Overview Day 4:

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

1 Detection with local PCA

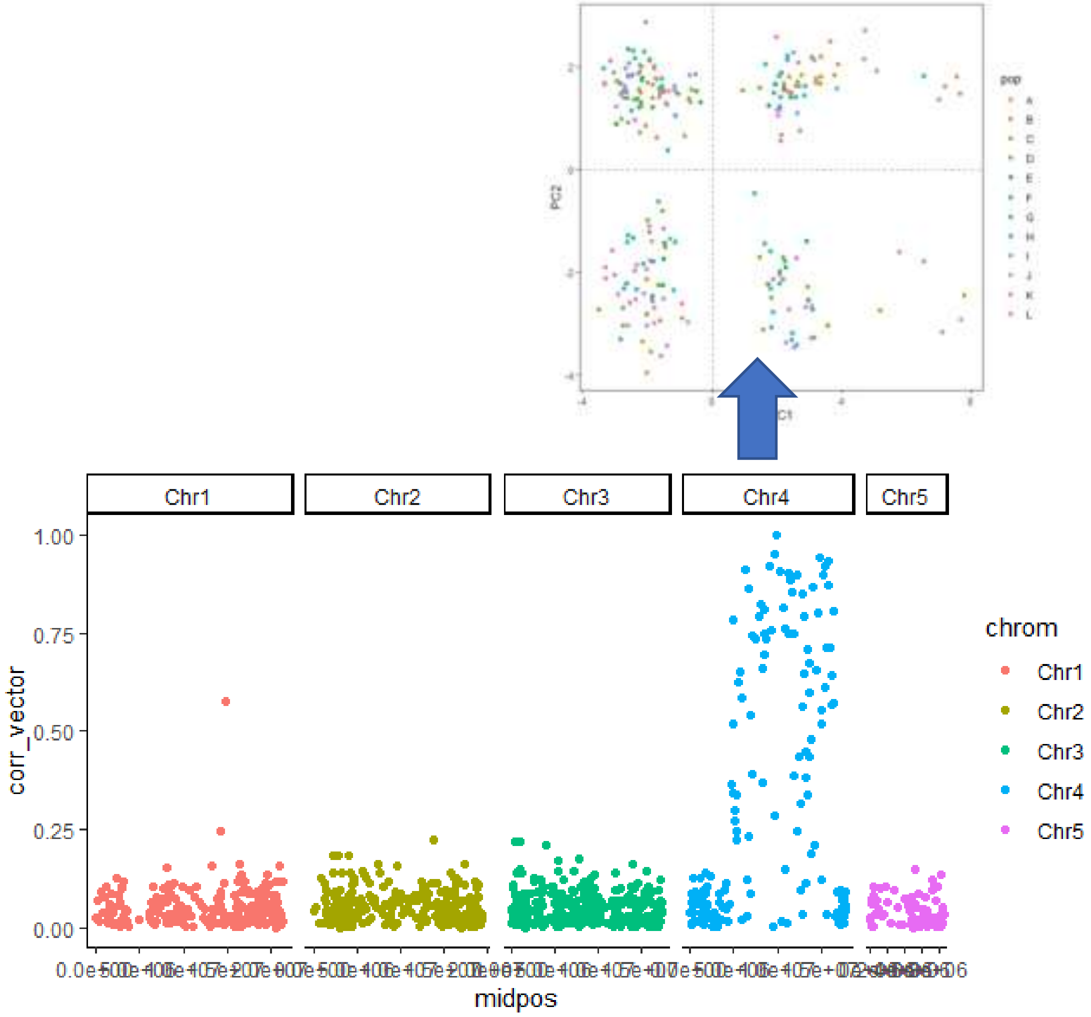
2 Exploration of the haploblocks (genotype, LD, Fst, Hobs)

