GWAS script JRL 2021

C.LE BOUAR ; C.MELOT ; L.RIZZI

20/06/2021

# Context

Creation of the directory place

knitr::opts\_chunk$set(echo = TRUE, cache.lazy = FALSE)  
  
#library("rstudioapi") # mettre la directory en place  
#setwd(dirname(getActiveDocumentContext()$path)) # Set working directory to source file location  
#getwd()

rm(list = ls())  
  
library(anyLib)  
anyLib(c("data.table", "apercu", "mlmm.gwas", "corpcor","plyr"))

## Loading required package: data.table

## Loading required package: apercu

## Loading required package: mlmm.gwas

## Loading required package: corpcor

## Loading required package: plyr

## data.table apercu mlmm.gwas corpcor plyr   
## TRUE TRUE TRUE TRUE TRUE

library(stringr)

# Introduction

The number of aphid mummies (terminal, living or both) are studied and the link with one or several SNP is observed.

## Model used

BLUEs and BLUPs of the first statistical analysis were used to make a GWAS analysis. Phenotypic values are modeled by this equation:

is the mean. Y contains phenotypic values. X gives the number of copies of the reference allele (for example A) . X could take the value 0 ( alleles aa) ; 1 (alleles aA) or 2 (alleles AA). 𝛽 is the regression slope that modifies the genotype according to the number of copies of the reference allele A. The genetic effect is modeled as fixed and additive. The null hypothesis H0 is 𝛽=0, which means that there is no genetic effect on phenotypic variation. Z is a design matrix which assigns a multigenic effect to each line. It is equivalent to the sum of all additive effects of other genes that affect the studied character. u is declared as a random effect, following a normal distribution of variance 2Kσ² where σ² is the additive polygenic variance of the character and K an apparent matrix between lines, calculated with markers. 𝜺 is the residual random of variance σ². A stepwise analysis is calculated and significative markers were successively added to see if 𝜺 is reduced. There are 96 000 markers but they are not all independant. With an estimation of the number of independent markers and by application of the Bonferroni correction, the threshold of 10^-5 is taken. First, a threshold for the allelic frequency is applied. Only allelic frequencies superior to 5% are kept. The apparent matrix K is calculated with the Van Raden method. The regression is calculated with the package mlmm.gwas of Rstudio. Manhattan plots are created and only the best model is kept. At the end, allelic effect and the number of each class (00, 01, 11) is calculated.

## Genotype data

### SNP location on chromosomes

Markers located on the wheat genome are loaded.

# physical positions of SNP on the Zavitan WEW2 version  
load("../data/GENOTYPES/BREEDWHEAT\_on\_durum\_physic\_WEW2.Rdata")  
BLAST[1,]

## Marker\_ID Chr Mstart Mend Chrlen start end v8 long simil  
## 1 AX-89309359 chr7A 1 71 71 742440137 742440067 747227478 71 97.18  
## evalue  
## 1 2e-25

Here it is possible to place 242 382 SNP on the wheat genom.

dim(BLAST)

## [1] 242382 11

# toutes les infos dispo sur les snp  
load("../data/GENOTYPES/caract\_SNP\_ALL\_PHYS.Rdata")

## Genotypic matrix

Lines was genotyped with a DNA chip on 420 000 SNP. But only a part of this data will be used. Usefull data are loaded here.

# breed wheat genotypes  
file="../data/GENOTYPES/SG\_EPO\_complet.Rdata"  
load(file)  
  
dim(SG)

## [1] 476 168725

There are 476 lines on 168 725 SNP

Then only the SNP with an assumed physical position are kept.

# optional Keep sNP only with an assumed physical position  
liste<-which(colnames(SG) %in% BLAST[,1])  
length(liste)

## [1] 112819

ap(liste)

## [1] 1 4 7 8 101

SG<-SG[,liste]  
dim(SG)

## [1] 476 112819

# Phenotypes

## Data

BLUPS and BLUE for total aphid mummies are loaded here

file1="../data/PHENOTYPES/BlueBlupS2.csv"  
  
myY <-read.table(file1, head = TRUE, sep=";", dec=",")  
  
names(myY)[1]<-"Taxa"  
  
dim(myY)

## [1] 184 3

myY[1:10, 1]

## [1] C495 ELAX\_100 ELAX\_101 ELAX\_104 ELAX\_114 ELAX\_117 ELAX\_118 ELAX\_120  
## [9] ELAX\_122 ELAX\_124  
## 184 Levels: C495 ELAX\_100 ELAX\_101 ELAX\_104 ELAX\_114 ELAX\_117 ... MariaB6R

