Wrap-up
Day 3: Local adaptation

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

-> works on windows and remove linked SNPs in that window (adjust window size and nb of SNPs to your needs and data: RAD-seq vs. WGS)

-> VIF (variance inlfexion factor) is a measure of non-independance (used here between SNPs). You could also use R^2

Plink manual https://www.cog-genomics.org/plink/2.0/ld

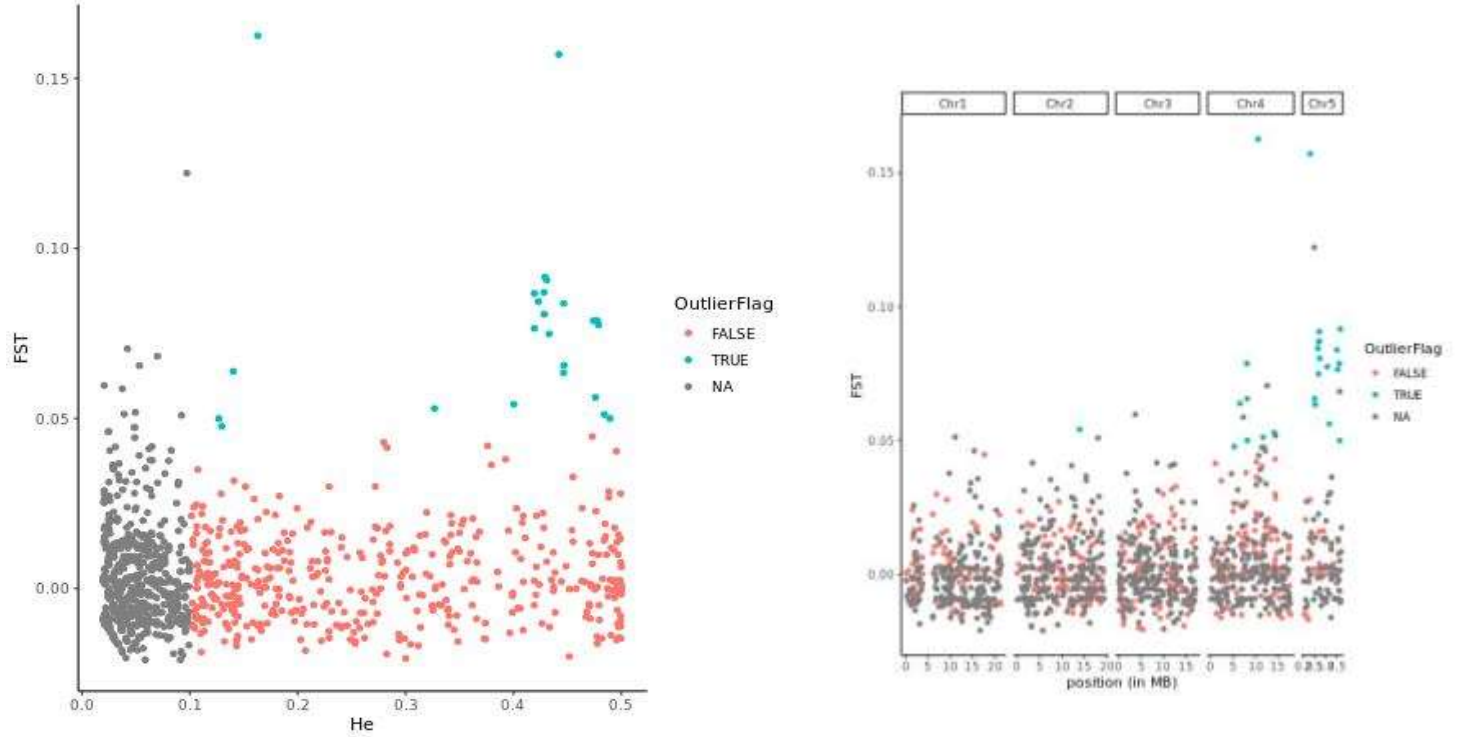
WINDOW=100
SNP=100
VIF=2

```
plink --bed 02_data/canada.bed
\ --bim 02_data/canada.bim \ --
fam 02_data/canada.fam \ --
indep $WINDOW['kb'] $SNP $VIF
--allow-extra-chr \ --out
02_data/canada
```

3-2 Outlier detection
-> with OutFlank

Based on Fst outliers across all pairs of populations

We will follow the best practices recommended by the authors:
•remove SNPs with very low heterozygosity (options: Hmin = 0.1)
•use the FSt uncorrected for population size (options: NoCorr = TRUE) (anyway, here all pop have 20 individuals)
•Compare the FSt against a distribution based on independant SNPs (pruned for short-distance and long-distance LD) We will use the list of pruned SNPs extracted with PLINK earlier.

