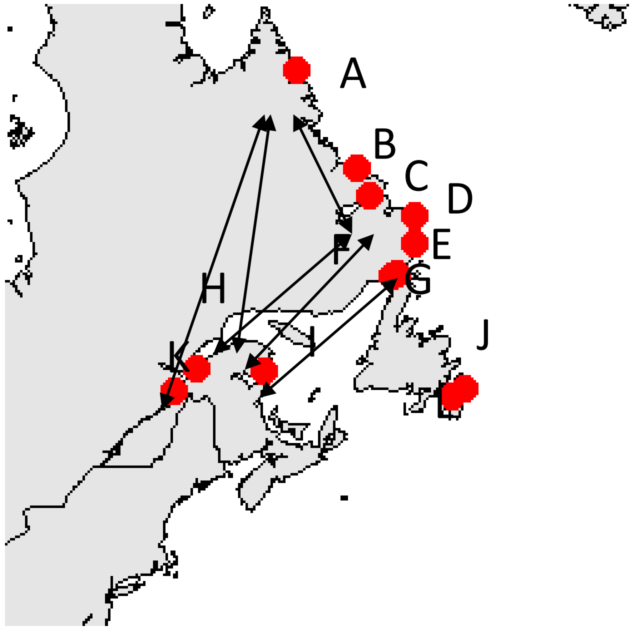
Medium values (Fst = 0.025)?
Lots of heterogeneity…

⇒ 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)!!

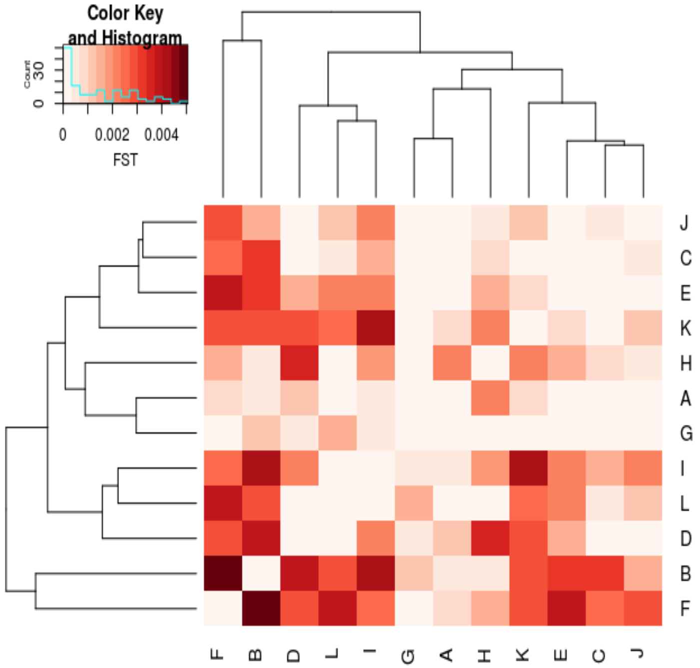
Fst



Do we observe genetic structure ?

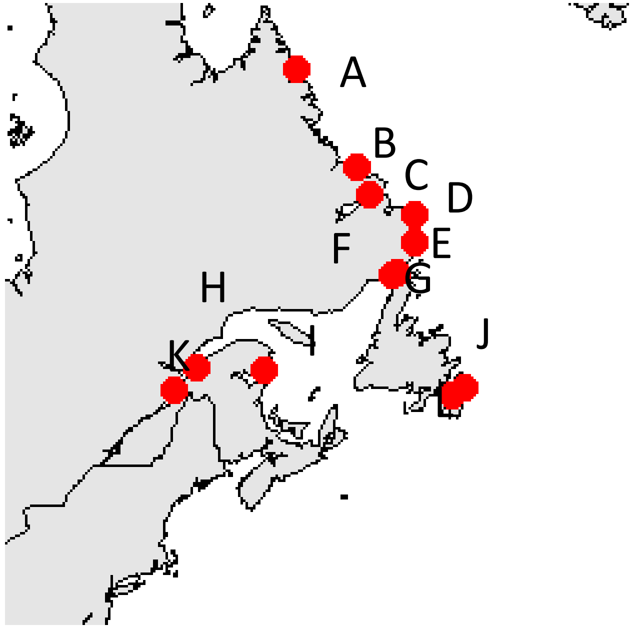
Fst between all poulations
Excluding chromosme 4 (inversion) & chromosome 5 (sex)

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

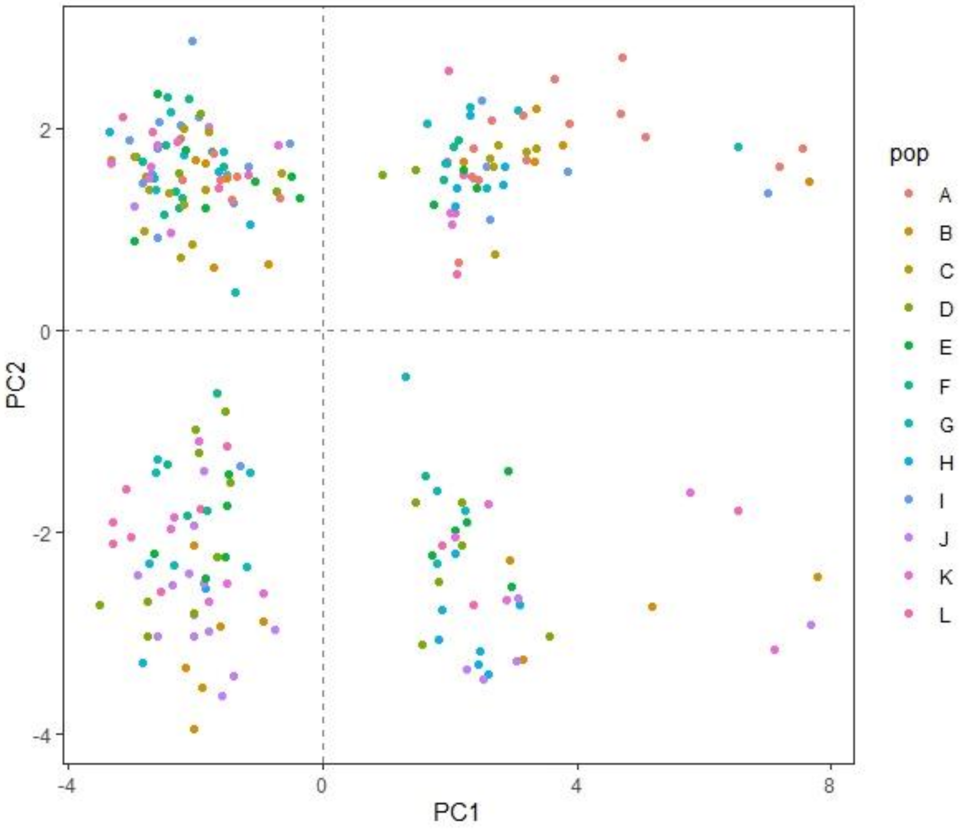


⇒ Almost no geographic structure…

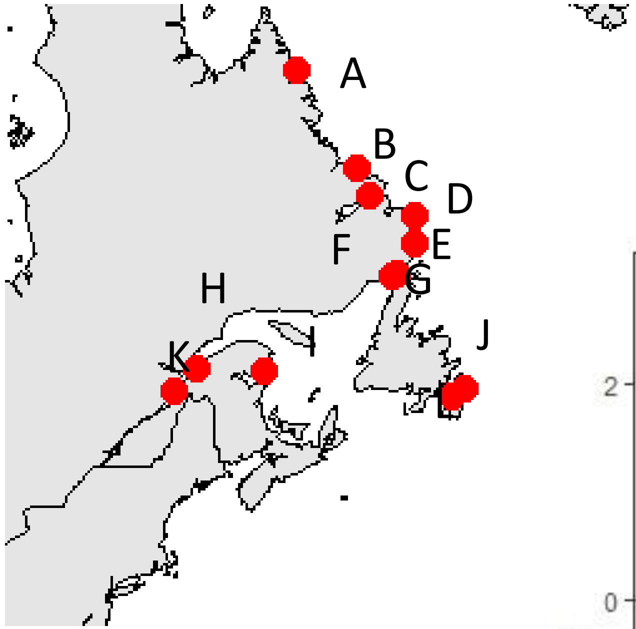
PCA



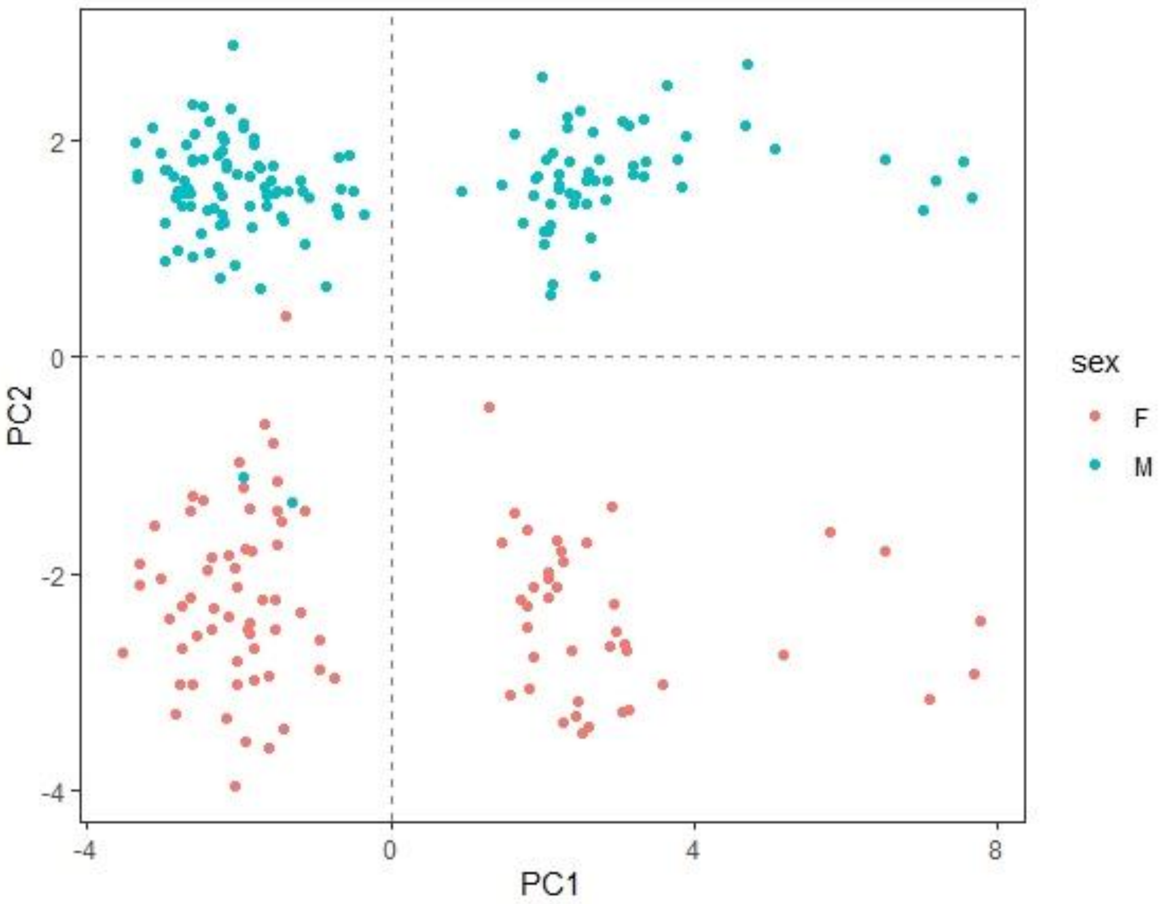
Canada (12 pop)