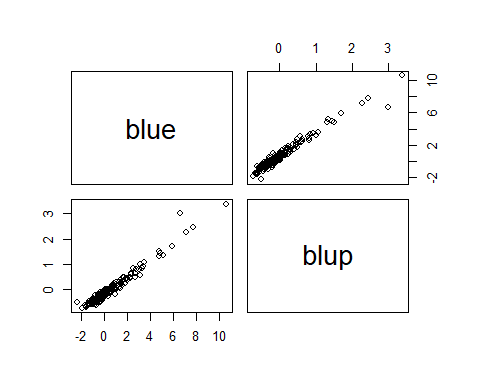
# modification of the data because of a spelling mistake.   
myY[,1]<-str\_replace(as.character(myY[,1]), "A", "4")

## Links between variable

names(myY)

## [1] "Taxa" "blue" "blup"

pairs(myY[,c(2,3)])



cor(myY[,c(2,3)])\*\*2

## blue blup  
## blue 1.0000000 0.9571475  
## blup 0.9571475 1.0000000

BLUE and BLUPS are correlated.

# Keep wheat lines only if phenotyped

Here, we only keep the wheat lines that were phenotyped on the experiment field.

SG<-SG[which(rownames(SG) %in% myY[,1]),]  
ap(SG)

## AX-89309359 AX-89309363 AX-89309376 AX-89309378 AX-89309726  
## EL4X\_240 "2" "2" "2" "2" "2"   
## EL4X\_356 "2" "2" "0" "2" "2"   
## EL4X\_476 "2" "2" "2" "2" "0"   
## EL4X\_241 "2" "2" "2" "2" "2"   
## EL4X\_16 "2" "2" "0" "2" "0"

dim(SG)

## [1] 181 112819

myY<-myY[which(myY[,1] %in% rownames(SG) ),]  
dim(myY)

## [1] 181 3

dim(SG)

## [1] 181 112819

To check of there is the same number of wheat lines in the selected genetic library (SG) as in our raw data (myY),the two number (dim(SG) and dim(myY)) have to be identical.

## Data pretreatment

genot<-SG  
  
class(genot)

## [1] "matrix"

dim(genot)

## [1] 181 112819

ap(genot)

## AX-89309359 AX-89309363 AX-89309376 AX-89309378 AX-89309726  
## EL4X\_240 "2" "2" "2" "2" "2"   
## EL4X\_356 "2" "2" "0" "2" "2"   
## EL4X\_476 "2" "2" "2" "2" "0"   
## EL4X\_241 "2" "2" "2" "2" "2"   
## EL4X\_16 "2" "2" "0" "2" "0"

### Imputation of Missing data

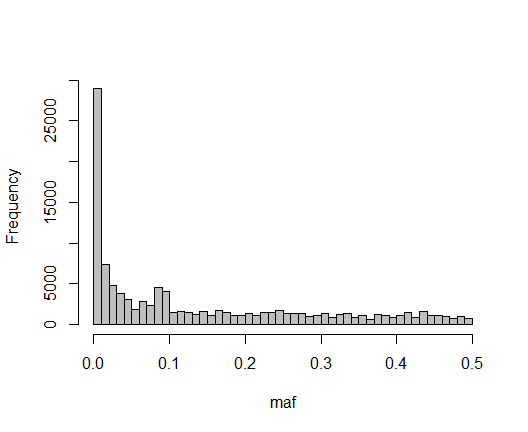
* simple binomial Imputation

noms<-rownames(genot)  
genot<-apply(genot,2,as.numeric)  
rownames(genot)<-noms  
  
genot.imp <- apply(genot, 2, function(x){  
 freq <- table(x)  
 x[is.na(x)] <- as.integer(names(which.max(freq)))  
 return(x)  
})

### Minimal allelic frequency

#Calcul of allelic frequency  
p <- colMeans(genot.imp) / 2  
q <- 1 - p

#Spatial representation of allelic frequency  
maf <- apply(cbind(p, q), 1, min)  
hist(maf, col = "grey", main = "", breaks = 50, xlim = c(0, 0.5))



Only allelic frequencies superior to 5% are kept.

sum(maf < 0.5)

## [1] 112764

genot.ok <- genot.imp[, maf >= 0.05]  
dim(genot.ok)

## [1] 181 64713

Only 64 713 markers are kept.