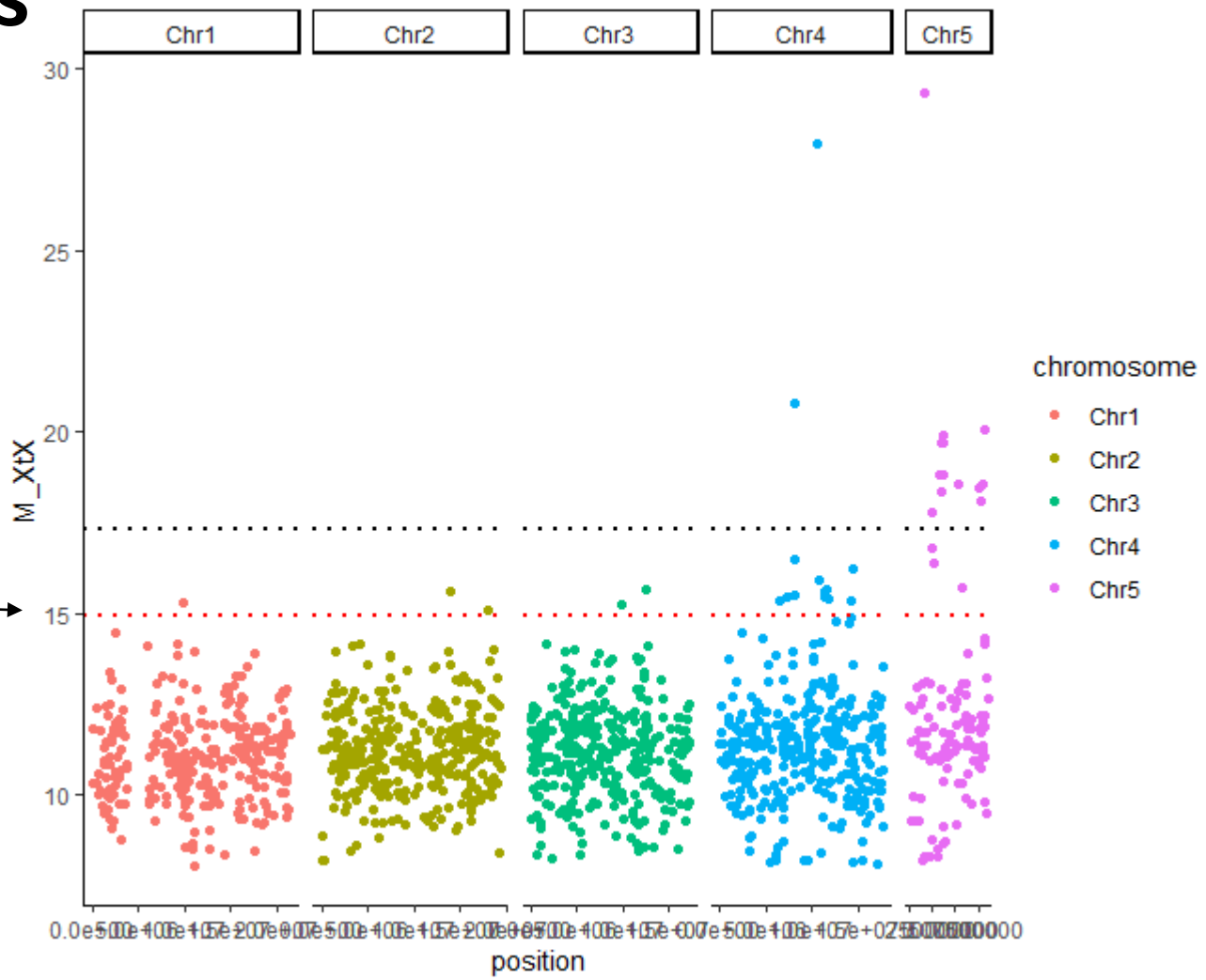


3-2 Outlier detection
-> with Baypass

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



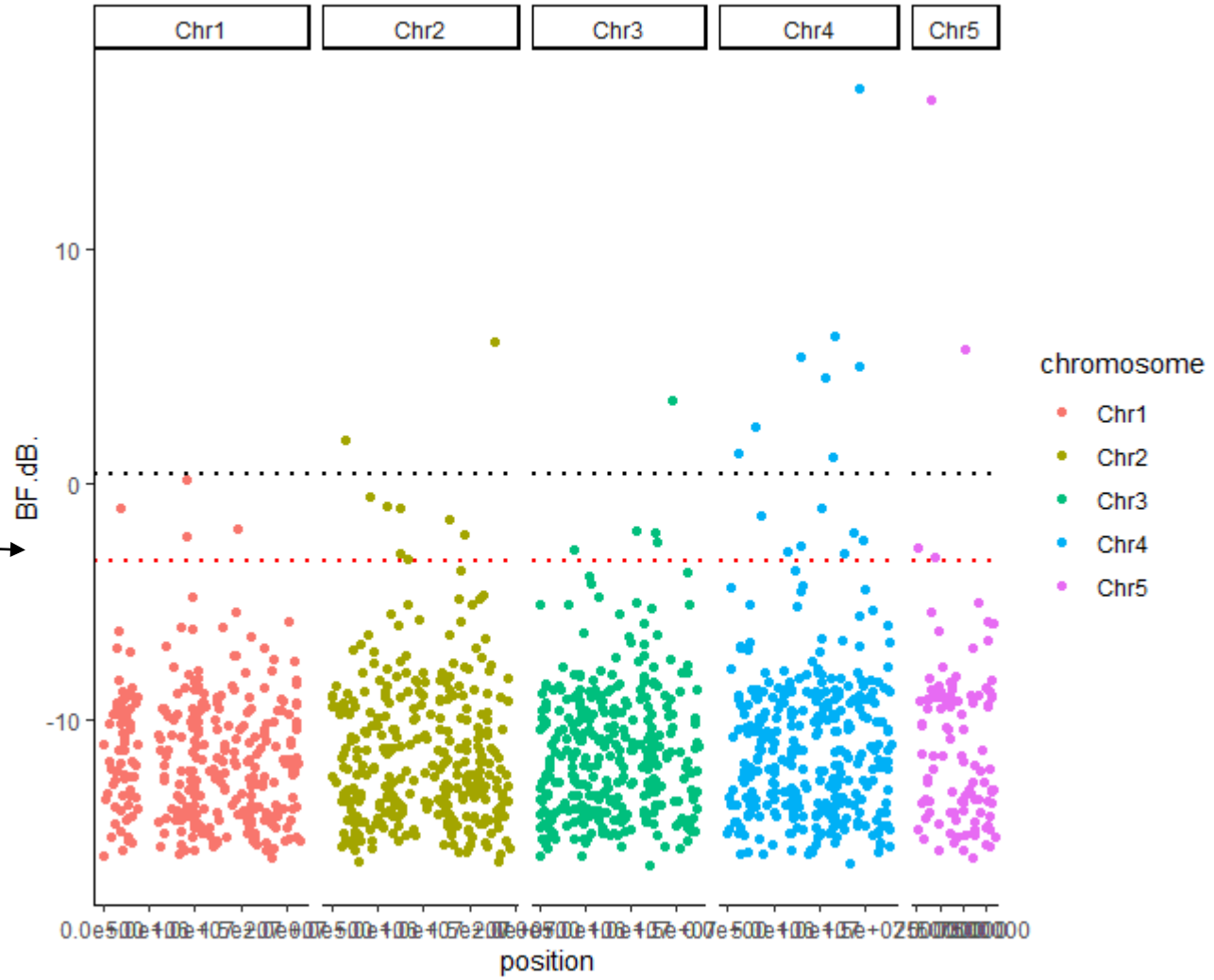
3-3 Environmental associations -> with Baypass

