Assignment 2

Krishna Kalyan

Question 1

The File CHBChr3-2000.rda contains genotype information (2000 SNPs) of individuals from a Chinese population of unrelated individuals. Load this data into the R environment. The file contains three data objects, X (genotype information), Allels (the possible alleles for each marker) and Pos(the position of each marker in base pairs)

```
## [1] "Alleles" "X"
```

Question 2

Remove all SNPs that consist of missing values only from the database. Also remove all monomorphic SNPs from the data bases. Apply a chi-square test without continuity correction for Hardy-Weinberg equilibrium to each SNP. How many SNPs remain? How many SNPs are significant (use alpha = 0.05)?

```
X = t(X)
missing_col = which(colMeans(is.na(X)) == 1)
remove_mono = which(apply(X,2,function(x) length(unique(na.omit(unlist(x))))) == 1)
remove_cols =  c(missing_col,remove_mono)
Alleles = t(Alleles)
Alleles = Alleles[-remove_cols]
X = X[,-remove_cols]
paste("snps remaining", length(Alleles))
```

```
## [1] "snps remaining 987"
```

```
hw_pval <- function(x){
  snpg <- genotype(x,sep="")
  hw <- HWE.chisq(snpg)$p.value
  return(hw)
}
hw <- apply(X,2,hw_pval)

paste("significant snps",sum(hw > 0.05))
```

```
## [1] "significant snps 941"
```

Question 3

Do you think the monomorphic markers are in Hardy-Weinberg equilibrium? Argue your answer? Since monomorphic markers have a Chi Square value of 1, they are significant. However monomorphic SNPs are uninformative in genetic association studies as there is no genotypic difference.

Question 4

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

```
## [1] "absolute effect 46"
```

```
## [1] "effect of random chance 49.35"
```

We would expect around 5% of the non-monomorphic markers to be significant, which is around 50 markers in this case.

Question 5

Apply an Exact test for Hardy-Weinberg equilibrium to each SNP. Use the pvaluetype="selome" option. How many SNPs are significant (use alpha = 0.05). Is the result consistent with the chi-square test?

```
snpg <- genotype(X,sep="")
geno_count = MakeCounts(X,Alleles)
chisel = apply(unlist(geno_count[,1:3]), 1, function(x){
  HWExact(x,pvaluetype="selome",verbose=F)$pval
  })
sig_snp = sum(chisel>0.5)
paste("significant snps HW",sig_snp)
```

```
## [1] "significant snps HW 637"
```

The result is not consistent with the Chi-square test without continuity, when doing the Exact test. We see a differnce of `304` snps and Exact Test has a lower number of significant SNPs.

Question 6

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

```
ml <- apply(unlist(geno_count[,1:3]),1,function(x){
  HWLratio(x, verbose = F, x.linked = FALSE)$pval
  })
signif_ml <- sum(ml > 0.05)
paste("significant snps ML", signif_ml)
```

```
## [1] "significant snps ML 933"
```

Question 7

Apply a permutation test for Hardy-Weinberg equilibrium to each SNP, using the classical chisquare test (without continuity correction) as a test statistic. Reduce the number of permutations in order to keep computation time within reasonable limits (argument nperm). How many SNPs are significant (use alfa = 0.05). Are the result consistent with the chi-square or exact test?.

```
pvalperm <- apply(unlist(geno_count[,1:3]),1,function(x){
  HWPerm(x, nperm = 50, verbose = F)$pval
})
perm <- sum(pvalperm > 0.05)
paste("significant snps for using 50 permutations", perm)
```

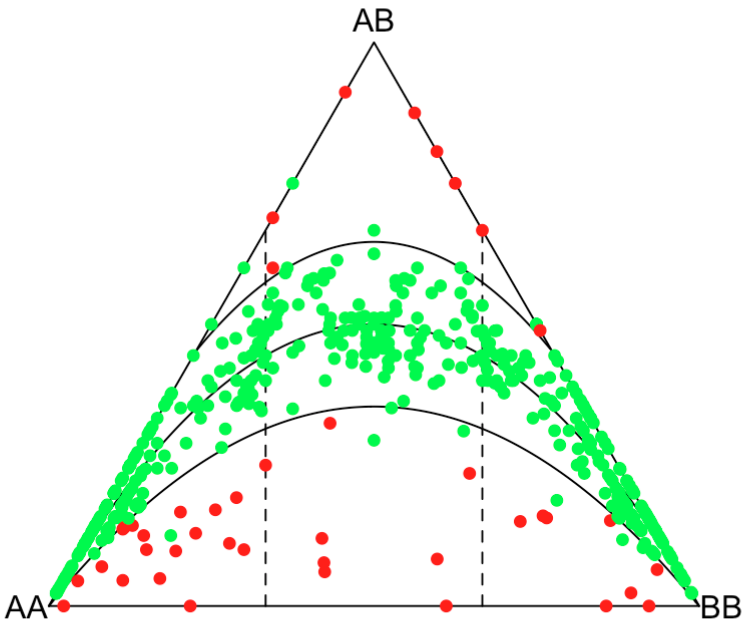
```
## [1] "significant snps for using 50 permutations 951"
```

The results obtained are higer than that obtained by Chi-Square/ Exact Test.

