Practical 02 SG: Hardy-Weinberg equilibrium

Quim Aguado, Marcel Cases

27-Nov-2021

Ex 1 Load *TSIChr22v4* data into R using `fread`. Ignore the first 6 columns.

```
filename <- "TSIChr22v4.raw"
dt <- fread(filename, header=TRUE) # class data.table
dt <- dt[, 7:ncol(dt)]
#dt <- dt[, 7:10006] # treballem amb un sample
#df <- as.data.frame(dt) # transform data.table into data.frame to work more efficiently
```

Ex 2 How many individuals does the database contain, and how many variants? What percentage of the variants is monomorphic? Remove all monomorphic SNPs from the database. How many variants remain in the database?

```
n <- nrow(dt) # number of individuals
n
```

```
## [1] 107
```

```
p <- ncol(dt) # number of variants
p
```

```
## [1] 1102156
```

The original database contains 107 individuals and 1102156 variants.

```
dt_poly = dt %>% # remove monomorphics efficiently
  select(where(~n_distinct(.) > 1))

#monom <- (dt[,2]==0 & dt[,1]==0) | (dt[,2]==0 & dt[,3]==0)
#dt_poly <- dt[!monom,] # exclude monomorphics

monomorphic.num <- ncol(dt)-ncol(dt_poly)
perc.monomorphic <- 100*monomorphic.num/(ncol(dt))
perc.monomorphic
```

```
## [1] 81.03045
```

```
ncol(dt_poly) #remaining (polymorphic) variants
```

```
## [1] 209074
```

The 81.03% of the original database contain monomorphic variants. After removing them, the database contains 209074 variants.

At this time, we recode (0,1,2) into (AA,AB,BB).

```
dt_poly <- as.data.table(lapply(dt_poly, function(x){replace(x, x == 0, "AA")}))
dt_poly <- as.data.table(lapply(dt_poly, function(x){replace(x, x == 1, "AB")}))
```

```
dt_poly <- as.data.table(lapply(dt_poly, function(x){replace(x, x == 2, "BB")}))
dt_poly <- as.data.frame(dt_poly)
```

Ex 3 Extract polymorphism rs587756191_T from the datamatrix, and determine its genotype counts. Apply a chi-square test for Hardy-Weinberg equilibrium, with and without continuity correction. Also try an exact test, and a permutation test. You can use function HWChisq, HWExact and HWPerm for this purpose. Do you think this variant is in equilibrium? Argue your answer.

```
rs587756191_T <- dt_poly[,c("rs587756191_T")]
rs587756191_T.count <- c( sum(rs587756191_T=="AA"),
                          sum(rs587756191_T=="AB"),
                          sum(rs587756191_T=="BB")
                          )
names(rs587756191_T.count) <- c("AA", "AB", "BB")
rs587756191_T.count
```

```
##  AA  AB  BB
## 106   1   0
```

Genotype counts: * AA -> 106 * AB -> 1

```
results.chi <- HWChisq(rs587756191_T.count)
```

```
## Warning in HWChisq(rs587756191_T.count): Expected counts below 5: chi-square
## approximation may be incorrect
```

```
## Chi-square test with continuity correction for Hardy-Weinberg equilibrium (autosomal)
## Chi2 =  106.2512 DF =  1 p-value =  6.495738e-25 D =  0.002336449 f =  -0.004694836
```

```
results.chi.nocor <- HWChisq(rs587756191_T.count,cc=0) # test without correction
```

```
## Warning in HWChisq(rs587756191_T.count, cc = 0): Expected counts below 5: chi-
## square approximation may be incorrect
```

```
## Chi-square test for Hardy-Weinberg equilibrium (autosomal)
## Chi2 =  0.002358439 DF =  1 p-value =  0.961267 D =  0.002336449 f =  -0.004694836
```

```
results.exact <- HWExact(rs587756191_T.count)
```

```
## Haldane Exact test for Hardy-Weinberg equilibrium (autosomal)
## using SELOME p-value
## sample counts: nAA =  106 nAB =  1 nBB =  0
## H0: HWE (D==0), H1: D <> 0
## D =  0.002336449 p-value =  1
```

```
results.perm <- HWPerm(rs587756191_T.count)
```

```
## Permutation test for Hardy-Weinberg equilibrium
## Observed statistic: 0.002358439   17000 permutations. p-value: 1
```

```
#HWAlltests(x,include.permutation.test=TRUE) # consistent values -> cannot reject equilibrium
```

We cannot reject equilibrium of the variant rs587756191_T, given that all the tests (except for HWChisq without correction) compute a p-value around the same value.

