Tutorials overview and wrap-up

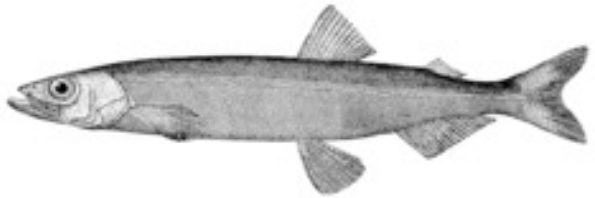
Adaptation Genomics Course

Mafalda Ferreria, PhD & Angela Fuentes Pardo, PhD

June 24 - 28, 2024

(adapted from Claire Mérot & Anna Tigano's slides)

Capelin dataset



DOI: 10.1111/mec.15499

ORIGINAL ARTICLE

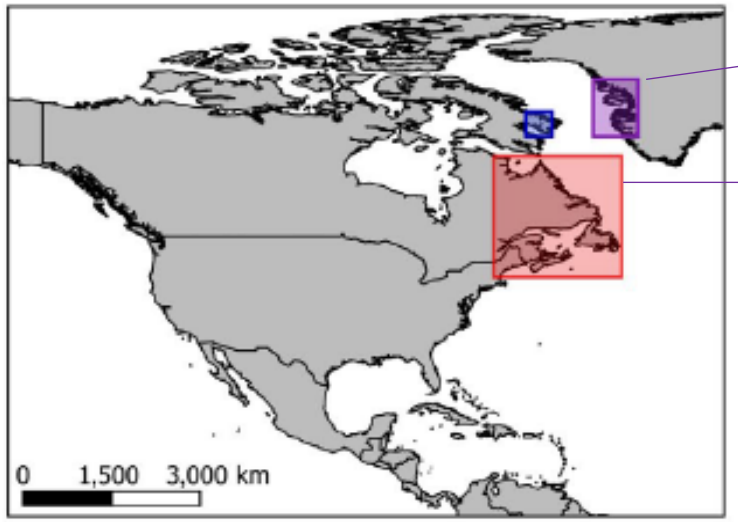
MOLECULAR ECOLOGY WILEY

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

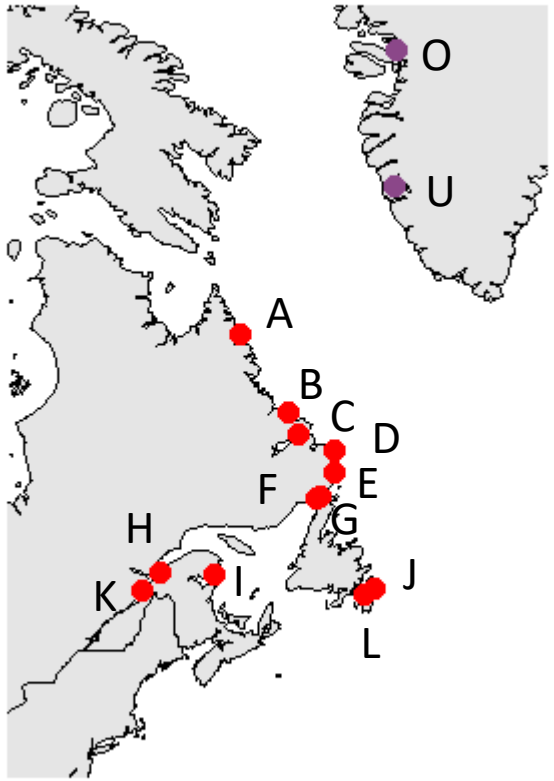


Greenland lineage

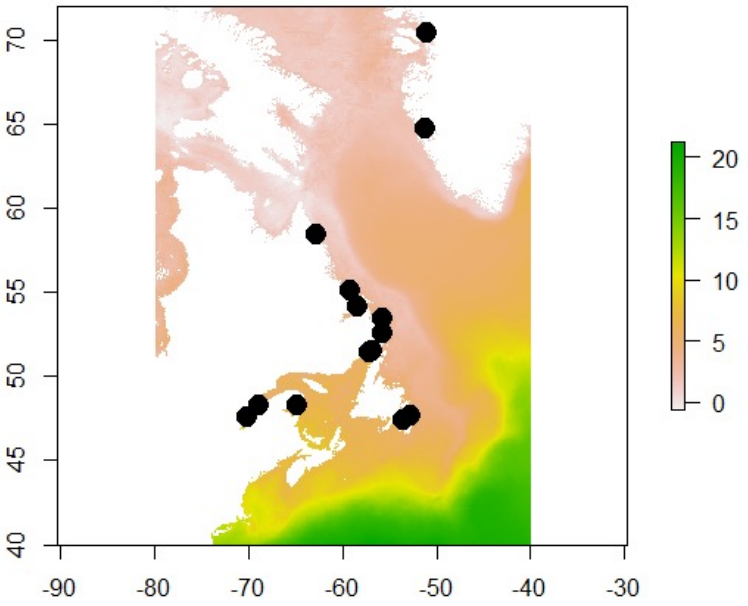
North American lineage

2 populations
N= 40

12 populations
N= 240 (20/pop)



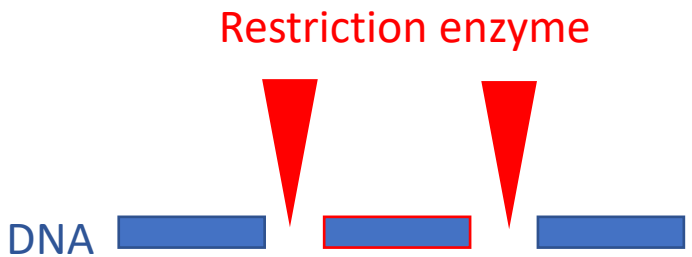
Sea temperature (from MARSPEC)



Capelin dataset

fish

DNA



Double-digest
RAD librairies

Reads
(single-end – 80 bp)

De-multiplex
Trim?