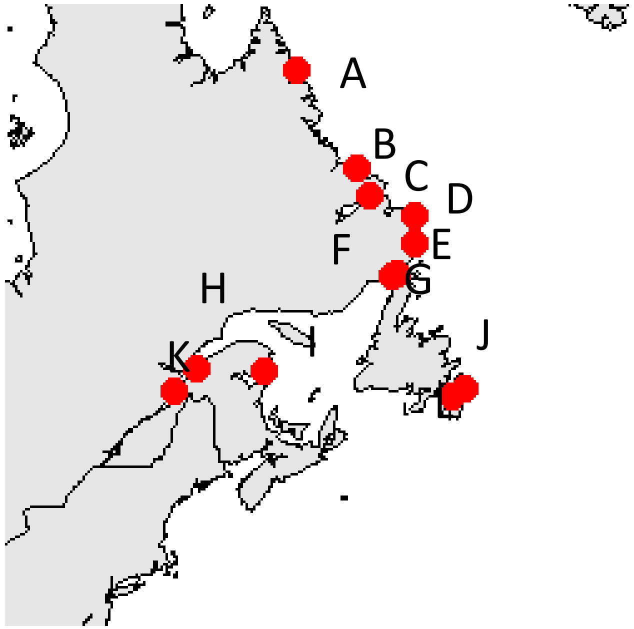
PCA



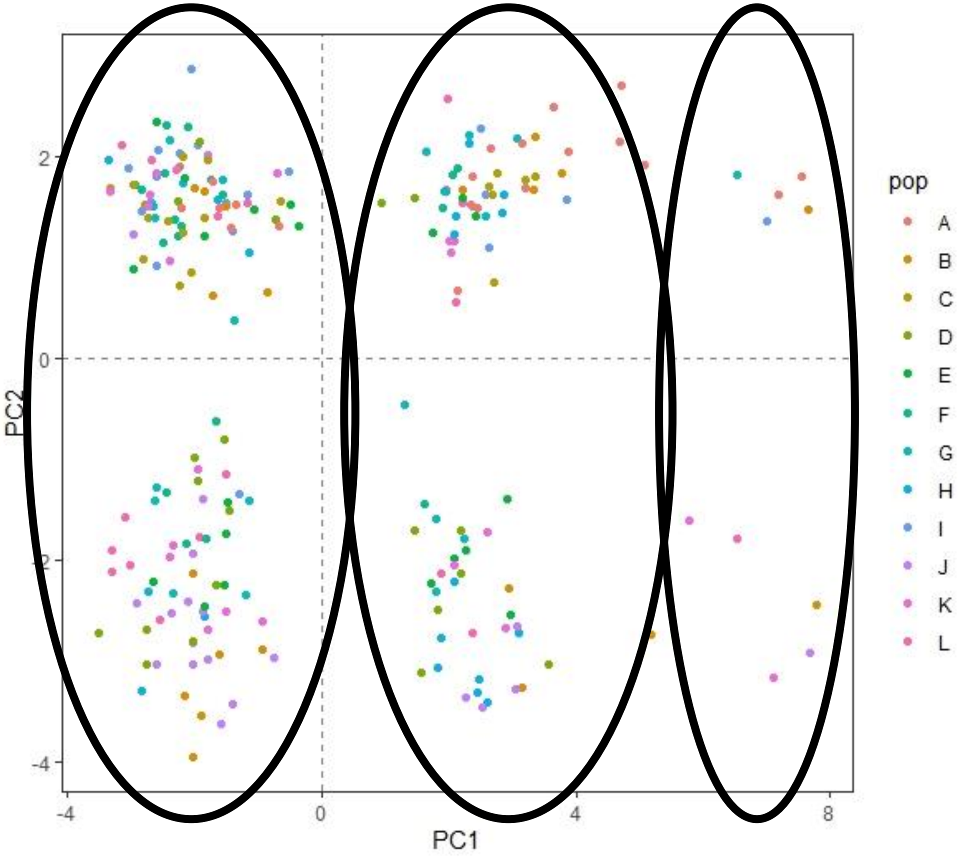
Canada (12 pop)



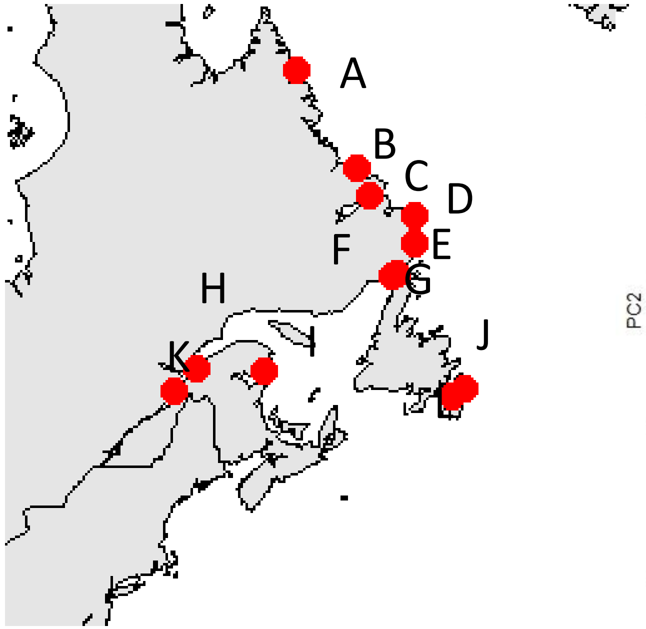
PCA

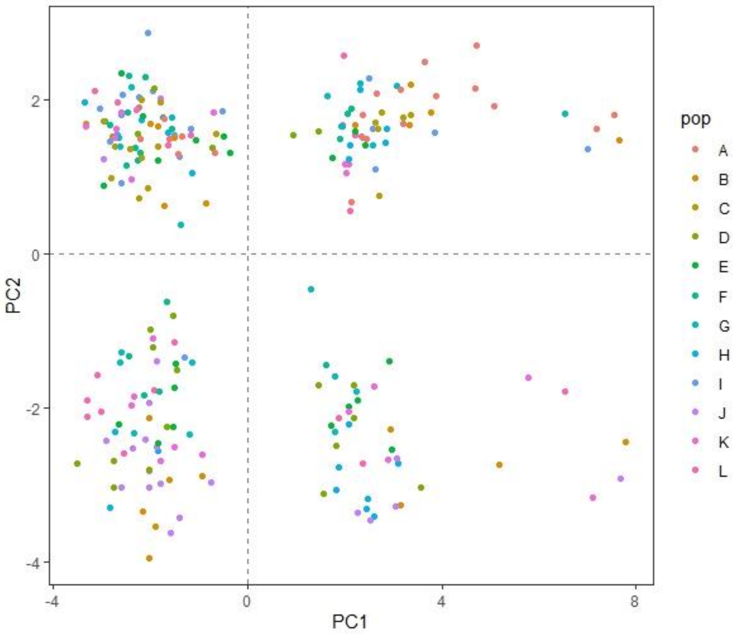


Canada (12 pop)



PCA

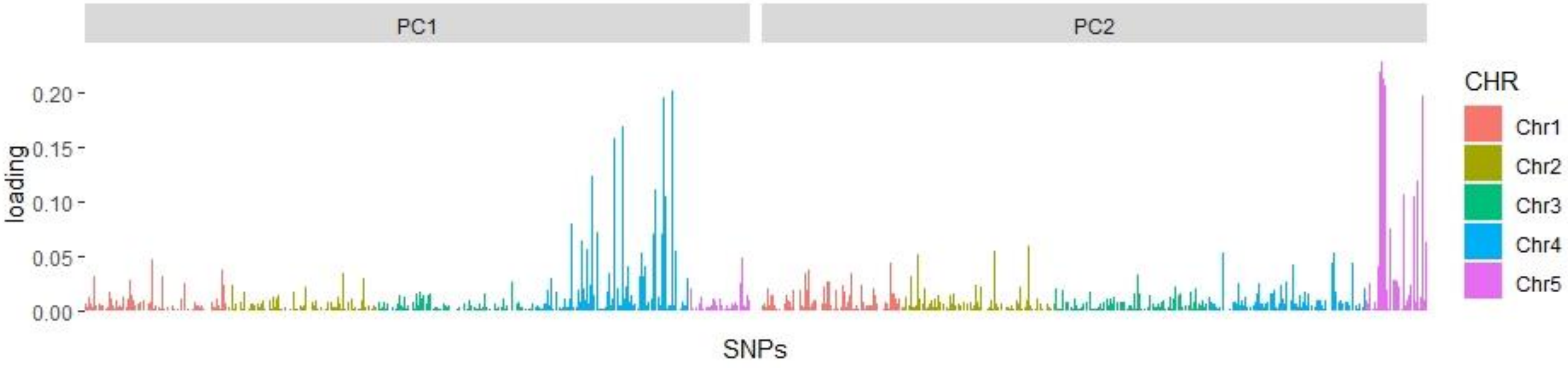




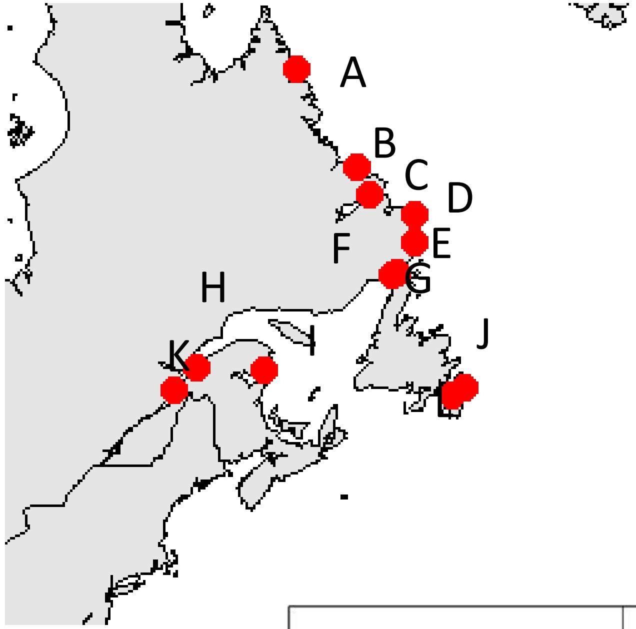
12 pop)