Ex 4 Determine the genotype counts for all these variants, and store them in a p $\times$ 3 matrix.

```
geno.matrix <- matrix(nrow=ncol(dt_poly), ncol=3)

for(i in 1:ncol(dt_poly)) {
  geno.matrix[i,1] <- sum(dt_poly[,i]=="AA")
```

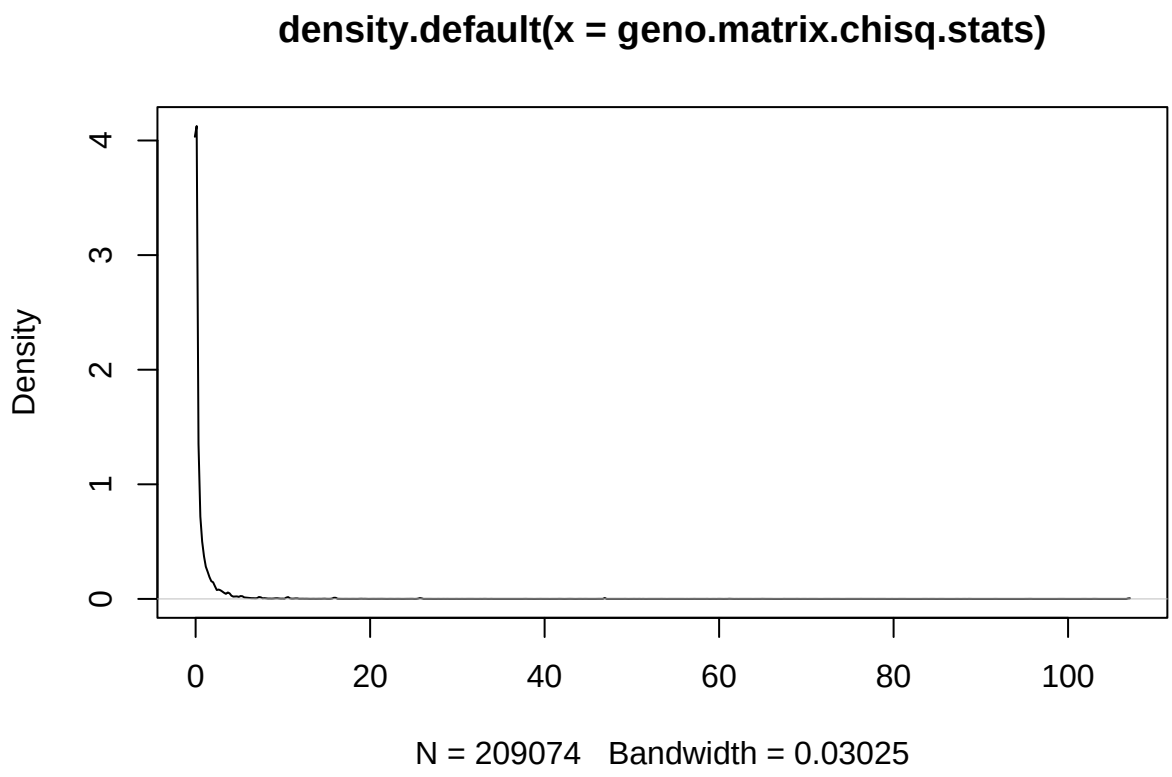
```
  geno.matrix[i,2] <- sum(dt_poly[,i]=="AB")
  geno.matrix[i,3] <- sum(dt_poly[,i]=="BB")
}

#geno.matrix
```

Ex 5 Apply a chi-square test without continuity correction for Hardy-Weinberg equilibrium to each SNP. You can use HWChisqStats for this purpose. How many SNPs are significant (use alpha = 0.05)?

```
geno.matrix.chisq.stats <- HWChisqStats(geno.matrix,pvalues = FALSE)
geno.matrix.chisq.pval <- HWChisqStats(geno.matrix,pvalues = TRUE)
significant.snp.num.chisq <- sum(geno.matrix.chisq.pval<0.05) # number of significant SNPs
significant.snp.num.chisq
```

```
## [1] 8162
```

```
plot(density(geno.matrix.chisq.stats))
```



8152 SNPs are significant.

Ex 6 How many markers of the remaining non-monomorphic markers would you expect to be out of equilibrium by the effect of chance alone?

