Script "GWAS - Analysis of association of genotype and phenotype –

Basic – Tanzania 22-27 january 2024"

Written by Y Vigouroux

Data samples of plants from study of pearl millet (Nat Biotech 2017), we restricted to SNP data on CHR5 and plant height traits in three conditions CONTROL (CONTROL_PH)/early drought stress (EARLY_PH)/late drought stress (LATE_PH)

We will only analyzed the CONTROL_PH flowering time. A total of 4 groups (populations) are considered: EARLY / LATE 4 and on a single traits/field trial YPP (yield per plot).

Installation of package

```
Put your working direction in the folder “TD_GWAS_basic”


Example:

base_dir <- getwd()
setwd("base_dir/TD_GWAS_basic")
```

```
#CODE

####Code for GWAS using a simple AOV analysis of variance

####Put a link to read genotypic data of chromosome 5
#### Read Phenotype data
inputFile <- "chr5_5association.txt"
outputFile2<- "resultat_CHR5.txt"
Genotypes  <- file(inputFile, open = "r")

Phenotype<-read.table("blup_PH.txt",header=TRUE)


### Visualisation of phenotypes

Phenotype
head(Phenotype)
attach(Phenotype)
Control_PH
Early_PH
Late_PH

hist(Control_PH)
plot(Control_PH~Early_PH)
plot(Control_PH~Late_PH)



### Analysis of a single SNP
# We read the first line of the file
oneLine <- readLines(Genotypes, n = 1, warn = FALSE)

#We modify the line to have the different genotype
oneLine<-unlist(strsplit(oneLine, split=" "))
oneLine<-unlist(strsplit(oneLine, split="\t"))
oneLine[ oneLine == "NA" ] = NA

#We performed an analysis of variance explaining the phenotype CONTROL_PH
#We control by the effect of population, 4 groups/populations: EA, ME, LA, ME
res2<-aov(Control_PH~Population+oneLine[4:253])

#Show the result of the analysis of variance
res2
summary(res2)

#Retrieve the pvalue
summary(res2)[[1]][["Pr(>F)"]][[2]]
```

```
### Do the analysis for each line of the file
l <- 0
while (length(oneLine <- readLines(Genotypes, n = 1, warn = FALSE)) > 0) {

  oneLine<-unlist(strsplit(oneLine, split=" "))
  oneLine<-unlist(strsplit(oneLine, split="\t"))
  if((length(which(oneLine[] == "NA"))/250)<=0.20) {
       oneLine[ oneLine == "NA" ] = NA
       resultat2<-""

       res2<-aov(Control_PH~Population+oneLine[4:253])
       resultat2<-c(resultat2,summary(res2)[[1]][["Pr(>F)"]][[2]])


write.table( t(c(oneLine[1:3],resultat2[])) , file=outputFile2, sep="\t",
quote=FALSE, col.names=F, row.names=F, append=TRUE)

}
}

data<-read.table("resultat_CHR5.txt",header=F)



### Data graphic between position pvalue
attach(data)
plot(V4~V2)
plot(-log10(V4)~V2)


#qq plot

nb<-length(V4)
pvalueexpected<-c(1:nb)/nb

plot(sort((V4))~sort((pvalueexpected)), main="QQplot", xlab="Expected",
ylab="Observed")



plot(sort(-log10(V4))~sort(-log10(pvalueexpected)), main="QQplot log scale",
xlab="Expected", ylab="Observed")
```