Rearrangement on Chr 4

Sex-determining locus on Chr 5

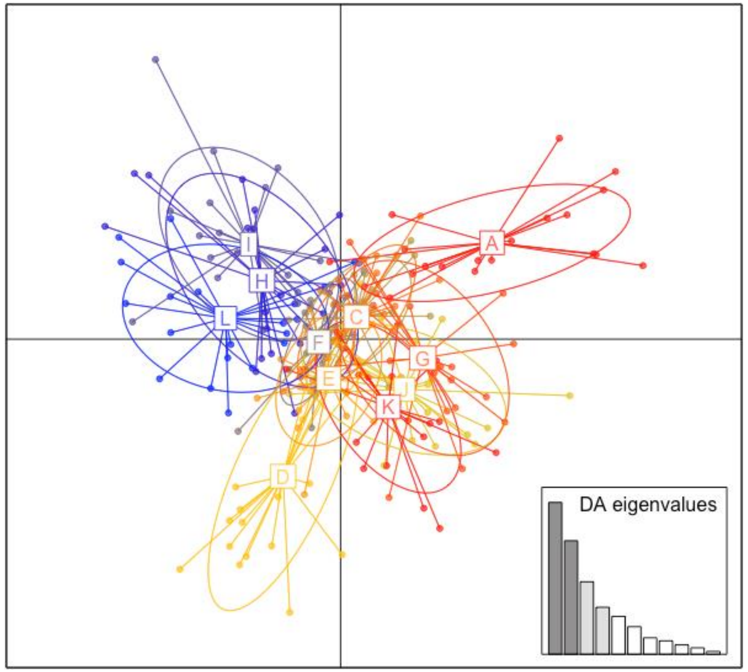


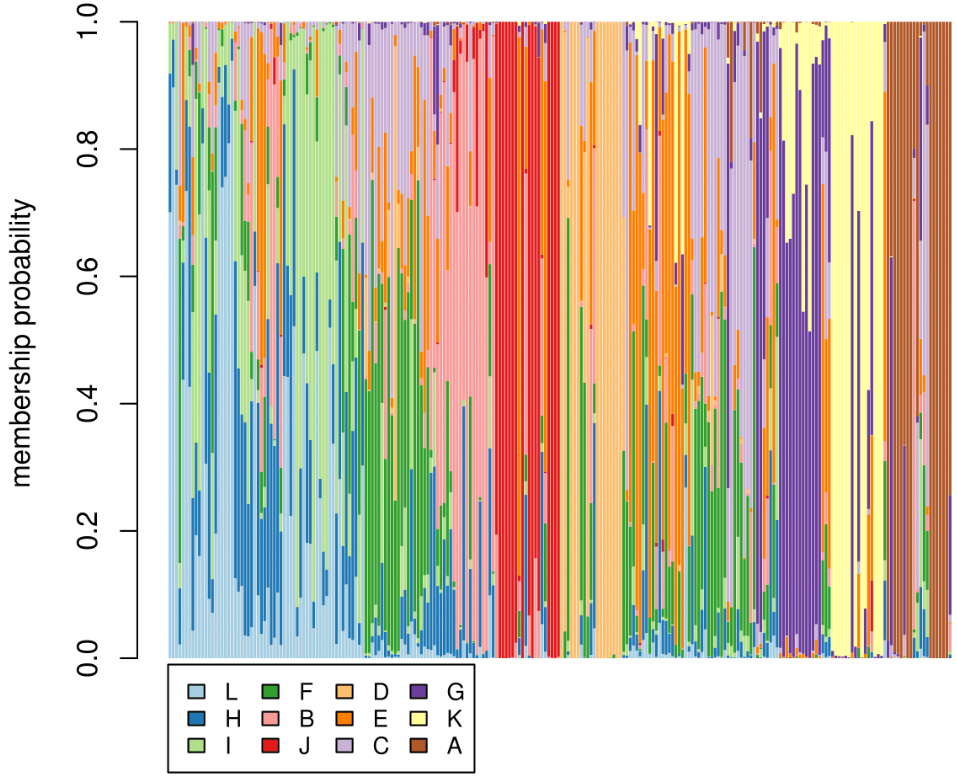
DAPC



Do we observe genetic structure ?

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





Day3 – Outlier detection and Environemental associations

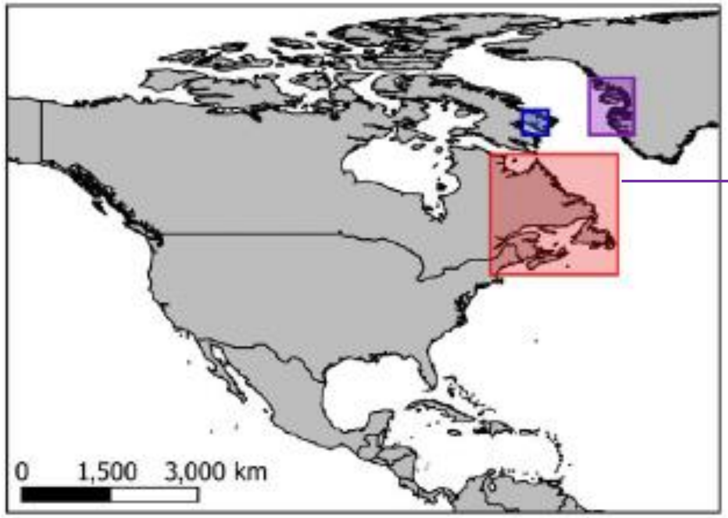
Disentangle population structure & putative signature of adaptation…

3-1 Fst statistics & geography

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

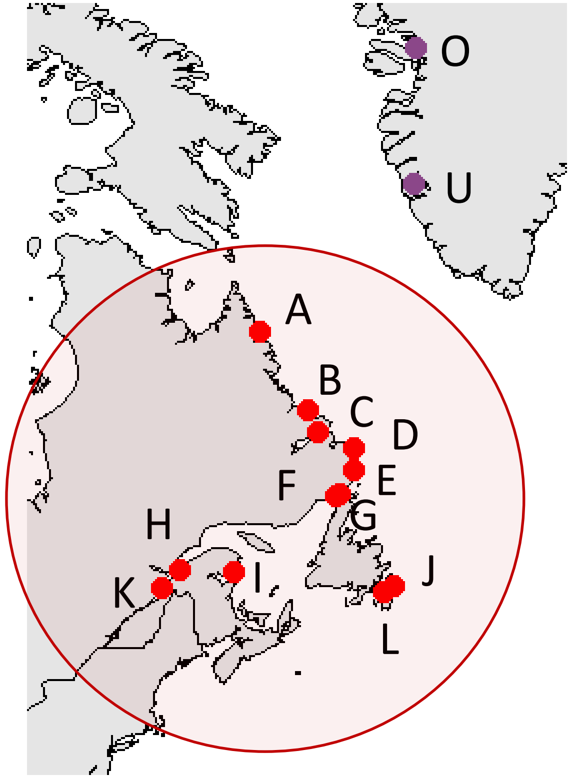
3-2 Outliers of differentiation

3-3 Genotype-environnement assocations

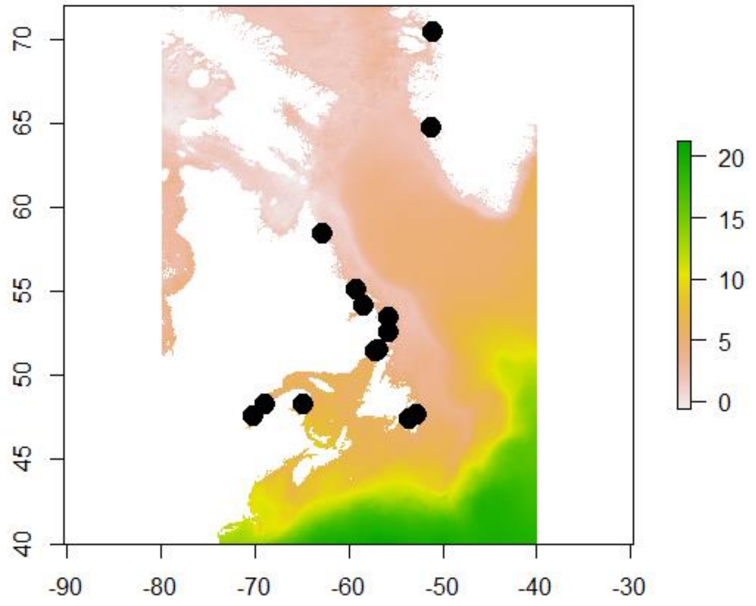


North American species

12 populations
N= 240 (20/pop)



Sea temperature
(from MARSPEC)

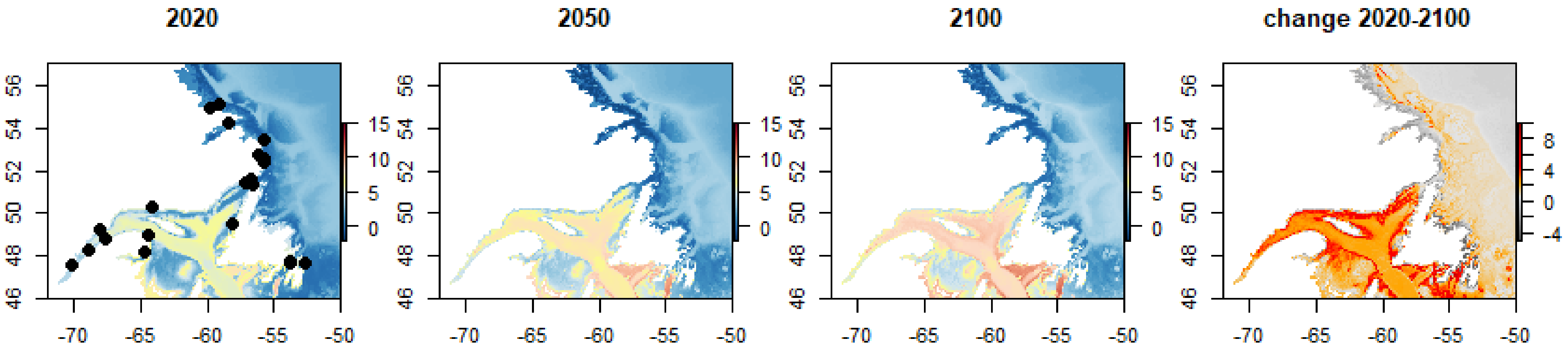


Climatic Variables
how to extract them from databases?