N fastQ files

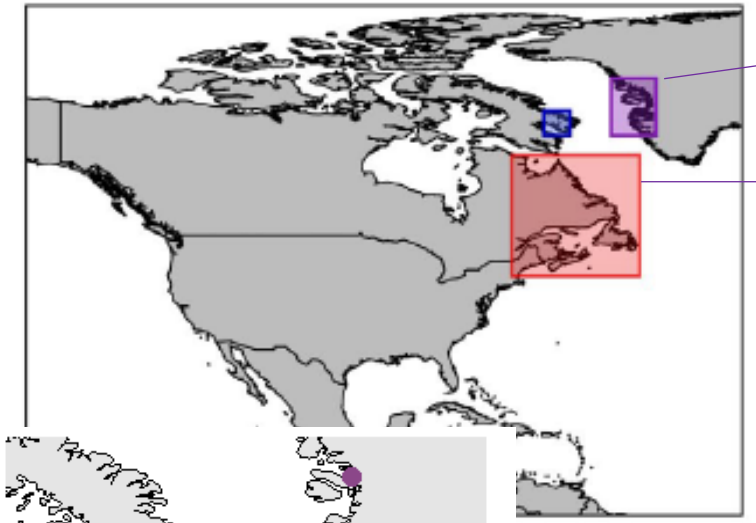
```
@70ZFD:01332:11598/1
TGCATCAACTTTAAGATACGCTATTGGAGCTGGAATTACCGCGGCTGCTGGCACCCAGACTTGCCCTCCAATGGATCCTC
+
7<<=<;<4676*115345::=;<=6;5<;<;;7<1918<199<6<<::9:5:556+38469166=3;<6<655-477-4/
@70ZFD:01334:11636/1
TGCATCCTGTGGAAGTAGCTGCACACCTGCTCATGCTGTGCCAGGAAGGGAGGGTGGGATCAGCCAATCGGGGAACAGAG
+
5;?;;;5855;4:4<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.

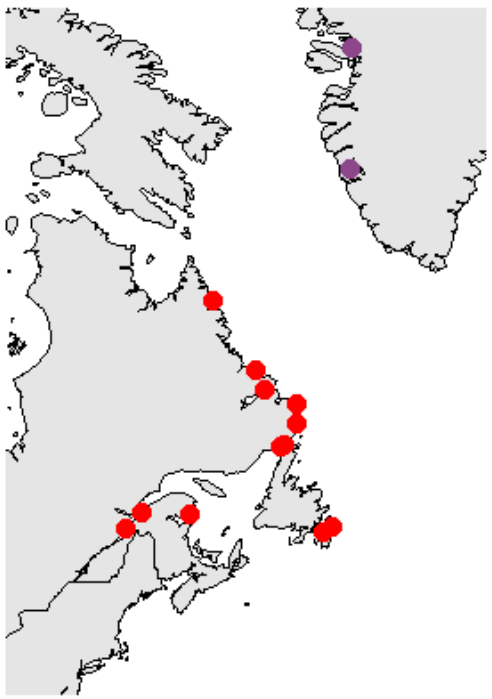


Greenland lineage

2 populations
N= 40

North American lineage

12 populations
N= 240 (20/pop)

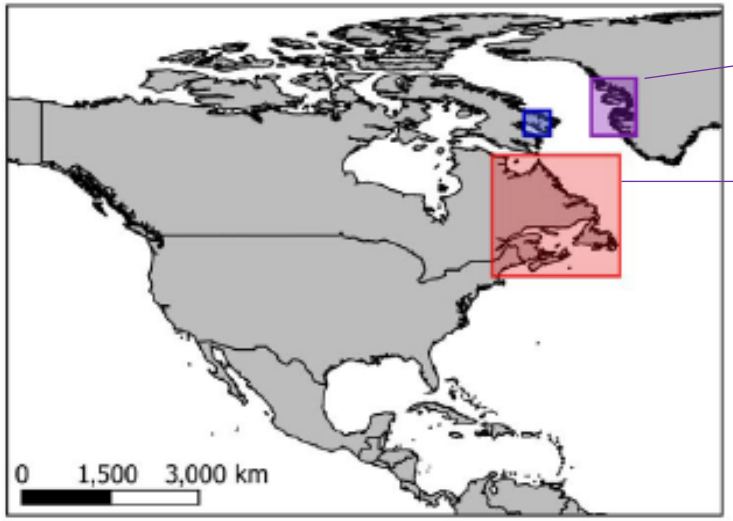


Dataset 1	Dataset 2	Dataset 3
« 2_lin »	« all »	« canada »
4 populations (2 greenland /2 canadian) => 80 samples	14 populations (2 greenland /12 canadian) => 280 samples	12 populations (12 canadian) => 240 samples
F_{st} (vcftools) PCA Optional (F_{st} with Stacks)	Faststructure DAPC	PCA DAPC Optional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

Day1 SNP calling with STACKS

Day2 Population structure

F_{st}



Greenland lineage

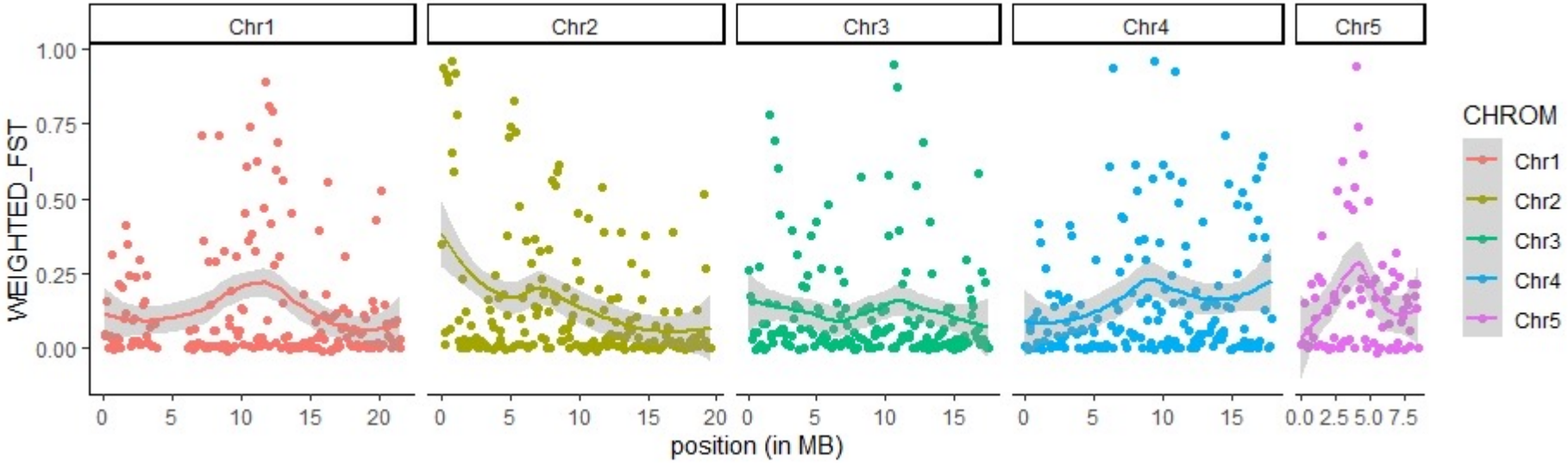
North American lineage

2 populations
N= 40

12 populations
N= 240 (20/pop)

F_{ST} = 0.23

Mean F_{ST} 0.03 / weighted F_{ST} 0.23

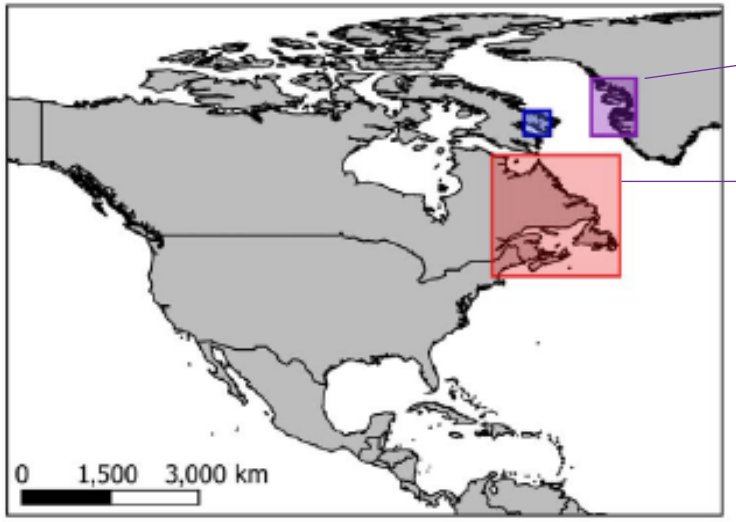


- Why do you think is necessary to impute missing values when performing a PCA? Would you follow the same approach for reduced representation sequence (RRS) and whole genome sequence (WGS) data?

PCA

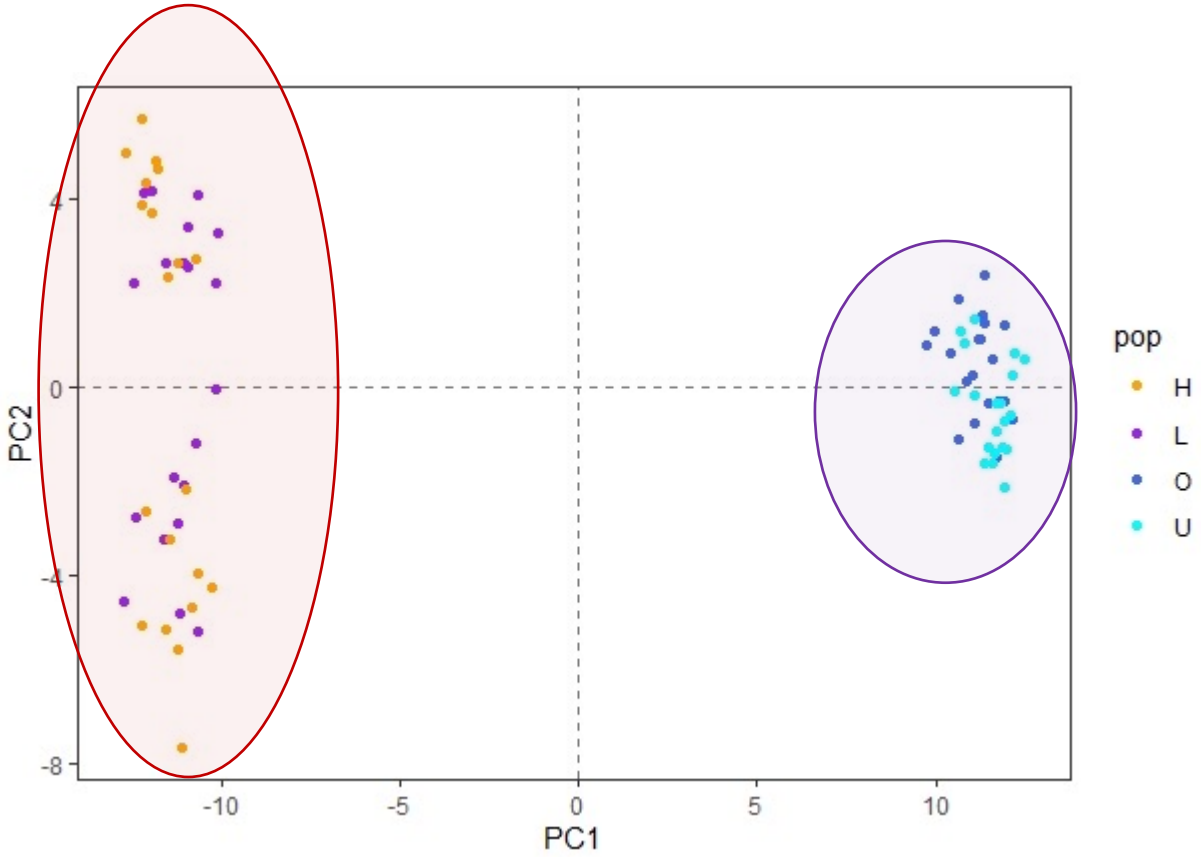
2 populations
N= 40

12 populations
N= 240 (20/pop)



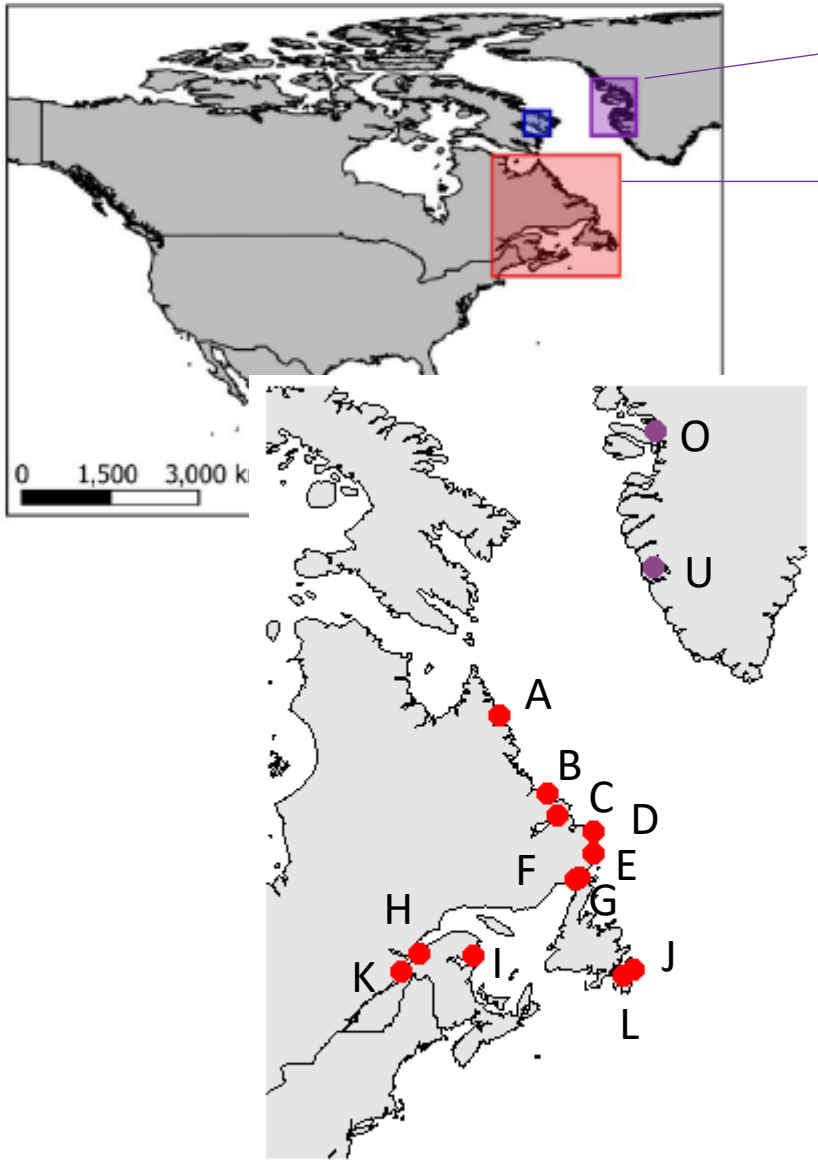
Greenland lineage

North American lineage



Q: What does it mean?

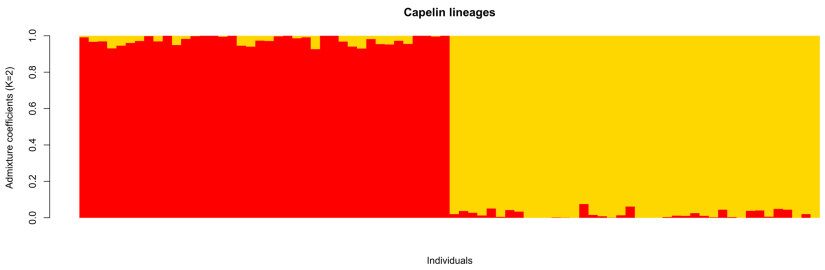
LEA – Clustering methods



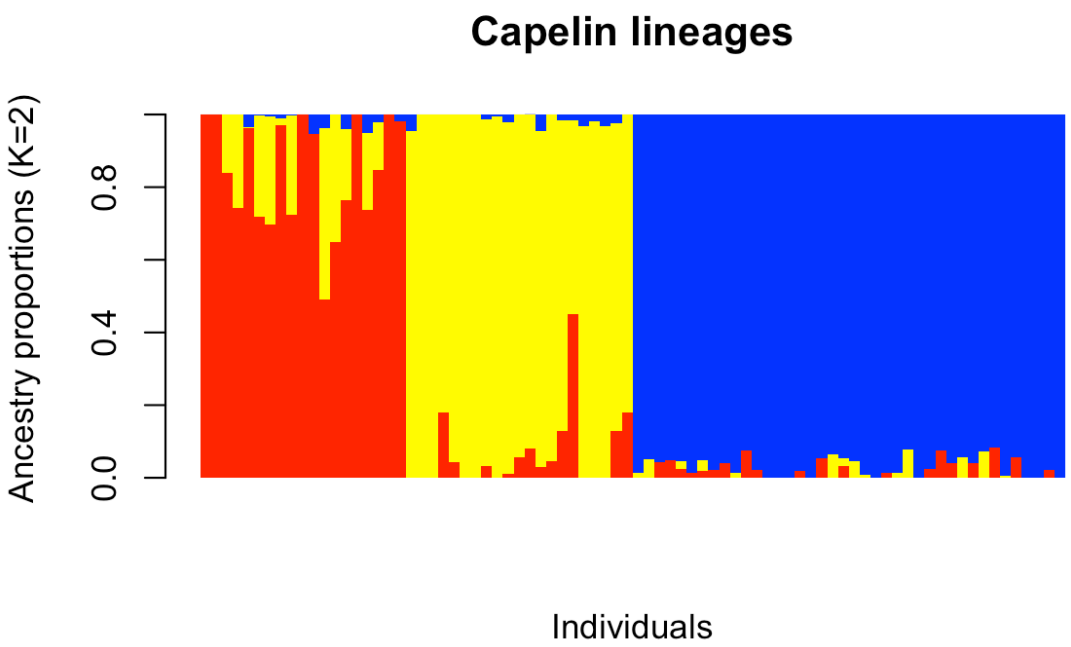
Greenland lineage

North American lineage

2 lineages



All