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



Baypass
about making independant runs

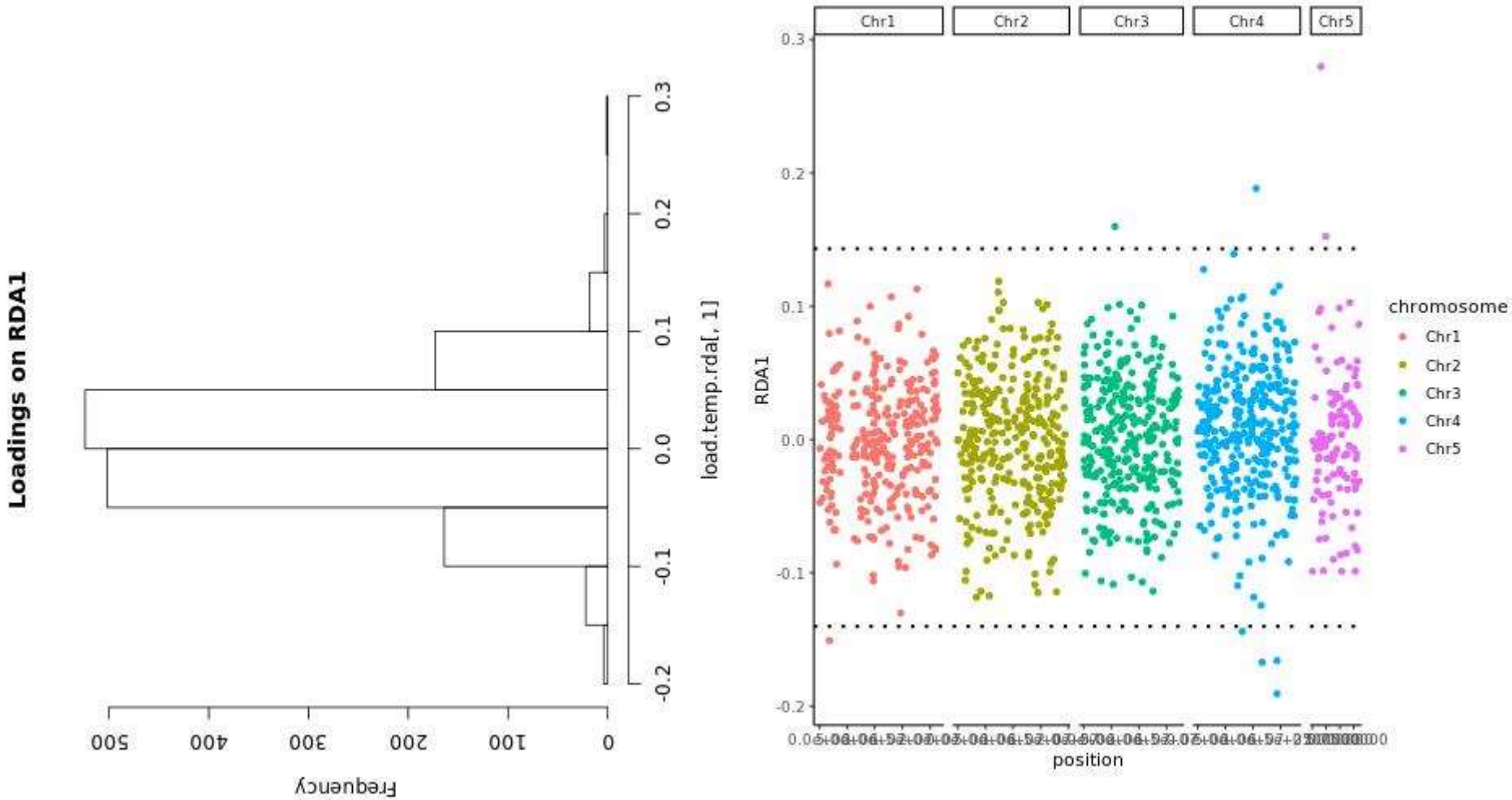
What we did

- 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)*

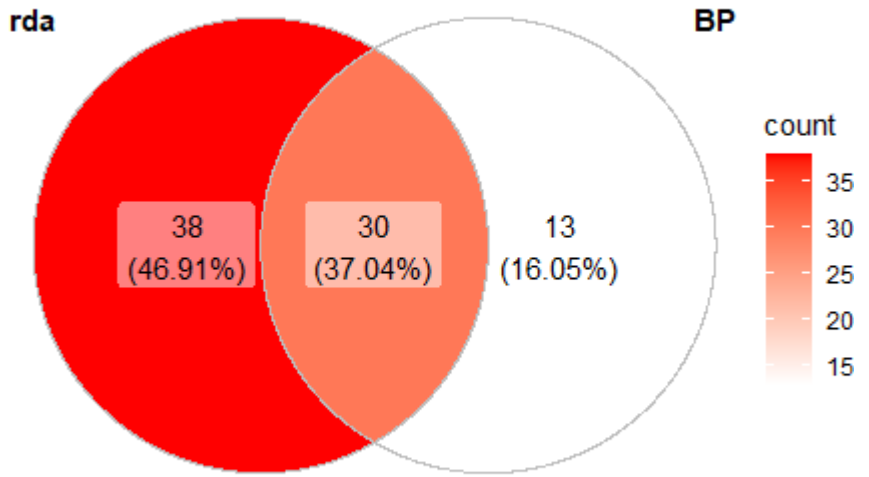
Recommended Practices for your dataset

- 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
-> with RDA



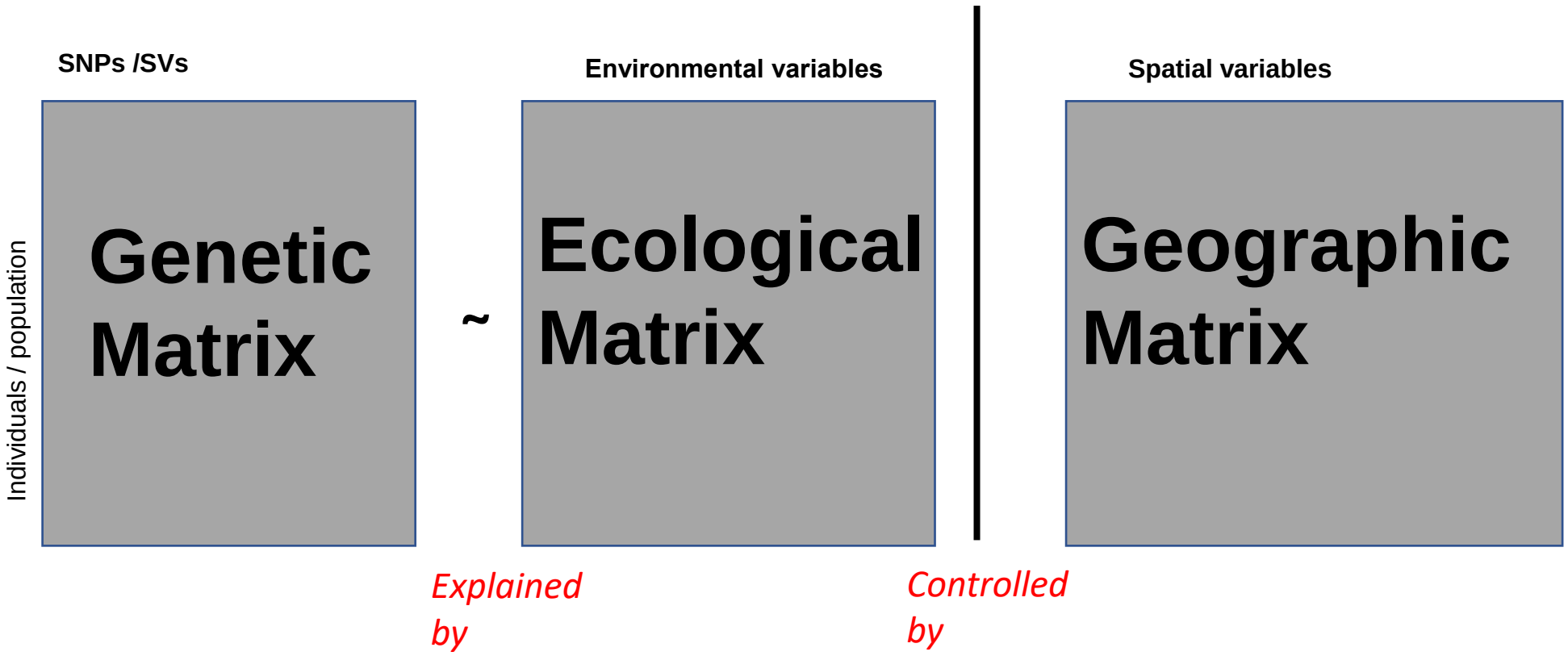
3-3 Environmental associations
-> Overlap



RDA
advanced options

(prepared with the help of
Dr. Martin Laporte)

RDA

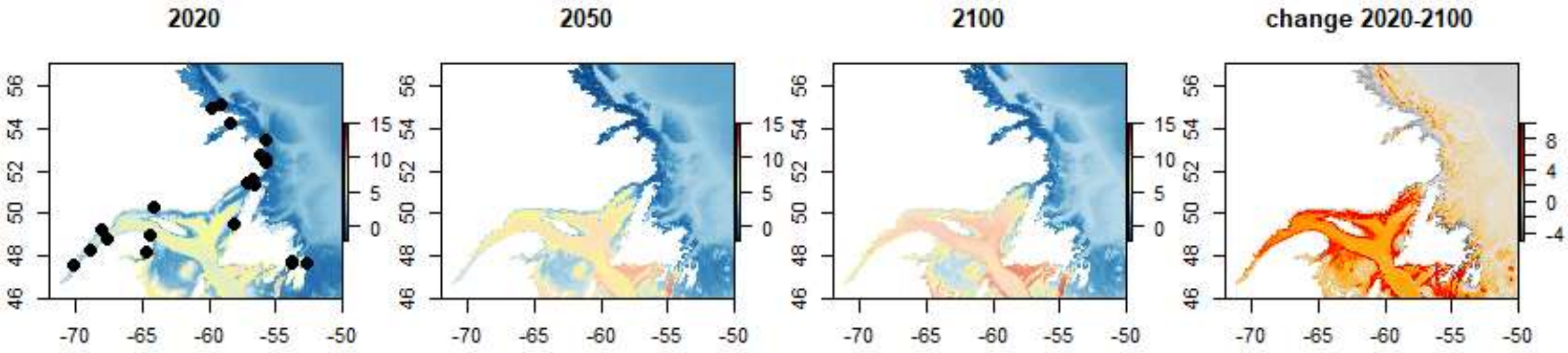

SNPs /SVs
Environmental variables
Spatial variables
Individuals / population
Genetic Matrix
~
Ecological Matrix
Geographic Matrix
Explained by
Controlled by