### Physical map

Creation of the physical map

map<-BLAST[which(BLAST[,1] %in% colnames(genot.ok)),c(1,2,6)]  
names(map)<-c("SNP","Chr","Pos")  
head(map)

## SNP Chr Pos  
## 8 AX-89309369 chr4B 631948705  
## 10 AX-89309371 chr1B 43228153  
## 18 AX-89309376 chr7A 605815984  
## 27 AX-89309722 chr5A 578939825  
## 32 AX-89309726 chr2A 7484084  
## 124 AX-89309734 chr3B 762124628

Only selected SNP are kept and chromosome are called by number between 1 and 14.

SNP are classified according to their location on chromosome

map <- map[order(map$Pos), ]  
map <- map[order(map$Chr), ]  
head(map)

## SNP Chr Pos  
## 524641 AX-89519629 chr1A 45012  
## 1071353 AX-89726564 chr1A 98264  
## 345944 AX-89445498 chr1A 100862  
## 247647 AX-89400264 chr1A 151786  
## 828617 AX-89635280 chr1A 152065  
## 303025 AX-89423540 chr1A 152603

tail(map)

## SNP Chr Pos  
## 614821 AX-89556573 chr7B 777503232  
## 452971 AX-89489367 chr7B 777547462  
## 957509 AX-89680733 chr7B 777547777  
## 713797 AX-89591720 chr7B 777548581  
## 1019901 AX-89707787 chr7B 777549644  
## 941809 AX-89673999 chr7B 777552033

#clean the memory  
rm(genot, genot.imp, maf, p, q)

# Choice of the analysed variable

## Relation exploration

The number 3 is choosen to take the BLUPs variables.

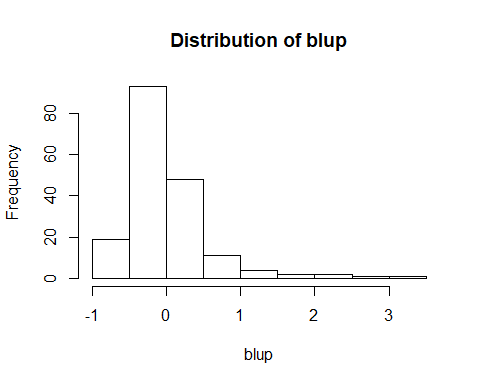
names(myY)

## [1] "Taxa" "blue" "blup"

#variable name i<-1  
# BLUP i<-3  
# BLUE i<-2  
i<-3

## Creation of the phenotype vector

hist(myY[,i], main=paste("Distribution of", names(myY)[i]), xlab=colnames(myY)[i])



y <- myY[,i]  
names(y) <- myY$Taxa  
summary(y)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## -0.714156 -0.355857 -0.154154 -0.001829 0.126770 3.370548

# Data merge and vizualisation

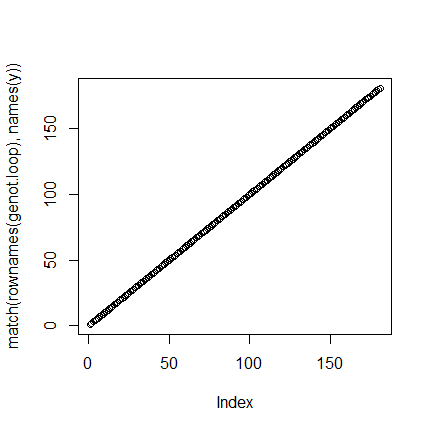
## Data check

# genotype are put in the same order as the list of phenotypes in myY.   
genot.ok<-genot.ok[order(rownames(genot.ok)),]  
y<-y[order(names(y))]  
genot.loop<-genot.ok  
  
# if they are missing data  
if (length(which(is.na(y)))>0 ) {   
 liste<- which(is.na(y))  
 y <- y[-liste]   
 genot.loop<-genot.loop[-liste,]  
 #rownames(genot.loop)<-rownames(genot.ok)[-which(is.na(y))]  
 }  
  
dim(genot.loop)

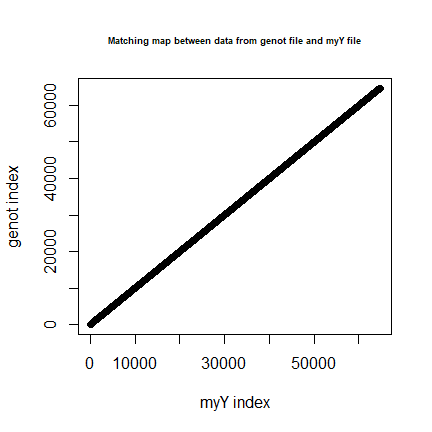
## [1] 181 64713

The two files must be in the same order and must create a straight line.

plot(match(rownames(genot.loop), names(y)))



# markers are putting in the same order in the map  
  
genot.loop<-genot.loop[,map$SNP]  
plot(match(map$SNP, colnames(genot.loop)), main= "Matching map between data from genot file and myY file", cex.main=0.55, xlab="myY index", ylab="genot index")



## Kinship matrix calcul

### K Van Raden

The apparent matrix of the model is calculated. The Van Raden method is used.