Ex 7 Which SNP is most significant according to the chi-square test results? Give its genotype counts. In which sense is this genotypic composition unusual?

```
geno.matrix.chisq.most <- max(geno.matrix.chisq.pval) # most significant SNP
geno.matrix.chisq.most
```

```
## [1] 0.9949094
```

```
chisq.index <- which(geno.matrix.chisq.pval==max(geno.matrix.chisq.pval)) # index of the most significa
names(dt_poly[chisq.index[1]]) # SNP name
```

```
## [1] "rs5748532_T"
```

```
geno.matrix[chisq.index[1],] # counts of the most significant SNP
```

```
## [1] 32 53 22
```

```
#...
```

```
sum(geno.matrix.chisq.pval==max(geno.matrix.chisq.pval)) # repetitions of the most significant SNPs
```

```
## [1] 117
```

This genotypic composition is unusual because it only appears 117 times in the whole the dataset.

Ex 8 Apply an Exact test for Hardy-Weinberg equilibrium to each SNP. You can use function HWExactStats for fast computation. How many SNPs are significant (use alpha = 0.05). Is the result consistent with the chi-square test?

```
geno.matrix.exact.stats <- HWExactStats(geno.matrix, pvalues = FALSE)
geno.matrix.exact.pval <- HWExactStats(geno.matrix, pvalues = TRUE)
significant.snp.num.exact <- sum(geno.matrix.exact.pval<0.05)
#plot(density(geno.matrix.exact.stats))

significant.snp.num.chisq
```

```
## [1] 8162
```

```
significant.snp.num.exact
```

```
## [1] 5793
```

The amount of significant p-values/SNPs under Chi Square and Exact tests are similar considering the larg size of the dataset, so the results are consistent.

Ex 9 Which SNP is most significant according to the exact test results? Give its genotype counts. In which sense is this genotypic composition unusual?

```
geno.matrix.exact.most <- max(geno.matrix.exact.pval) # most significant SNP
geno.matrix.exact.most
```

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

```
exact.index <- which(geno.matrix.exact.pval==max(geno.matrix.exact.pval)) # index of the most significa
names(dt_poly[exact.index[1]])
```

```
## [1] "rs587720402_A"
```

```
geno.matrix[exact.index[1],] # counts of the most significant SNP
```

```
## [1] 106   1   0
```

```
#...
```

```
sum(geno.matrix.exact.pval==max(geno.matrix.exact.pval)) # repetitions of the most significant SNPs
```

```
## [1] 135407
```

This genotype composition is not unusual. It appears many times in the dataset.

Ex 10 Apply a likelihood ratio test for Hardy-Weinberg equilibrium to each SNP, using the HWLratio function. How many SNPs are significant (use alpha = 0.05). Is the result consistent with the chi-square test?

```
geno.matrix.like.pval <- c()
for(i in 1:ncol(dt_poly)) {
  snp <- geno.matrix[i,]
  names(snp) <- c("AA", "AB", "BB")
  geno.matrix.like.pval <- c(geno.matrix.like.pval, HWLratio(snp,verbose = FALSE)$pval)
}
significant.snp.num.like <- sum(geno.matrix.like.pval<0.05)

significant.snp.num.chisq
```

```
## [1] 8162
```

```
significant.snp.num.like
```

```
## [1] 7955
```

The amount of significant p-values/SNPs under Chi Square and Likelihood tests are similar, so the results are consistent.

Ex 11 Apply a permutation test for Hardy-Weinberg equilibrium to the first 10 SNPs, using the classical chi-square test (without continuity correction) as a test statistic. List the 10 p-values, together with the 10 p-values of the exact tests. Are the result consistent?

```
geno.matrix.perm.pval = c()
for(i in 1:10) {
  snp <- geno.matrix[i,]
  names(snp) <- c("AA", "AB", "BB")
  geno.matrix.perm.pval <- c(geno.matrix.perm.pval,HWPerm(snp, verbose = FALSE)$pval)
}

geno.matrix.perm.pval
```

```
##  [1] 1.000000000 1.000000000 1.000000000 1.000000000 0.646764706 1.000000000
##  [7] 1.000000000 1.000000000 0.126470588 0.008705882
```