Option 2: Explore duplicated loci in RAD-seq data

1 Detection and filtering of duplicated loci

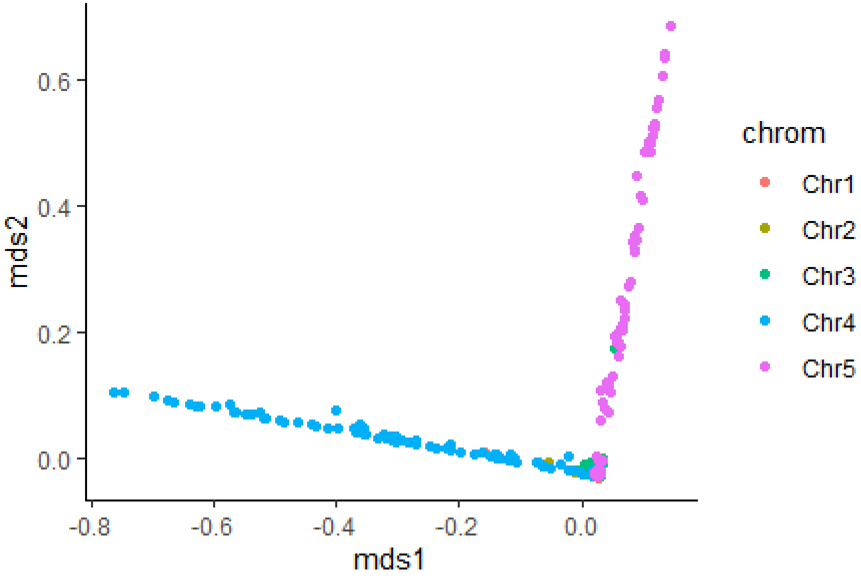
2 Analysis of those CNVs in pop G

4-1 Detection with local PCA



Correlation between local PCA and global PCA

MDS looking at similar windows accross the genome

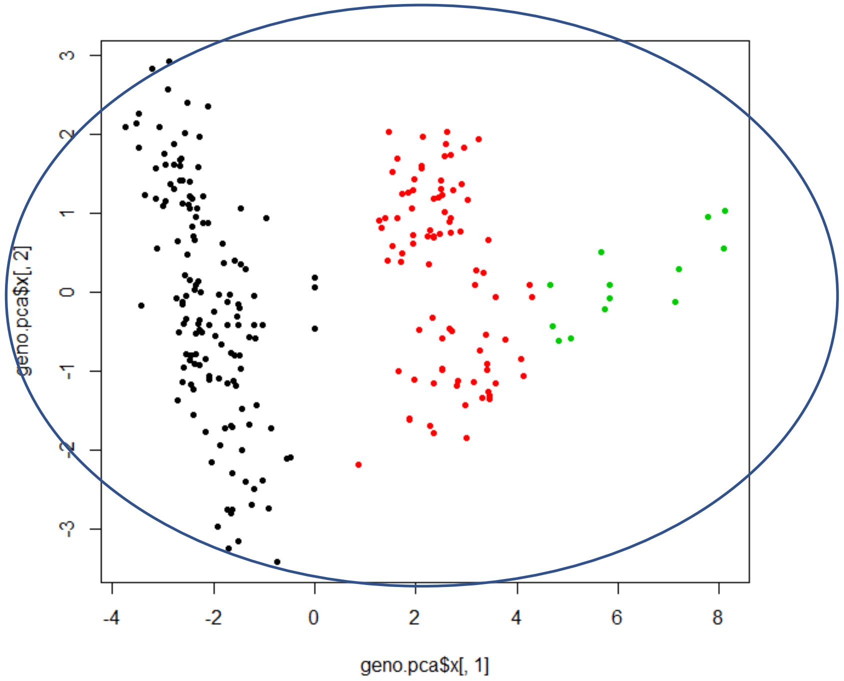


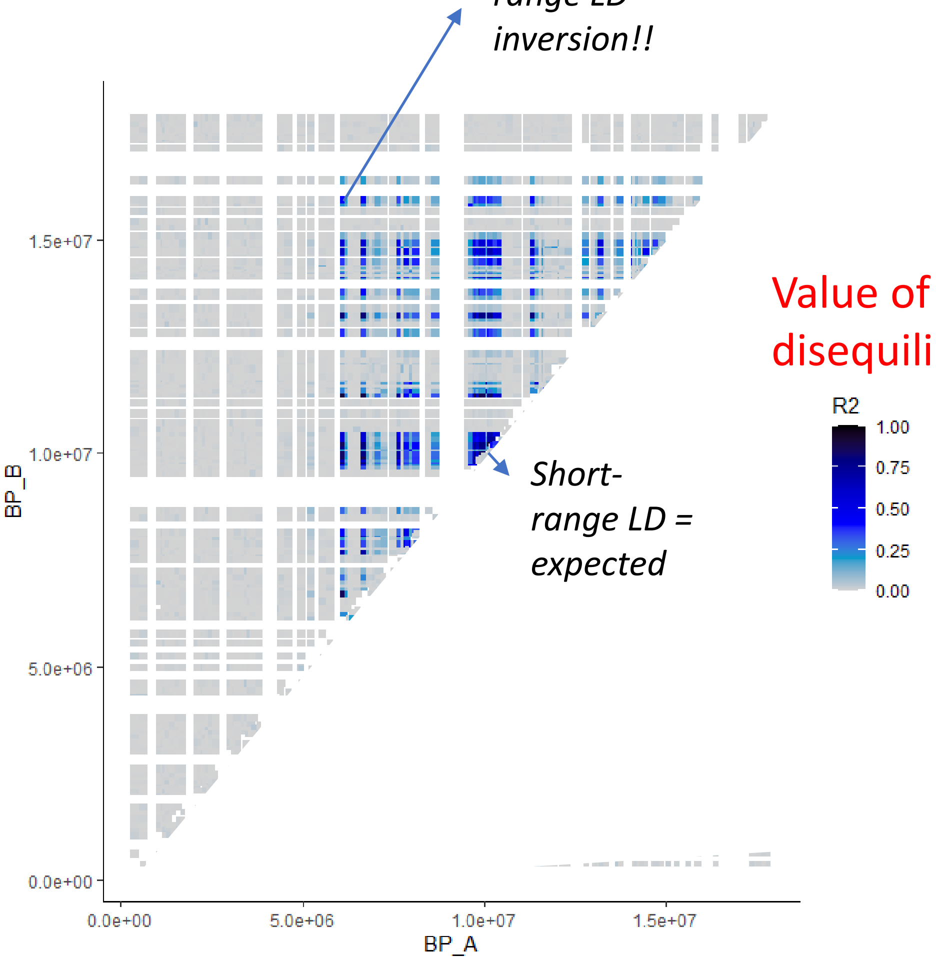
4-1 Exploration of the haploblocks

-> Genotype

-> Linkage disequilibrium

All samples





Long-range LD = inversion!!