Climatic Variables
how to extract them from databases?

https://www.worldclim.org/

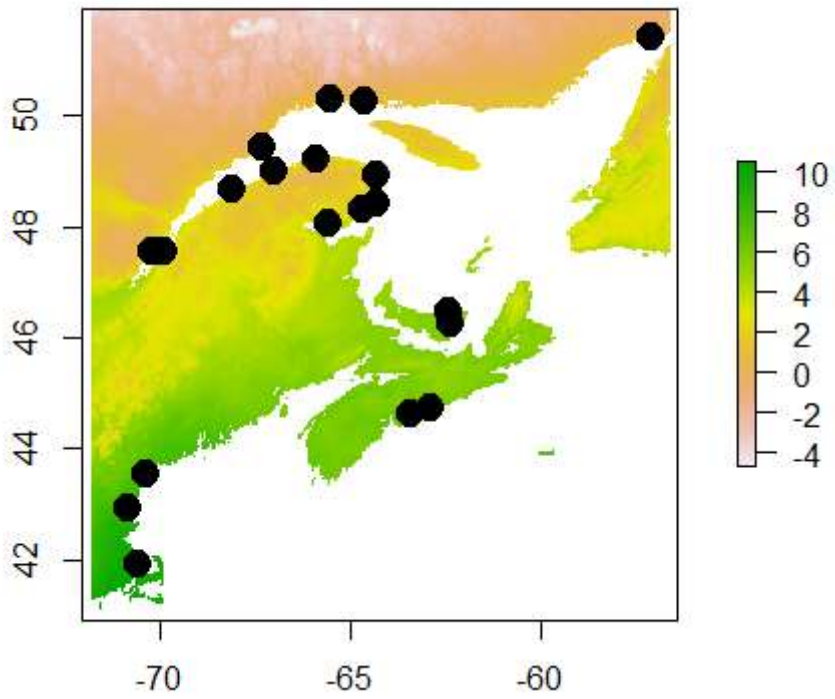
http://www.marspec.org/
(with useful tutorials)

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

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)

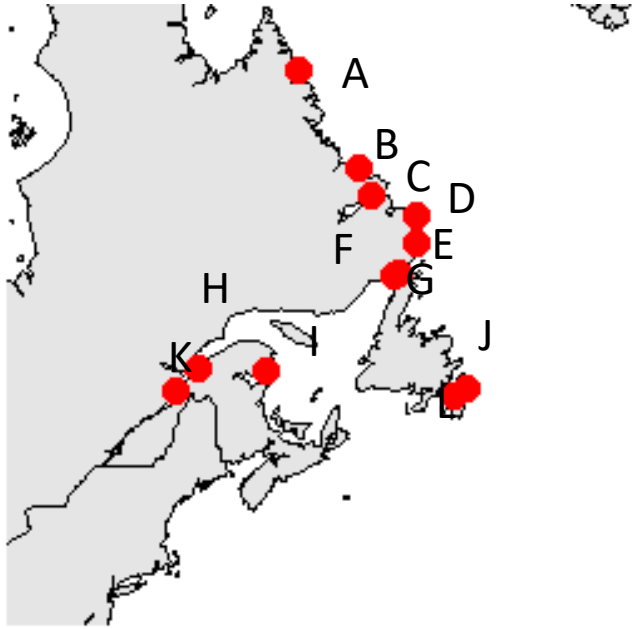
Option 2: Explore duplicated loci in RAD-seq data

1 Detection and filtering of duplicated loci

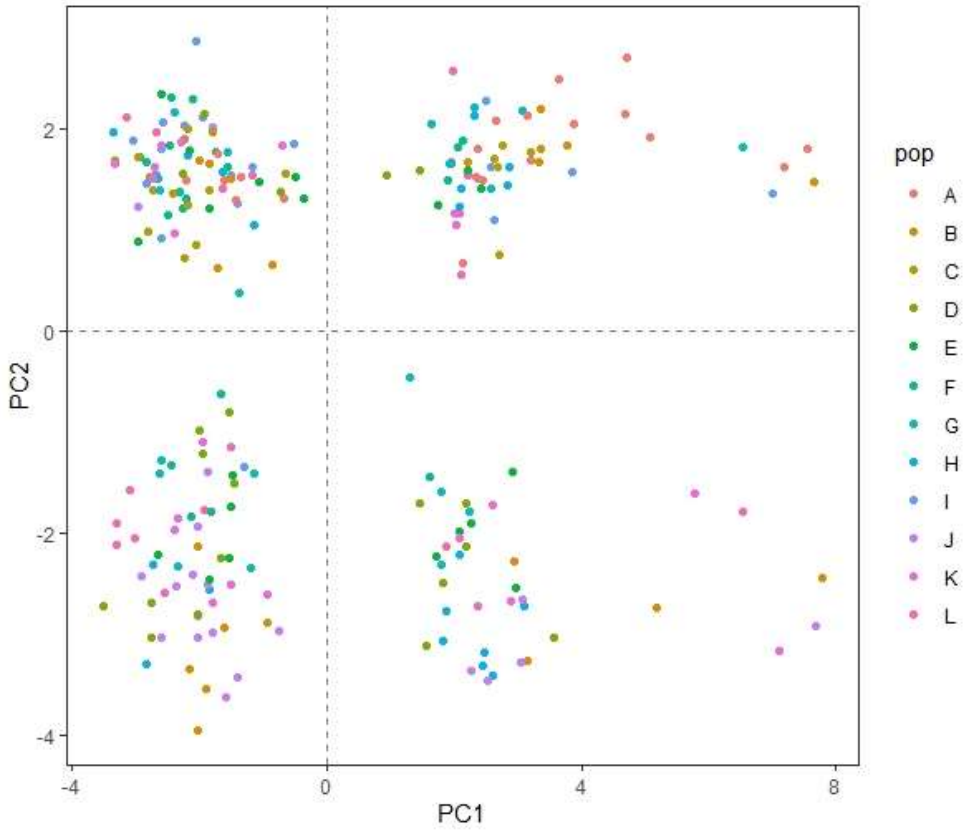
2 Analysis of those CNVs in pop G

Why?

On day 2, we observed a strong structure 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)