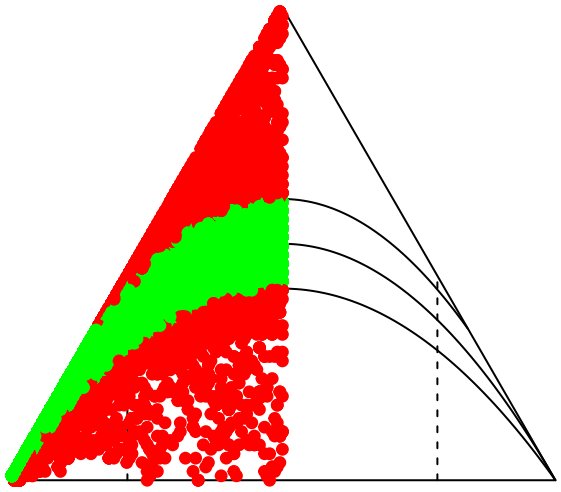
```
geno.matrix.exact.pval[1:10]
```

```
##  [1] 1.000000000 1.000000000 1.000000000 1.000000000 1.000000000 1.000000000
##  [7] 1.000000000 1.000000000 0.214715301 0.008643867
```

The results are quite consistent in most of the tests. P-values are similar except for the 5th SNP.

Ex 12 Depict all SNPs simultaeneously in a ternary plot with function HWTernaryPlot and comment on your result (because many genotype counts repeat, you may use UniqueGenotypeCounts to speed up the computations)

```
#geno.matrix.unique <- UniqueGenotypeCounts(geno.matrix)
#geno.matrix.unique
HWTernaryPlot(geno.matrix)
```



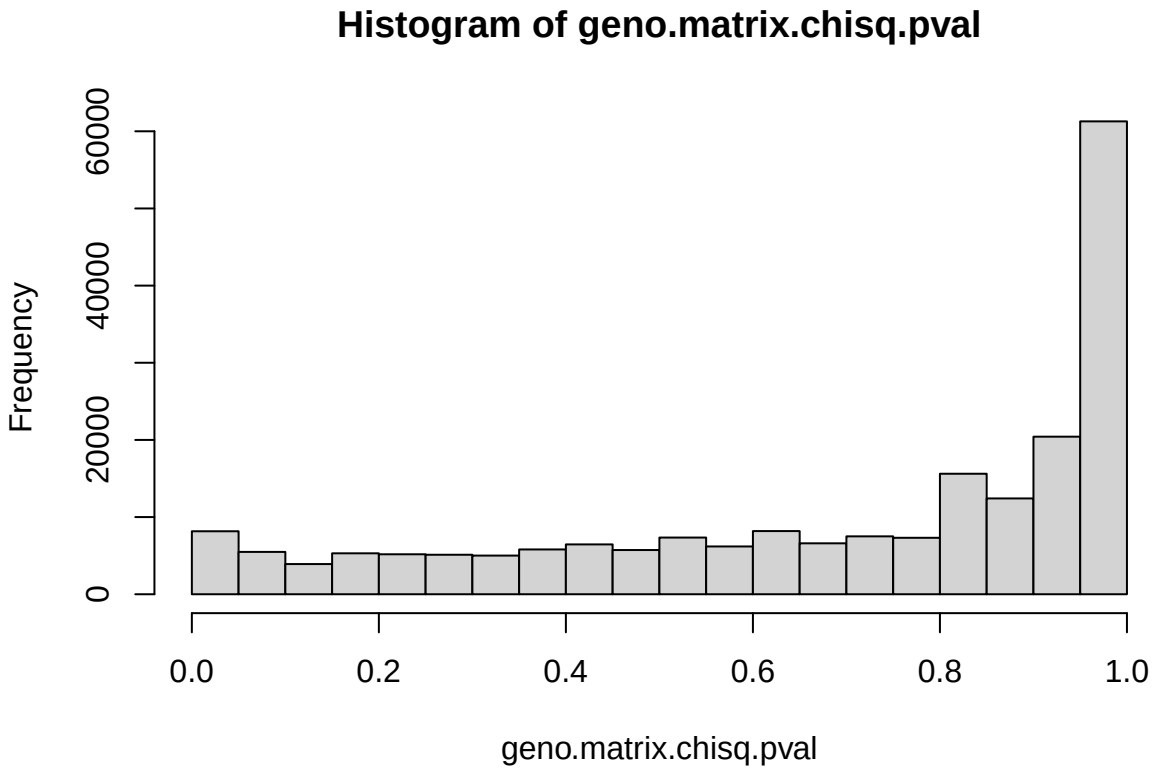
Some SNPs are within the acceptance region (equilibrium), but others are in disequilibrium. The SNP dataset seems to be in disequilibrium.

Ex 13 Can you explain why half of the ternary diagram is empty?