Short-range LD = expected

Value of Linkage disequilibrium

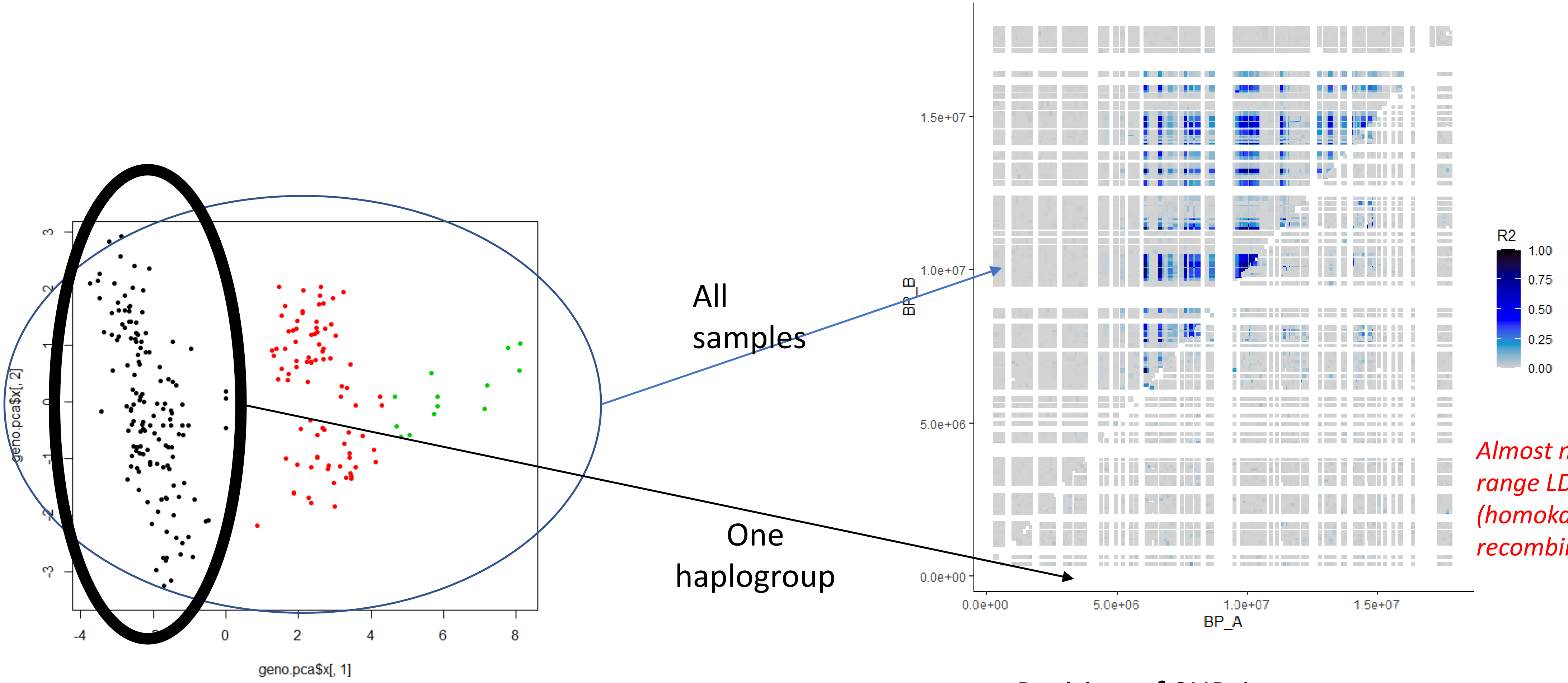
Position of SNP 2

Position of SNP 1

4-1 Exploration of the haploblocks

-> Genotype

-> Linkage disequilibrium



All samples

One haplogroup

Almost no long-range LD (homokaryotes recombine)

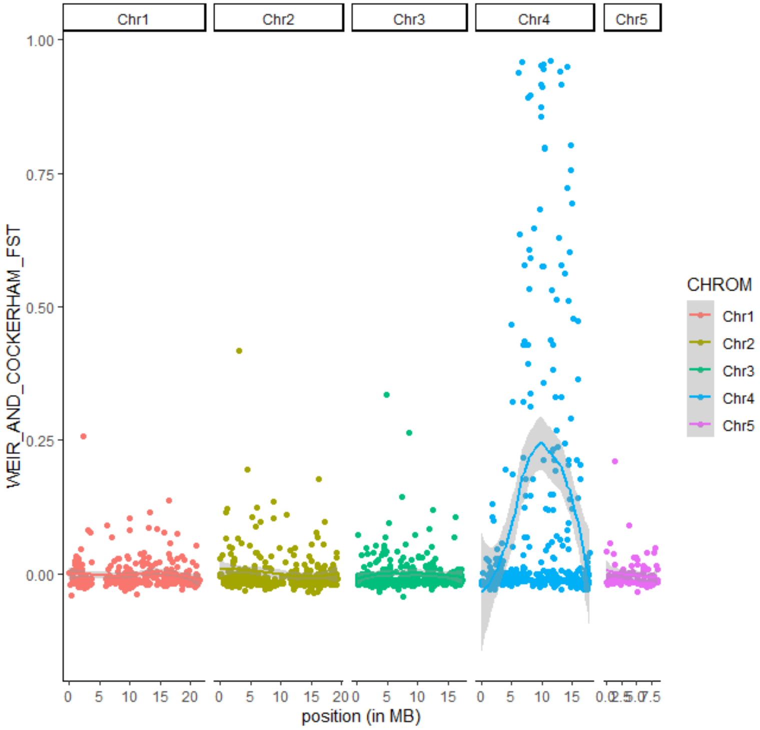
Position of SNP 1

4-1 Exploration of the haploblocks

-> Genotype

-> Linkage disequilibrium

-> Fst between haplogroups (optional)