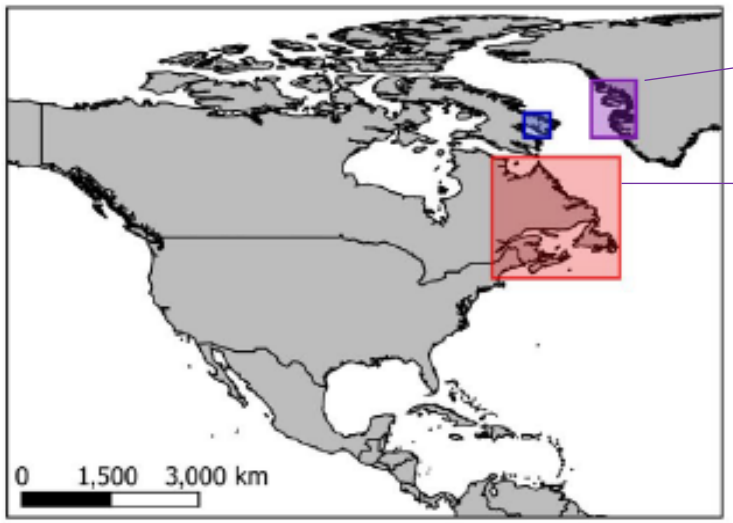


K=3

DAPC

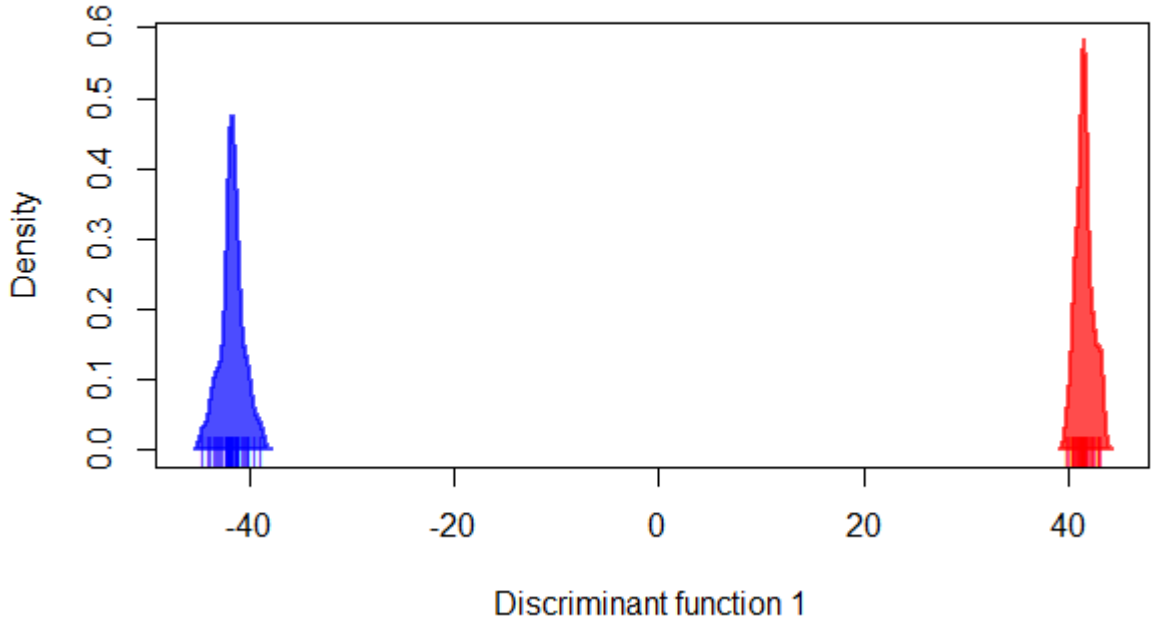
2 populations
N= 40

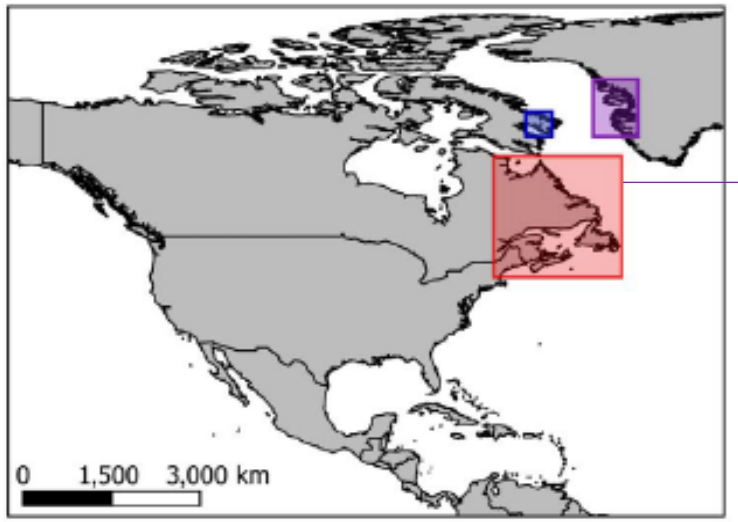
12 populations
N= 240 (20/pop)



Greenland lineages

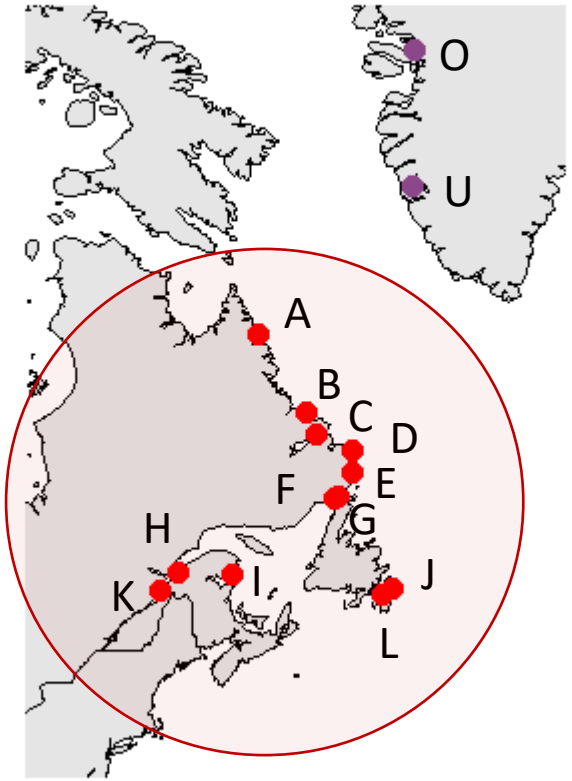
North American lineages



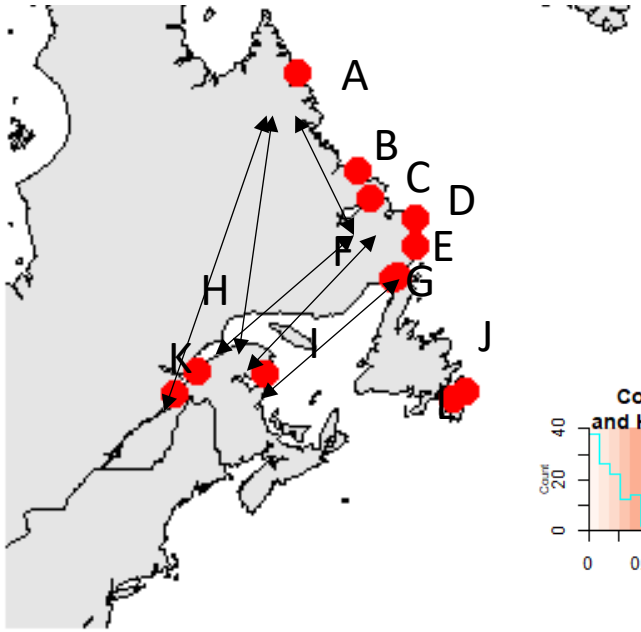


North American lineage

12 populations
N= 240 (20/pop)

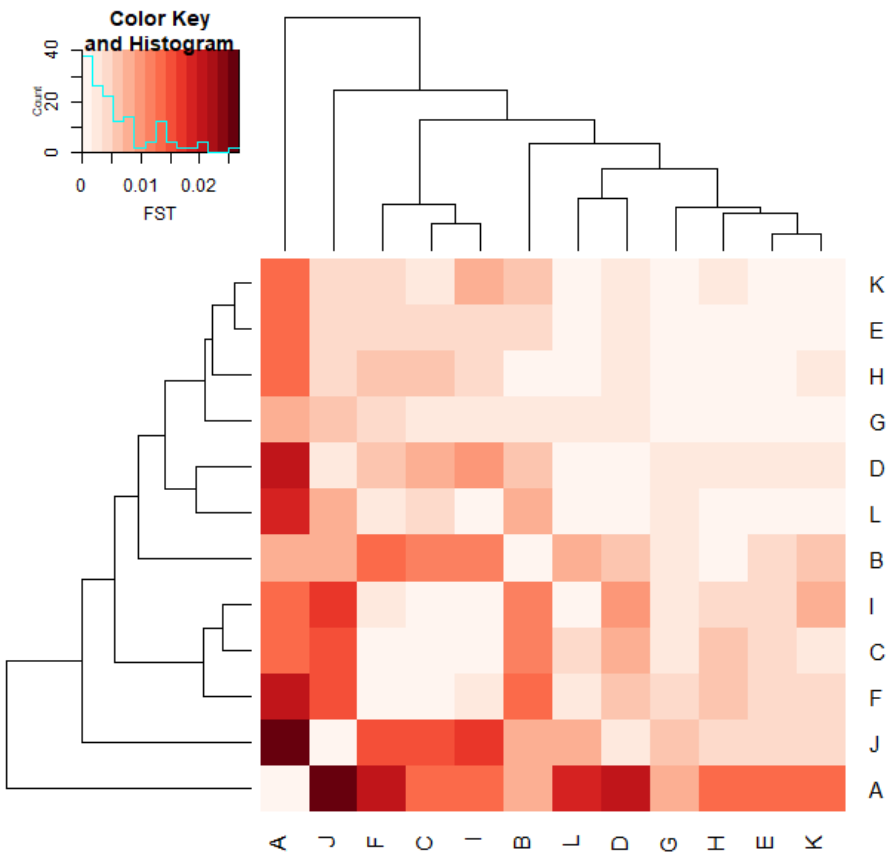


F_{ST}



Do we observe genetic structure ?

F_{ST} between all populations



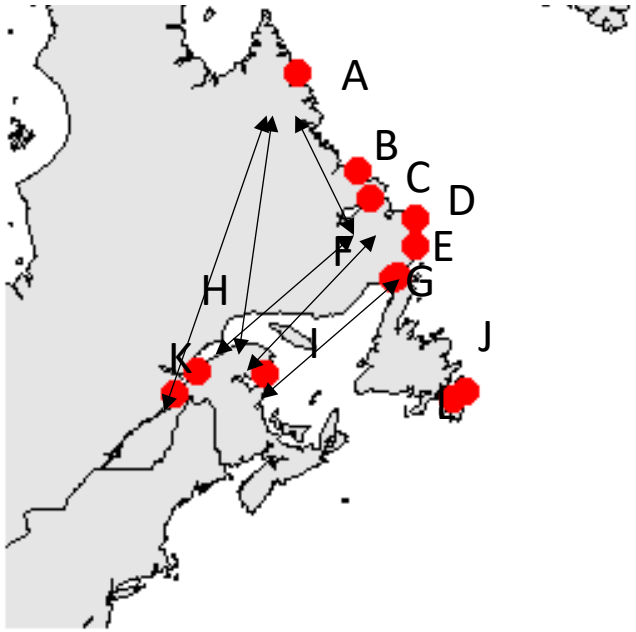
Medium values (F_{ST} = 0.025)?
Lots of heterogeneity…

⇒ 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)!!

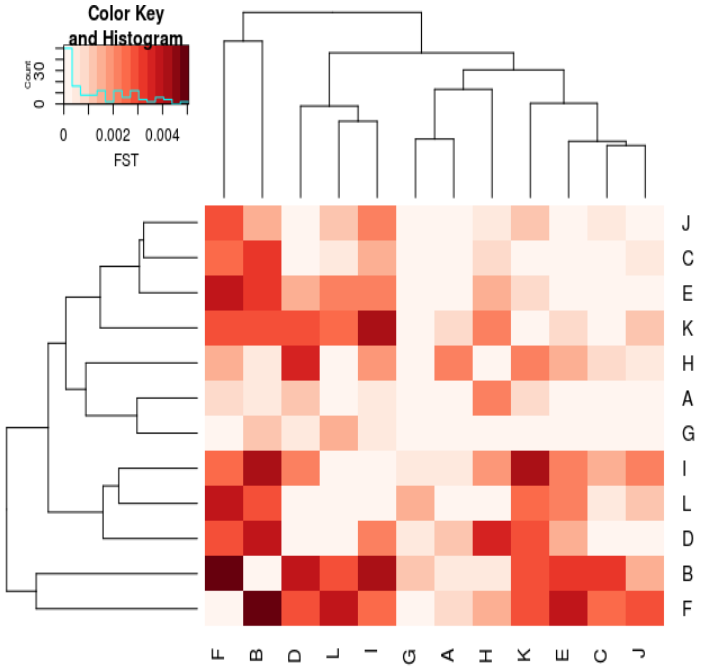
F_{ST}



Do we observe genetic structure ?

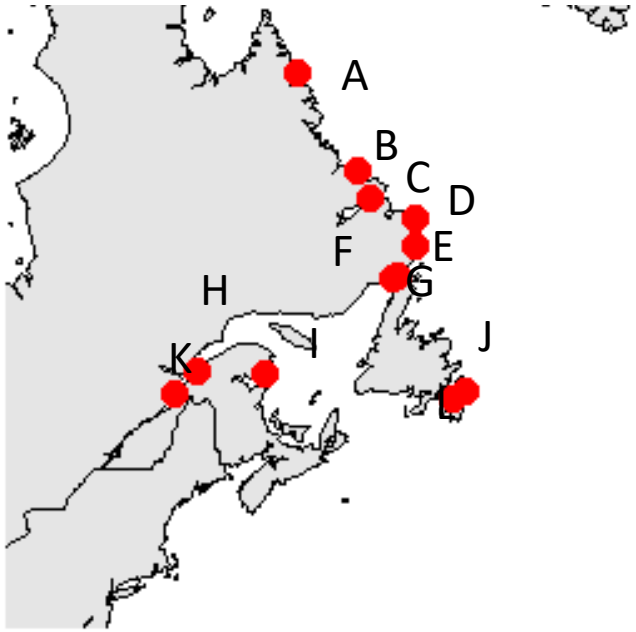
F_{ST} between all populations
Excluding chromosome 4 (inversion) & chromosome 5 (sex)

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

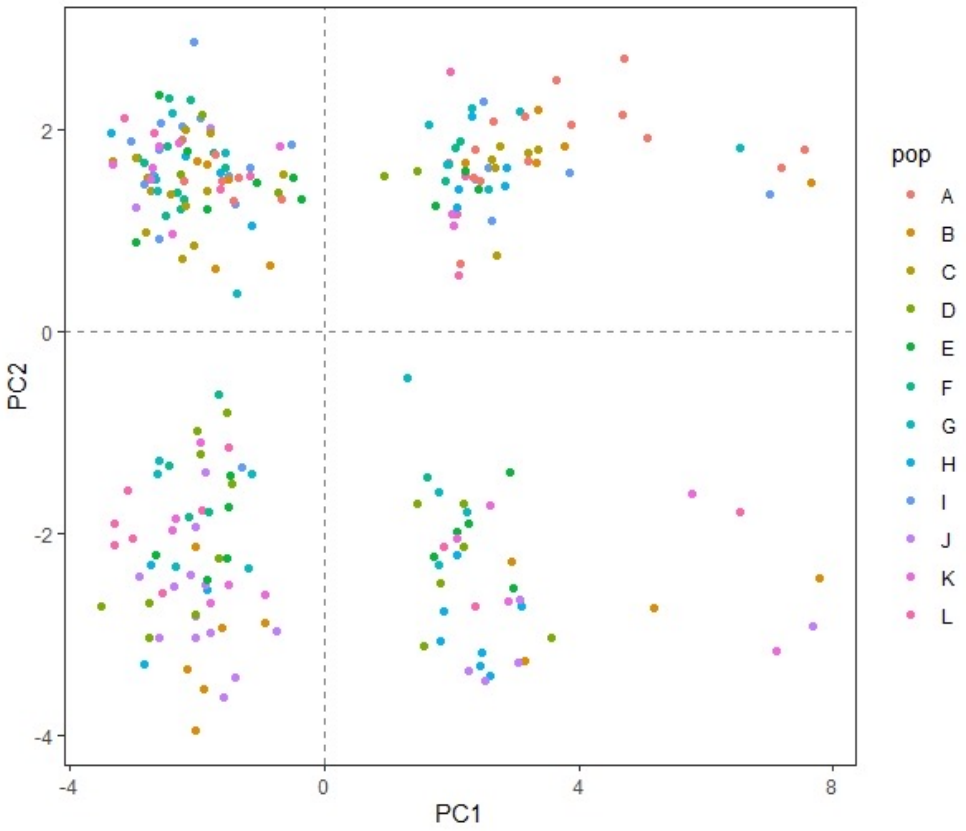


⇒ Almost no geographic structure…

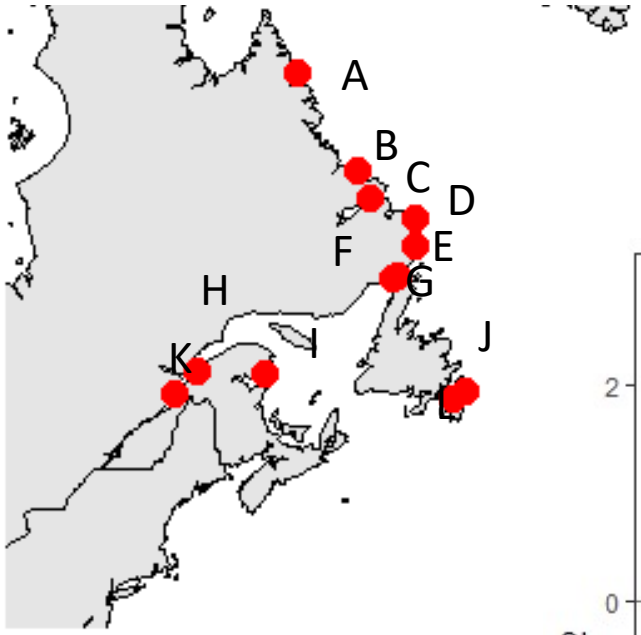
PCA



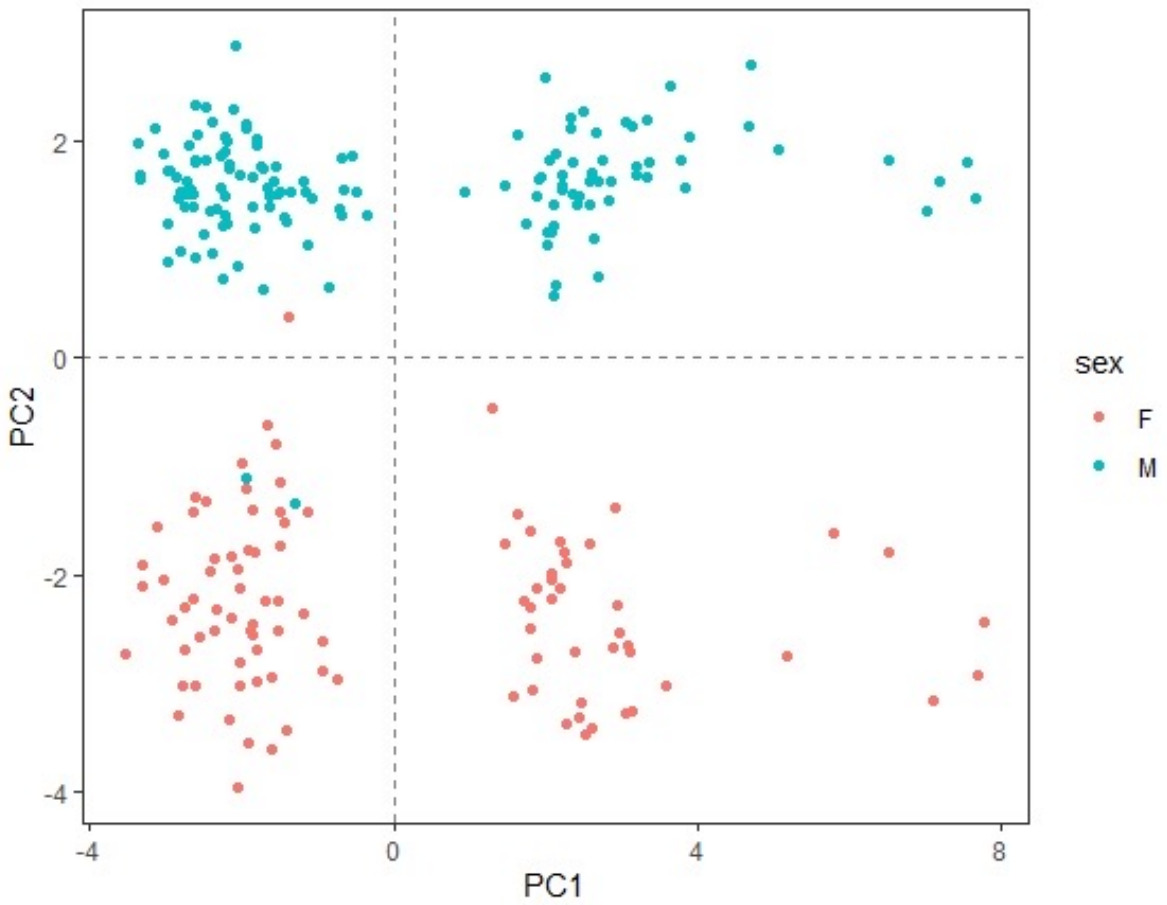
Canada (12 populations)



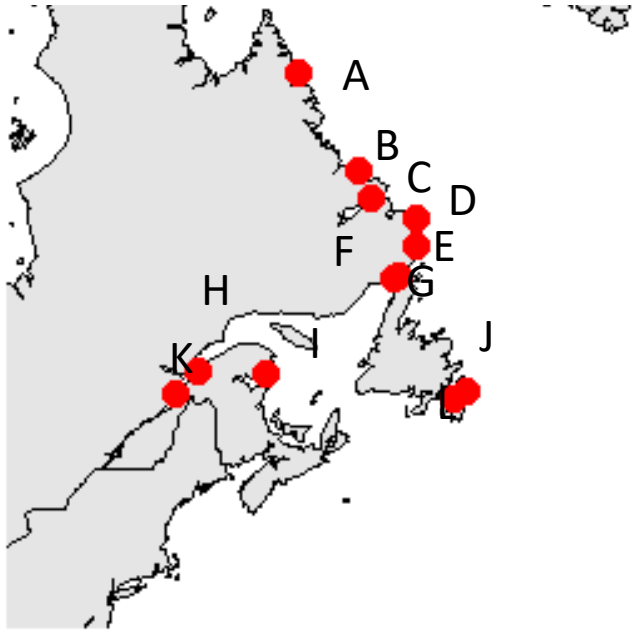
PCA



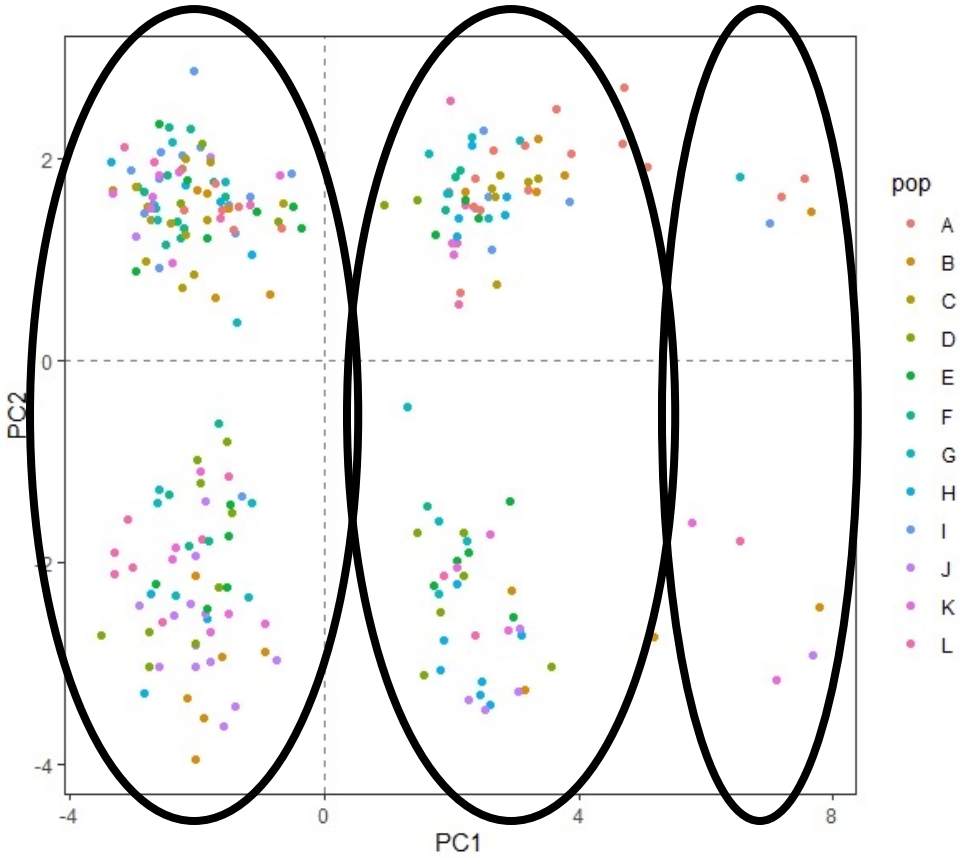
Canada (12 populations)



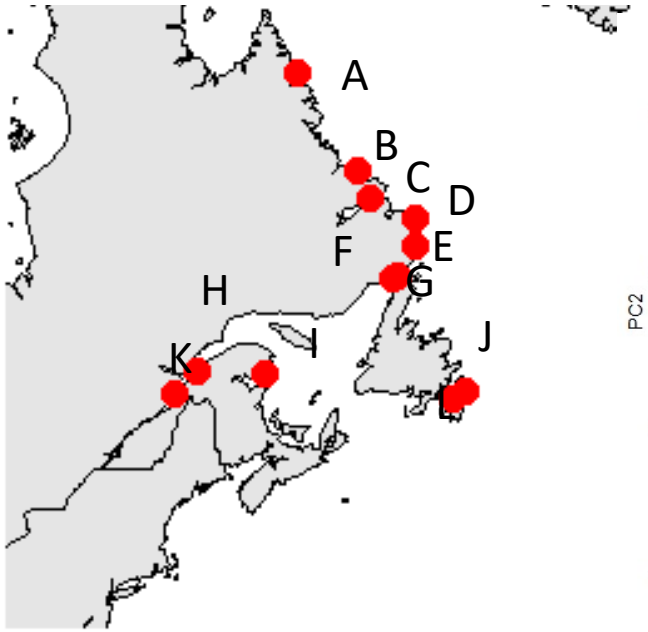
PCA

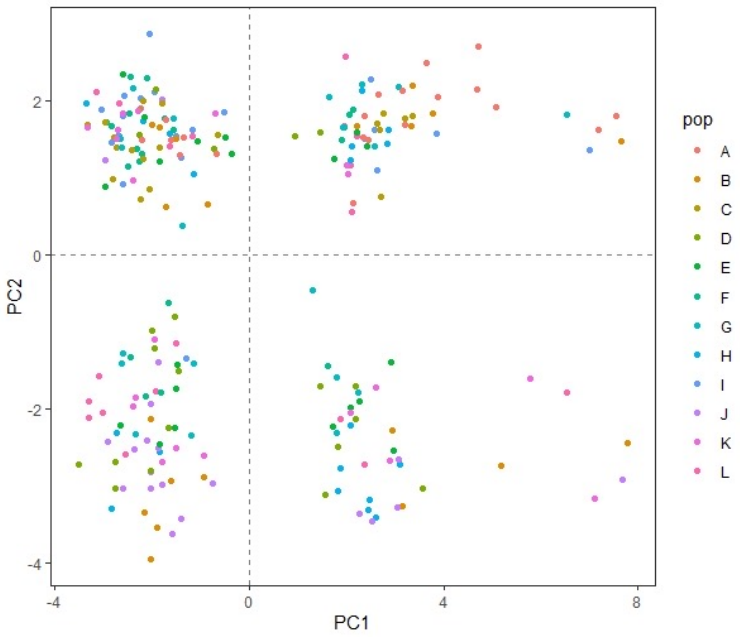


Canada (12 populations)



PCA

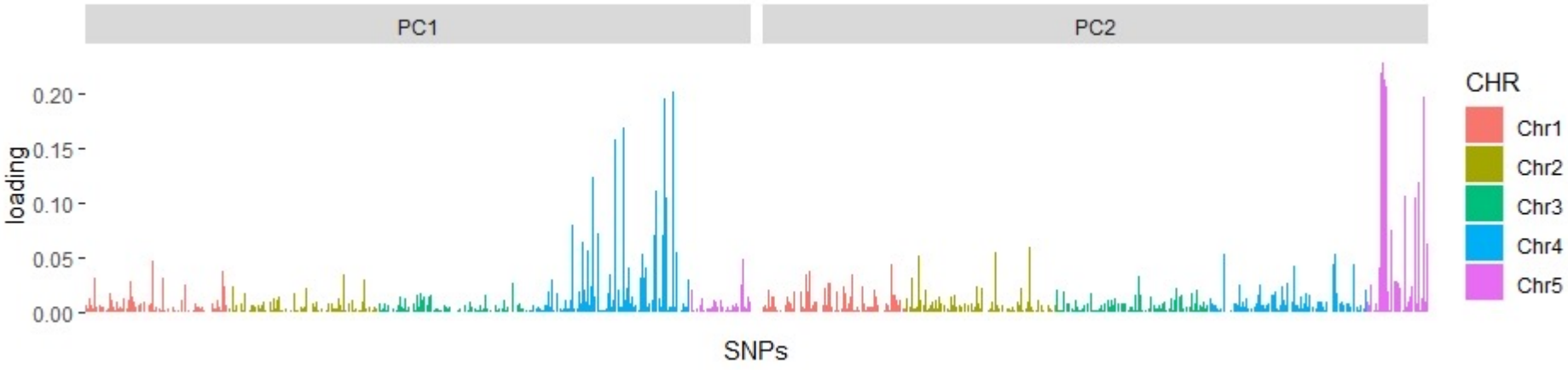




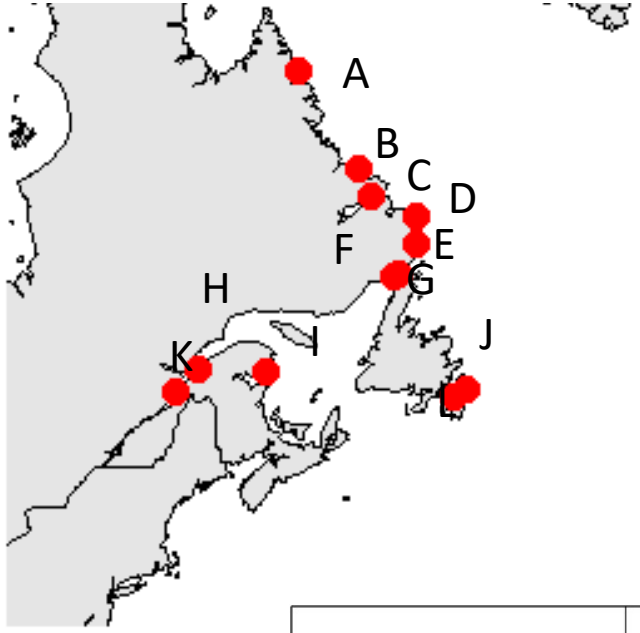
Canada (12 populations)

Rearrangement on Chr 4

Sex-determining locus on Chr 5

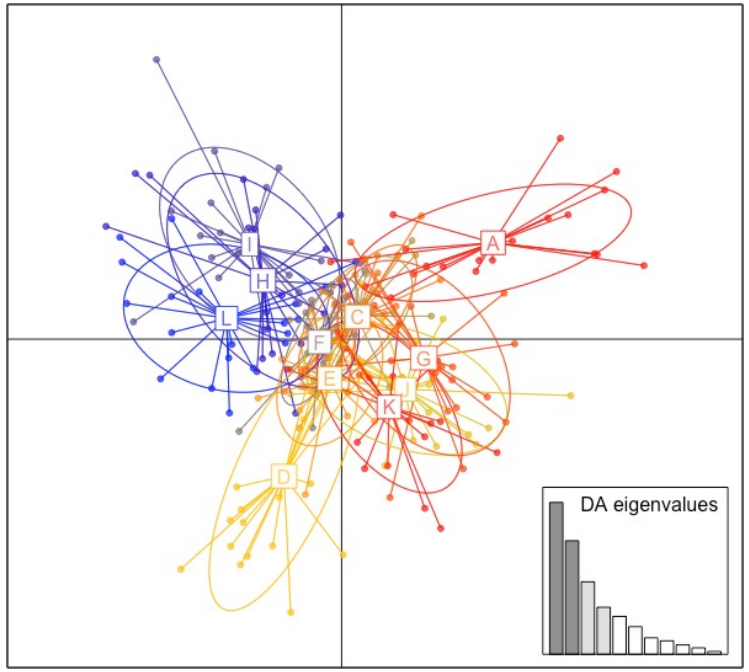


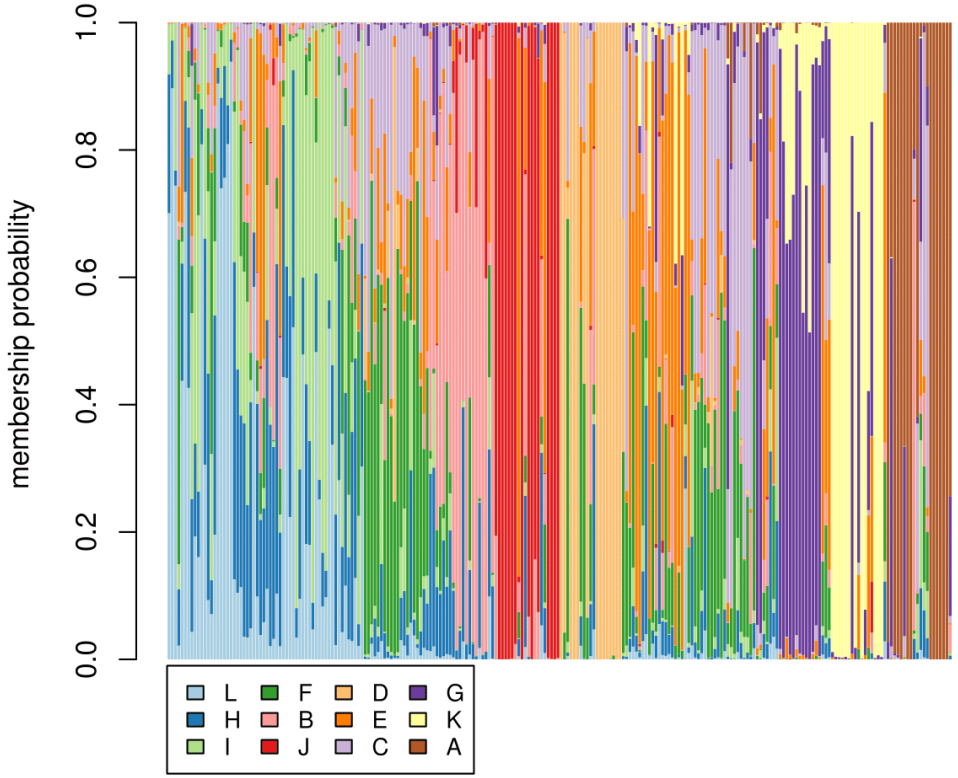
DAPC



Do we observe genetic structure ?

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





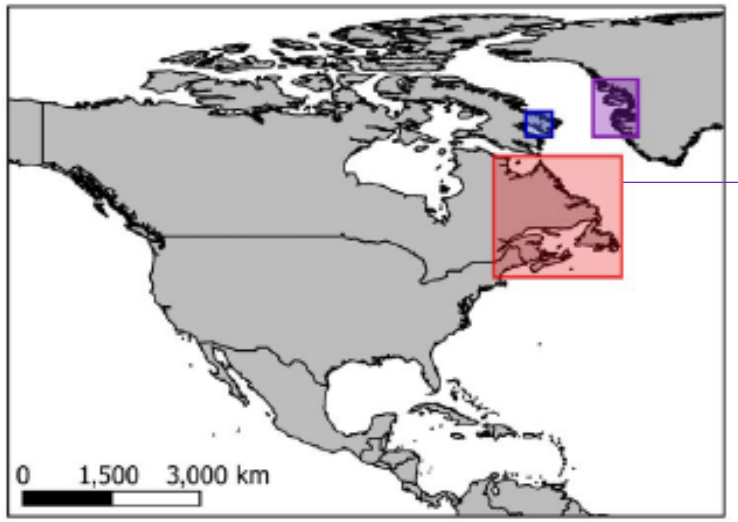
Day3 – Outlier detection and Environmental associations