The ternary diagram is half-empty, which means that the allele B does not influence the frequency/ratio in which the SNPs appear (AA or AB).

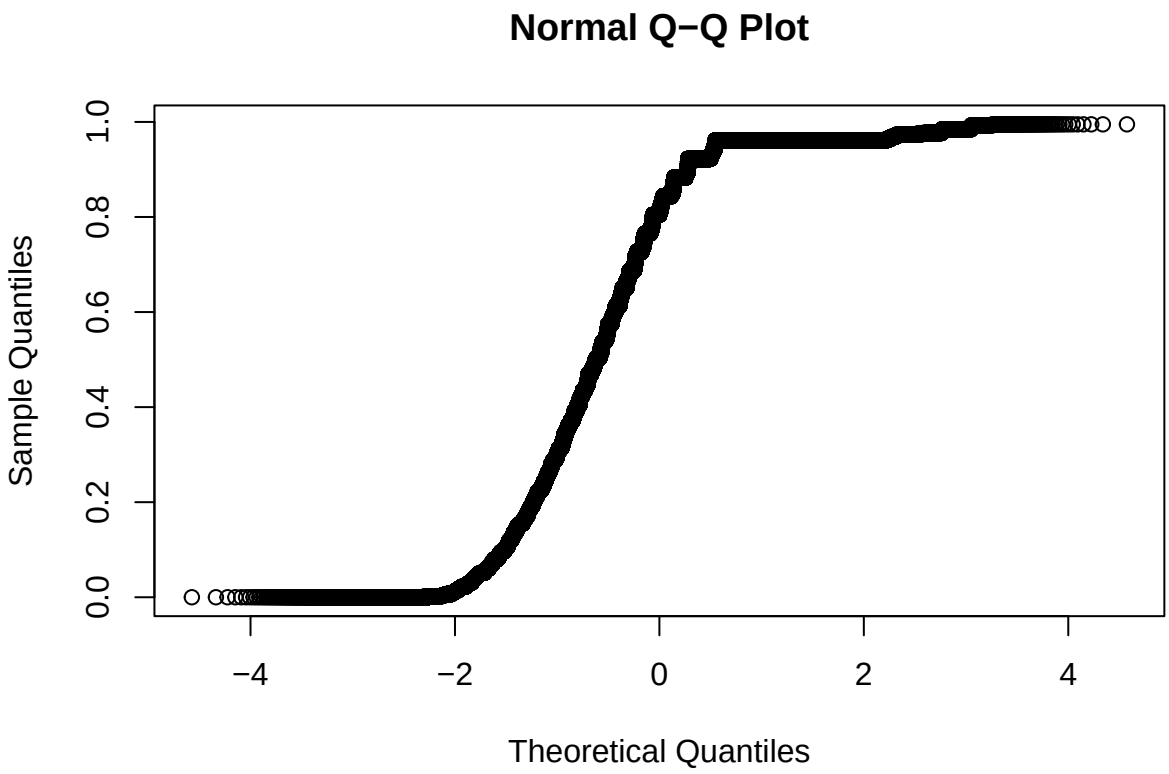
Ex 14 Make a histogram of the p-values obtained in the chi-square test. What distribution would you expect if HWE would hold for the data set? Make a Q-Q plot of the p values obtained in the chi-square test against the quantiles of the distribution that you consider relevant. What is your conclusion?

```
hist(geno.matrix.chisq.pval)
```



If the dataset was in HW equilibrium, we would obtain a normal distribution.

```
#plot(density(geno.matrix.chisq.pval))
#hist(geno.matrix.chisq.pval,breaks=30)
qqnorm(geno.matrix.chisq.pval)
```



Ex 15 Imagine that for a particular marker the counts of the two homozygotes are accidentally interchanged. Would this affect the statistical tests for HWE? Try it on the computer if you want. Argue your answer.

```
x <- geno.matrix[9,]
names(x) <- c("AA", "AB", "BB")
HWAlltests(x,include.permutation.test=TRUE)
```

```
## Warning in HWChisq(x, cc = 0, x.linked = x.linked, verbose = FALSE): Expected
## counts below 5: chi-square approximation may be incorrect
```

```
## Warning in HWChisq(x, cc = 0.5, x.linked = x.linked, verbose = FALSE): Expected
## counts below 5: chi-square approximation may be incorrect
```

```
##                                           Statistic    p-value
## Chi-square test:                           2.844400 0.09169283
## Chi-square test with continuity correction:  1.786515 0.18135141
## Likelihood-ratio test:                     4.913209 0.02665208
## Exact test with selome p-value:                  NA 0.21471530
## Exact test with dost p-value:                    NA 0.18736330
## Exact test with mid p-value:                     NA 0.16787448
## Permutation test:                          2.844400 0.12958824
```

```
x <- rev(geno.matrix[9,])
names(x) <- c("AA", "AB", "BB")
HWAlltests(x,include.permutation.test=TRUE)
```

```
## Warning in HWChisq(x, cc = 0, x.linked = x.linked, verbose = FALSE): Expected
## counts below 5: chi-square approximation may be incorrect
```