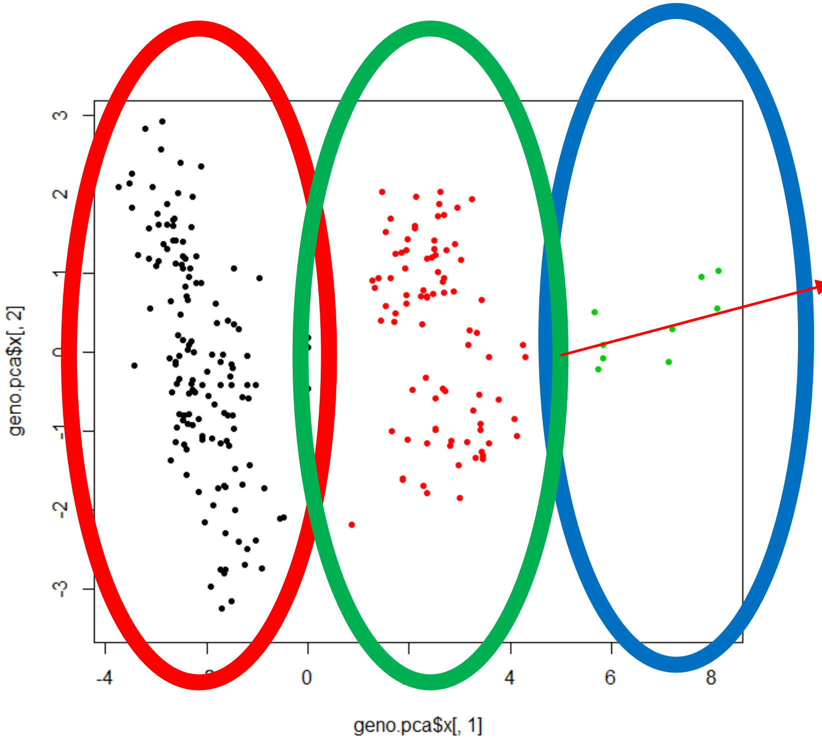
4-1 Exploration of the haploblocks

-> Genotype

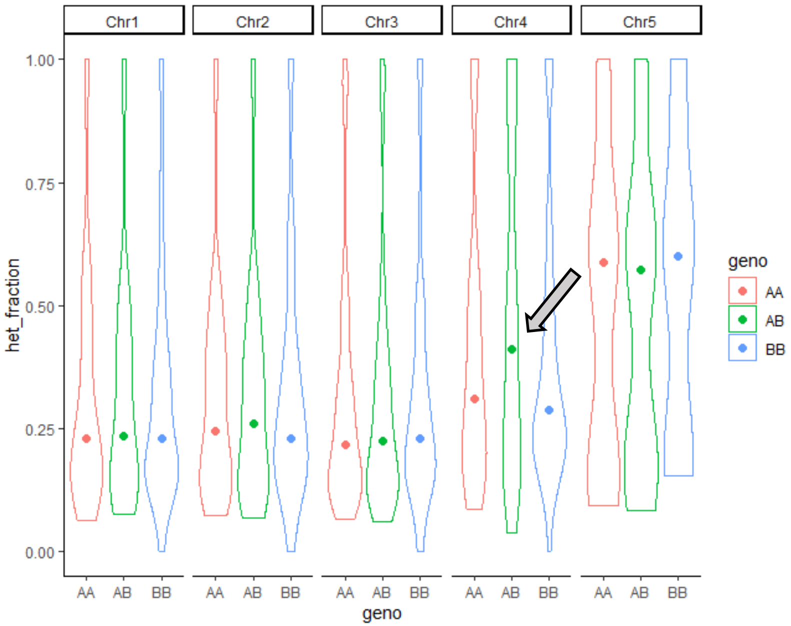
-> Linkage disequilibrium

-> Fst between haplogroups (optional)

-> Observed fraction of heterozygotes (optional)



More heterozygotes in that region ?



Day 4:

Option 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)

Option 2: Explore duplicated loci in RAD-seq data

1 Detection and filtering of duplicated loci
2 Analysis of those CNVs in pop G

Option 2: Explore duplicated loci in RAD-seq data

Main recap of the tutorial

- Use a lax filtered vcf file (the higher the number SNPs and samples, the better CNV detection is)

- Each dataset is unique and the characterization of singletons/duplicated require settings adjustement for each datasets

- Use the read count info embeded in the vcf format (vcftools --geno-depth)

- Normalize the read count data in R with edgeR

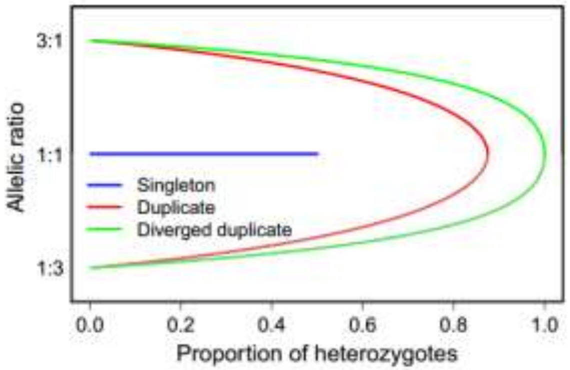
- Remove sex related loci (it depends what is your question)

- Fill missing data

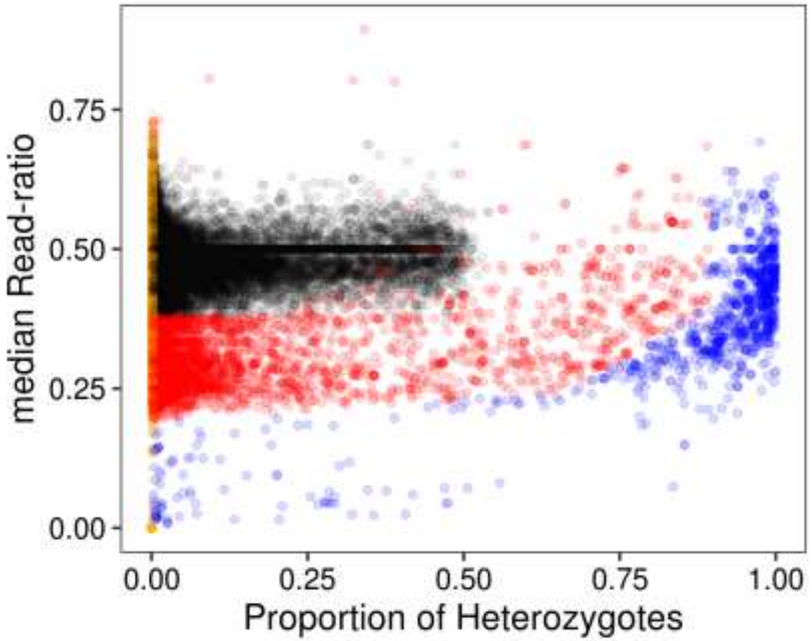
- Use RDA for CNV-Environment association (or other ways such as GLMMs)