Question 8

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

```
ug <- UniqueGenotypeCounts(unlist(geno_count[,1:3]),verbose=F)
ternary <- HWTernaryPlot(ug[,c(1:3)])
```




```
ternary$percinrange
```

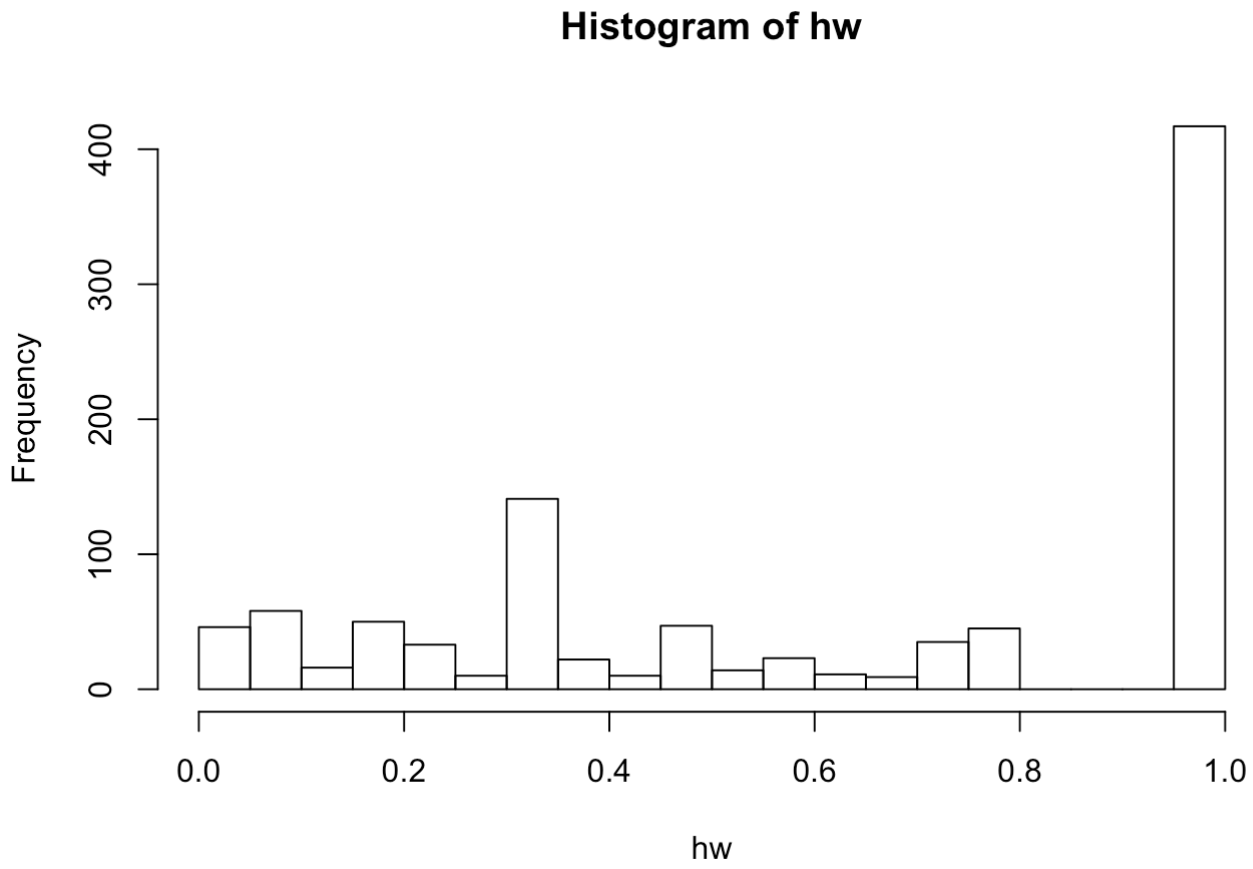
```
## [1] 29.14
```

HWTernaryPlot is a routine that draws a ternary plot for three-way genotypic compositions (AA,AB,BB), and represents the acceptance region for different tests for Hardy-Weinberg equilibrium (HWE) in the plot. By taking unique gene count we reduce the size from 987 to . We can see all significant marker along the edge and the conic region of the Ternary Plot.