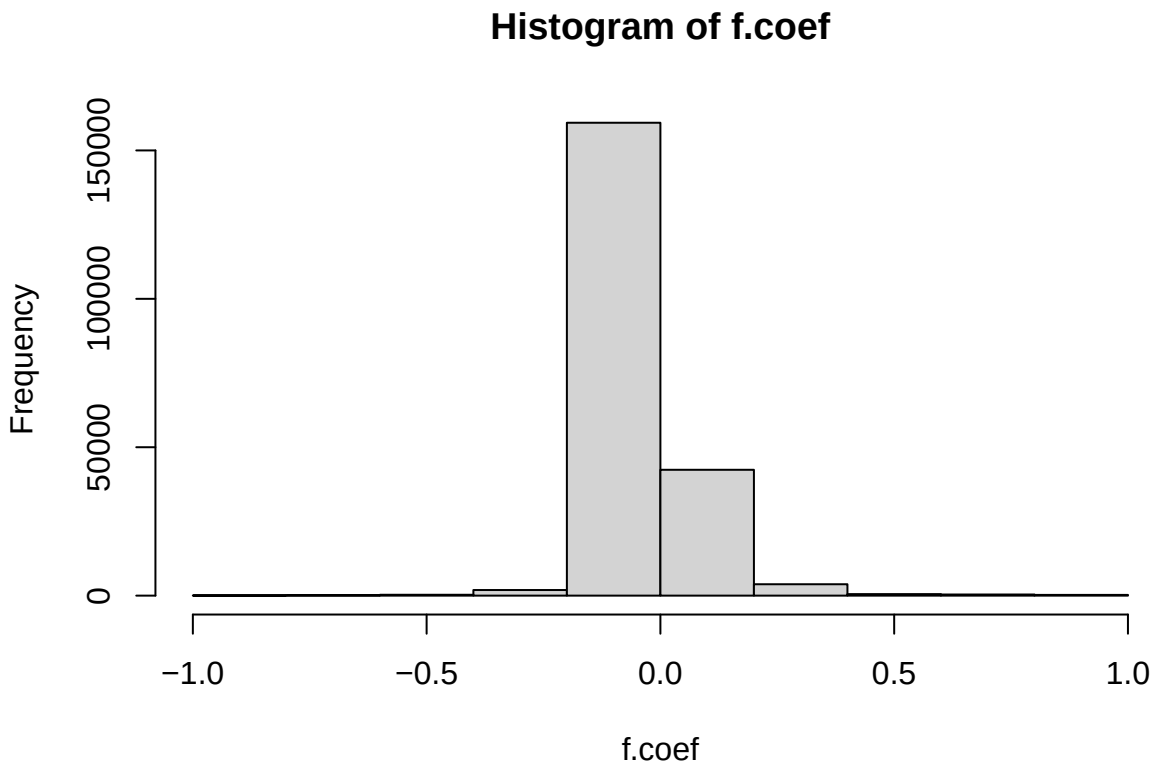
```
## Warning in HWChisq(x, cc = 0.5, x.linked = x.linked, verbose = FALSE): Expected
## counts below 5: chi-square approximation may be incorrect
```

```
##                                            Statistic    p-value
## Chi-square test:                            2.844400 0.09169283
## Chi-square test with continuity correction:  1.786515 0.18135141
## Likelihood-ratio test:                      4.913209 0.02665208
## Exact test with selome p-value:                   NA 0.21471530
## Exact test with dost p-value:                     NA 0.18736330
## Exact test with mid p-value:                      NA 0.16787448
## Permutation test:                           2.844400 0.12335294
```

Swapping the homozygotes does not change significantly the results of the tests. Homozygotes are assigned names randomly, no matter the order.