# apparent matrix calcul  
p <- colMeans(genot.loop) / 2  
q <- 1 - p  
  
# The Van Raden matrix is centred  
genot.scaled <- scale(genot.loop, center = 2 \* p, scale = sqrt(2 \* p \* q))  
  
K <- tcrossprod(genot.scaled) / ncol(genot.scaled)  
K <- make.positive.definite(K)

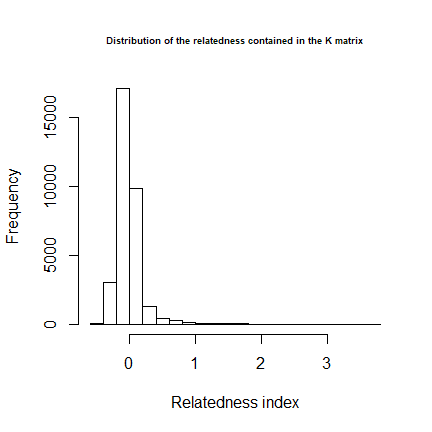
The K matrix gives the relation between two wheat lines.

#dimension of the matrix  
dim(K)

## [1] 181 181

Here is the distribution of the values. Most lines have very low relatedness.

hist(K, main="Distribution of the relatedness contained in the K matrix", cex.main=0.55, xlab="Relatedness index")



# Regression

The package mlmm.gwas. proposed by V.Segura and his co-workers is used. ## package mlmm.gwas

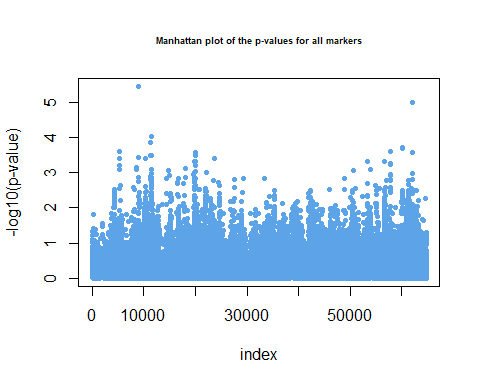
# from https://cran.r-project.org/web/packages/mlmm.gwas/vignettes/gwas-manual.html  
  
# genotypes must be centered  
genot.scaled <- scale(genot.loop, center = 2 \* p)   
  
#In the function, y is the analysed phenotype ; K the Kinship matrix, genot.scaled is the matrix with X values (0,1 or 2) ; the threshold value controls the risk level; maxsteps gives the number of cofactor  
  
mygwas.gwas <- mlmm\_allmodels(y, list(genot.scaled), list(K),   
 maxsteps = 4, threshold=1e-4)

## iteration LogLik wall cpu(sec) restrained  
## 1 -89.7162 9:28:39 0 0  
## 2 -89.6008 9:28:39 0 0  
## 3 -89.5484 9:28:39 0 0  
## 4 -89.5382 9:28:39 0 0  
## 5 -89.5376 9:28:39 0 0  
## null model done! pseudo-h= 0.092   
## iteration LogLik wall cpu(sec) restrained  
## 1 -79.3956 9:28:48 0 0  
## 2 -78.6903 9:28:48 0 0  
## 3 -78.5447 9:28:49 1 0  
## 4 -78.543 9:28:49 1 0  
## 5 -78.5429 9:28:49 1 0  
## step 1 done! pseudo-h= 0.036 model: Y ~ mu + AX\_89312681   
## iteration LogLik wall cpu(sec) restrained  
## 1 -71.3066 9:28:58 0 0  
## 2 -69.7149 9:28:58 0 1  
## 3 -69.323 9:28:58 0 1  
## 4 -69.323 9:28:58 0 1  
## step 2 done! pseudo-h= 0 model: Y ~ mu + AX\_89312681 + AX\_89346847

res\_mlmm <- mygwas.gwas

P-values are calculated for the 4 models and for all markers. A graph of -log10(Pvalues) for all markers is created.

# manhattan plot without map  
manhattan.plot(res\_mlmm, main="Manhattan plot of the p-values for all markers", cex.main=0.55)



Here, two markers are above the threshold.