https://www.worldclim.org/

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

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

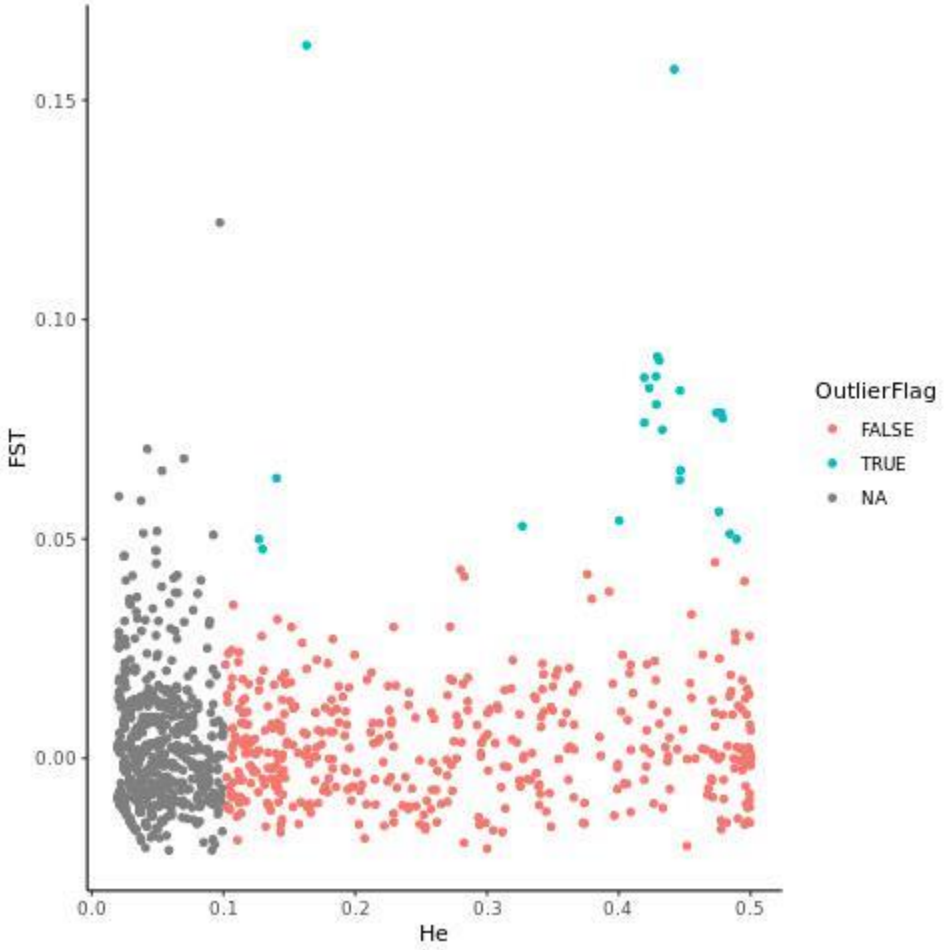


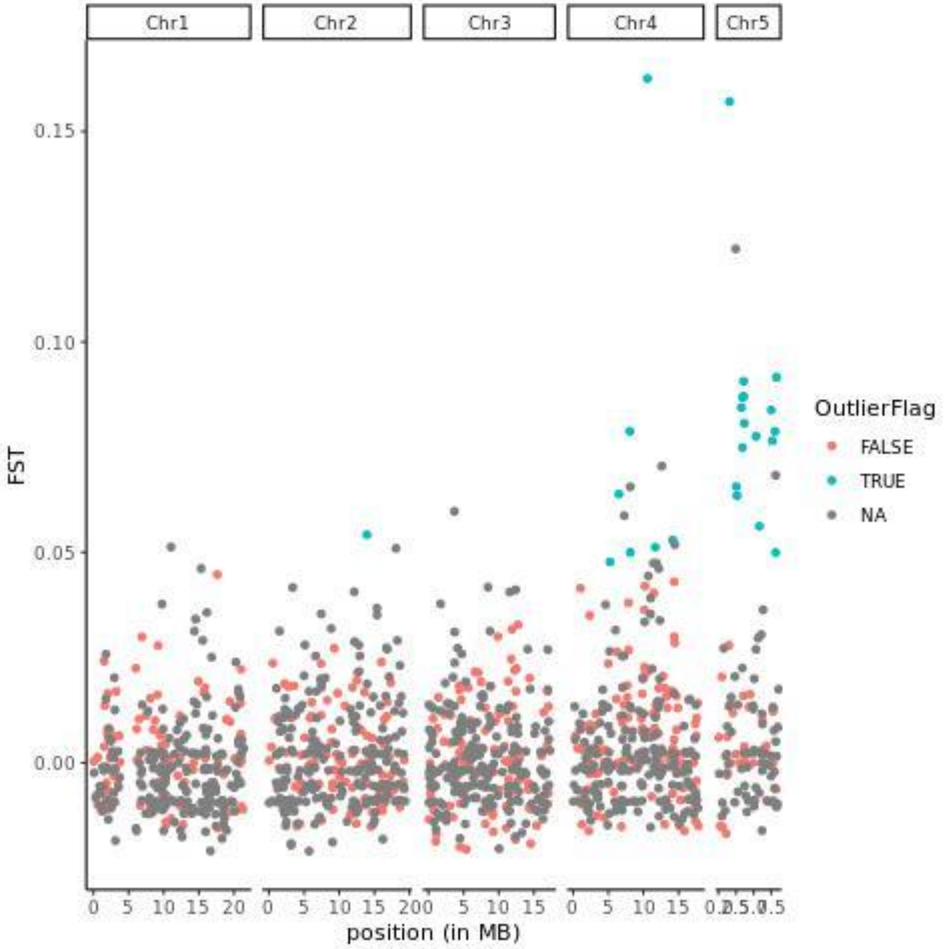
3-1 Create a subset of LD-pruned SNPs
-> we use plink