Ex 16 Compute the inbreeding coefficient (ˆf) for each SNP, and make a histogram of ˆf. You can use function HWf for this purpose. Give descriptive statistics (mean, standard deviation, etc) of ˆf calculated over the set of SNPs. What distribution do you expect ˆf to follow theoretically? Use a probability plot to confirm your idea.

```
f.coef <- c()
for(i in 1:ncol(dt_poly)) {
  x <- geno.matrix[i,]
  names(x) <- c("AA", "AB", "BB")
  f.coef <- c(f.coef,HWf(x))
}

hist(f.coef,breaks=10)
```



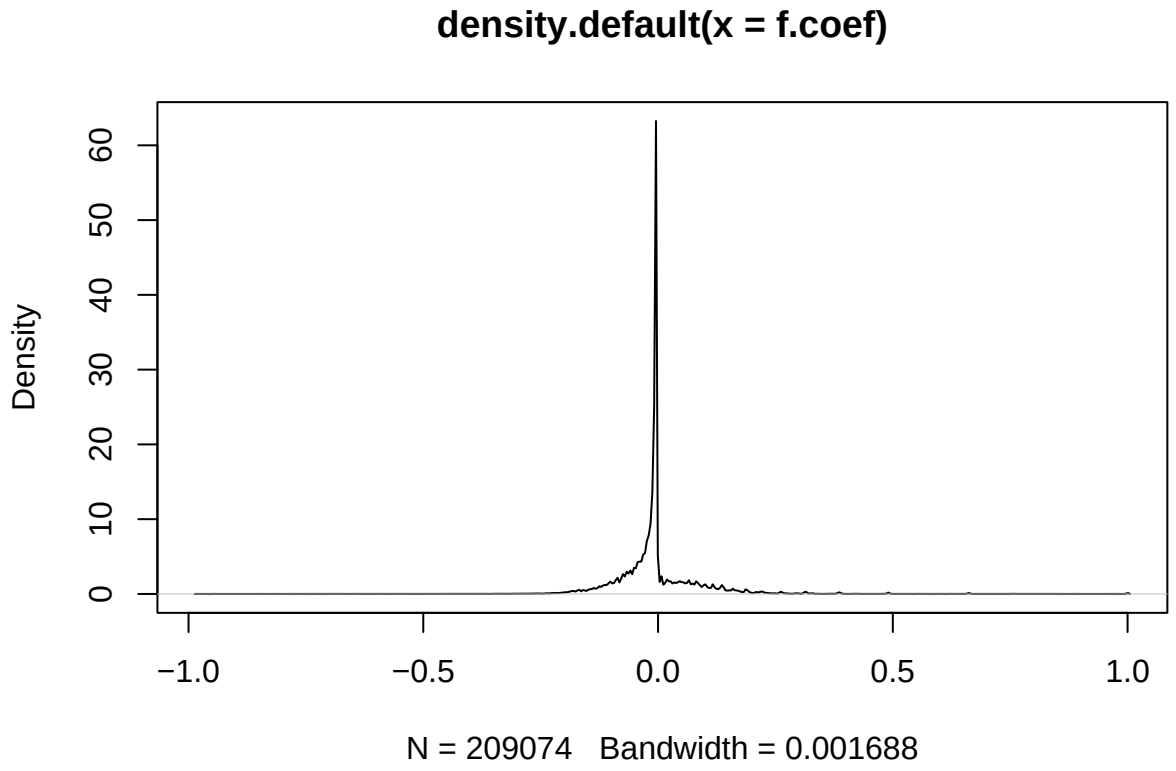
```
f.coef.stats <- sapply(f.coef,mean)
summary(f.coef.stats)
```

```
##       Min.    1st Qu.     Median       Mean    3rd Qu.       Max.
## -0.981482 -0.033816 -0.004695 -0.004668 -0.004695  1.000000
```

```
sd(f.coef)
```

```
## [1] 0.095012
```