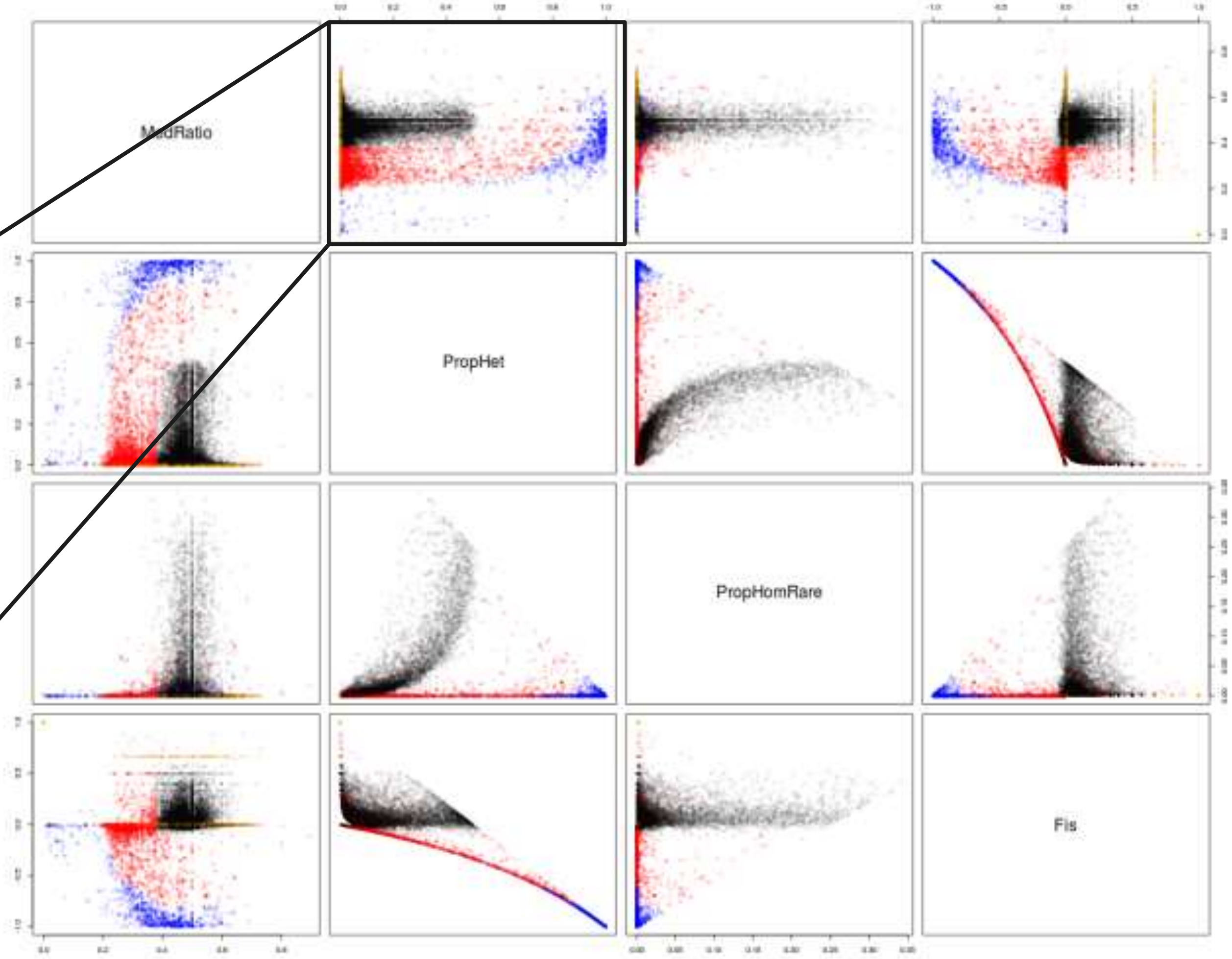
- Explore the results as you wish (e.g. PCA, BrayCurtis trees...)

Discover and split SNPs categories



Therorical pattern (McKinney et al., 2017)





● Singletons

● Duplicated

● Duplicated & diverged

● Low confidence

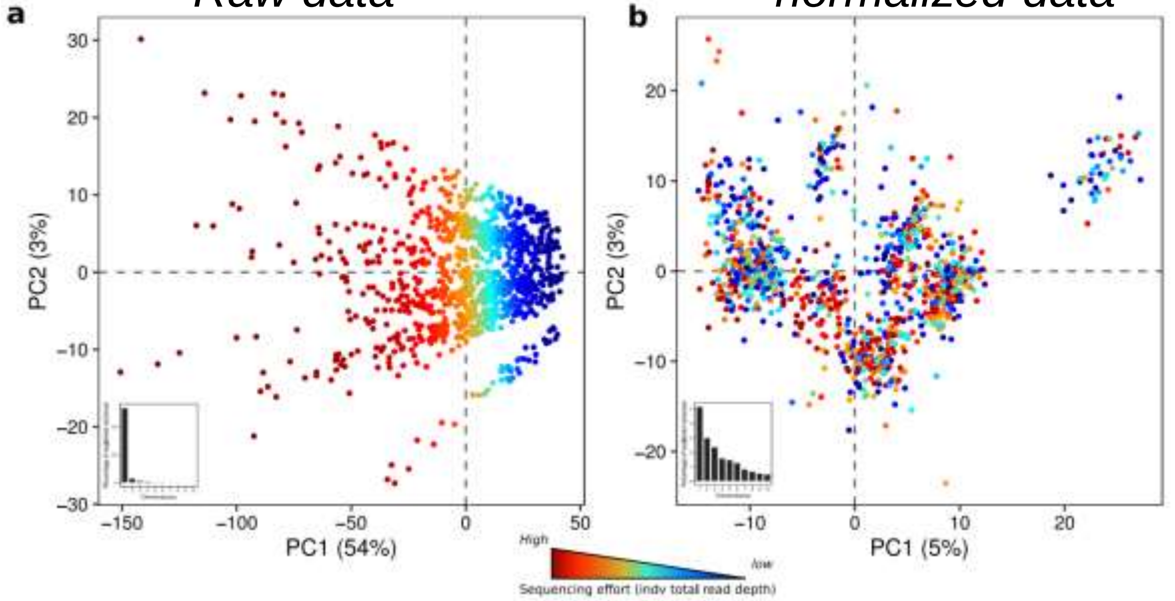
Dorant et al., 2020

Use duplicated loci to explore CNVs variants

1. Use locus read depth as a proxy of Copy Number Variation among samples.

- ➔ Read depth normalization using RNAseq methods.