HIGHLIGHTED ARTICLE
GENETICS | INVESTIGATION

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

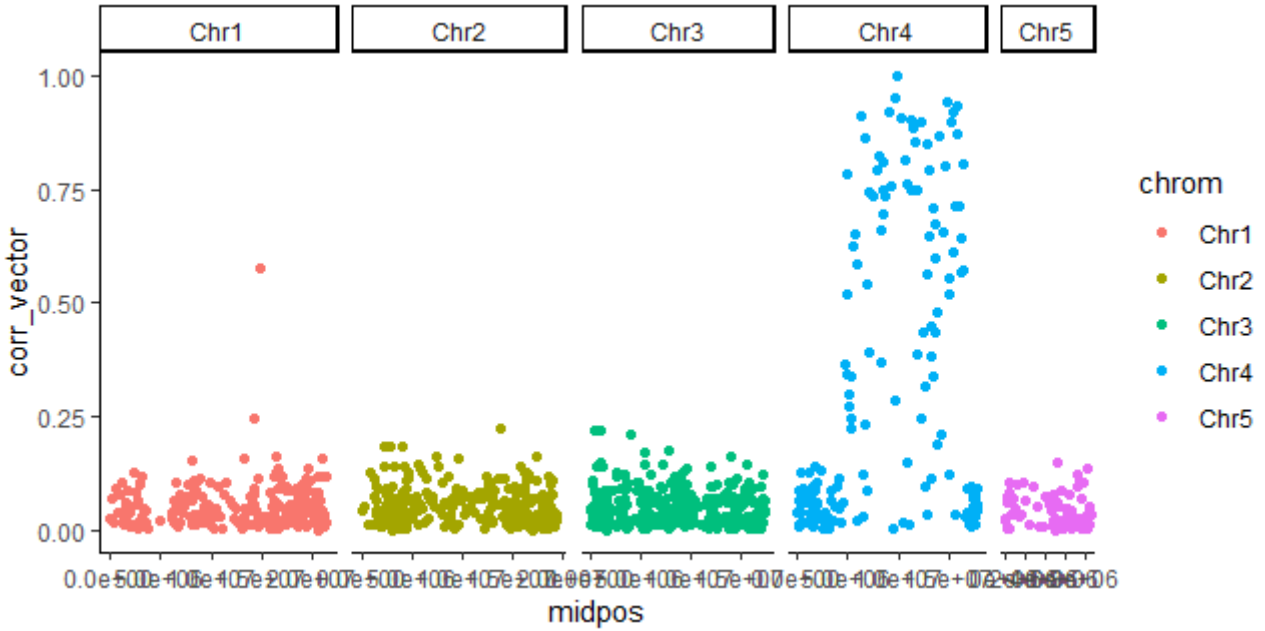
Han Li* and Peter Ralph*,†,‡,1

*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

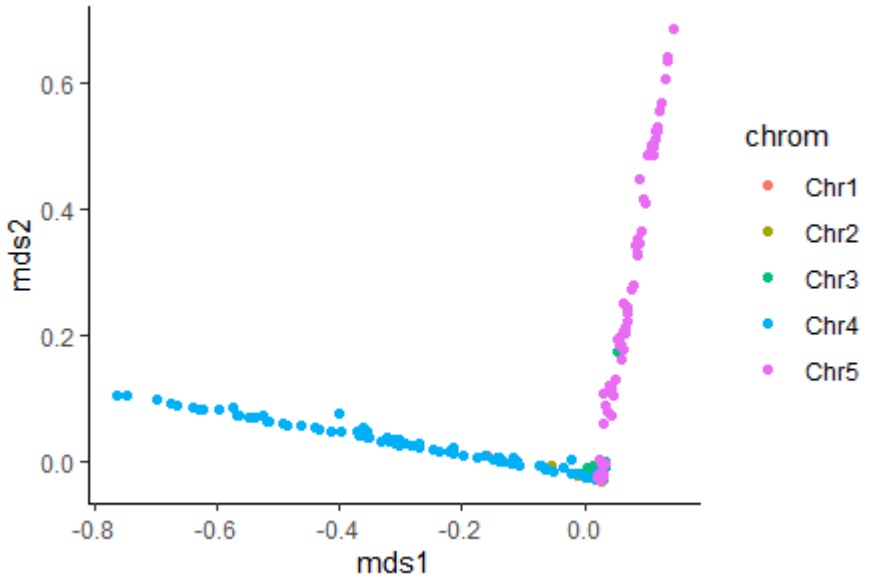
ORCID ID: 0000-0002-9459-6866 (P.R.)

4-1 Detection with local PCA
-> We will use the package lostruct

Correlation between local PCA and global PCA



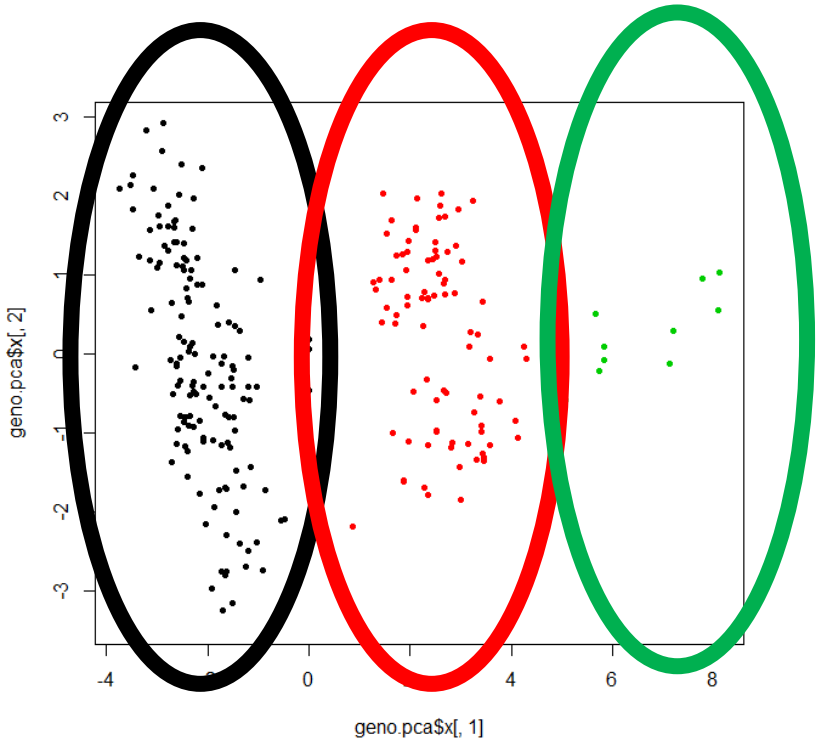
MDS looking at similar windows accross the genome