## Best model selection

Then the best model is selected

# model selection  
sel\_XX = frommlmm\_toebic(list(genot.scaled), res\_mlmm)  
res\_eBIC = eBIC\_allmodels(y, sel\_XX, list(K), ncol(genot.scaled))  
  
res\_eBIC

## BIC ajout eBIC\_0.75 LogL  
## mu 527.2261 0.00000 527.2261 -255.8153  
## AX-89312681 510.2779 16.61658 526.8945 -244.7419  
## AX-89346847 496.6108 32.19341 528.8043 -235.3092

The best model is the one with the smallest eBIC. Here the model with the marker AX-89312681 has the smallest eBIC. It is significative.

The same significative results are obtained with the analysis of BLUES

## Best model selection under a fixed treshold

The threshold is fixed at 1e-5:

res\_threshold <- threshold\_allmodels(threshold=1e-5, res\_mlmm)

## Warning in rbind(res, c(names(tab), as.vector(tab), (i - 1))): number of columns  
## of result is not a multiple of vector length (arg 2)

## [1] "No p-value below the threshold at step 2"

res\_threshold

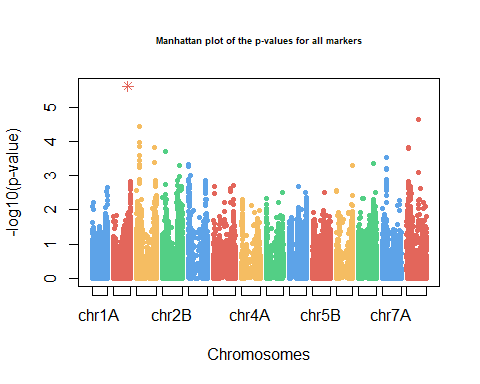
## SNP p-value MLMM\_Step  
## 1 AX-89312681 AX-89493607 3.57574485007076e-06

The SNP AX-89312681 is the only one selected after we fixed the treshold at 1e-5.

## Manhattan plot with chromosom mapping

A manhattan plot map showing the distribution of p-values on the genome is created.

mip<-map  
mip$Chr<-sprintf("%s",mip$Chr)  
  
manhattan.plot(res\_mlmm, map = mip, steps = 2, hideCofactors = FALSE, chrToPlot = "all", unit = "bp", main="Manhattan plot of the p-values for all markers", cex.main=0.55)

 There is a peak on the chromosome 1B, for the SNP AX-89312681.

# Visualization of allelic effects

The allelic effect (slope value) is calculated

sel\_XXclass = fromeBICtoEstimation(sel\_XX, res\_eBIC, res\_threshold)  
  
effects = Estimation\_allmodels(y, sel\_XXclass, list(K))  
effects

## BLUE Tukey.Class Frequency  
## AX-89312681\_00 0.92329878 a 0.05524862  
## AX-89312681\_11 0.00000000 b 0.94475138  
## mu -0.05283973 <NA> NA

At the locus AX-89312681, the genotype 00(aa) has a difference of 0.9232 aphid mummies compared to the genotype 11(AA).

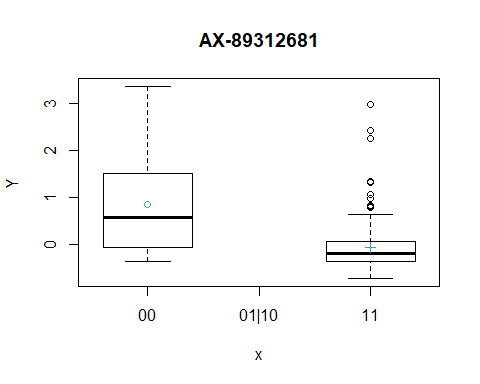
# Informations on the SNP selected

#Identification of the SNP  
mark\_list<-names(sel\_XXclass)  
mark\_list[1]

## [1] "AX-89312681"

A boxplot showing the difference between the two homozygous is created.

# m is the number of the selected marker in the list  
m<-1  
genotypes.boxplot(genot.scaled, y , mark\_list[m], effects)



The frequence of each class is calculated

table(genot.loop[,mark\_list[m]])

##   
## 0 2   
## 10 171

Allele 00 with more aphid mummies is uncommon.

map[which(map$SNP == mark\_list[m]),]

## SNP Chr Pos  
## 12000 AX-89312681 chr1B 594671851

## Analysis of close markers

C<-cor(genot.ok[,mark\_list[m]], genot.ok)\*\*2  
  
liste<-which(C>0.8)  
liste<-colnames(C)[liste]  
  
liste<-map[which(map$SNP %in% liste),]  
liste

## SNP Chr Pos  
## 12000 AX-89312681 chr1B 594671851  
## 461422 AX-89493607 chr1B 595787838

## Effect of close loci

genotypes.boxplot(genot.scaled, y , liste[2,1], effects)