Disentangle population structure & putative signature of adaptation

3-1 F_{ST} statistics (we did this yesterday, today we generated a LD-pruned VCF file)

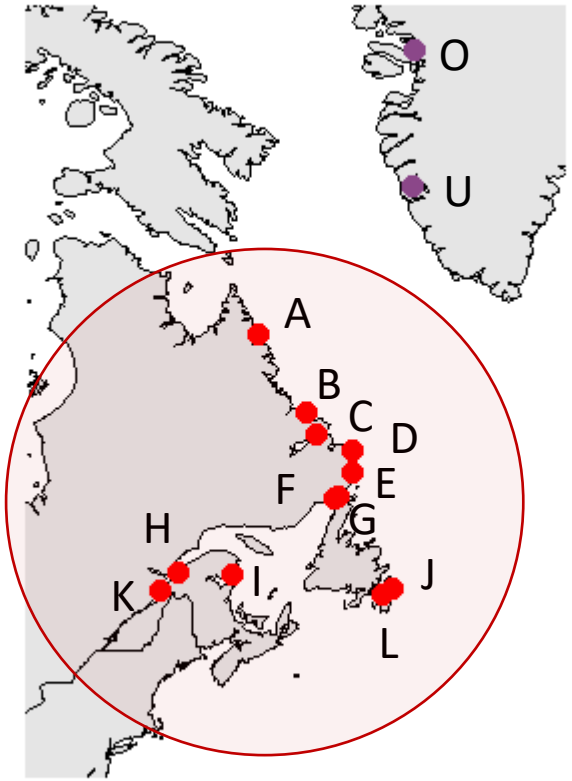
3-2 Outliers of differentiation

3-3 Genotype-environment associations

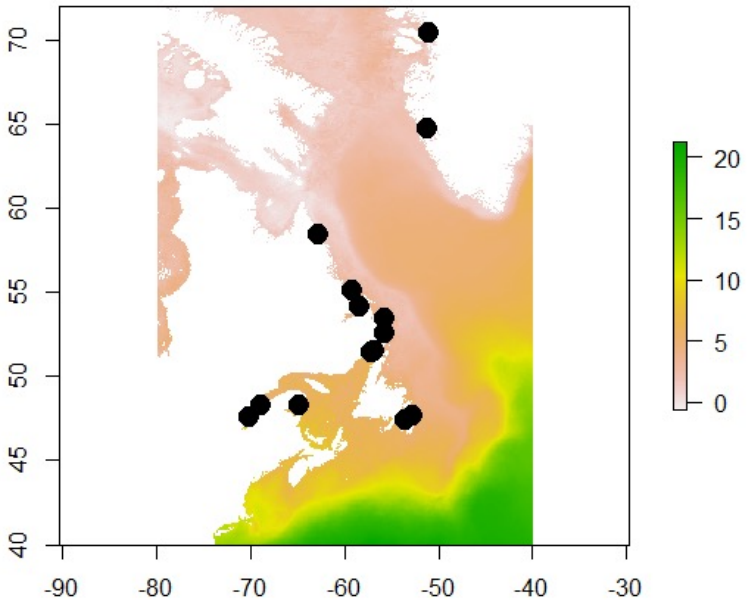


North American lineage

12 populations
N= 240 (20/pop)



Sea temperature (from MARSPEC)



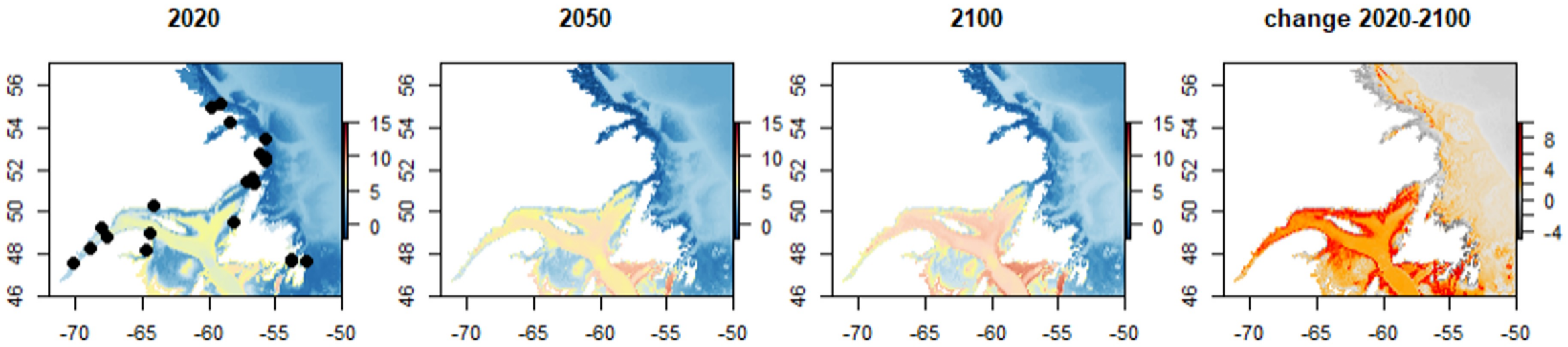
Climatic variables

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)

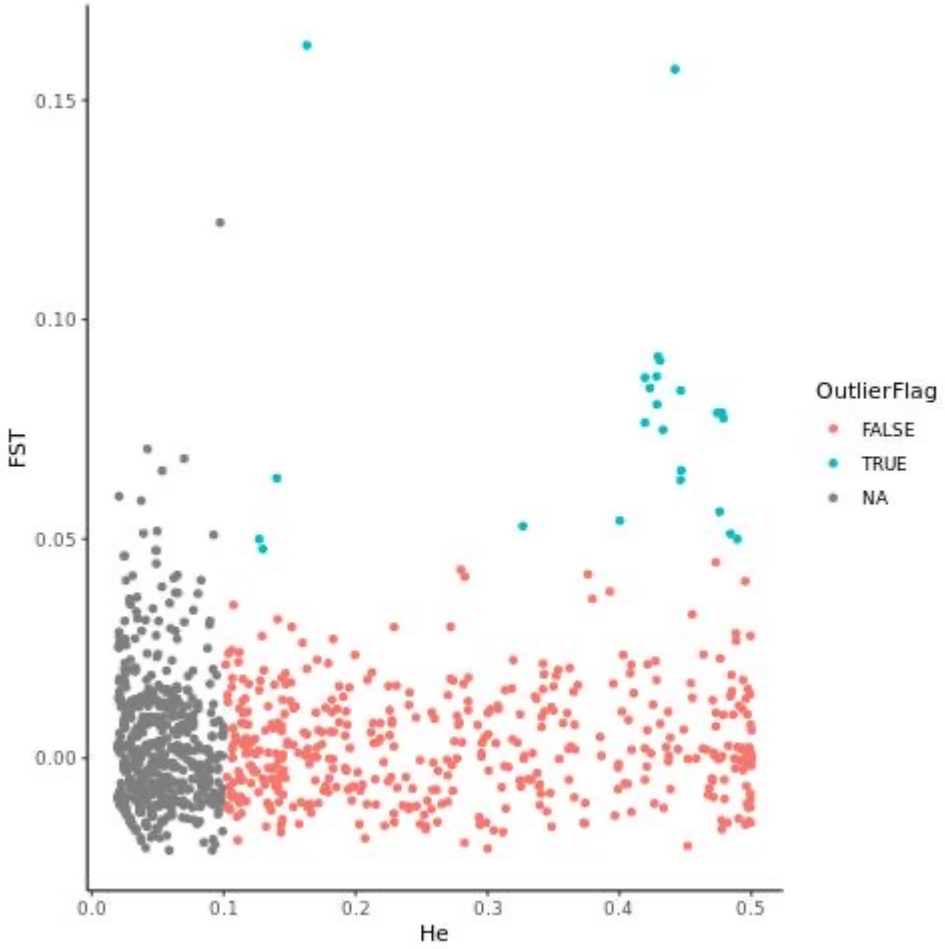


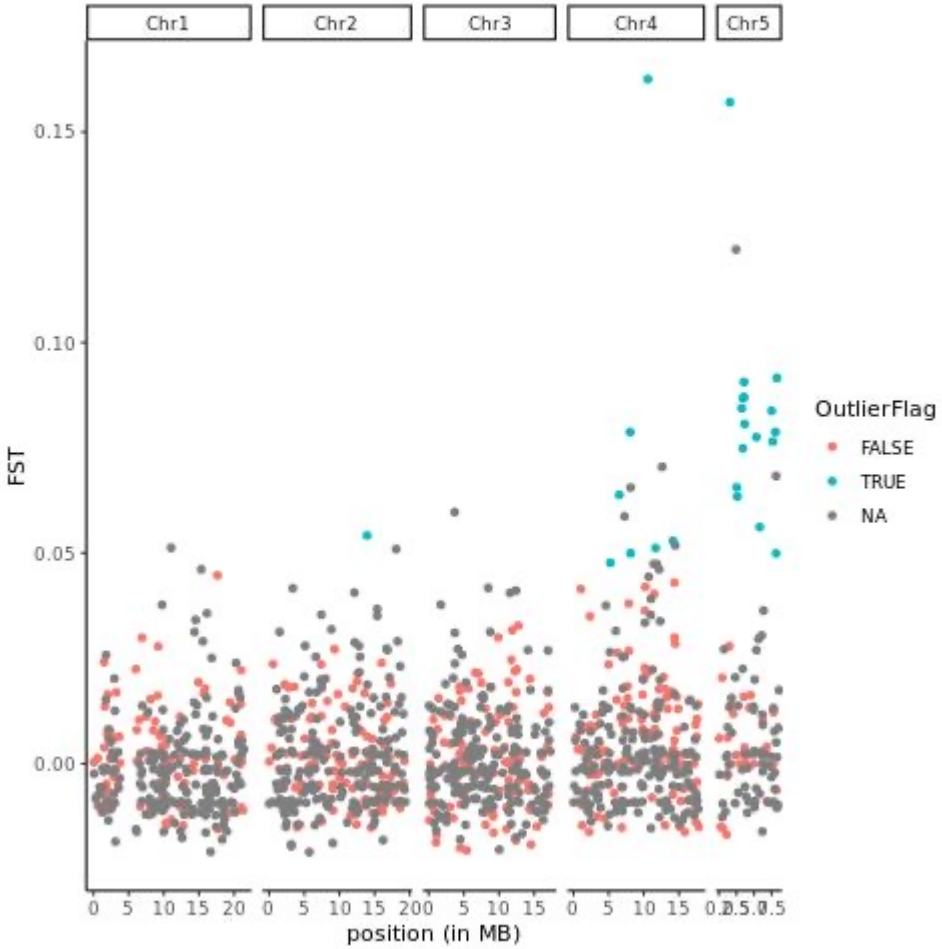
3-1 Create a subset of LD-pruned SNPs (with plink)

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

3-2 Outlier detection (with OutFlank)

Based on F_{st} outliers across all pairs of populations



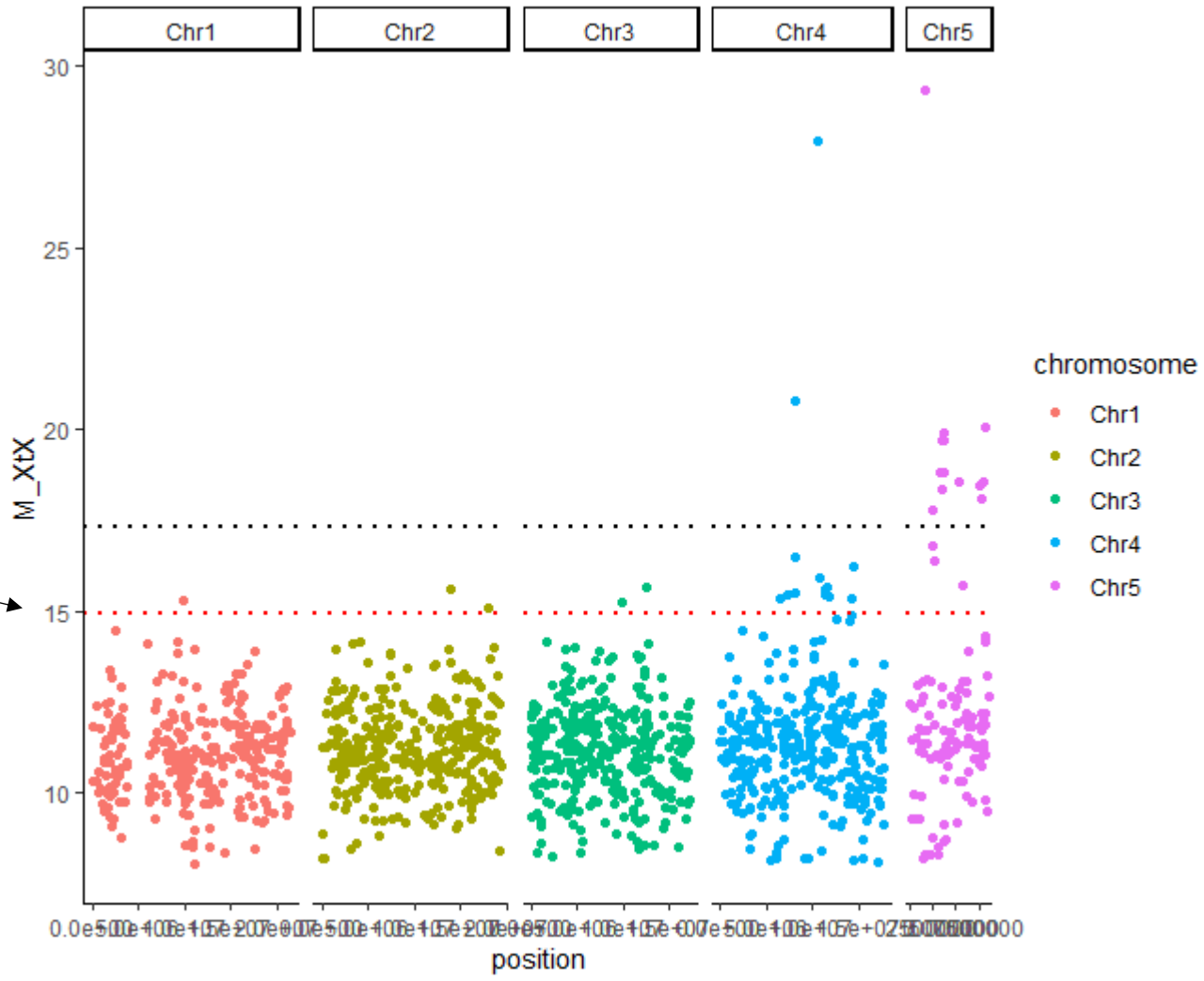


3-2 Outlier detection (with Baypass)

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

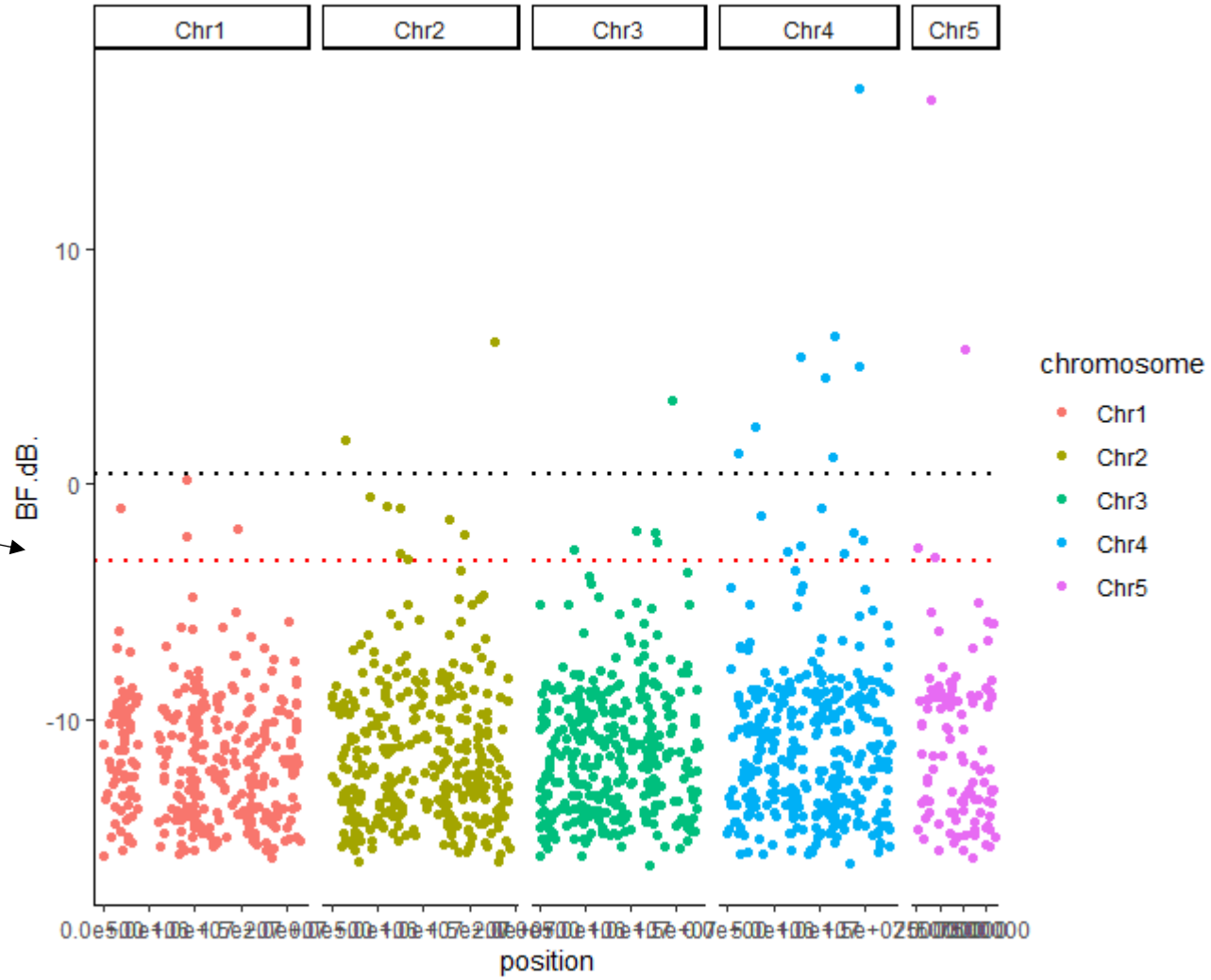


3-3 Environmental associations (Baypass)

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

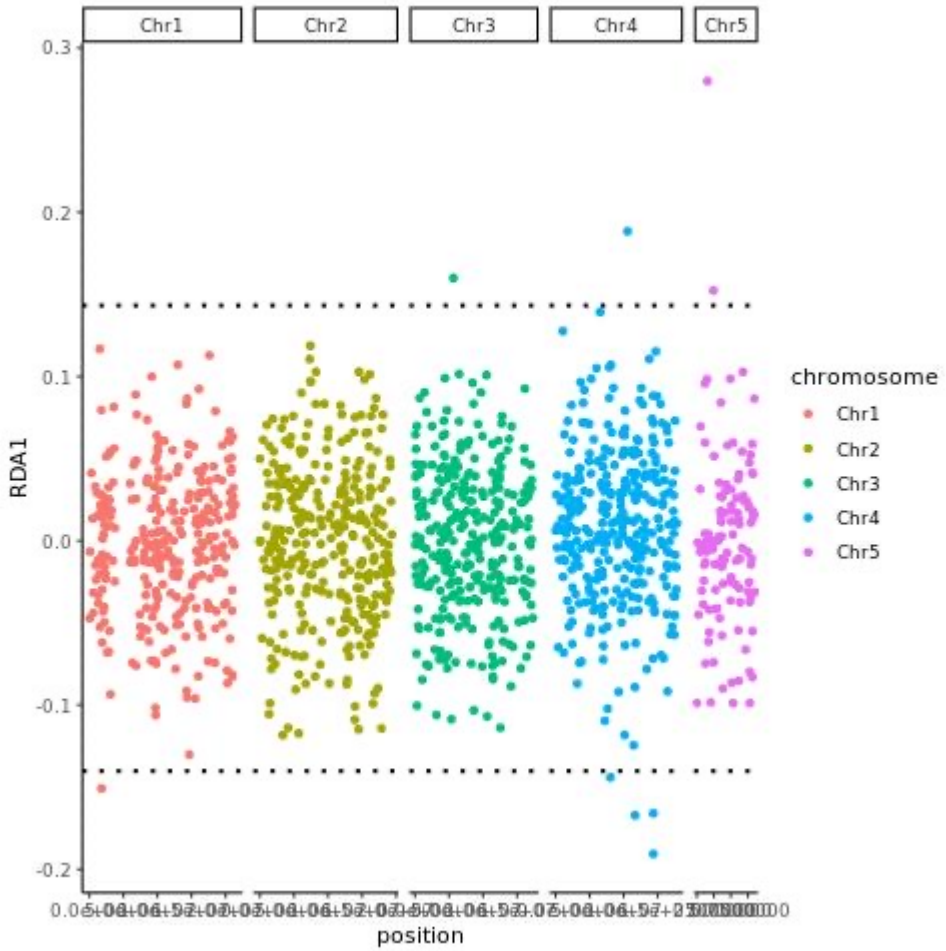


Simply add a co-variable matrix describing environmental variations between pop

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

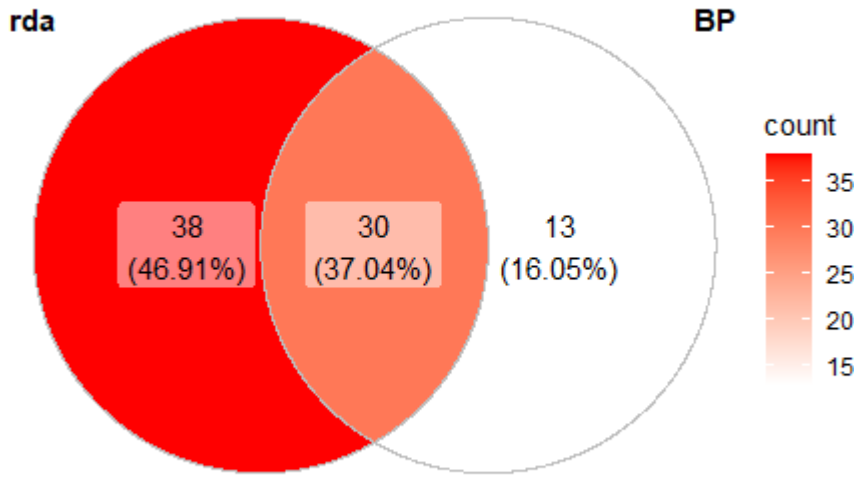
What we did

- 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 (n threads) 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
- examine overlap



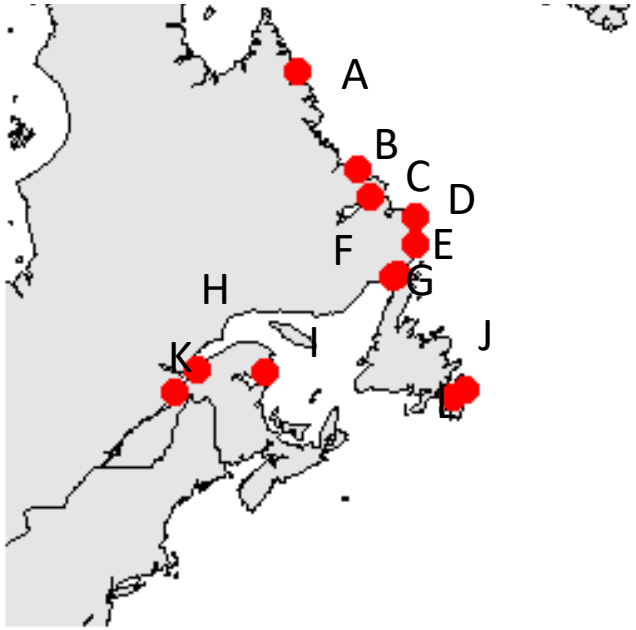
Day 4: Detecting structural variants

Detection of haplotype blocks (putative inversions, young sex chromosomes, etc.)

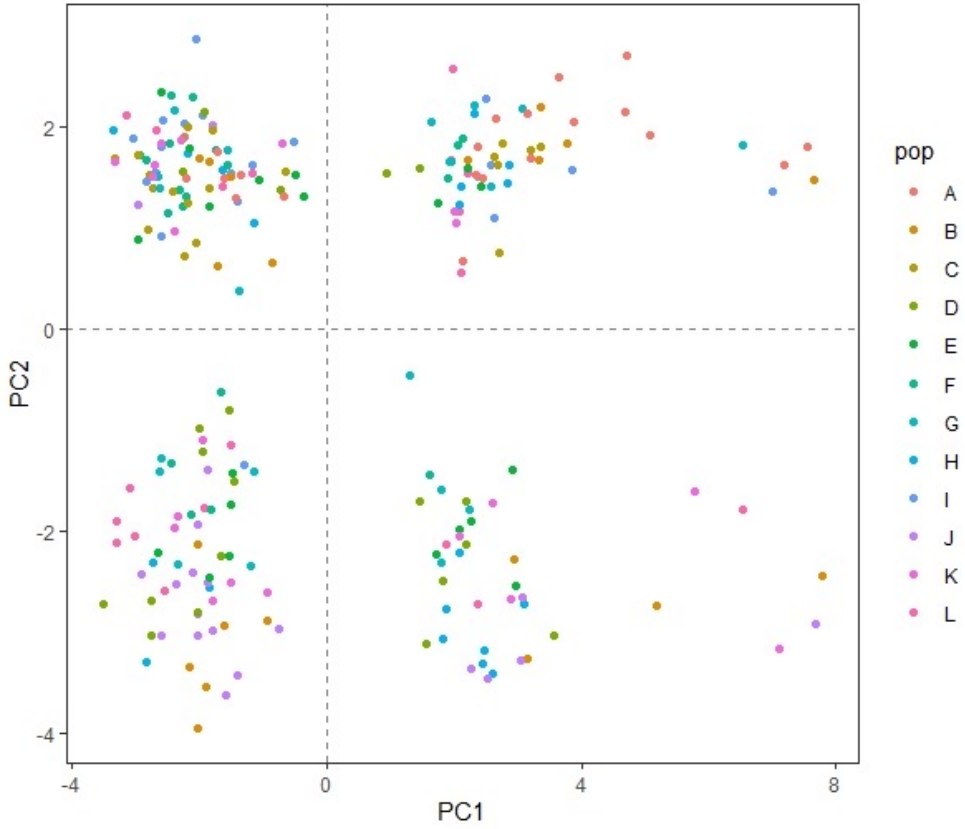
1. Detection with local PCA
2. Exploration of the haploblocks (genotype, LD, F_{st}, H obs)

Why?

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

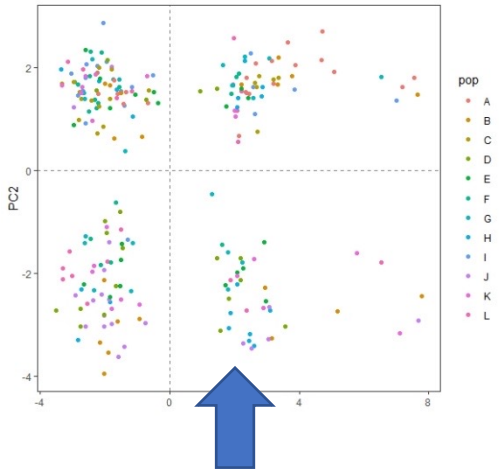


