LD_SNPS_RESISTIN

KIRABO GLORIA

2023-05-17

```r
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.2 --
## v ggplot2 3.4.0      v purrr   1.0.1
## v tibble  3.1.8      v dplyr   1.0.10
## v tidyr   1.2.1      v stringr 1.5.0
## v readr   2.1.3      v forcats 0.5.2
```

```
## Warning: package 'stringr' was built under R version 4.2.3
```

```
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```r
library(pheatmap)
```

```
## Warning: package 'pheatmap' was built under R version 4.2.3
```

```r
library(haplo.stats)
```

```
## Warning: package 'haplo.stats' was built under R version 4.2.3
```

```
## Loading required package: arsenal
```

```
## Warning: package 'arsenal' was built under R version 4.2.3
```

```r
library(dplyr)
library(magrittr)
```

```
##
## Attaching package: 'magrittr'
##
## The following object is masked from 'package:arsenal':
##
##     set_attr
##
## The following object is masked from 'package:purrr':
##
```

```
##     set_names
##
## The following object is masked from 'package:tidyr':
##
##     extract
```

```
library(genetics)
```

```
## Warning: package 'genetics' was built under R version 4.2.3
```

```
## Loading required package: combinat
##
## Attaching package: 'combinat'
##
## The following object is masked from 'package:utils':
##
##     combn
##
## Loading required package: gdata
```

```
## Warning in system(cmd, intern = intern, wait = wait | intern,
## show.output.on.console = wait, : running command 'C:\WINDOWS\system32\cmd.exe /c
## ftype perl' had status 2
```

```
## Warning in system(cmd, intern = intern, wait = wait | intern,
## show.output.on.console = wait, : running command 'C:\WINDOWS\system32\cmd.exe /c
## ftype perl' had status 2
```

```
## gdata: read.xls support for 'XLS' (Excel 97-2004) files ENABLED.
##
## gdata: Unable to load perl libaries needed by read.xls()
## gdata: to support 'XLSX' (Excel 2007+) files.
##
## gdata: Run the function 'installXLSXsupport()'
## gdata: to automatically download and install the perl
## gdata: libaries needed to support Excel XLS and XLSX formats.
##
## Attaching package: 'gdata'
##
## The following objects are masked from 'package:dplyr':
##
##     combine, first, last
##
## The following object is masked from 'package:purrr':
##
##     keep
##
## The following object is masked from 'package:stats':
##
##     nobs
##
## The following object is masked from 'package:utils':
```

```
##
##     object.size
##
## The following object is masked from 'package:base':
##
##     startsWith
##
## Loading required package: gtools

## Warning: package 'gtools' was built under R version 4.2.3

## Loading required package: MASS

## Warning: package 'MASS' was built under R version 4.2.3

##
## Attaching package: 'MASS'
##
## The following object is masked from 'package:dplyr':
##
##     select
##
## Loading required package: mvtnorm
##
##
## NOTE: THIS PACKAGE IS NOW OBSOLETE.
##
##
##
##   The R-Genetics project has developed an set of enhanced genetics
##
##   packages to replace 'genetics'. Please visit the project homepage
##
##   at http://rgenetics.org for informtion.
##
##
##
##
## Attaching package: 'genetics'
##
## The following object is masked from 'package:haplo.stats':
##
##     locus
##
## The following objects are masked from 'package:base':
##
##     %in%, as.factor, order
```

```
##STEP ONE-READING IN THE DATASET FROM THE LINK
fams <- read.delim("http://www.biostat.umn.edu/~cavanr/FMS_data.txt",header = T,
               sep = "\t")
#Vieing the first five rows and 10columns of the dataset

fams[1:5,1:10]
```

```
##        id acdc_rs1501299 ace_id actn3_r577x actn3_rs540874 actn3_rs1815739
## 1 FA-1801             CA     DD          CC             GG              CC
## 2 FA-1802             CA     ID          CT             GA              TC
## 3 FA-1803             CA     ID          CT             GA              TC
## 4 FA-1804             CC     DD          CT             GA              TC
## 5 FA-1805             CA     ID          CC             GG              CC
##   actn3_1671064 ardb1_1801253 adrb2_1042713 adrb2_1042714
## 1            AA          <NA>            GA            CG
## 2            GA          <NA>            GA            CC
## 3            GA          <NA>            GA            CG
## 4            GA          <NA>            AA            CC
## 5            AA          <NA>            GA            CG
```