4-1 Exploration of the haploblocks

-> Genotype

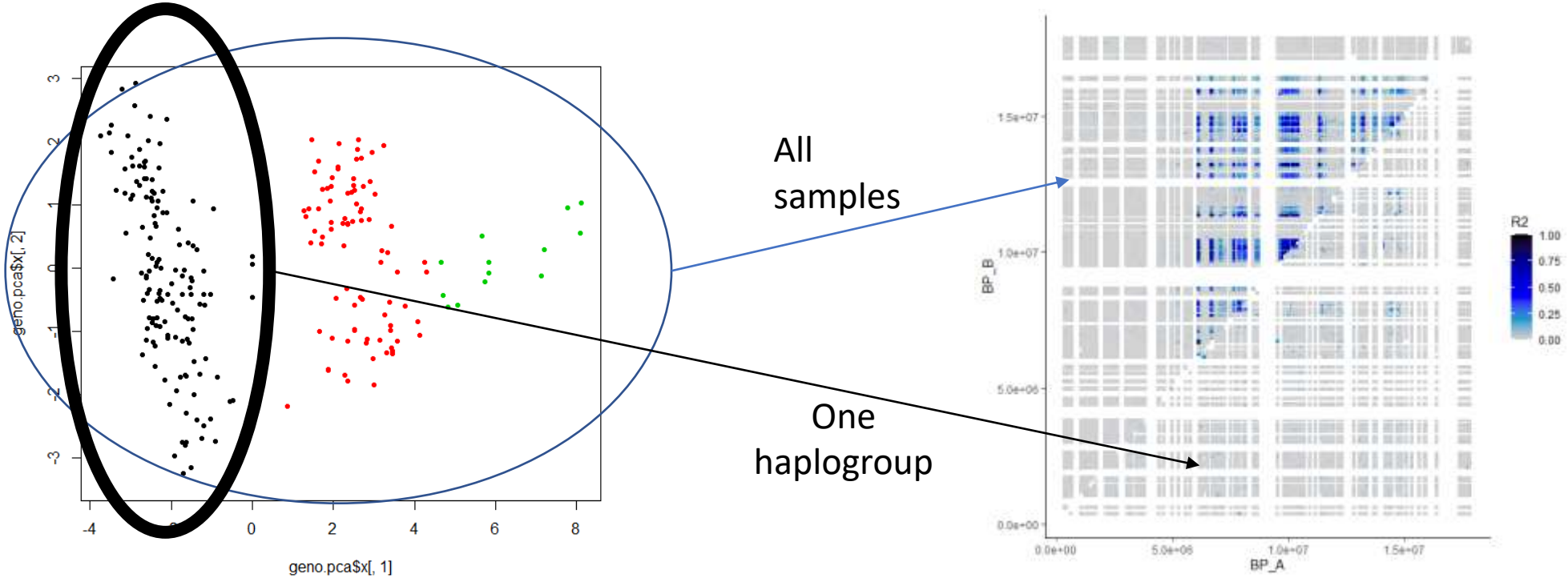
-



4-1 Exploration of the haploblocks

-> Genotype

-> Linkage disequilibrium

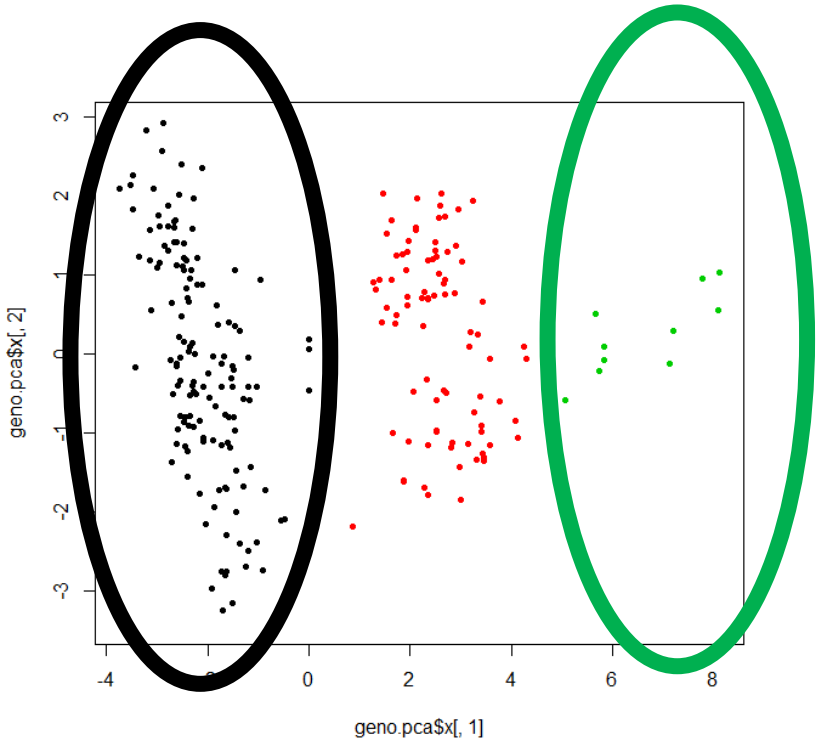


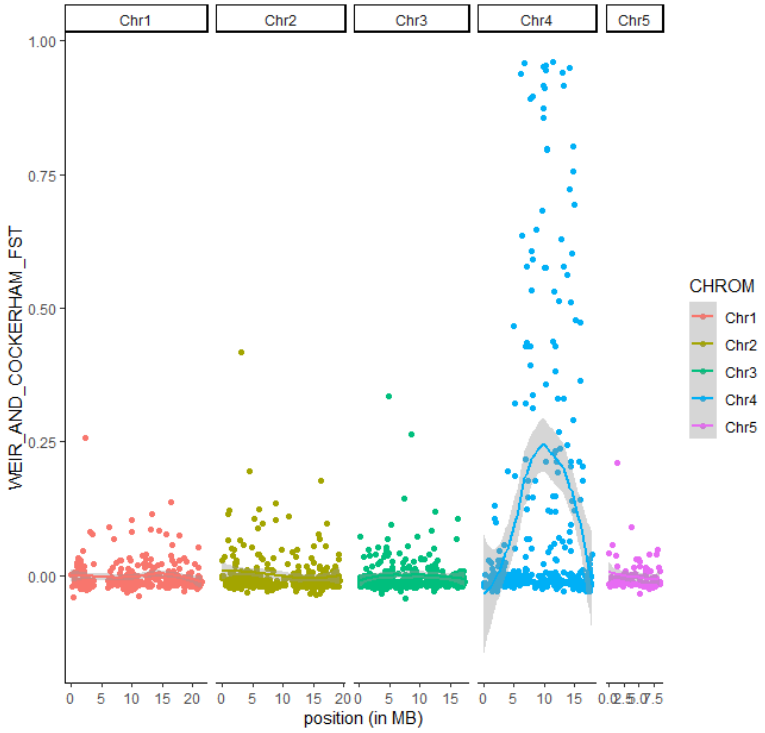
4-1 Exploration of the haploblocks

-> Genotype

-> Linkage disequilibrium

-> Fst between haplogroups (optional)





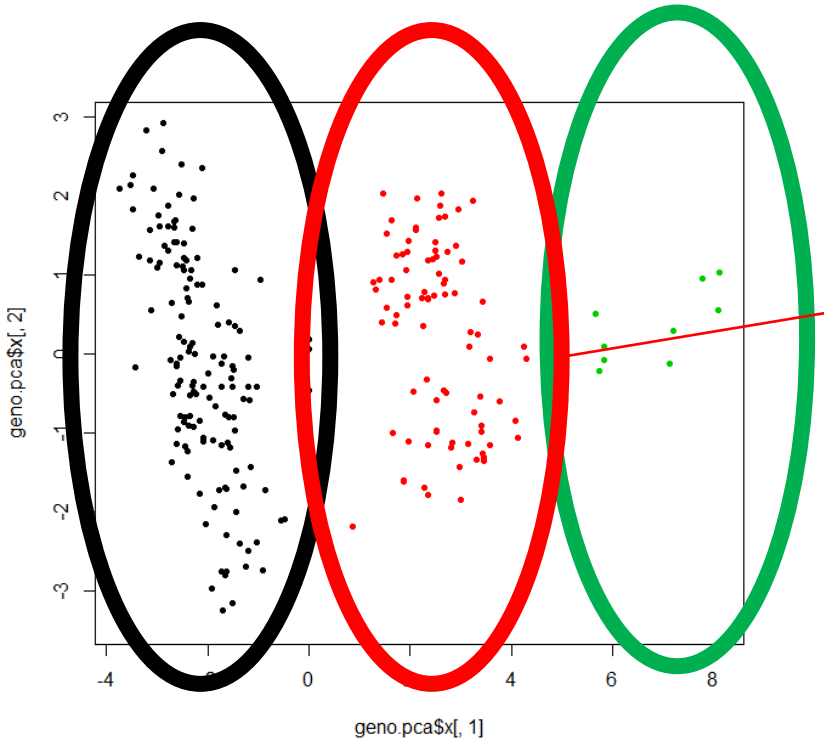
4-1 Exploration of the haploblocks

-> Genotype

-> Linkage disequilibrium

-> Fst between haplogroups (optional)

-> Observed fraction of heterozygotes (optional)



More heterozygotes in that region ?

Day 4:

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

Why?