Canada (12 pop)

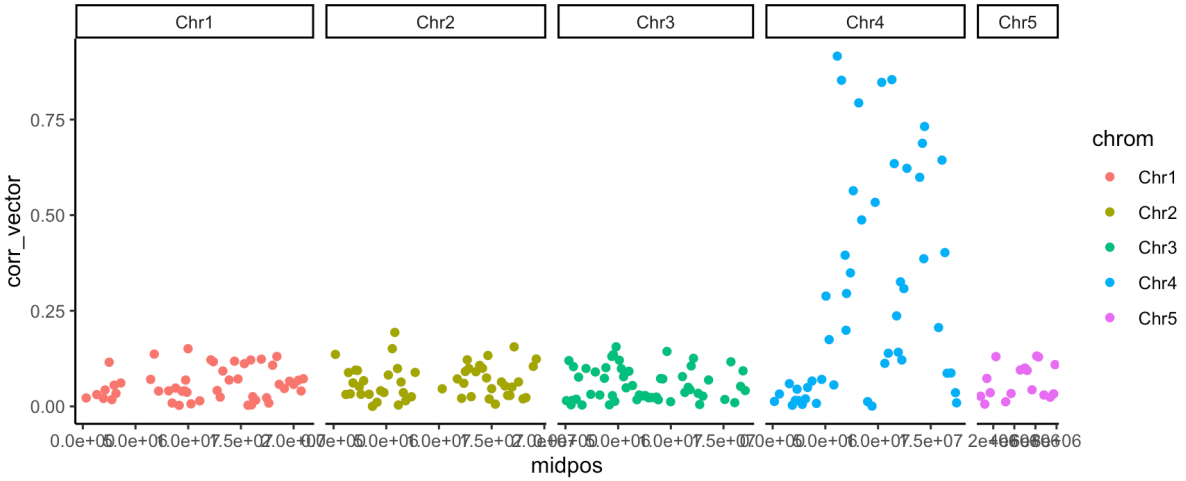


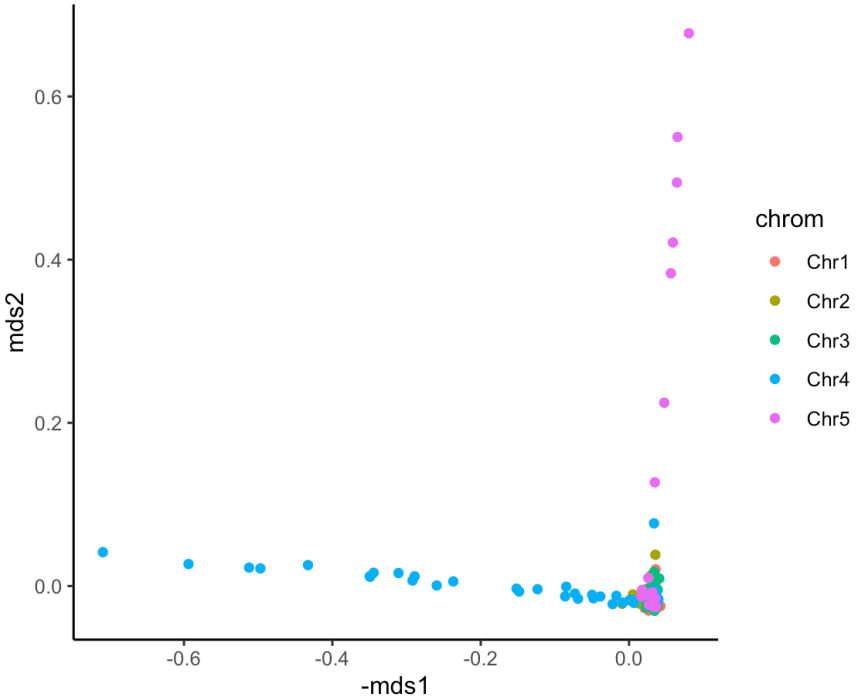
⇒ More generally, structural rearrangements and sex-linked regions may bias populations structure inference when left unknown (particularly in species with high gene flow)

4-1 Detection of SVs with local PCA



MDS looking at similar windows accross the genome

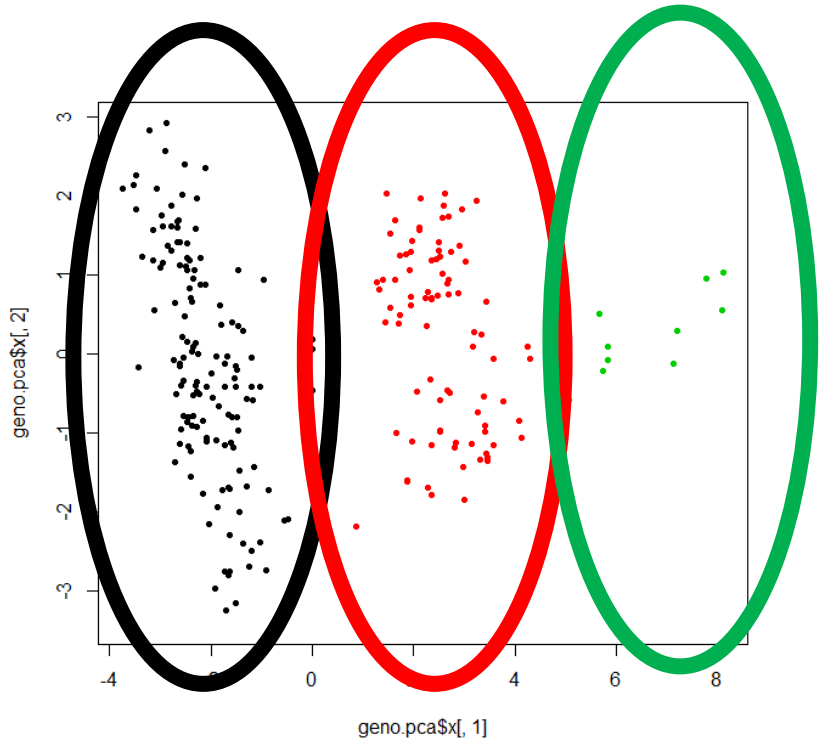




Correlation between local PCA and global PCA

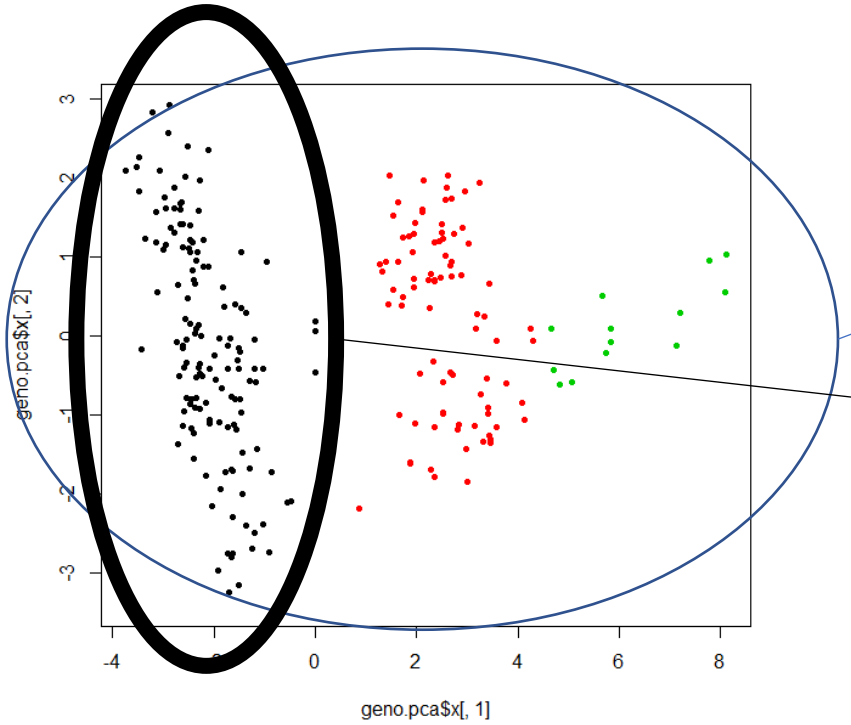
4-1 Exploration of the haploblocks

- Genotype



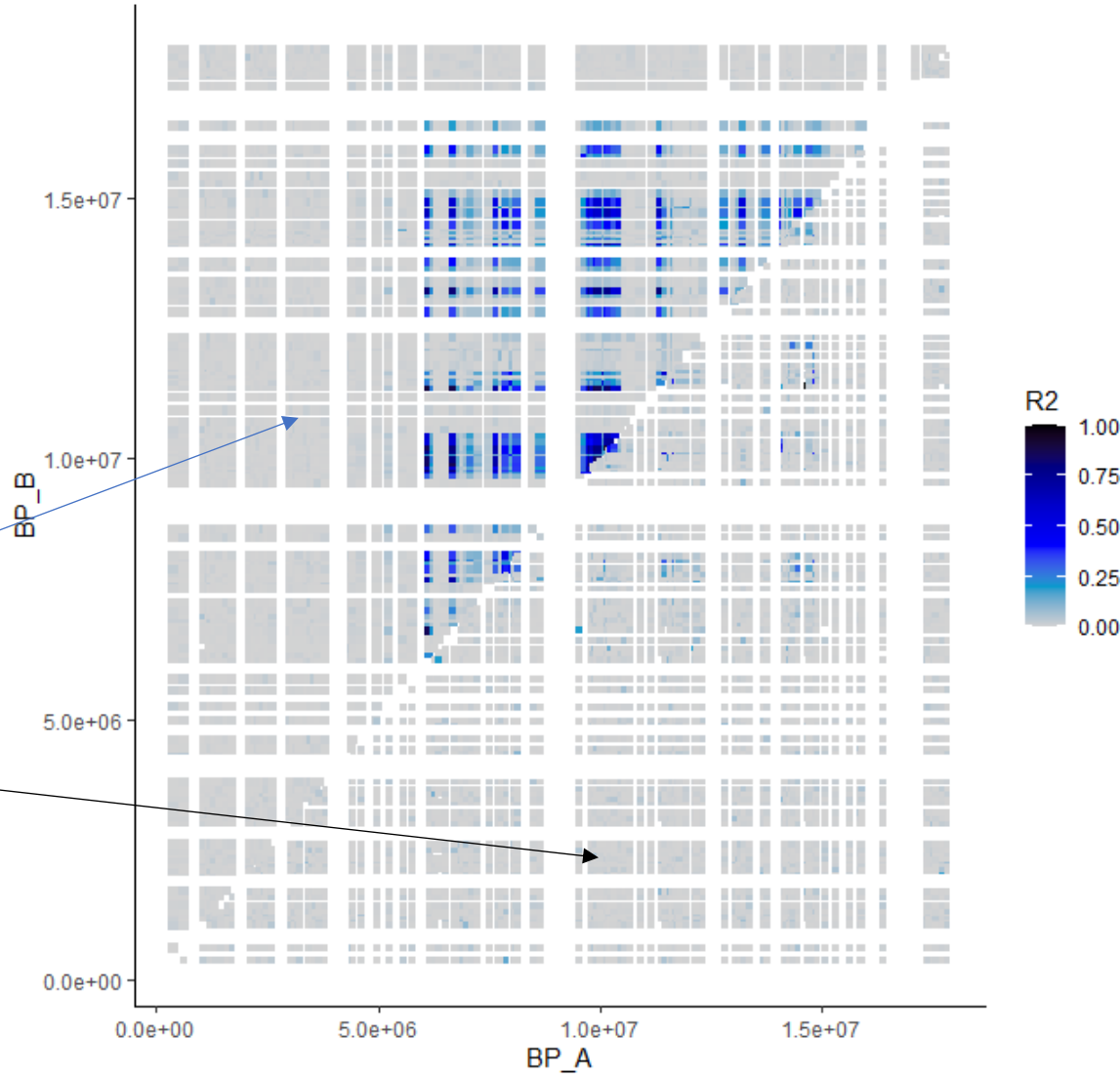
4-1 Exploration of the haploblocks

- Genotype
- Linkage disequilibrium

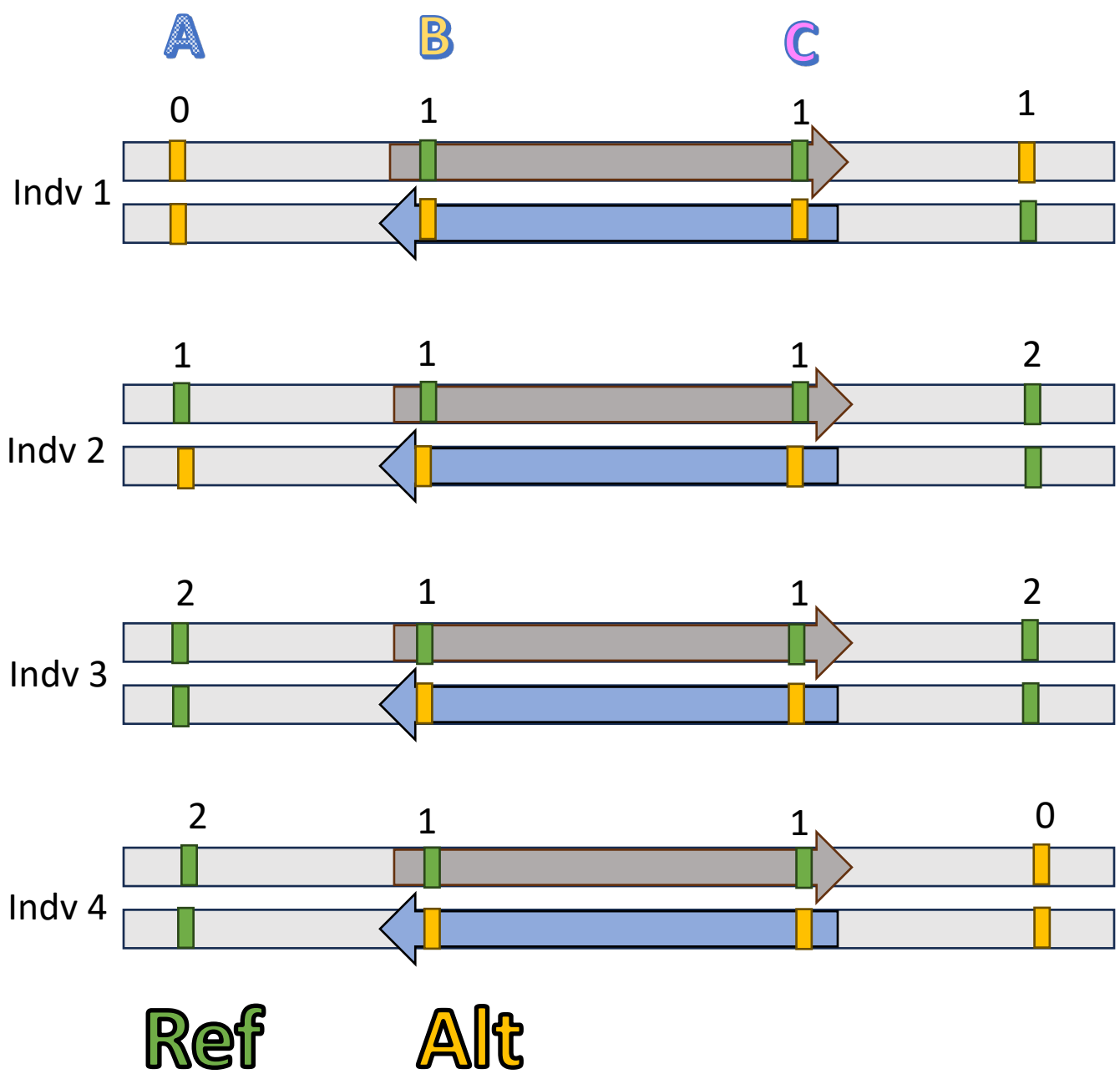


All samples

One haplogroup AA



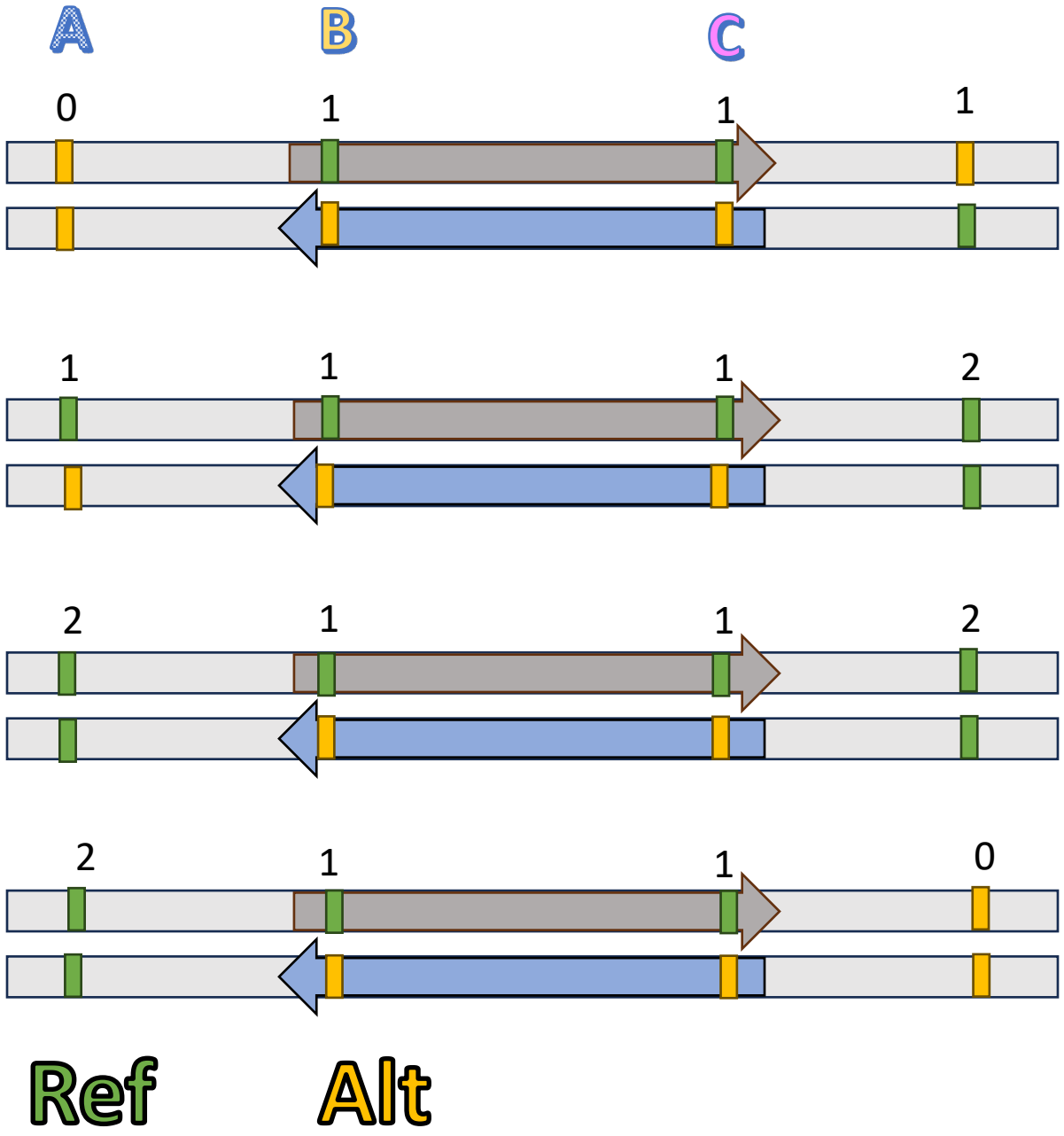
Heterozygote for inversion



$$D = p_{A_1B_1} - p_{A_1}p_{B_1} \qquad (1)$$

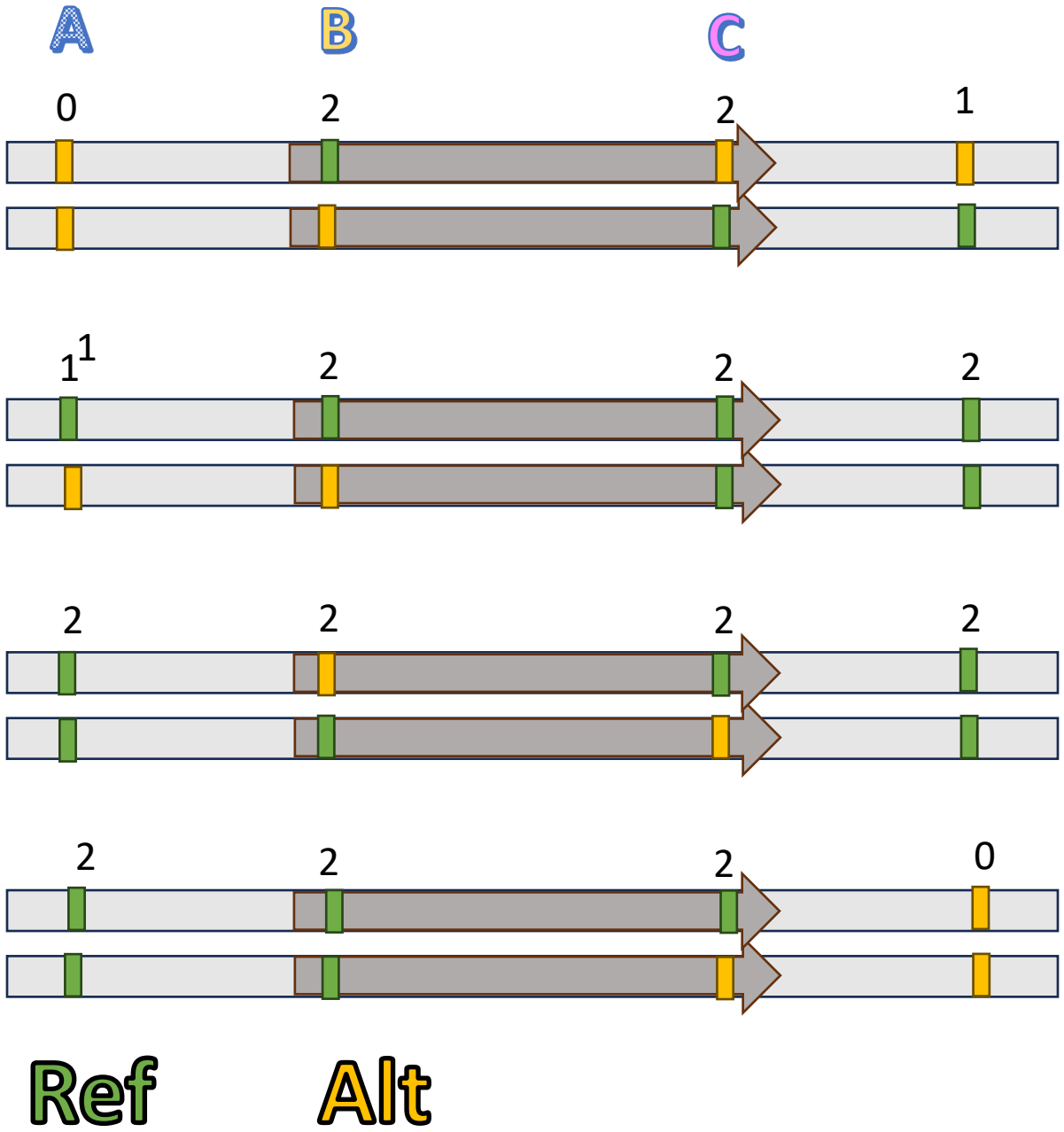
$$r^2 = \frac{D^2}{p_{A_1}(1-p_{A_1})p_{B_1}(1-p_{B_1})} \qquad (2)$$

pA_R Frequency of allele R in locus A
pB_R Frequency of allele R in locus B
pA_RB_R Frequency of haplotype A_RB_R



	Locus A	Locus B
Freq of allele R	0.625	0.5
Freq Haplotype $A_R B_R$	0.375	
D	0.0625	
R2	0.06666667	

	Locus B	Locus C
Freq of allele R	0.5	0.5
Freq Haplotype $B_R C_R$	0.5	
D	0.25	
R2	1	



	Locus A	Locus B
Freq of allele R	0.625	0.5
Freq Haplotype $A_R B_R$	0.375	
D	0.0625	
R2	0.06666667	

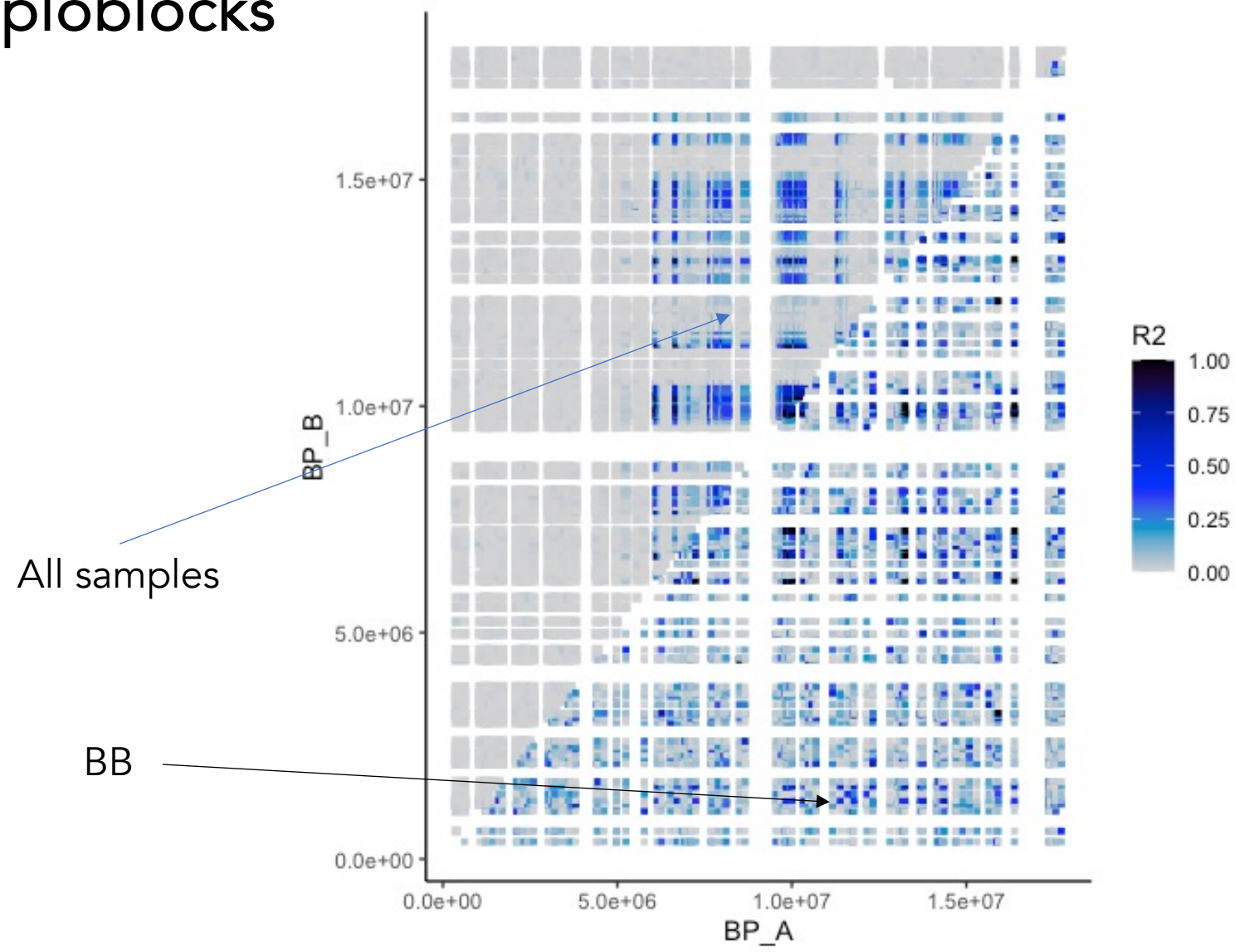
	Locus B	Locus C
Freq of allele R	0.625	0.625
Freq Haplotype $B_R C_R$	0.25	
D	-0.14	
R2	0.36	

4-1 Exploration of the haploblocks

- Genotype
- Linkage disequilibrium

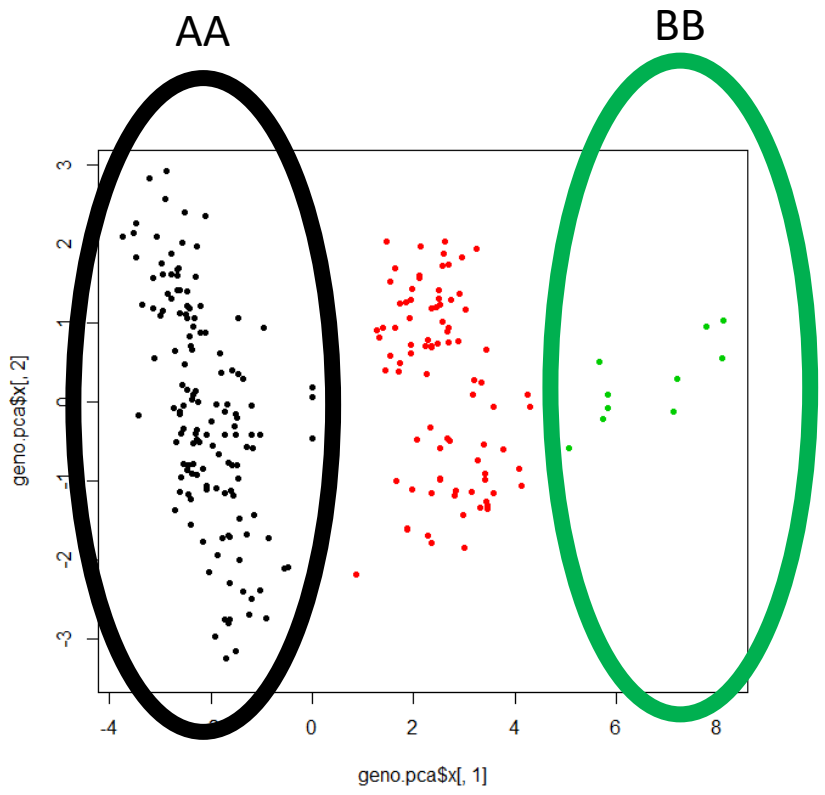
You don't observe the same in the BB group.

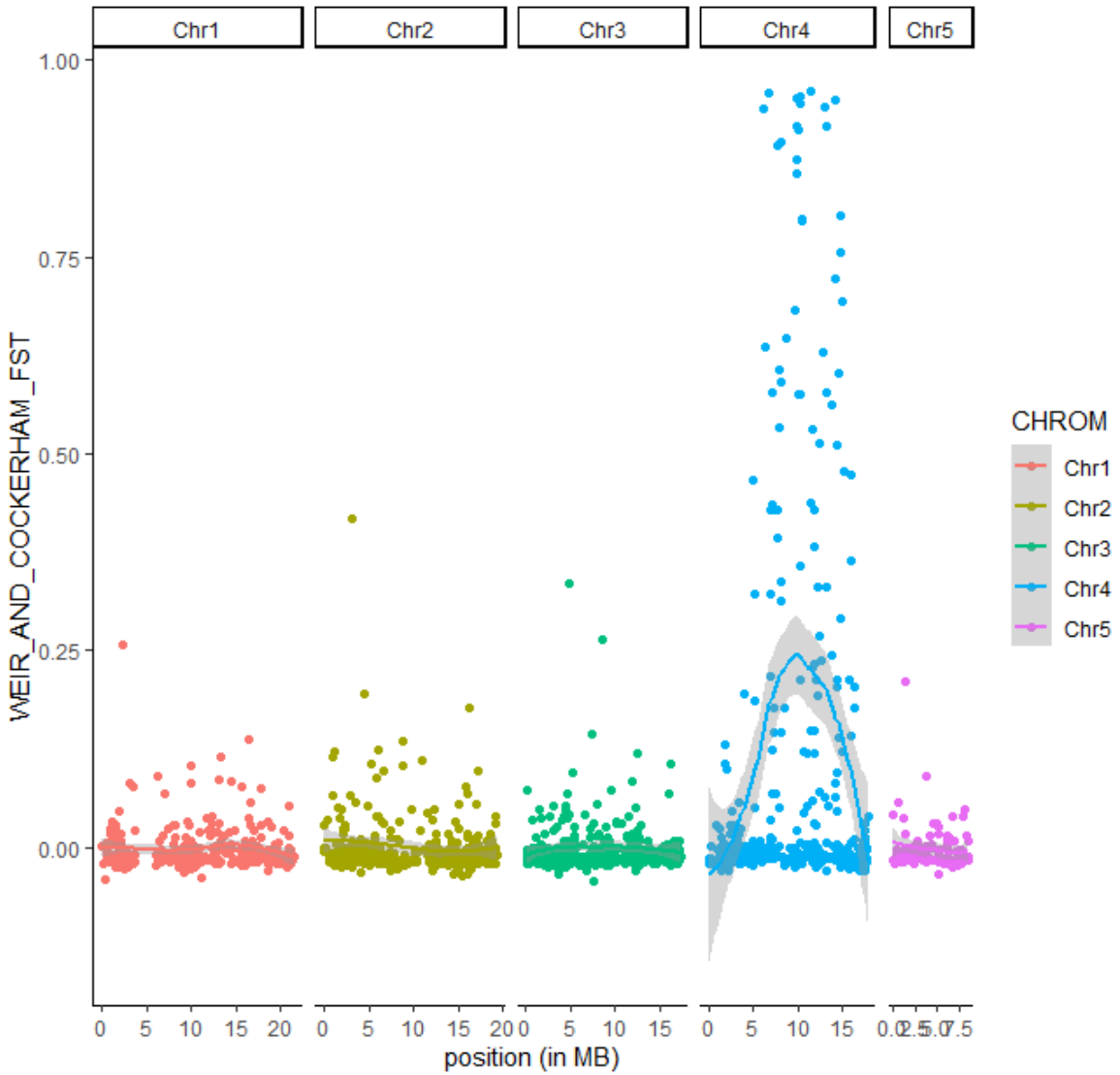
It seems that the BB group has higher linkage overall (possibly due to low sample size).



4-1 Exploration of the haploblocks