Raw data *normalized data*



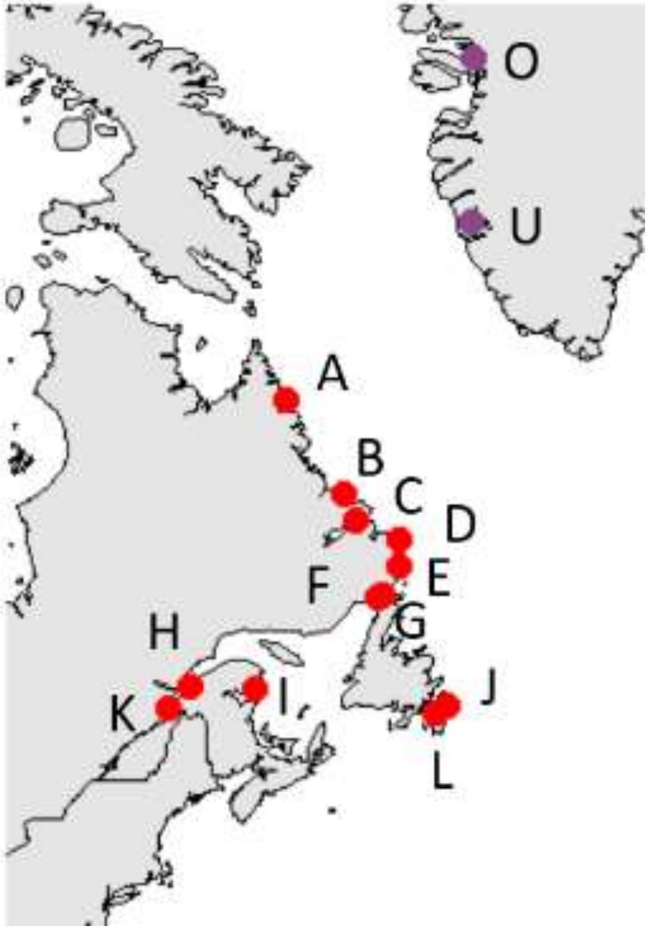
Sequencing effort (indv total read depth)

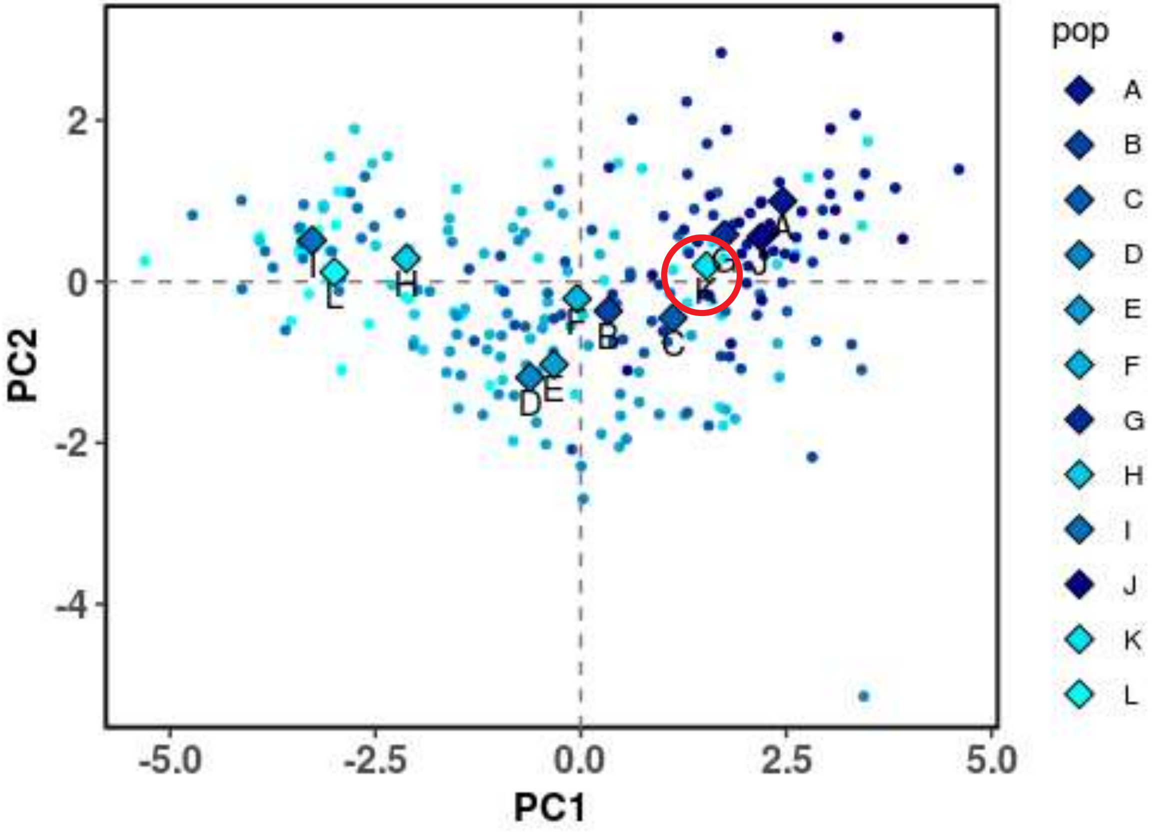
1.2. Use the normalized read depth martix of CNVs loci to explore population genomics

- Environment X CNVs associations (RDA)
- Basic genetic structure with PCA

Plot the adaptive CNVs information → CNVs putatively associated with the temperature