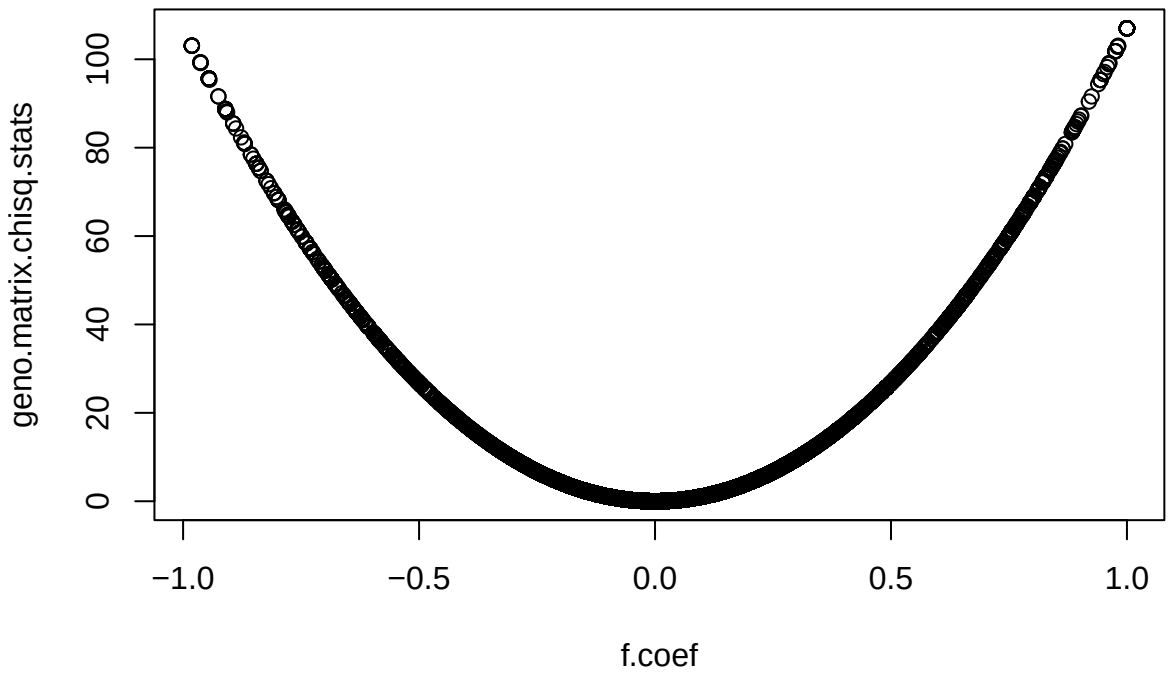
```
plot(density(f.coef))
```



We expect a normal distribution as we got (standard deviation is low).

Ex 17 Make a plot of the observed chi-square statistics against the inbreeding coefficient (ˆf). What do you observe? Can you give an equation that relates the two statistics?

```
fit<-lm(geno.matrix.chisq.stats~poly(f.coef,2,raw=TRUE))
#summary(fit)

quadratic = fit$coefficient[3]*f.coefˆ2 + fit$coefficient[2]*f.coef + fit$coefficient[1]
plot(f.coef,geno.matrix.chisq.stats)
```




```
summary(fit)$coefficient[3]
```

```
## [1] 107
```

```
summary(fit)$coefficient[2]
```

```
## [1] -1.782353e-11
```

```
summary(fit)$coefficient[1]
```

```
## [1] -5.472794e-13
```

The plot follows a quadratic correlation. The quadratic equation that relates the inbreeding coefficient with the observed chi-square is: f(x) = 107x^2, where the observed chi-square is a function of the the inbreeding coefficient.

Ex 18 We reconsider the exact test for HWE, using different significant levels. Report the number and percentage of significant variants using an exact test for HWE with alpha = 0.10, 0.05, 0.01 and 0.001. State your conclusions.

```
alpha_i <- c(0.10, 0.05, 0.01, 0.001)
significant.snp.num.exact.alphai <- c()

for(i in alpha_i) {
  significant.snp.num.exact.alphai <- c(significant.snp.num.exact.alphai, sum(geno.matrix.exact.pval<i))
}

significant.snp.perc.exact.alphai <- 100*significant.snp.num.exact.alphai/length(geno.matrix.exact.pval)
```

```
alpha_i
```

```
## [1] 0.100 0.050 0.010 0.001
```

```
significant.snp.num.exact.alphai
```

```
## [1] 10049  5793  2508  1485
```

```
significant.snp.perc.exact.alphai
```

```
## [1] 4.8064322 2.7707893 1.1995753 0.7102748
```

We consider: * p > 0.05 -> Not significant, no evidence against the null hypothesis * p <= 0.05 -> Significant, weak to moderate evidence of the null hypothesis * p <= 0.01 -> Very significant, good to strong evidence of the null hypothesis * p <= 0.001 -> Highly significant, very strong evidence of the null hypothesis

With the results obtained with the different significant levels, we conclude that ~2.77% of the SNPs in the dataset are significant, ~1.19% of the SNPs in the dataset are very significant, and ~0.71% of the SNPs in the dataset are highly significant.