```
##SELECTING OUT COLUMNS WITH RESISTIN GENES

fams_restn <- fams[grepl("^resistin", names(fams))]
head(fams_restn)
```

```
##   resistin_c30t resistin_c398t resistin_g540a resistin_c980g resistin_c180g
## 1            CC             CC             GG             GG             CC
## 2            CC             TT             AA             CG             GG
## 3            CC             CC             GG             CG             CC
## 4            CC             CC             GA             CG             CG
## 5            CT             CC             GG             CC             CC
## 6            CC             CC             GG             CG             CC
##   resistin_a537c
## 1             AA
## 2             AA
## 3             AA
## 4             AA
## 5             AA
## 6             AA
```

```
##Counting the number of generated columns

restn_col <- ncol(fams_restn)
cat("The columns with resistin gene are:", restn_col)
```

```
## The columns with resistin gene are: 6
```

```
##IDENTIFYING UNIQUE SNPS IN THE COLUMNS

##STEP ONE CREATE A VECTOR
restn_snp <- unlist(fams_restn)
head(restn_snp)
```

```
## resistin_c30t1 resistin_c30t2 resistin_c30t3 resistin_c30t4 resistin_c30t5
##           "CC"           "CC"           "CC"           "CC"           "CT"
## resistin_c30t6
##           "CC"
```

```r
##STEP TWO:IDENTIFYING UNIQUE VALUES IN THE VCECTOR
num_snps_restn <- length(unique(restn_snp))
cat("The total number of snps in the restitin genes is :",num_snps_restn)
```

```
## The total number of snps in the restitin genes is : 9
```

```r
##CREATING A GENOTYPE OBJECT FOR OUR GENES OF THE RESISTIN GENE

geno_restn <- as.data.frame(lapply(fams_restn, genotype, sep=""))
geno_restn[1:5,]
```

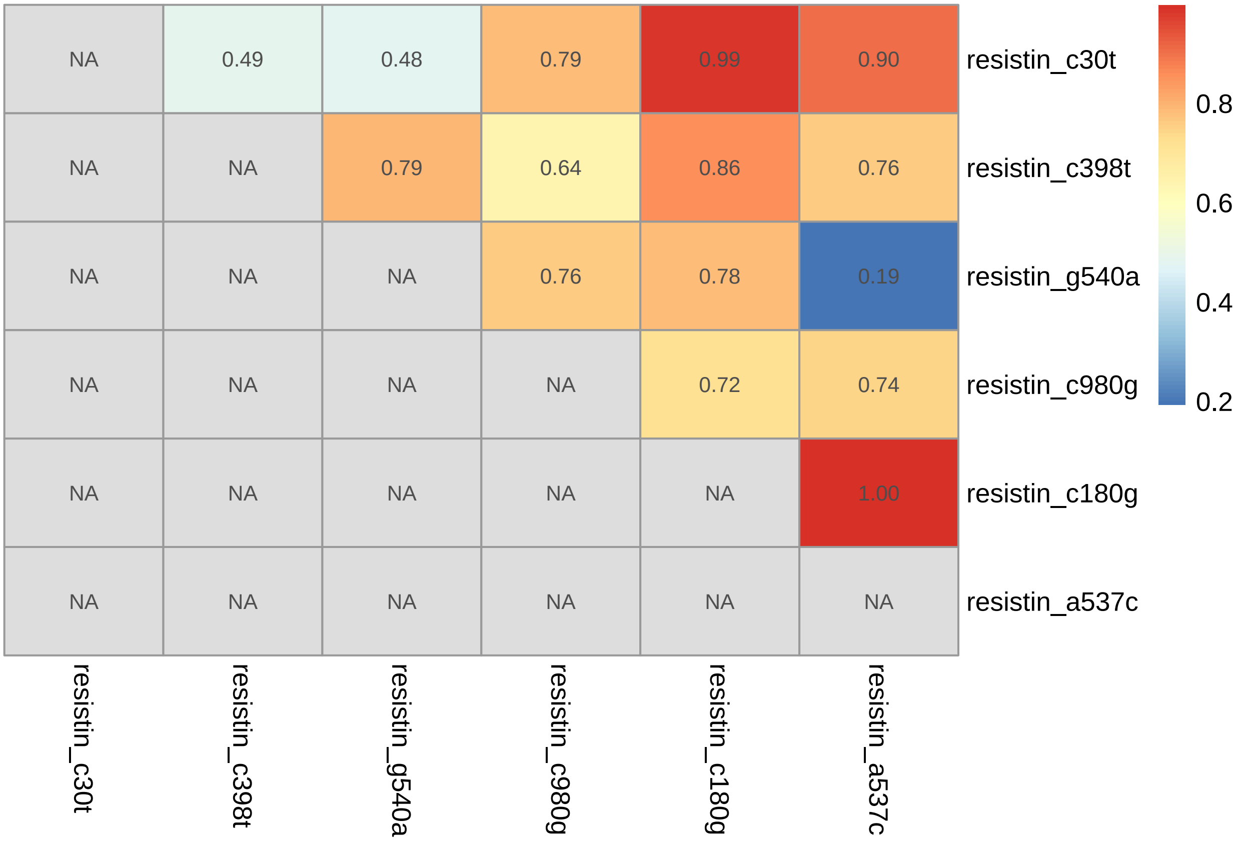
```
##   resistin_c30t resistin_c398t resistin_g540a resistin_c980g resistin_c180g
## 1           C/C            C/C            G/G            G/G            C/C
## 2           C/C            T/T            A/A            C/G            G/G
## 3           C/C            C/C            G/G            C/G            C/C
## 4           C/C            C/C            G/A            C/G            C/G
## 5           C/T            C/C            G/G            C/C            C/C
##   resistin_a537c
## 1            A/A
## 2            A/A
## 3            A/A
## 4            A/A
## 5            A/A
```

```r
##Calculating D' OF THE SNPS IN THE RESISTIN GENE

restn_D <- LD(geno_restn)$`D'`
restn_D
```

```
##                resistin_c30t resistin_c398t resistin_g540a resistin_c980g
## resistin_c30t             NA      0.4855449      0.4810285      0.7877728
## resistin_c398t            NA             NA      0.7944463      0.6403467
## resistin_g540a            NA             NA             NA      0.7586427
## resistin_c980g            NA             NA             NA             NA
## resistin_c180g            NA             NA             NA             NA
## resistin_a537c            NA             NA             NA             NA
##                resistin_c180g resistin_a537c
## resistin_c30t       0.9850864      0.9021448
## resistin_c398t      0.8581316      0.7633220
## resistin_g540a      0.7833792      0.1929656
## resistin_c980g      0.7183456      0.7441364
## resistin_c180g             NA      0.9983006
## resistin_a537c             NA             NA
```

```r
##GENERATING A HEATMAP TO SHOW THE D' OF SNPS
pheatmap(restn_D, cluster_cols = FALSE, cluster_rows = FALSE, display_numbers = TRUE)
```



```
##From the heatmap, it shows that the SNP at resistin_c180g is in high LD with resistin_a537c