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



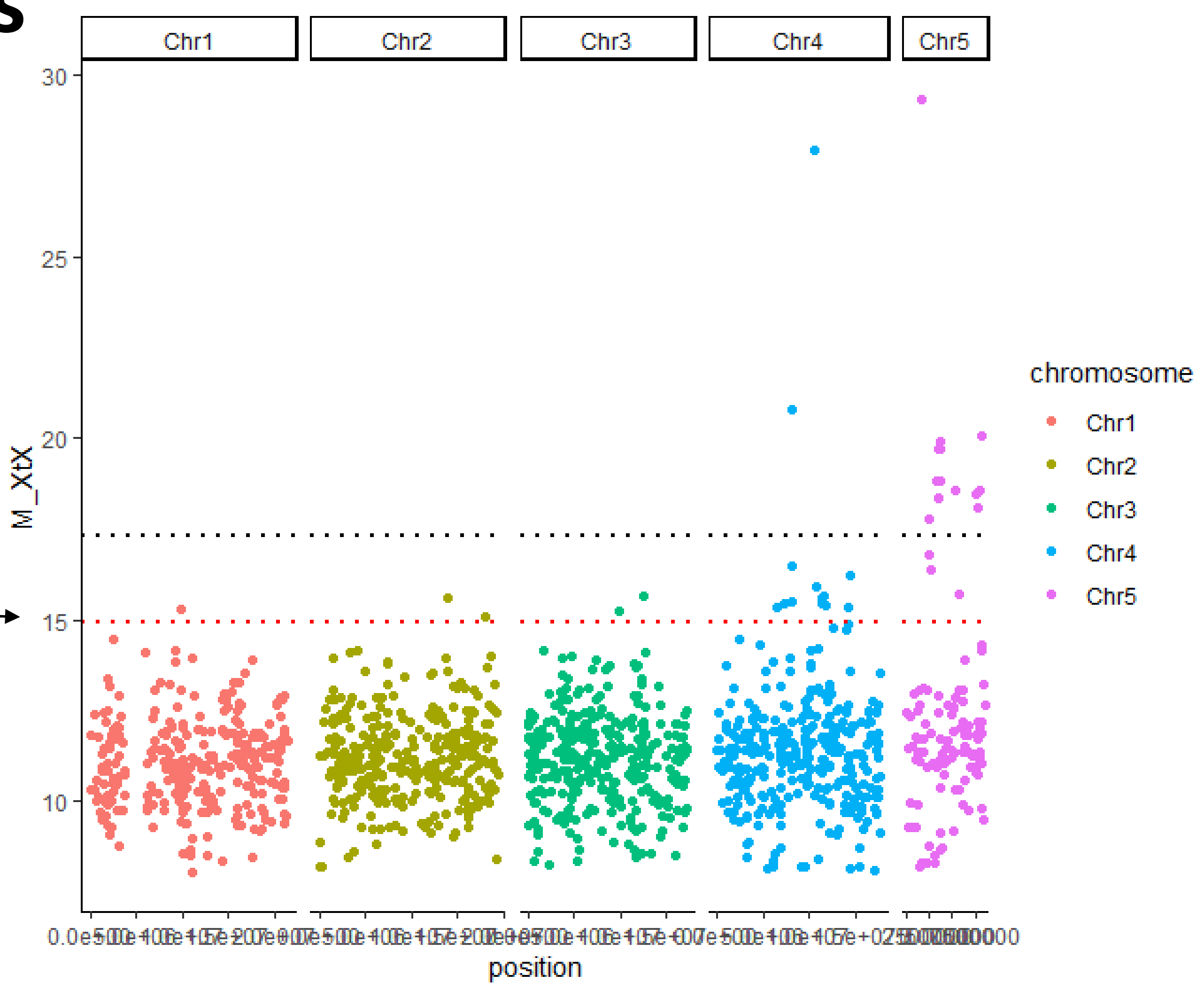


3-2 Outlier detection
-> with Baypass

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

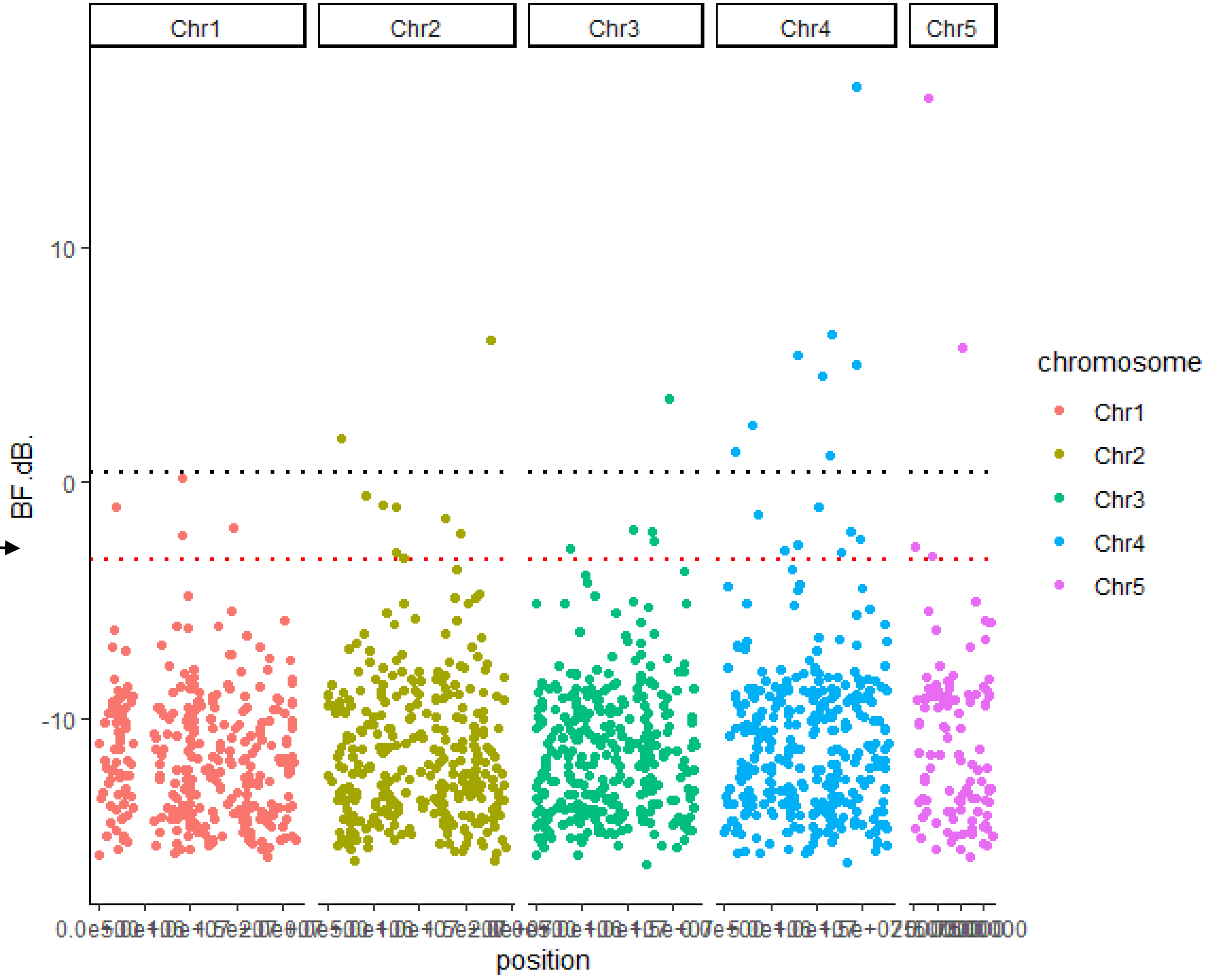


3-3 Environmental associations -> with Baypass

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

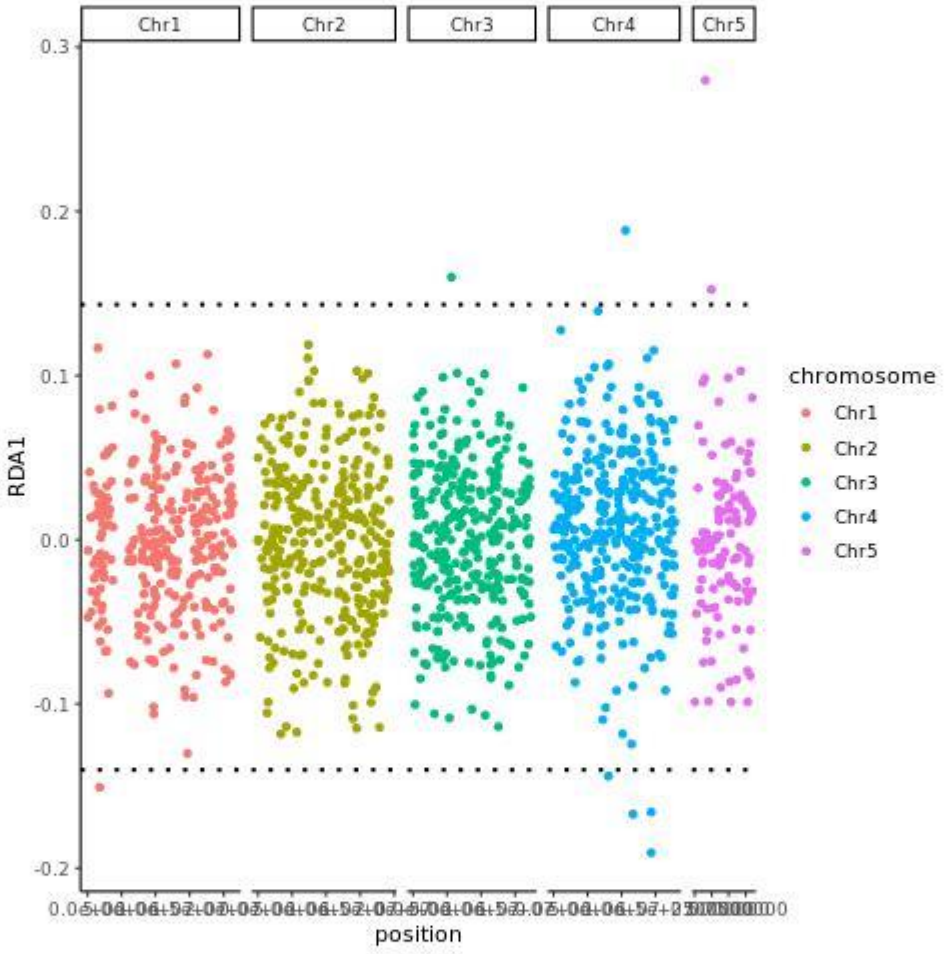


Simply add a co-variable matrix describing environmental variations between pop

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



Baypass
about making independant runs

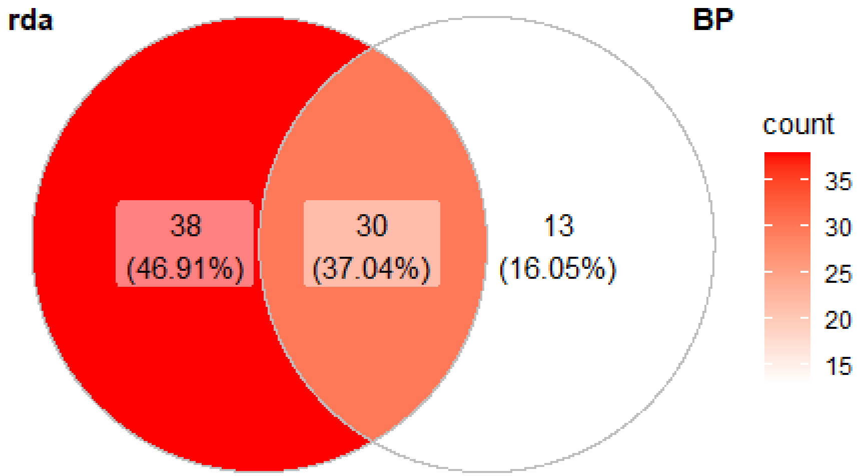
What we did

- 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

3-3 Environmental associations
-> Overlap



Day 4: Detecting structural variants

1: Detection of haplotypic blocks (putative inversions, young sex chromosomes, etc)

1 Detection with local PCA

2 Exploration of the haploblocks (genotype, LD, Fst, Hobs)

2: Whole-genome sequencing for SNPs and small/medium SVs

3: How to explore duplicated loci in RAD-seq data

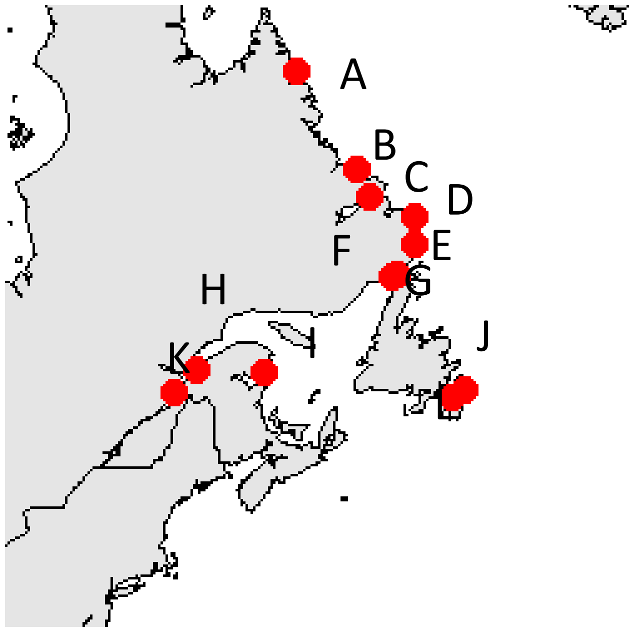
Demonstration by Yann

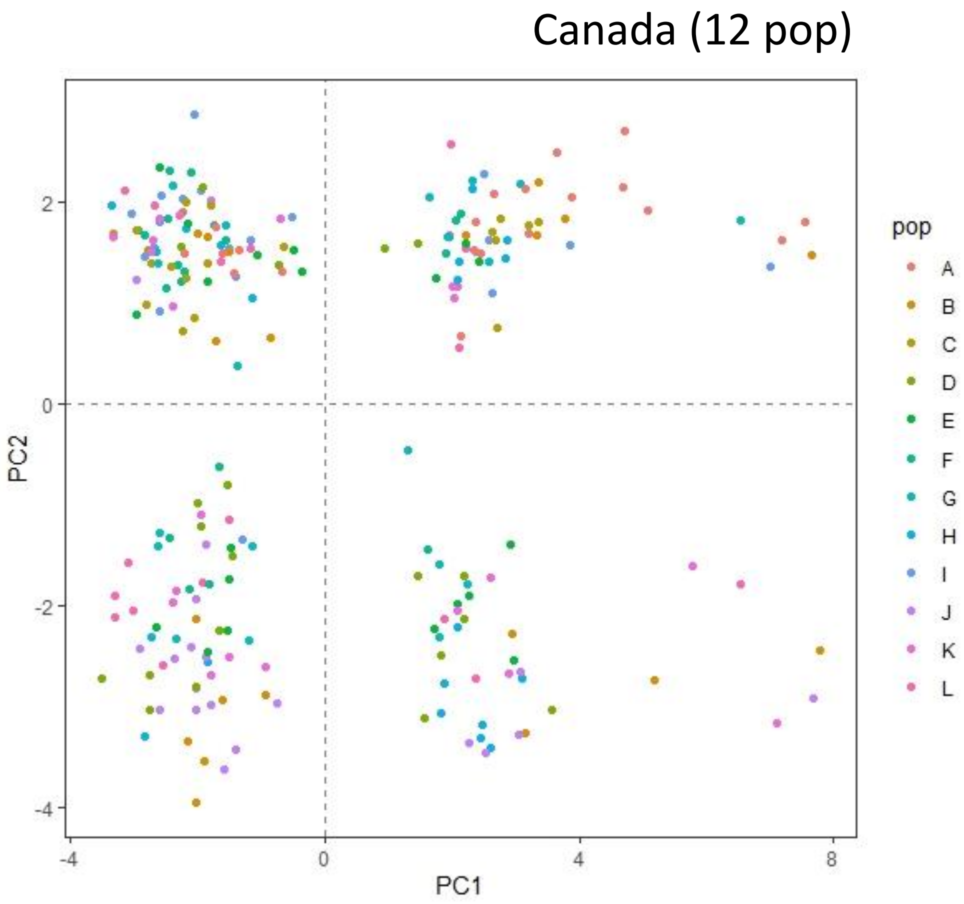
Detection and filtering of duplicated loci

Analysis of those CNVs in pop G

Why?