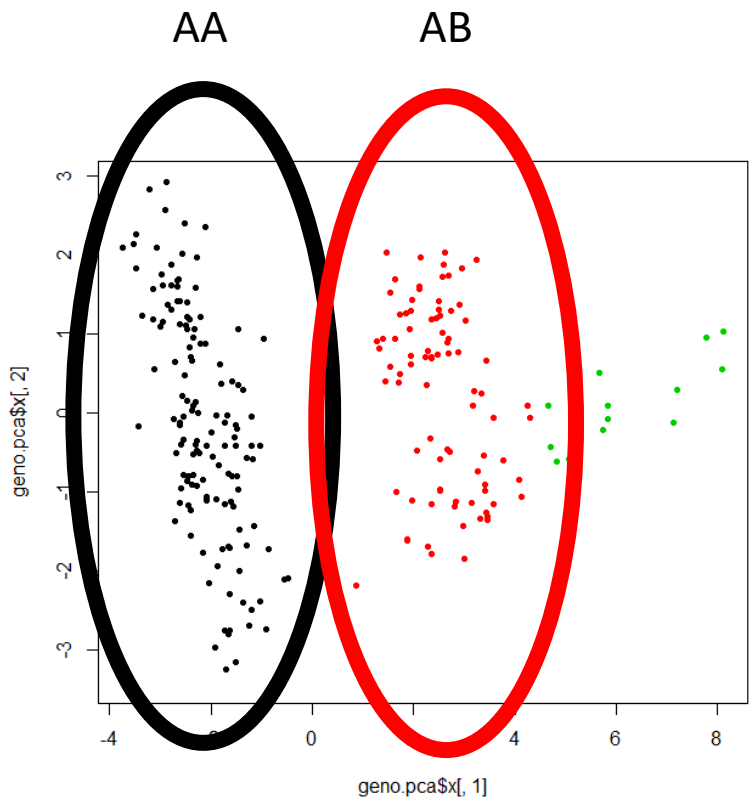
- Genotype
- Linkage disequilibrium
- FST between haplogroups

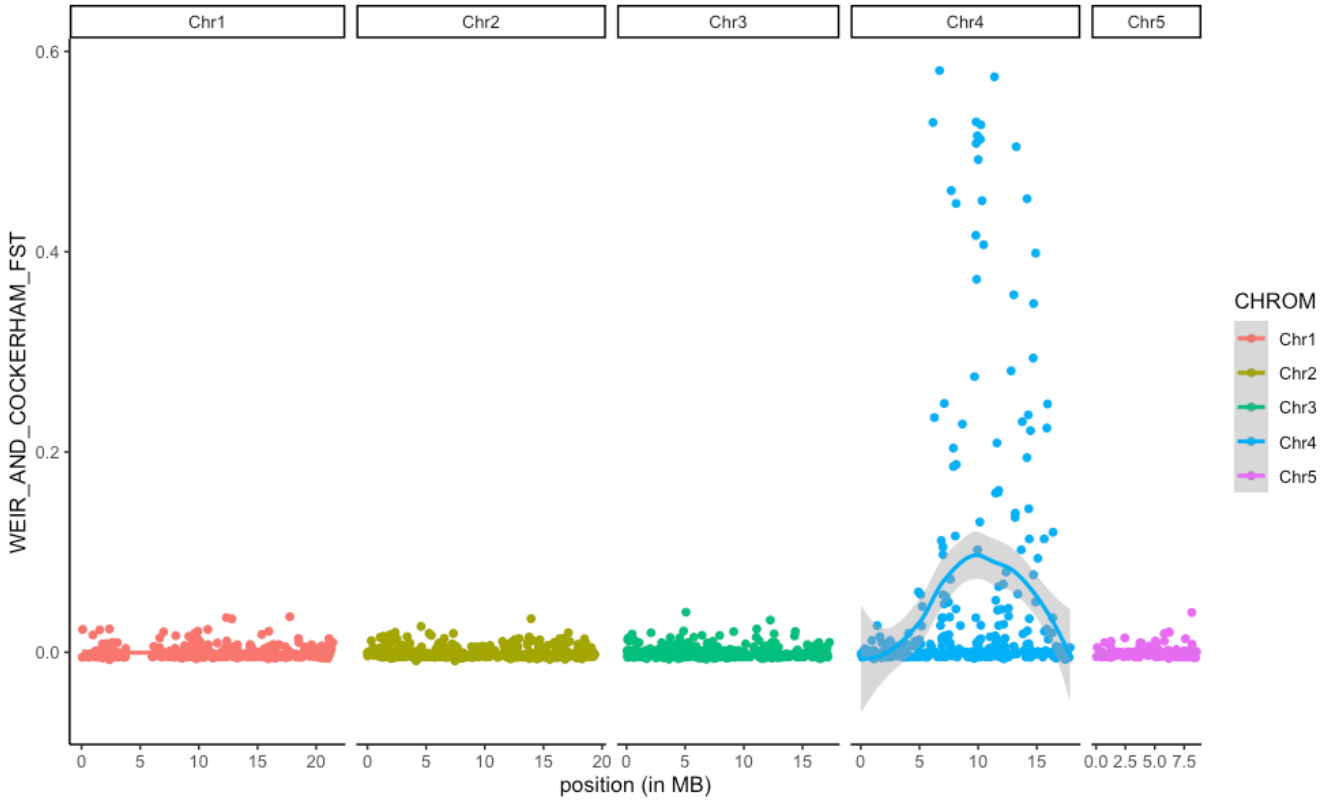




4-1 Exploration of the haploblocks

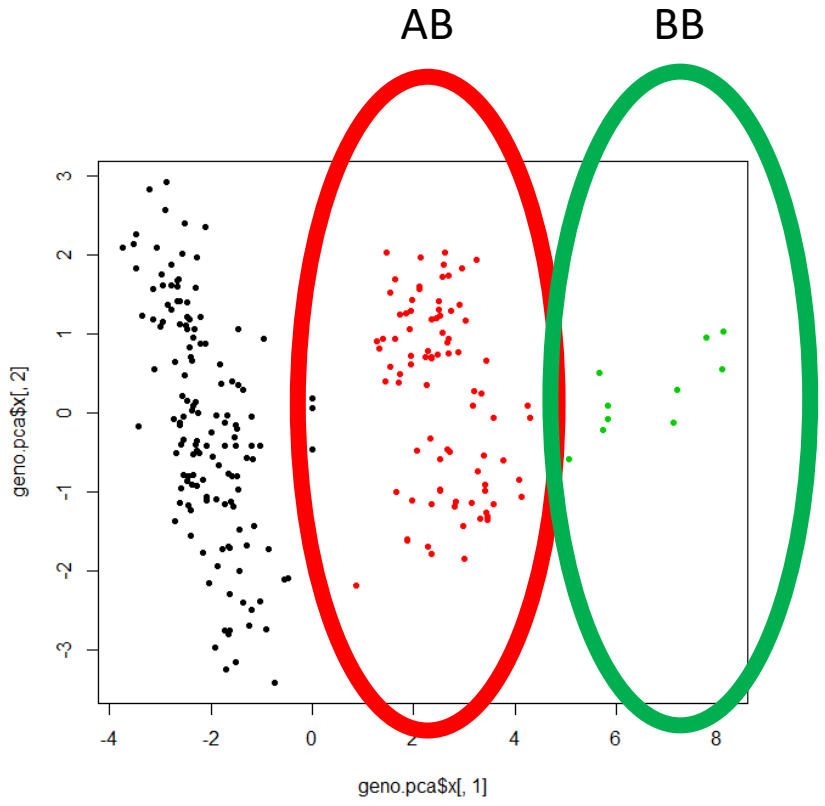
- Genotype
- Linkage disequilibrium
- FST between haplogroups



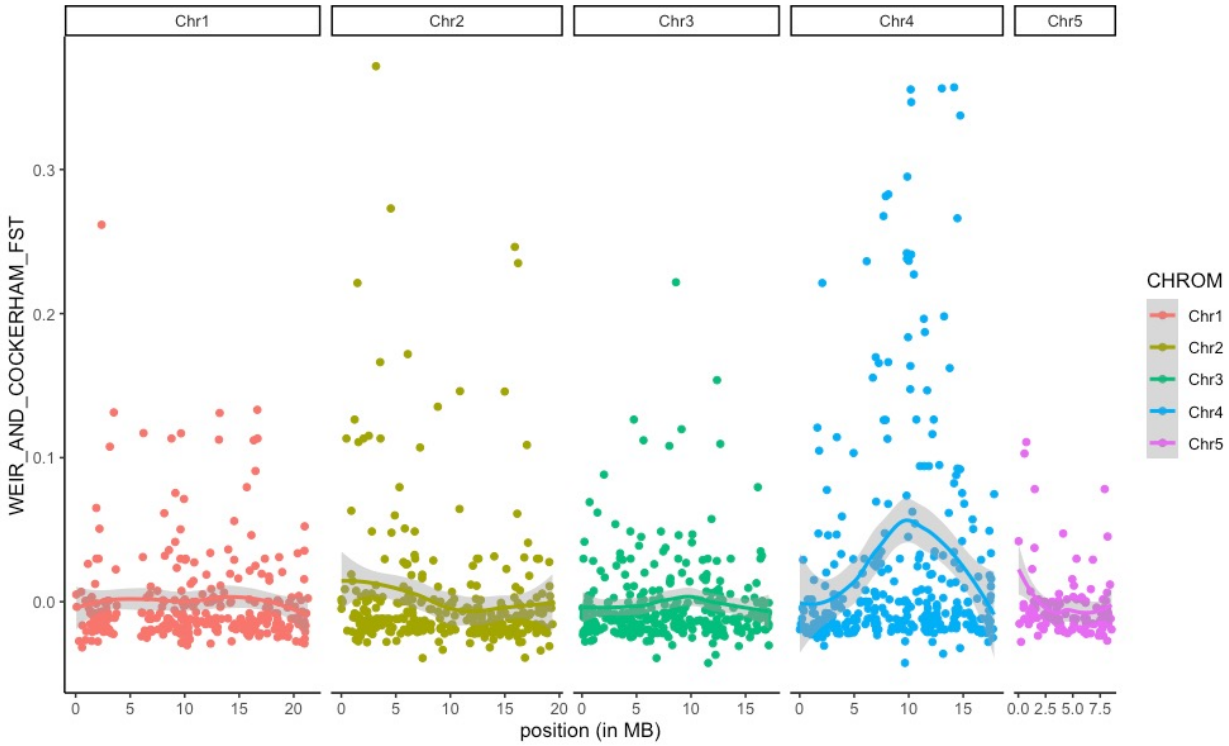


4-1 Exploration of the haploblocks

- Genotype
- Linkage disequilibrium
- FST between haplogroups

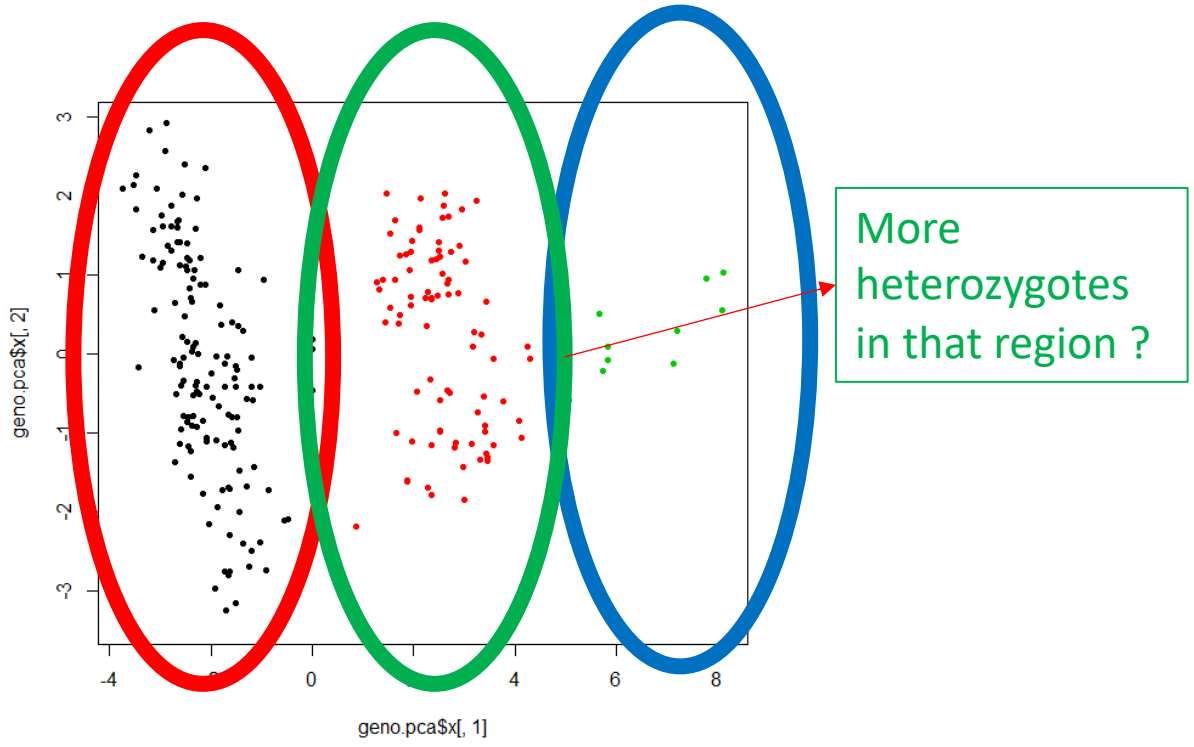


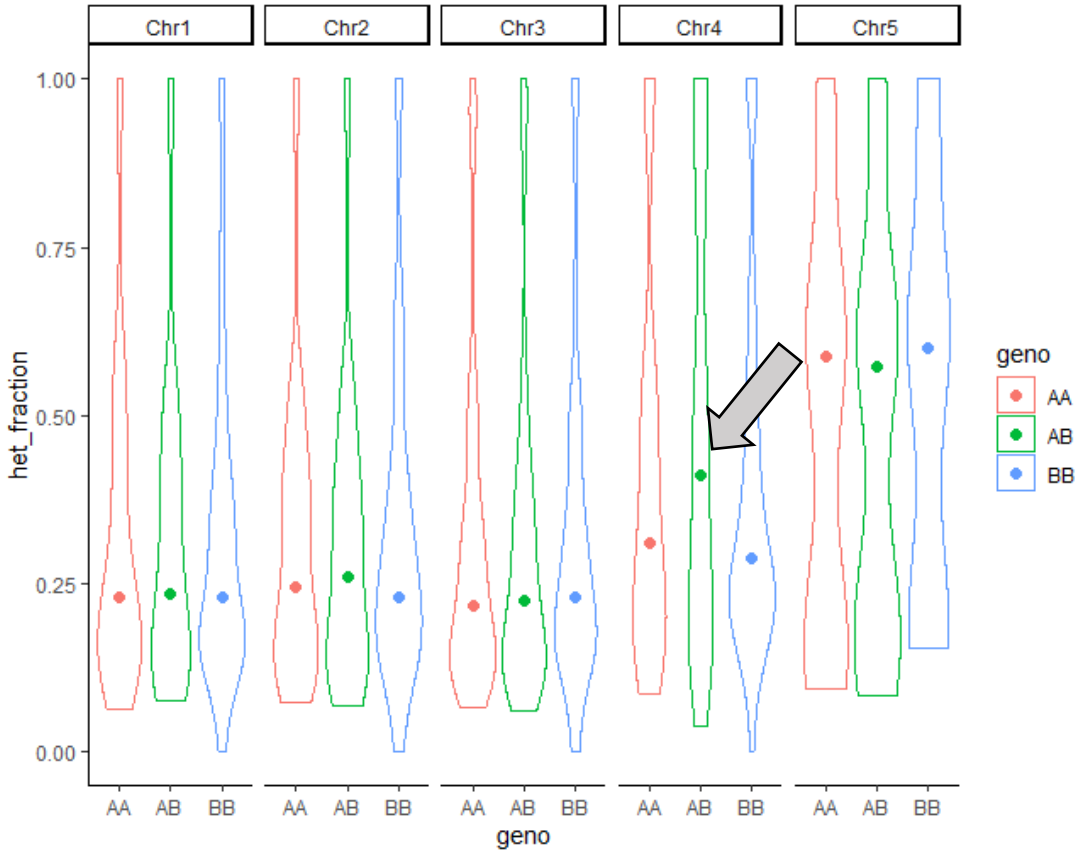
Higher variation in Fst since BB is small



4-1 Exploration of the haploblocks

- Genotype
- Linkage disequilibrium
- FST between haplogroups
- Observed fraction of heterozygotes





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)

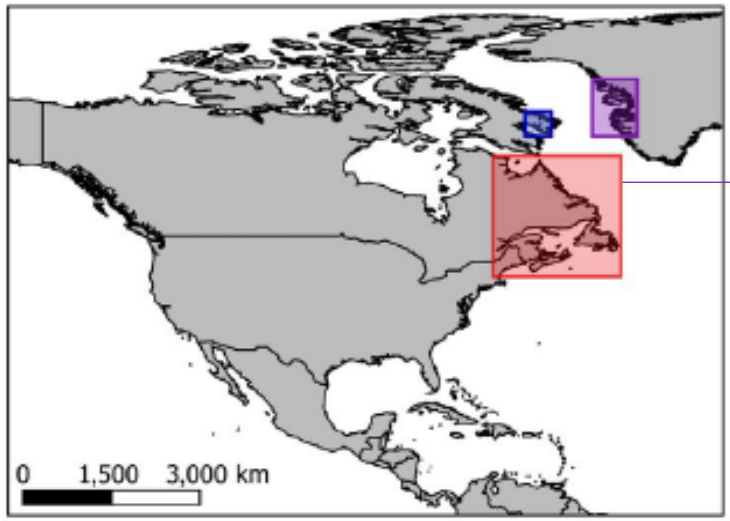
2: Whole-genome sequencing for SNPs and small/medium SVs

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

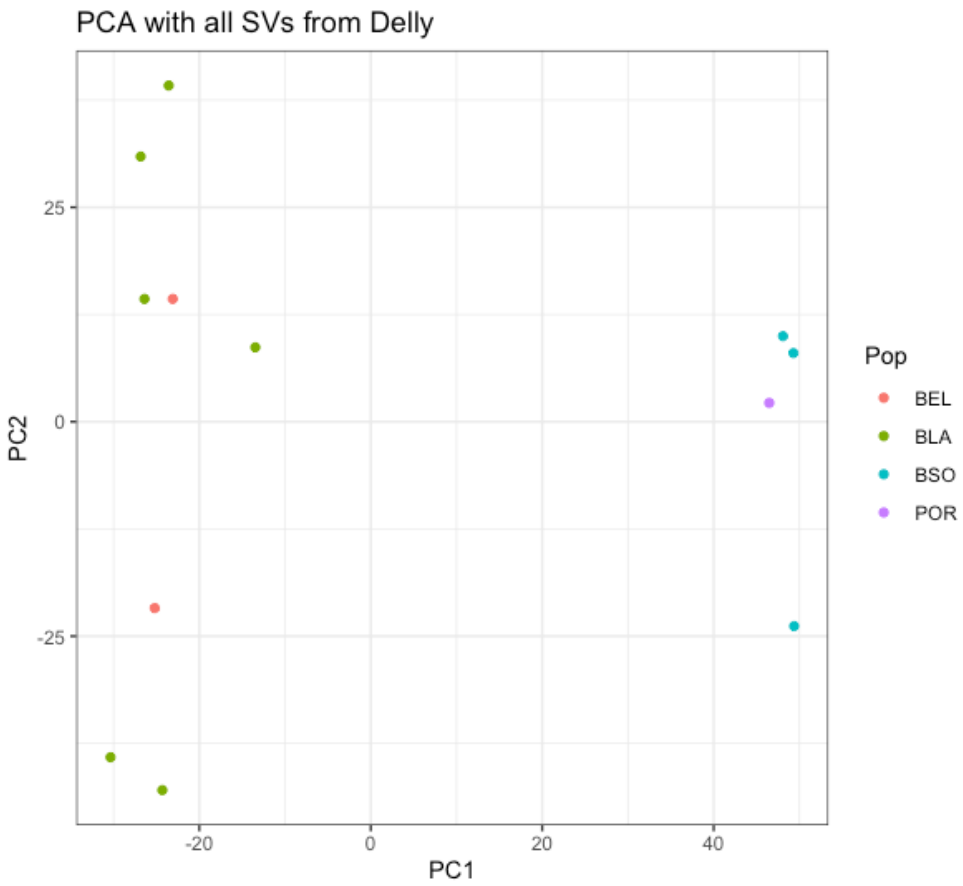


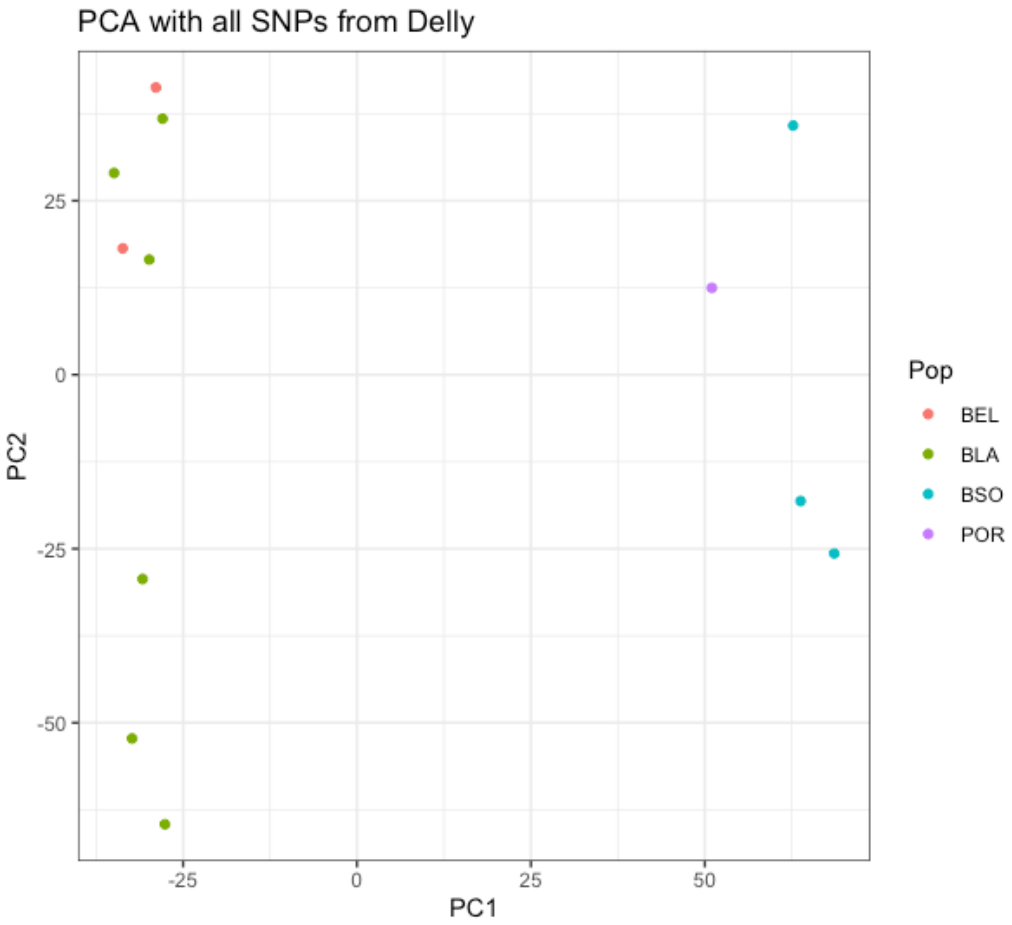
North American lineage

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!!


PCA with all SVs from Delly
PC2
PC1
25
0
-25
-20
0
20
40
Pop
BEL
BLA
BSO
POR


PCA with all SNPs from Delly
PC2
PC1
25
0
-25
-50
-25
0
25
50
Pop
BEL
BLA
BSO
POR

Tutorial day 5

Most methods that we saw during the week will provide

⇒ General knowledge about:

- ⇒ isolation-by-adaptation
- ⇒ genetic architecture of adaptation
- ⇒ genomic variance related to possible ecological variation, etc …

⇒ 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 gathered 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:

- 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

5-1 Annotate SNPs (with 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 (with 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.

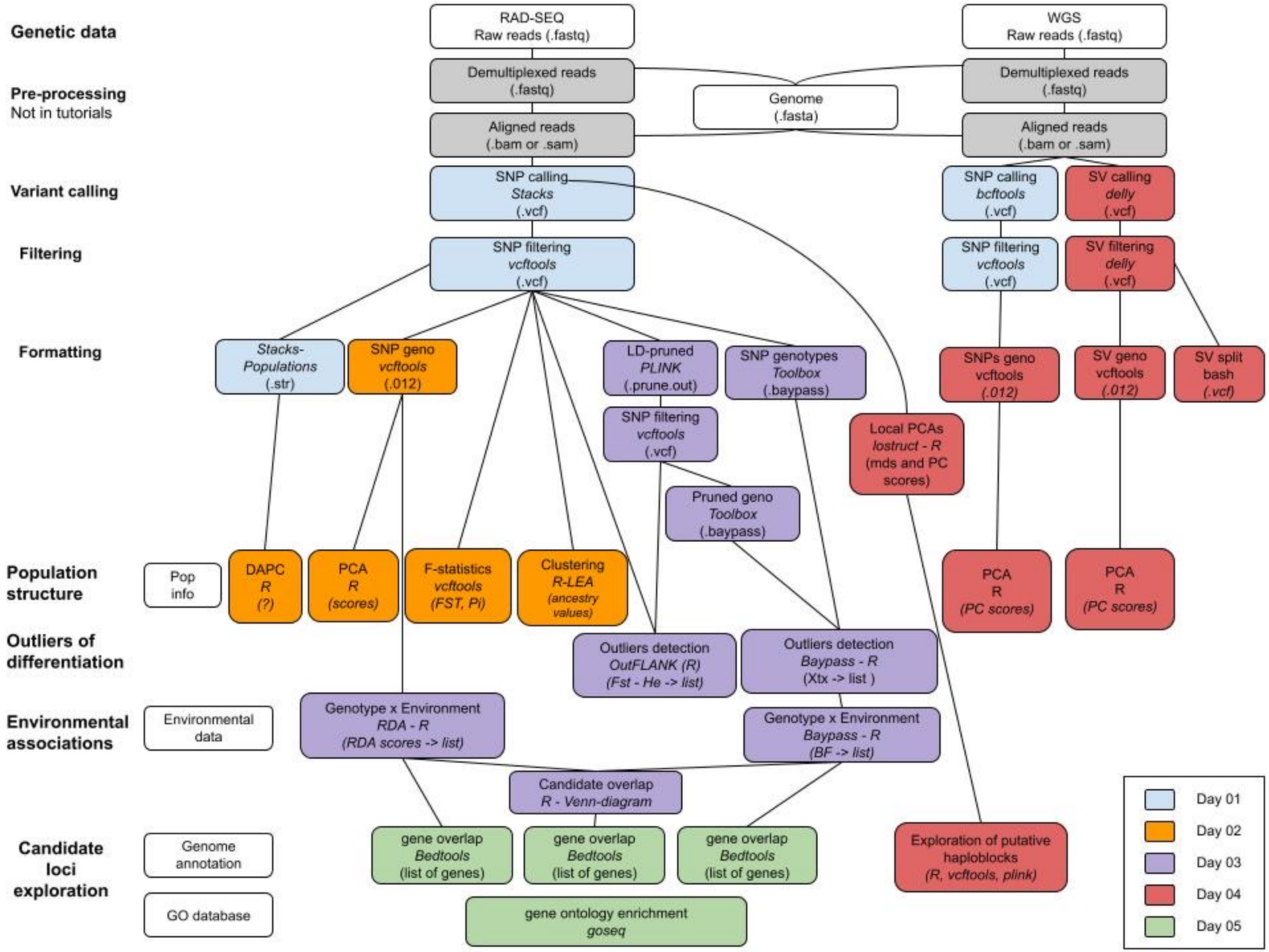