## Also resistin_c30t is in high LD with resistin_c180g.




###CALCULATING OF HARDY WEINBERG EQUILLIBRIUM(USING CHISQUARE AND FISCHER'S EXACT TEST)

#chi_pval <- vector() #create empty vector

#for (col in geno_restn){
  #chiq <- HWE.chisq(col)
#  chi_pval <- c(chi_pval,chisq$p.value)
#}
#sort(chi_pval)

##naming p-values
#names(chi_pval)=colnames(geno_restn) ##assigning names with those in the genotype object
 #sort(chi_pval)

#sum(chi_pval<0.05)##finding total pvalues <0.5

#names(chi_pval[chi_pval<0.05])##getting names of columns with pvalues < 0.5


##CALCULATING THE FISCHER'S EXACT
```

```
ext_pval <- vector()

for (col in geno_restn){
  exact <- HWE.exact(col)
  ext_pval <- c(ext_pval,exact$p.value)
}
sort(ext_pval)
```

```
## [1] 0.1166305 0.3153145 0.4105245 1.0000000 1.0000000 1.0000000
```

```
sum(ext_pval<0.05)
```

```
## [1] 0
```

```
names(ext_pval)=colnames(geno_restn)
sort(ext_pval)
```

```
## resistin_g540a resistin_c980g resistin_c180g resistin_a537c  resistin_c30t
##      0.1166305      0.3153145      0.4105245      1.0000000      1.0000000
## resistin_c398t
##      1.0000000
```

```
names(ext_pval[ext_pval<0.05])
```

```
## character(0)
```

```
##Adjusting using Bonferoni to cater for multiple testing

##Adjusting chi-square values
#set.seed(100)
#adj_pva <- p.adjust(chi_pval, method ="bonferroni")
#sort(adj_pva)
#sum(adj_pva<0.05)
#names(adj_pva[adj_pva<0.05])

##Adjusting exact p-values
set.seed(42)#for reproducibility
aj_val <- p.adjust(ext_pval,method = "bonferroni")
sort(aj_val)
```

```
## resistin_g540a  resistin_c30t resistin_c398t resistin_c980g resistin_c180g
##       0.699783       1.000000       1.000000       1.000000       1.000000
## resistin_a537c
##       1.000000
```

```
sum(aj_val<0.05)
```

```
## [1] 0
```

```
names(aj_val[aj_val<0.05])
```

```
## character(0)
```

```
##CALCULATING MINOR ALLELE FREQUENCY(MAF) OF SNPS

##We first identify missing values

miss_gen <- data.frame(summary(is.na(geno_restn))[3,])

names(miss_gen) <- c("Missing values")
miss_gen
```

```
##                   Missing values
## resistin_c30t  TRUE :664
## resistin_c398t TRUE :662
## resistin_g540a TRUE :661
## resistin_c980g TRUE :660
## resistin_c180g TRUE :658
## resistin_a537c TRUE :657
```

```
##Regardless, we take it that our allele frequencies will remain the same even if we had got the missin


##Calculating the MINOR ALLELE FREQUENCY(MAF) FOR AT EACH SNP

round(summary(geno_restn$resistin_c30t)$"allele.freq",1)
```

```
##    Count Proportion
## C   1445          1
## T     21          0
## NA  1328         NA
```

```
round(summary(geno_restn$resistin_c398t)$"allele.freq",1)
```

```
##    Count Proportion
## C   1165        0.8
## T    305        0.2
## NA  1324         NA
```

```
round(summary(geno_restn$resistin_g540a)$"allele.freq",1)
```

```
##    Count Proportion
## G   1025        0.7
## A    447        0.3
## NA  1322         NA
```

```
##MAF gives us a picture of frequency of a variation within a population
```

Note that the `echo = FALSE` parameter was added to the code chunk to prevent printing of the R code that generated the plot.