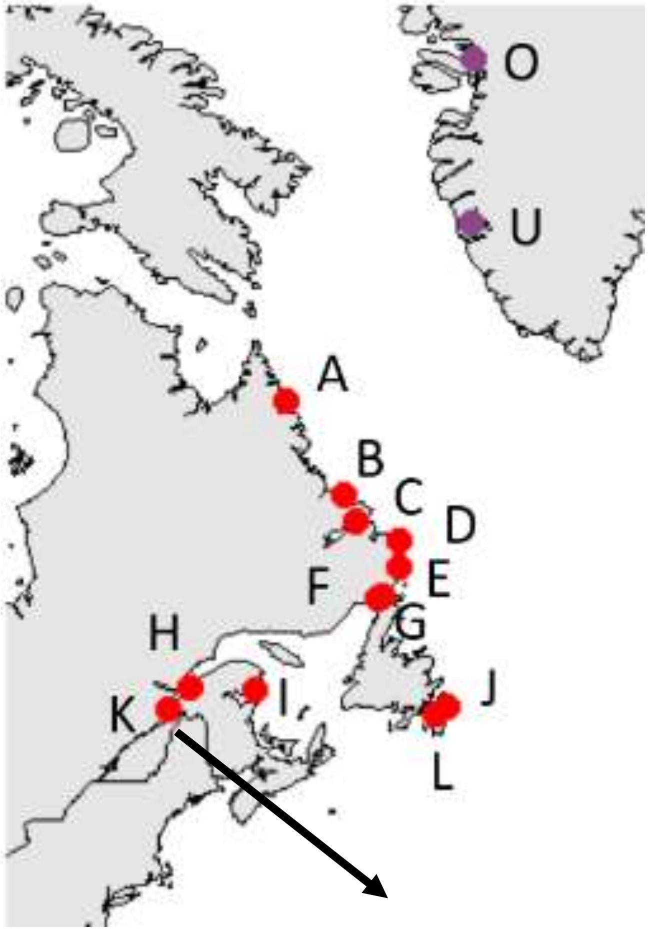
Note the position of the samplig site K along the PC1 !

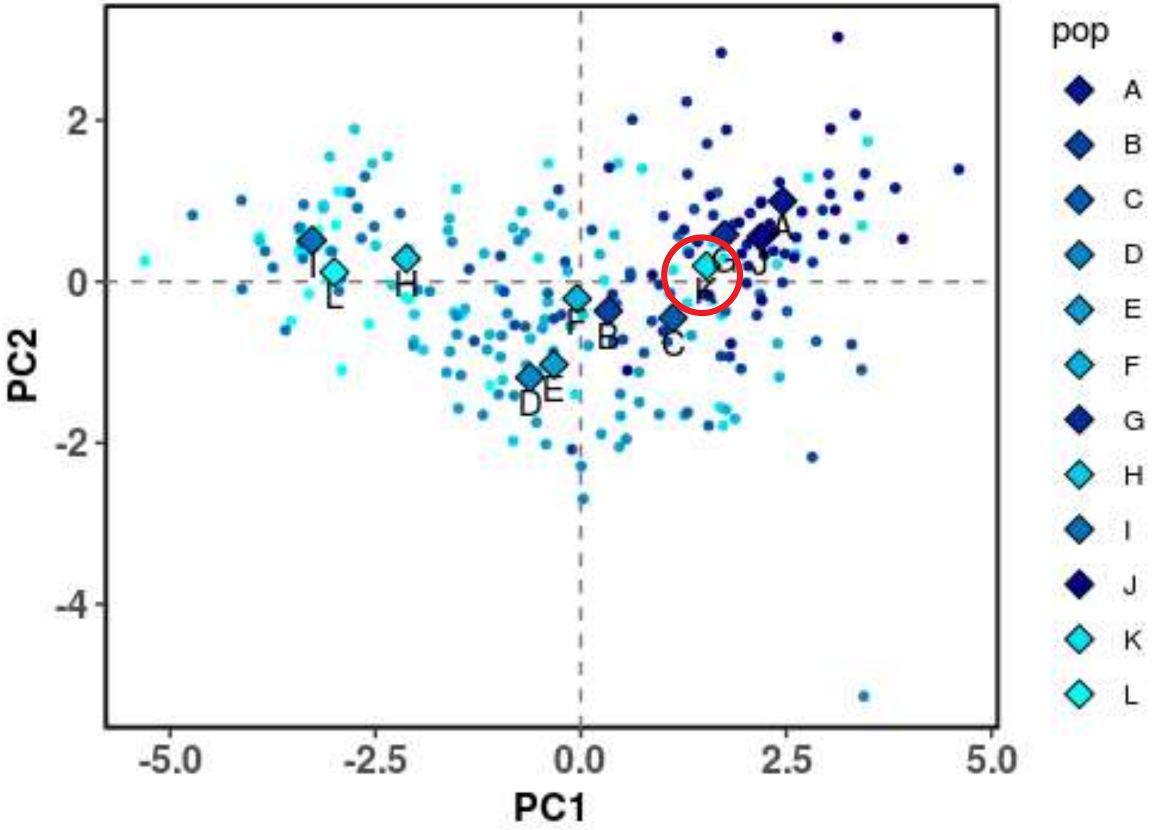




Plot the adaptive CNVs information → CNVs putatively associated with the temperature