Because some loci are duplicated but collapsed in a single loci by SNP callers tools

So instead of having a SNP that is A/A or A/T

It is a SNPs A/A/A/A or A/A/T/A, etc

Worse when there are multiple copies

$\Rightarrow$ Biais genotype estimation and allelic frequencies

$\Rightarrow$ But it is also a mine of gold: other variants = CNVs

Received: 18 September 2017 | Revised: 18 January 2018 | Accepted: 19 January 2018

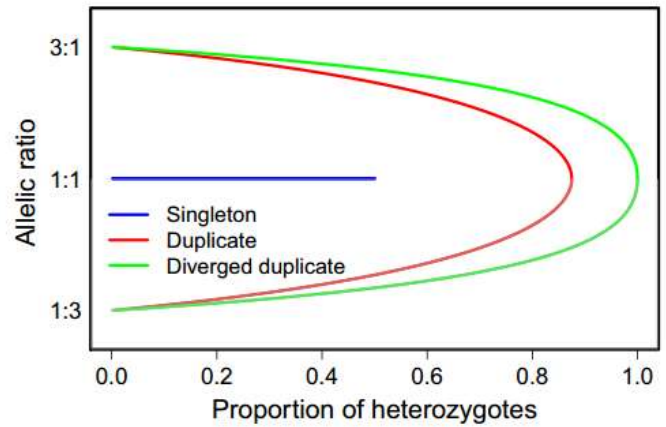
DOI: 10.1111/1755-0998.12763

RESOURCE ARTICLE

WILEY MOLECULAR ECOLOGY RESOURCES

Resolving allele dosage in duplicated loci using genotyping-by-sequencing data: A path forward for population genetic analysis

Garrett J. McKinney | Ryan K. Waples | Carita E. Pascal | Lisa W. Seeb | James E. Seeb



Received: 28 January 2020 | Revised: 16 July 2020 | Accepted: 21 July 2020

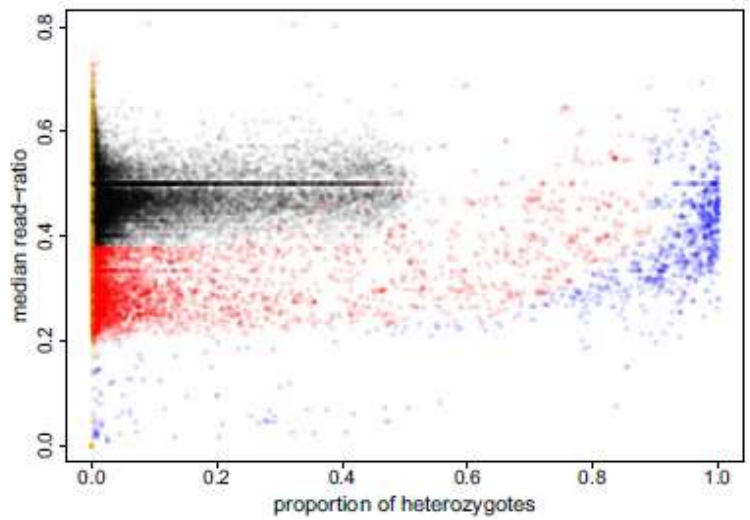
DOI: 10.1111/mec.15565

ORIGINAL ARTICLE

MOLECULAR ECOLOGY WILEY

Copy number variants outperform SNPs to reveal genotype–temperature association in a marine species

Yann Dorant[1] | Hugo Cayuela[1] | Kyle Wellband[1] | Martin Laporte[1] | Quentin Rougemont[1] | Claire Mérot[1] | Eric Normandeau[1] | Rémy Rochette[2] | Louis Bernatchez[1]



Received: 18 September 2017 | Revised: 18 January 2018 | Accepted: 19 January 2018

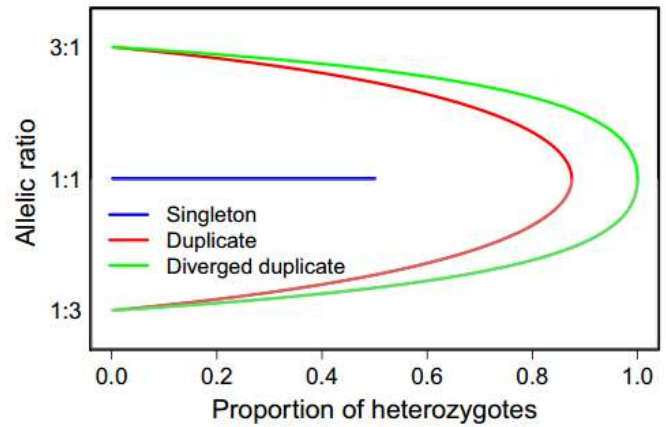
DOI: 10.1111/1755-0998.12763

RESOURCE ARTICLE

WILEY MOLECULAR ECOLOGY RESOURCES

Resolving allele dosage in duplicated loci using genotyping-by-sequencing data: A path forward for population genetic analysis

Garrett J. McKinney | Ryan K. Waples | Carita E. Pascal | Lisa W. Seeb | James E. Seeb



Received: 28 January 2020 | Revised: 16 July 2020 | Accepted: 21 July 2020

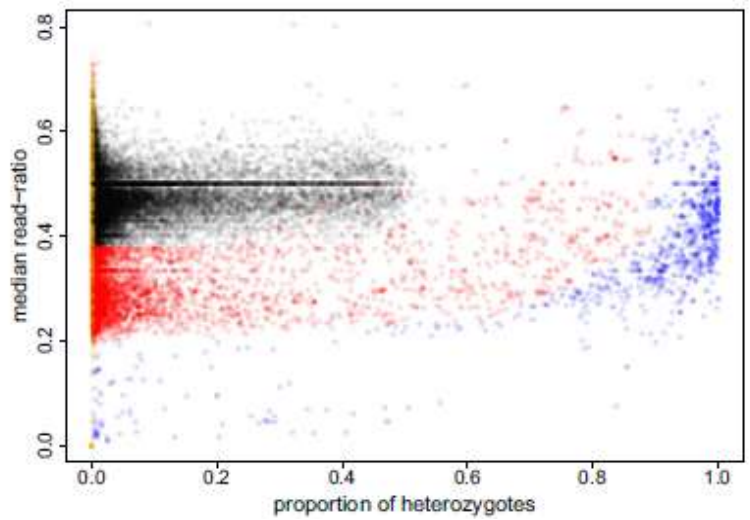
DOI: 10.1111/mec.15565

ORIGINAL ARTICLE

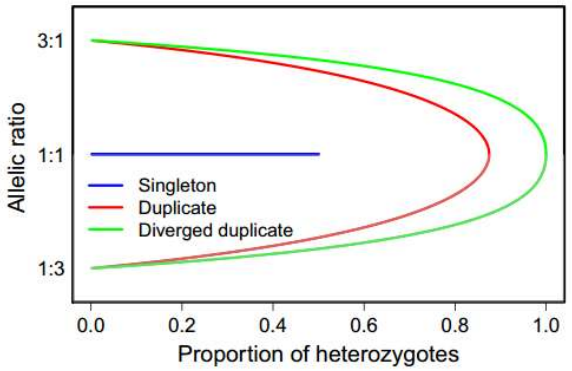
MOLECULAR ECOLOGY WILEY

Copy number variants outperform SNPs to reveal genotype–temperature association in a marine species