5-3 Gene ontology enrichment (with goseq library in R)

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 genes in the genome as some of them may simply not be covered).

	A	B	C	D	E	F	G	H	I	J
1		category	over_represented_pvalue	under_represented_pval	numDEInCat	numInCat	term	ontology	over_represented_padjust	
2	312	GO:0002084	0.000156082315798037	1	3	3	protein depalmitoylation	BP	0.285337979103912	
3	1464	GO:0008474	0.000156082315798037	1	3	3	palmitoyl-(protein) hydrolase activity	MF	0.285337979103912	
4	321	GO:0002116	0.000294654865200251	0.999994454136017	4	5	semaphorin receptor complex	CC	0.285337979103912	
5	2083	GO:0017154	0.000294654865200251	0.999994454136017	4	5	semaphorin receptor activity	MF	0.285337979103912	
6	5962	GO:1902287	0.000294654865200251	0.999994454136017	4	5	semaphorin-plexin signaling pathway involved in axon guidance	BP	0.285337979103912	
7	1216	GO:0007162	0.000296809418783432	0.999983768129723	5	9	negative regulation of cell adhesion	BP	0.285337979103912	
8	4373	GO:0050772	0.000310874062836947	0.999974869452678	6	12	positive regulation of axonogenesis	BP	0.285337979103912	
9	2441	GO:0030334	0.000486425168909856	0.999953366171224	6	14	regulation of cell migration	BP	0.352095652505383	
10	4744	GO:0060173	0.000493207917906374	0.999967242689428	5	10	limb development	BP	0.352095652505383	
11	2271	GO:0021915	0.000874038794384151	0.999926922864949	5	12	neural tube development	BP	0.561569925391817	
12	415	GO:0003184	0.000978441764940278	0.999961433037353	4	6	pulmonary valve morphogenesis	BP	0.571498939976481	
13	4251	GO:0048663	0.00172080373358469	0.99996543727975	3	4	neuron fate commitment	BP	0.914071591313777	
14	477	GO:0003677	0.00184948337542087	0.999358433585154	16	101	DNA binding	MF	0.914071591313777	
15	3795	GO:0044853	0.00209456701107637	0.999852868671463	4	9	plasma membrane raft	CC	0.961256646154693	