On day 2, we observed a strong structure on the PCA of the 12 Canadian populations…





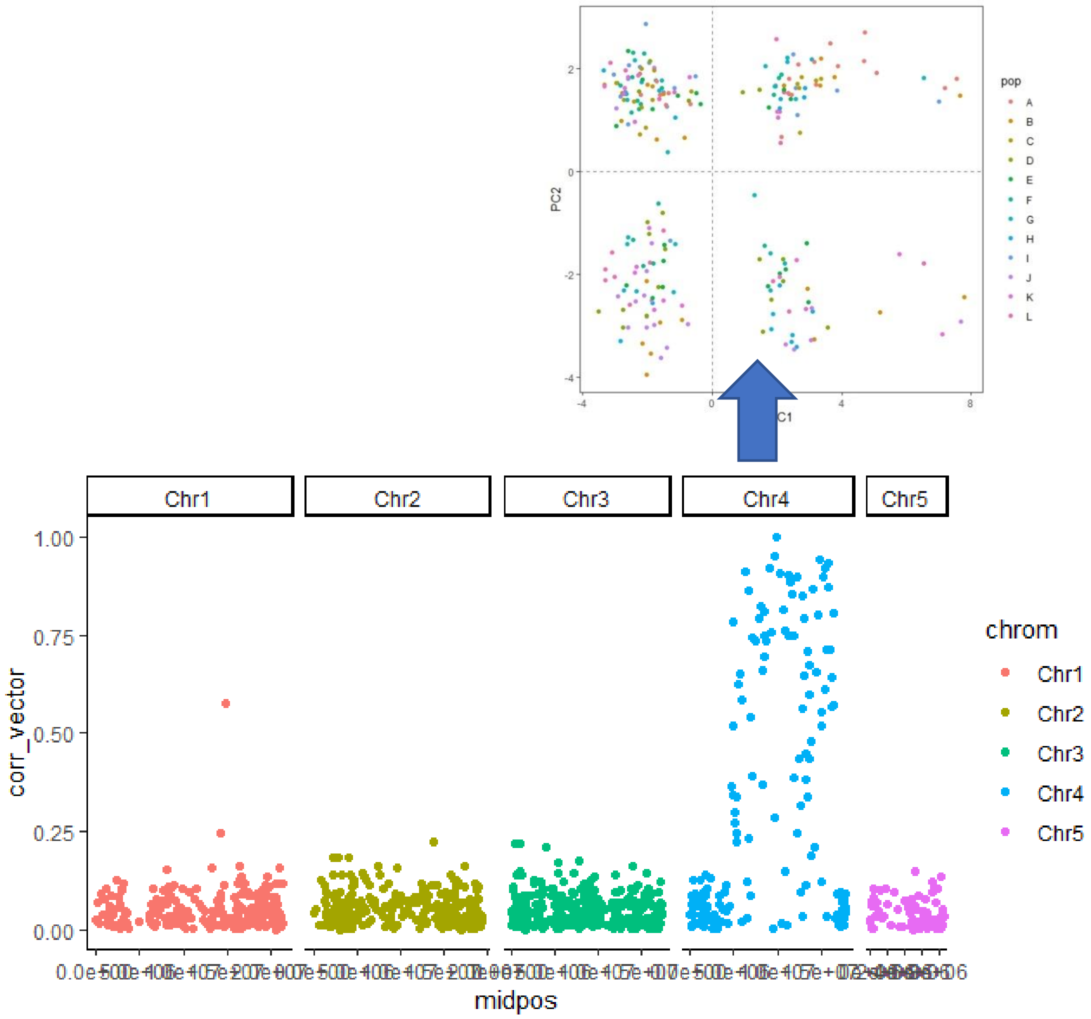
⇒ More generally, structural rearrangements and sex-linked regions may bias populations structure inference when left unknown (particularly in species with high gene flow)

HIGHLIGHTED ARTICLE
GENETICS | INVESTIGATION

Local PCA Shows How the Effect of Population Structure Differs Along the Genome

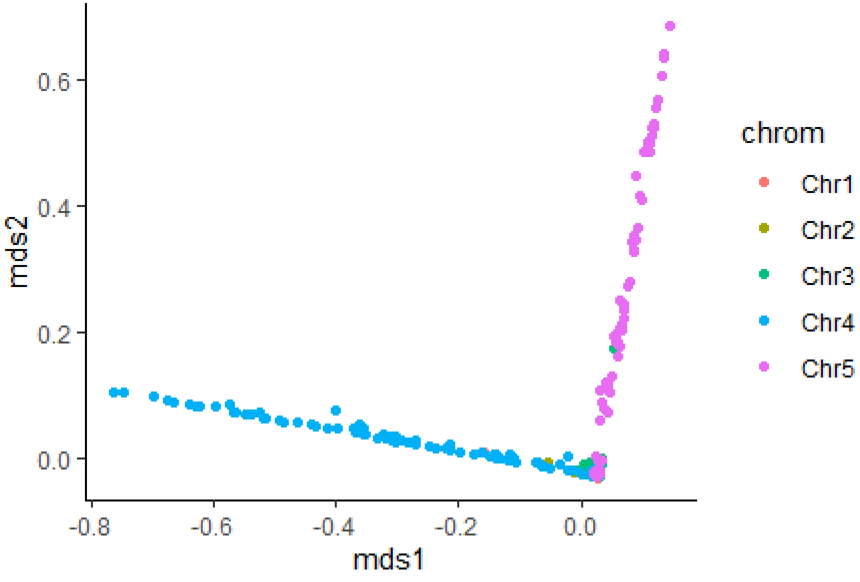
Han Li*[*] and Peter Ralph[*,†,‡,1]
[*]Department of Molecular and Computational Biology, University of Southern California, Los Angeles, California 90089 and [†]Institute of Ecology and Evolution and [‡]Department of Mathematics, University of Oregon, Eugene, Oregon 97403
ORCID ID: 0000-0002-9459-6866 (P.R.)

4-1 Detection with local PCA



Correlation between local PCA and global PCA

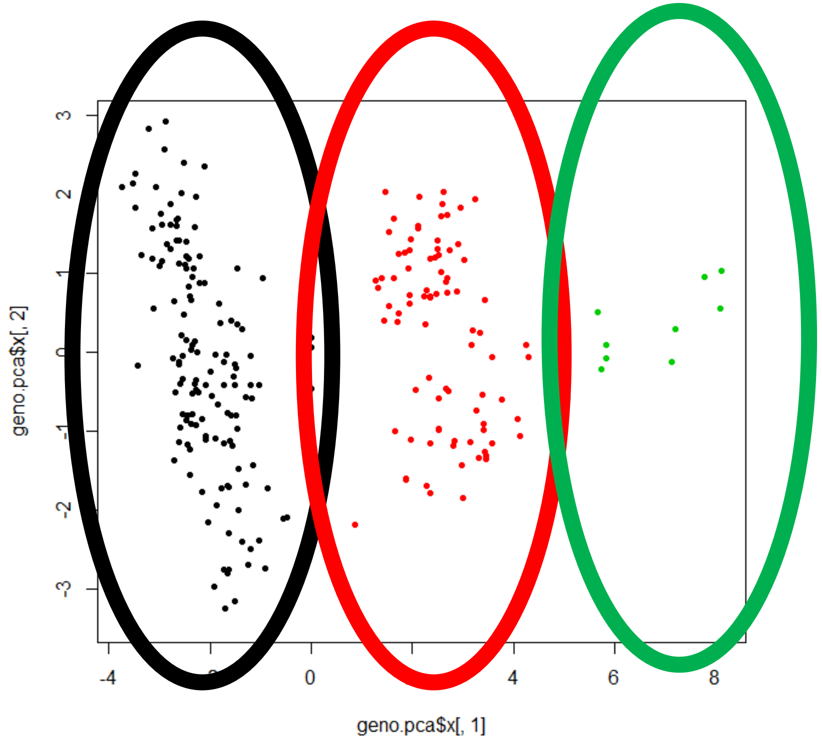
MDS looking at similar windows accross the genome