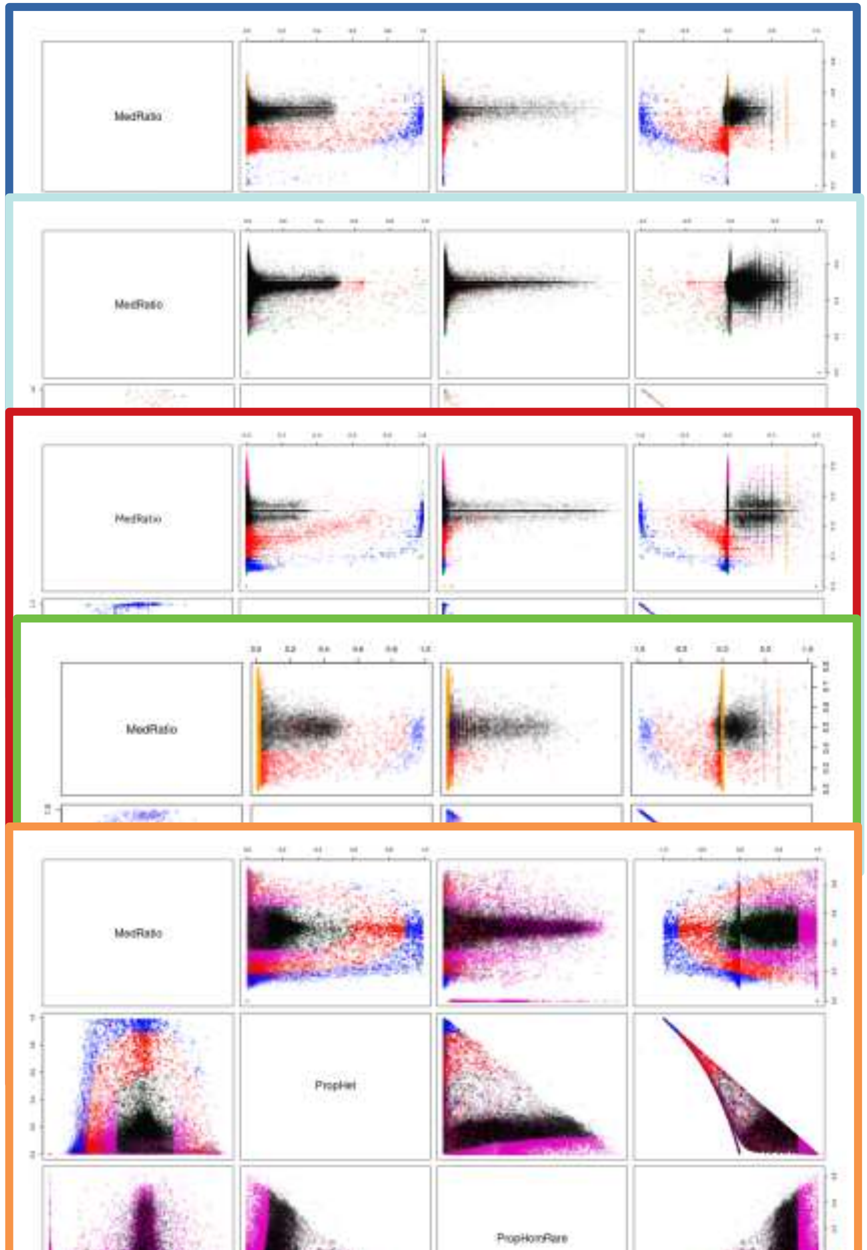
Note the position of the samplig site K along the PC1 !





Population K is in fact affiliated to Fjord region with much lower temperatures

Applicability of the CNVs approach



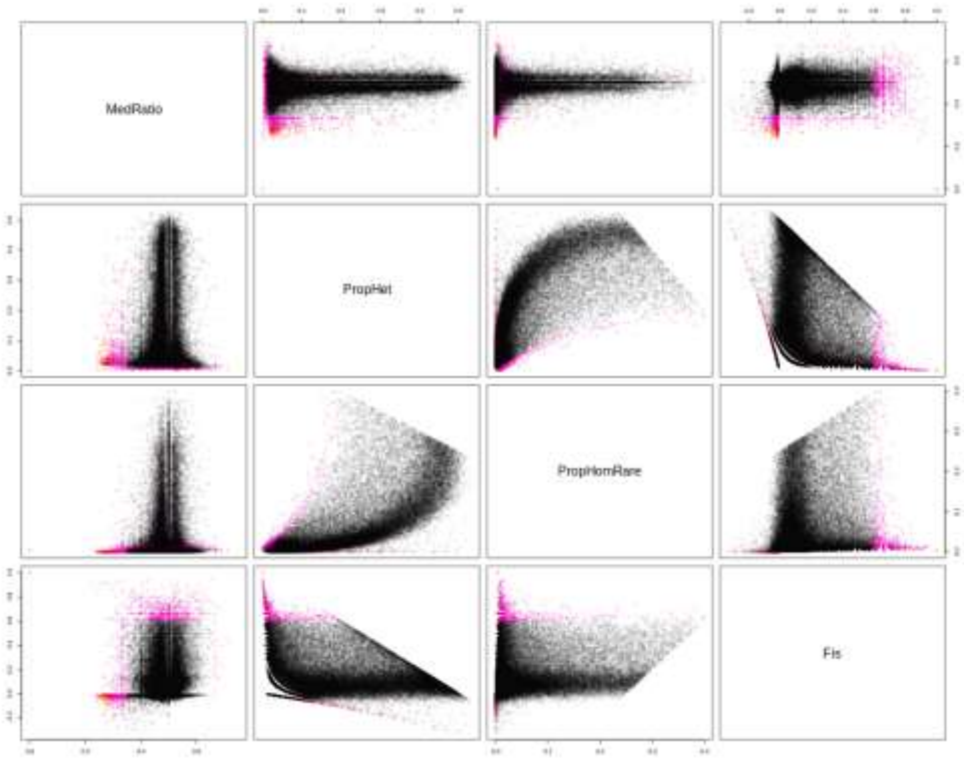
Homarus americanus

Mallotus villosus

Salvelinus fontinalis

Lithobates sylvaticus

Rana luteiventris



Reinhardtius hippoglossoides

Tutorial day 5

Most methods that we saw during the week will provide

- ⇒ **General knowledge about isolation-by-adaptation, the genetic architecture of adaptation, an idea of 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 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/

Day 5: follow-up and annotation

5-1 Annotate SNPs
5-2 Overlap SNPs/Genes
5-3 Gene Ontology Enrichment
5-4 (Optional) Overlap CNVs/Repeated elements

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
```

5-4 Overlap CNVs / repeats or TE

Optional!
It uses the annotation by repeatMasker to test if CNVs detected yesterday overlap with repeated regions or transposable elements.