16	2417	GO:0030279	0.00342223535358641	0.999871336014125	3	5	negative regulation of ossification	BP	1	
17	1788	GO:0014807	0.00364924867417417	1	2	2	regulation of somitogenesis	BP	1	
18	4372	GO:0050771	0.00432286945672922	0.999790741338814	3	6	negative regulation of axonogenesis	BP	1	
19	4466	GO:0051124	0.0043715948142503	1	2	2	synaptic assembly at neuromuscular junction	BP	1	
20	5209	GO:0071340	0.0043715948142503	1	2	2	skeletal muscle acetylcholine-gated channel clustering	BP	1	
21	5535	GO:0097105	0.0043715948142503	1	2	2	presynaptic membrane assembly	BP	1	


Genetic data
RAD-SEQ
Raw reads (.fastq)
WGS
Raw reads (.fastq)
Pre-processing
Not in tutorials
Demultiplexed reads
(.fastq)
Genome
(.fasta)
Demultiplexed reads
(.fastq)
Aligned reads
(.bam or .sam)
Aligned reads
(.bam or .sam)
Variant calling
SNP calling
Stacks
(.vcf)
SNP calling
bcftools
(.vcf)
SV calling
delly
(.vcf)
Filtering
SNP filtering
vcftools
(.vcf)
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)
SNP filtering
vcftools
(.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)
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
gene overlap
Bedtools
(list of genes)
gene overlap
Bedtools
(list of genes)
gene overlap
Bedtools
(list of genes)
Exploration of putative haploblocks
(R, vcftools, plink)
GO database
gene ontology enrichment
goseq
Day 01
Day 02
Day 03
Day 04
Day 05

How to extract environmental data

- Check out the new tutorial available in the GitHub repo https://github.com/MafaldaSFerreira/physalia_adaptation_course-2024/blob/main/03_day3/tutorial_bioOracle_optional.R

- We use the R package sdmpredictors to connect to the environmental database (bio-Oracle in this case). To learn which databases can be accessed with this package, see https://cran.r-project.org/web/packages/sdmpredictors/vignettes/quickstart.html