4-1 Exploration of the haploblocks

-> Genotype

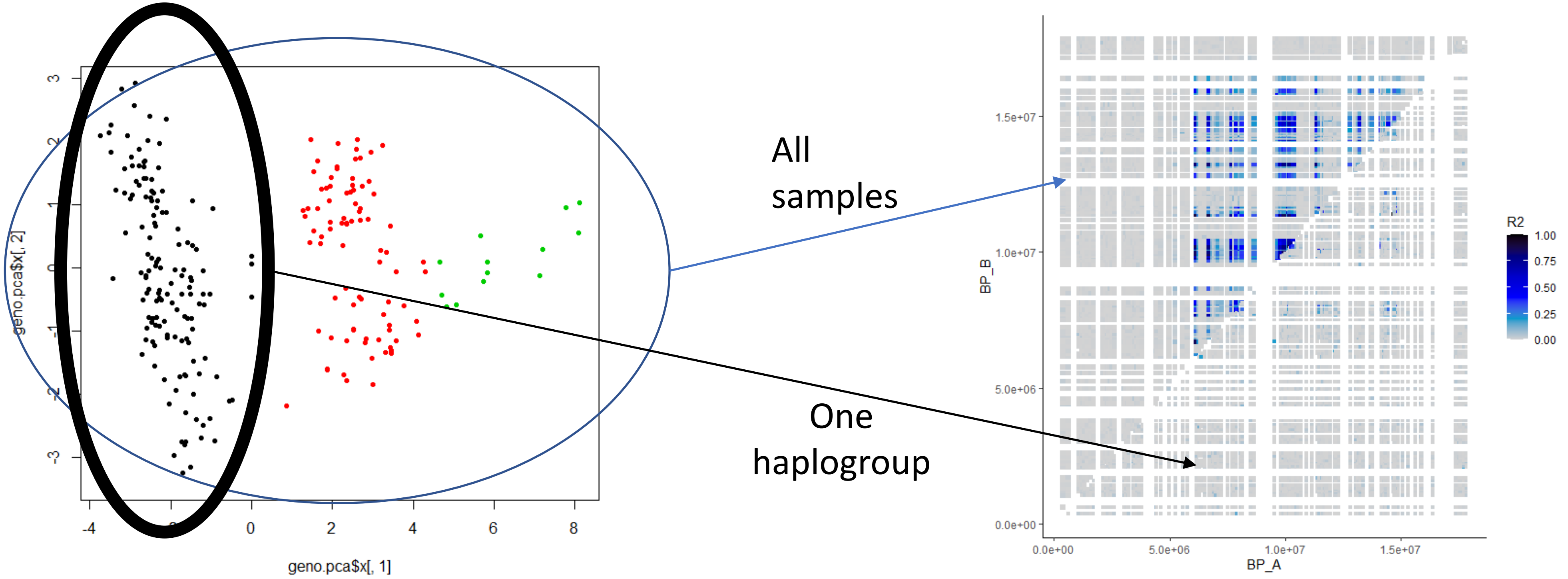
-



4-1 Exploration of the haploblocks

-> Genotype

-> Linkage disequilibrium

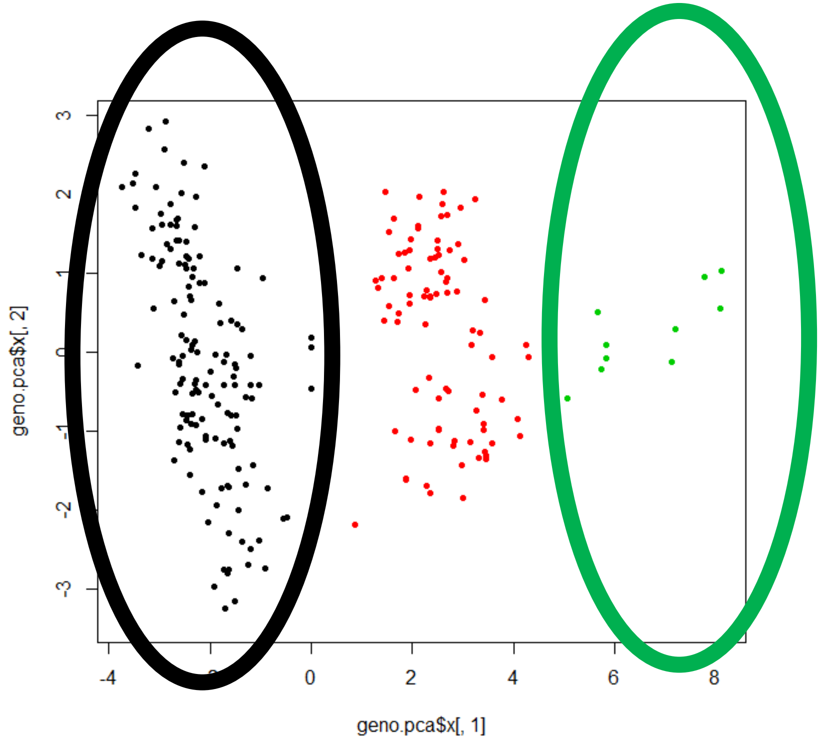


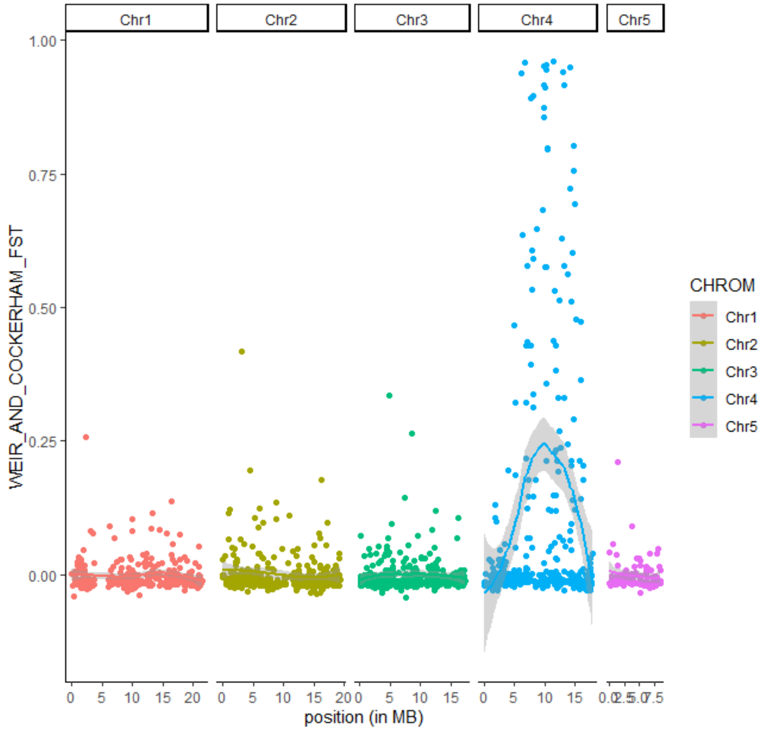
4-1 Exploration of the haploblocks

-> Genotype

-> Linkage disequilibrium

-> Fst between haplogroups (optional)





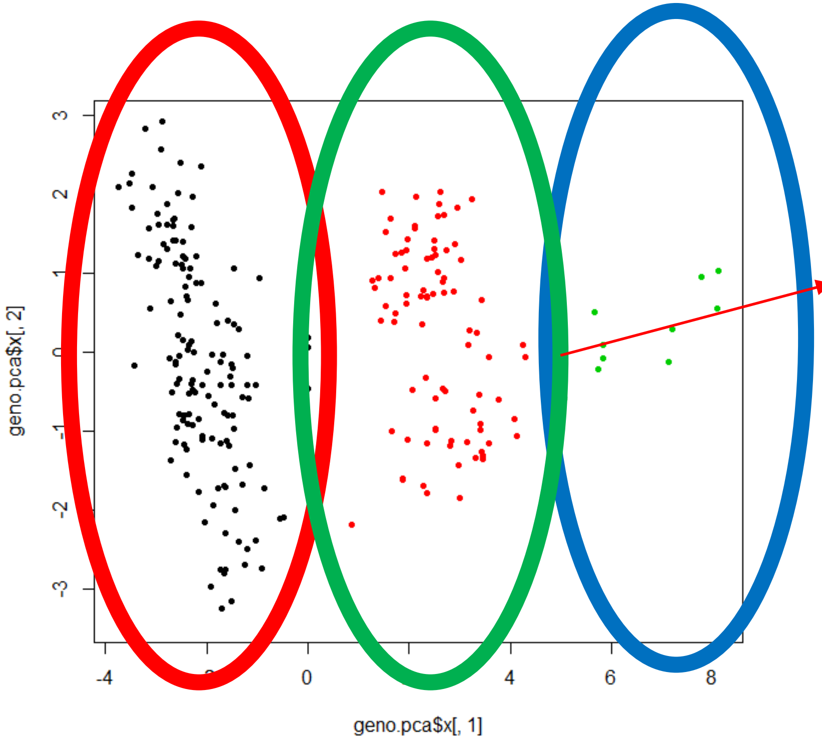
4-1 Exploration of the haploblocks

-> Genotype

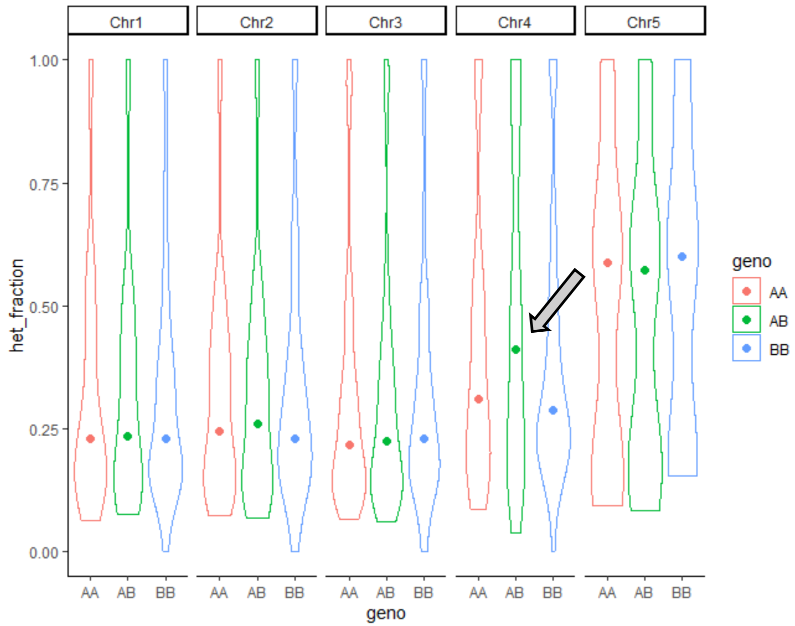
-> Linkage disequilibrium

-> Fst between haplogroups (optional)

-> Observed fraction of heterozygotes (optional)



More heterozygotes in that region ?



Day 4: Detecting structural variants

1: Detection of haplotypic blocks (putative inversions, young sex chromosomes, etc)

1 Detection with local PCA

2 Exploration of the haploblocks (genotype, LD, Fst, Hobs)

2: Whole-genome sequencing for SNPs and small/medium SVs

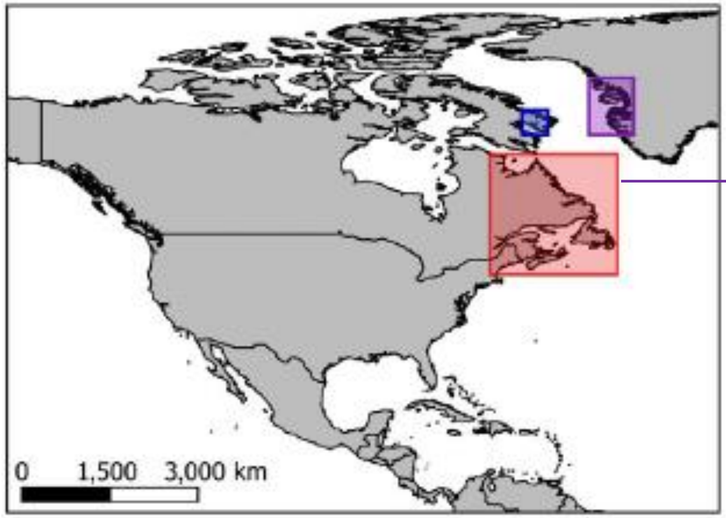
3: How to explore duplicated loci in RAD-seq data

Demonstration by Yann

Detection and filtering of duplicated loci

Analysis of those CNVs in pop G

For Day4: whole-genome sequencing



North American species

12 samples from different canadian populations

Whole-genome sequencing = much bigger files
BUT useful for SV detection or for a higher density of SNPs

Here we pick a very reduced dataset to make things run fast!!