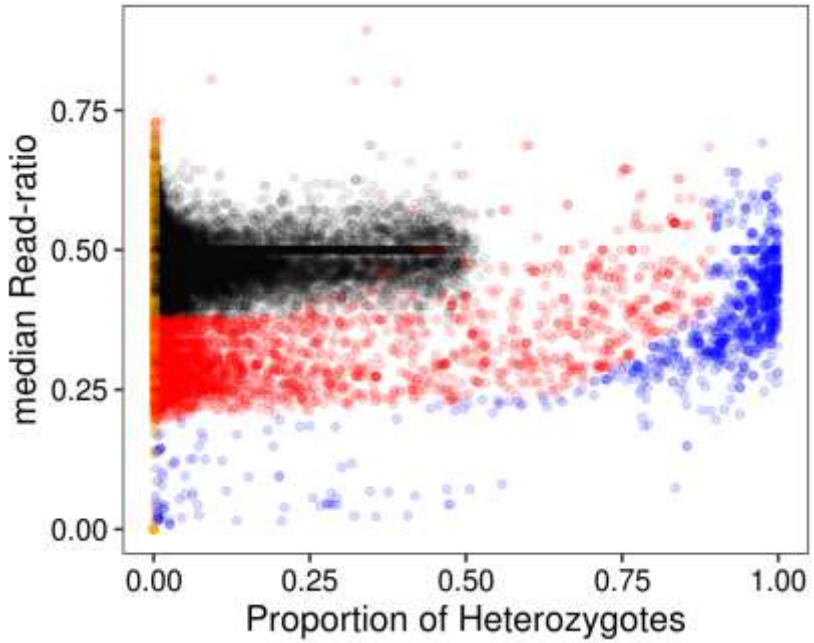
Yann Dorant[1] | Hugo Cayuela[1] | Kyle Wellband[1] | Martin Laporte[1] | Quentin Rougemont[1] | Claire Mérot[1] | Eric Normandeau[1] | Rémy Rochette[2] | Louis Bernatchez[1]

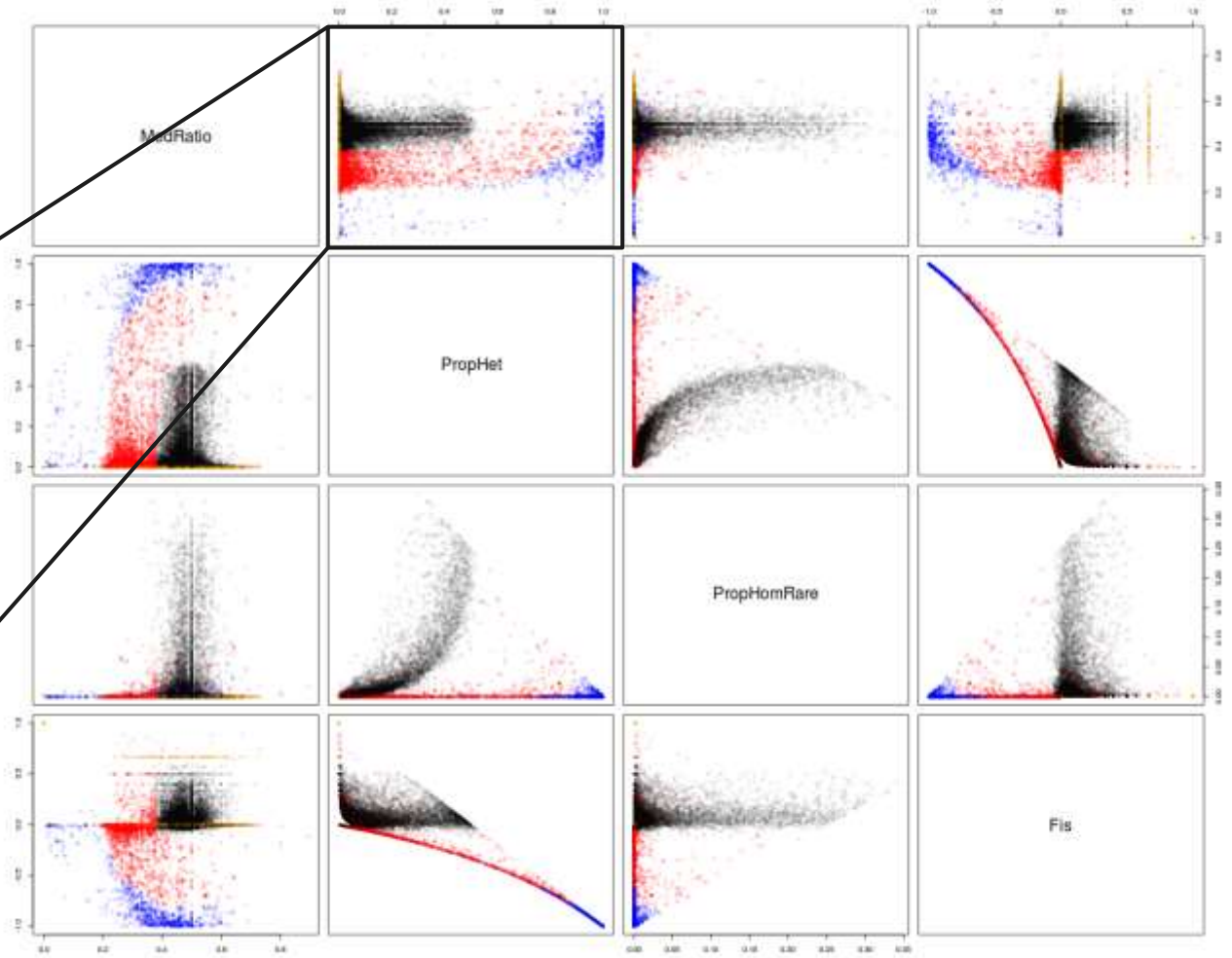


Statistical metrics from vcf file to assess duplicated loci



Therorical pattern (McKinney et al., 2017)





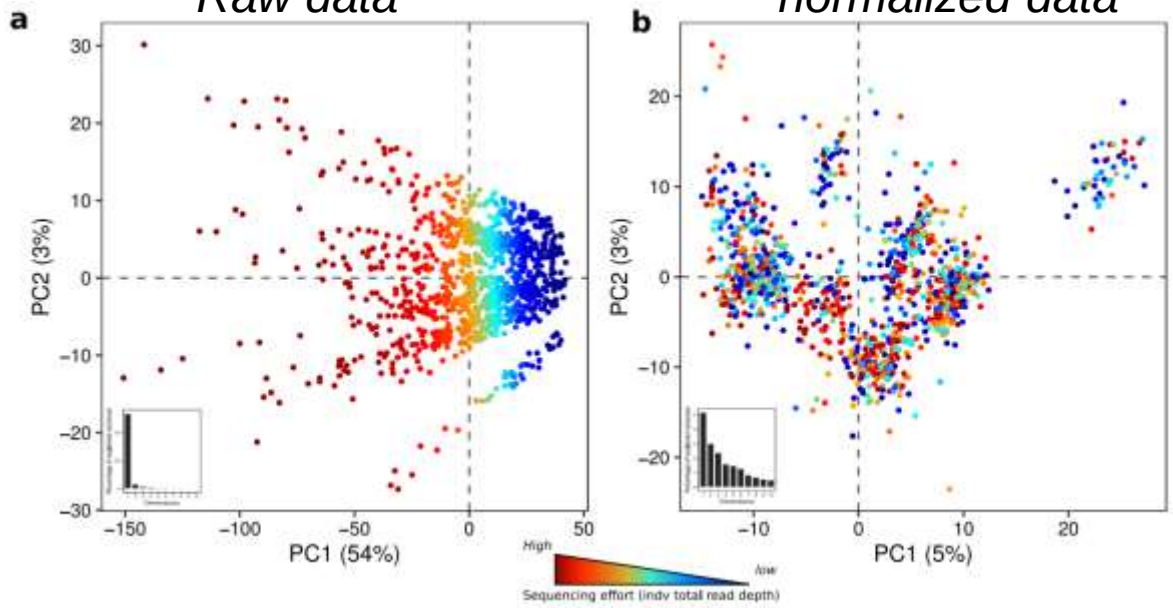
- Singletons
- Duplicated
- Duplicated & diverged
- Low confidence

Dorant et al., 2020

1. **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

- Basic structure with PCA
- Environment X CNVs associations (RDA, linear mixed models)