Tutorial day 5

Most methods that we saw during the week will provide

$\Rightarrow$ **General knowledge about isolation-by-adaptation, the genetic architecture of adaptation, an idea of genomic variance related to possible ecological variation, etc …**

$\Rightarrow$ **Putatively-adapted SNPs, SVs or genomic regions**

- Can we point towards causal candidate genes or pathways ?

Local adaptation / population genomics

Gene annotation, gene ontology, gene enrichment

Genome + transcriptome + protein databases + transposable elements databases

⇒ By aligning the transcriptome on the genome we can know gene positions (and exon, intron, etc…)

⇒ The transcriptome can be annotated thanks to protein databases (protein sequences usually more conserved than DNA sequences)

⇒ Genes/Proteins are gather into functional categories called « gene ontology »
http://geneontology.org/docs/ontology-documentation/

⇒ Thanks to TE databases and repeat detection, the genome can be annotated for interspersed reapeats.

Tutorial day 5

We will:

- **Annotate the SNPs to know whether they belong to exon, intron, regulatory regions**
- **Look for genes at the proximity of our outlier SNPs**
- **Test for enrichment in the outliers for particular GO categories**
- **Investigate whether some of the CNV are transposable elements or repeated regions**

http://geneontology.org/docs/ontology-documentation/

5-1 Annotate SNPs

-> We will use SNPeff

It uses genome annotation (Gff) to say whether SNPs belong to genes, intergenic region, introns, etc…

```
#CHROM  POS     ID      REF     ALT     QUAL    FILTER  INFO    FORMAT
Chr1    53559   49:9:-  C       G       .       PASS    ANN=G   upstream_gene_variant
Chr1    94208   95:21:+ A       G       .       PASS    ANN=G   intergenic_region
Chr1    308478  248:57:+        T       G       .       PASS    ANN=G   downstream_gene_variant
Chr1    510235  370:36:+        G       A       .       PASS    ANN=A   intergenic_region
Chr1    586674  438:51:-        T       A       .       PASS    ANN=A   splice_region_variant&intron_variant
```

We will do a small analysis to look whether outliers are enriched in one category

5-2 Overlap SNPs / Genes

-> We will use Bedtools

It takes bedfiles with position of the SNPs, position of the outliers, and position of the genes

```
Chr1    1518343 1528343 1262:33:-
Chr1    1785873 1795873 1582:14:+
Chr1    3100385 3110385 2846:22:+
Chr1    9138069 9148069 6032:68:+
```

Bed format is CHR START STOP and then 1 to 9 columns with informations

Bedtools function « intersect » is used to look for the overlap

5-3 Gene ontology enrichment
-> We will use goseq library in R

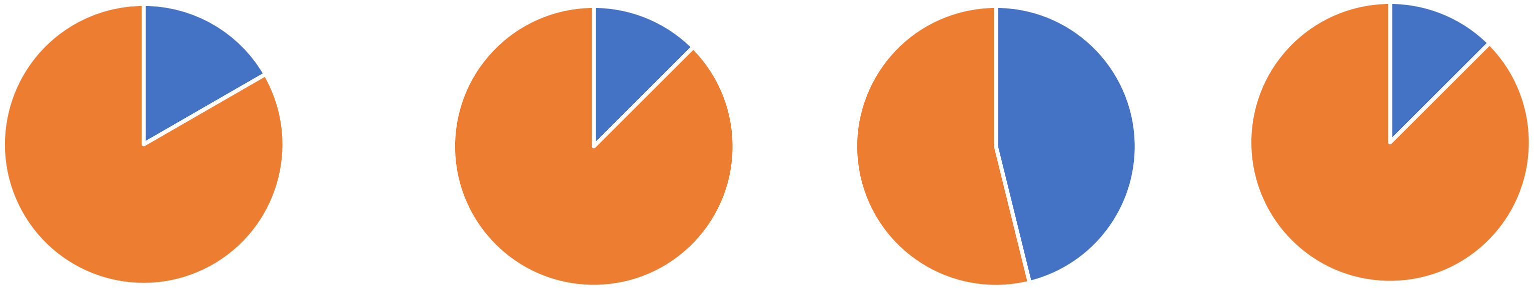
Warning: lots of the tutorial is about getting the good format!

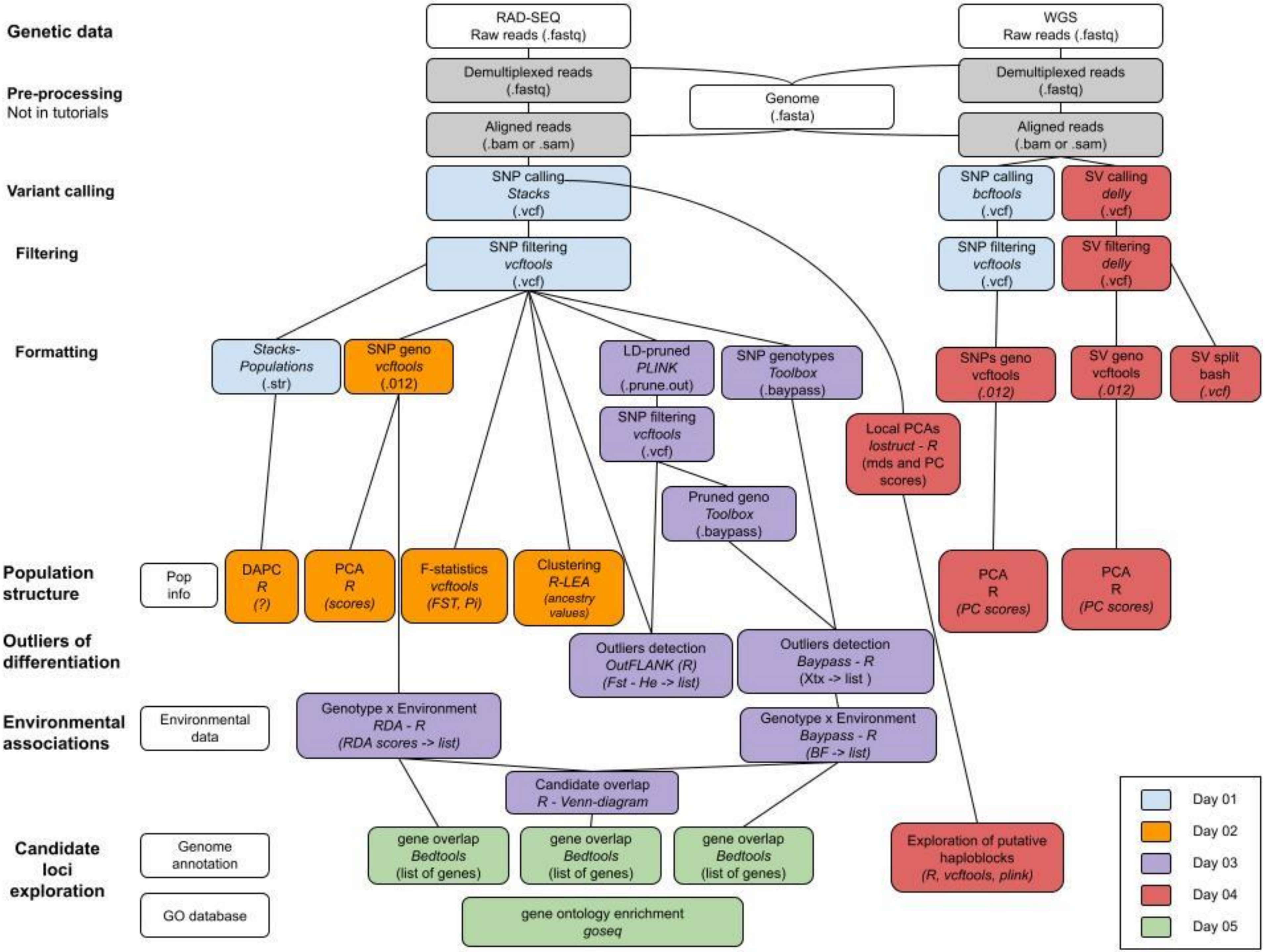
Warning: GO enrichment are more appropriate for RNAseq analysis & whole-genome analysis.

Warning: The genes overlapping with outliers should be contrasted against the pool of genes overlapping with SNPs (not with all the gnees in the genome as some of them may simply not be covered)

```
 category over_represented_pvalue under_represented_pvalue numDEInCat numInCat                                                  term ontology
GO:0002084            0.0001560823                1.0000000          3        3                              protein depalmitoylation       BP
GO:0008474            0.0001560823                1.0000000          3        3                  palmitoyl-(protein) hydrolase activity       MF
GO:0002116            0.0002946549                0.9999945          4        5                            semaphorin receptor complex       CC
GO:0017154            0.0002946549                0.9999945          4        5                           semaphorin receptor activity       MF
GO:1902287            0.0002946549                0.9999945          4        5 semaphorin-plexin signaling pathway involved in axon guidance       BP
GO:0007162            0.0002968094                0.9999838          5        9                    negative regulation of cell adhesion       BP
```

	genome	hit by sNPs	in outliers with env A	in outliers with env B
G0 code X	2000	100	30	5
all genes	10000	700	35	35




Genetic data
RAD-SEQ
Raw reads (.fastq)
WGS
Raw reads (.fastq)
Pre-processing
Not in tutorials
Demultiplexed reads
(.fastq)
Genome
(.fasta)
Aligned reads
(.bam or .sam)
Variant calling
SNP calling
Stacks
(.vcf)
SNP calling
bcftools
(.vcf)
SV calling
delly
(.vcf)
Filtering
SNP filtering
vcftools
(.vcf)
SV filtering
delly
(.vcf)
Formatting
Stacks-
Populations
(.str)
SNP geno
vcftools
(.012)
LD-pruned
PLINK
(.prune.out)
SNP genotypes
Toolbox
(.baypass)
SNPs geno
vcftools
(.012)
SV geno
vcftools
(.012)
SV split
bash
(.vcf)
Local PCAs
lostruct - R
(mds and PC
scores)
Pruned geno
Toolbox
(.baypass)
Population
structure
Pop
info
DAPC
R
(?)
PCA
R
(scores)
F-statistics
vcftools
(FST, Pi)
Clustering
R-LEA
(ancestry
values)
PCA
R
(PC scores)
Outliers of
differentiation
Outliers detection
OutFLANK (R)
(Fst - He -> list)
Outliers detection
Baypass - R
(Xtx -> list)
Environmental
associations
Environmental
data
Genotype x Environment
RDA - R
(RDA scores -> list)
Genotype x Environment
Baypass - R
(BF -> list)
Candidate overlap
R - Venn-diagram
Candidate
loci
exploration
Genome
annotation
GO database
gene overlap
Bedtools
(list of genes)
Exploration of putative
haploblocks
(R, vcftools, plink)
gene ontology enrichment
goseq
Day 01
Day 02
Day 03
Day 04
Day 05