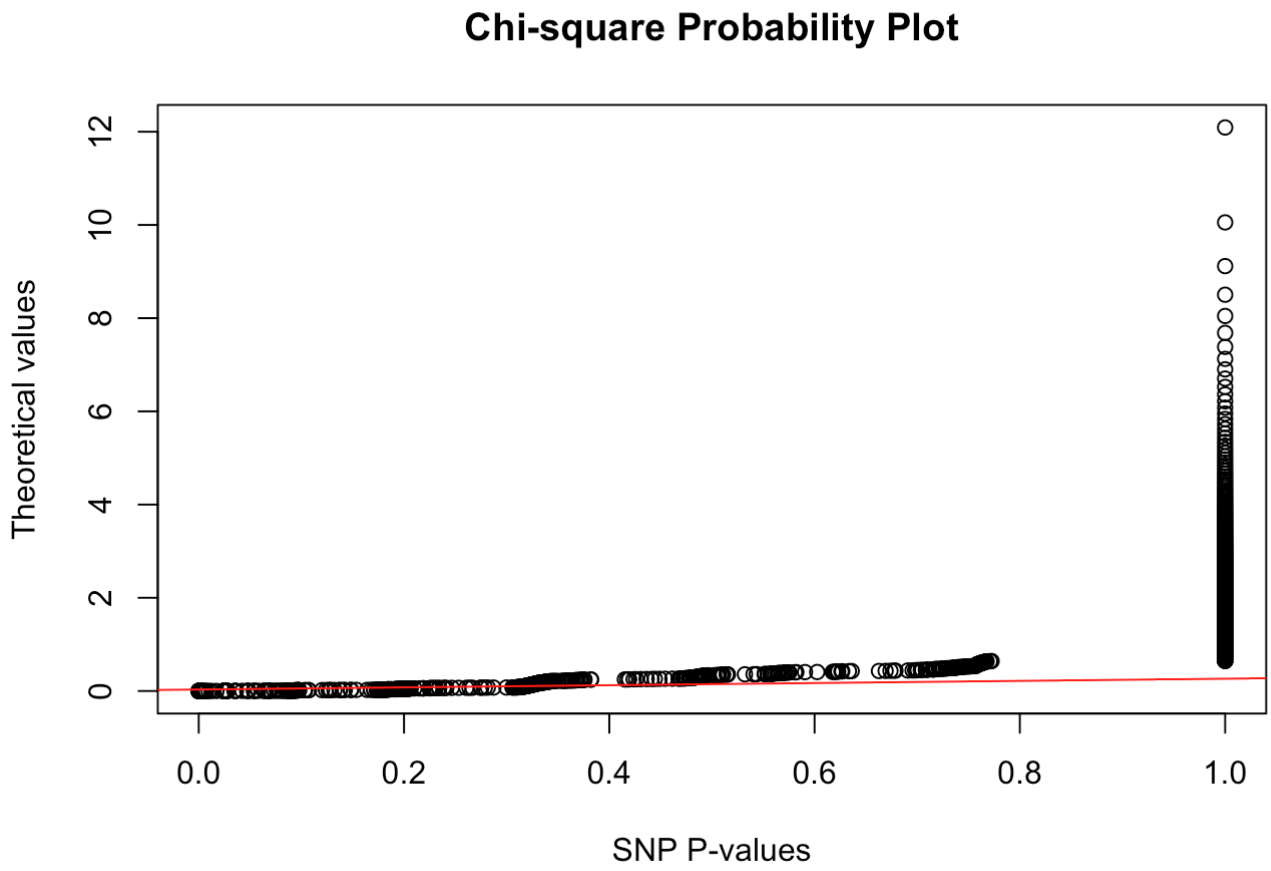
Question 9

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? What distribution do you observe? Also 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(hw,breaks = 20)
```




```
chi.sq <- qchisq(ppoints(ncol(X)), df = 1)
qqplot(hw,chi.sq, main="Chi-square Probability Plot",
       xlab= "SNP P-values", ylab="Theoretical values")
qqline(hw, distribution = function(p) qchisq(p,df = 3),col="red")
```



The distribution of the histogram peak as the values when pvalue are `1` and around `0.35`. It seems to have lot of values with with `1` as its `p-value`. For a qqplot we expect to see the observations along the straight line so as to confirm that they are normally distributed. Howver in this case it follows a straight line and peaks at the end.

Question 10

Imagine that for a particular marker the counts of the two homozygotes are accidentally interchanged. Would this affect the statistical tests for HWE? Argue your answer.

```
gene1 <- c(AA=10, AB=20, BB=30)
gene2 <- c(AA=30, AB=20, BB=10)
HWChisq(gene1,verbose = F)$pval == HWChisq(gene2,verbose = F)$pval
```

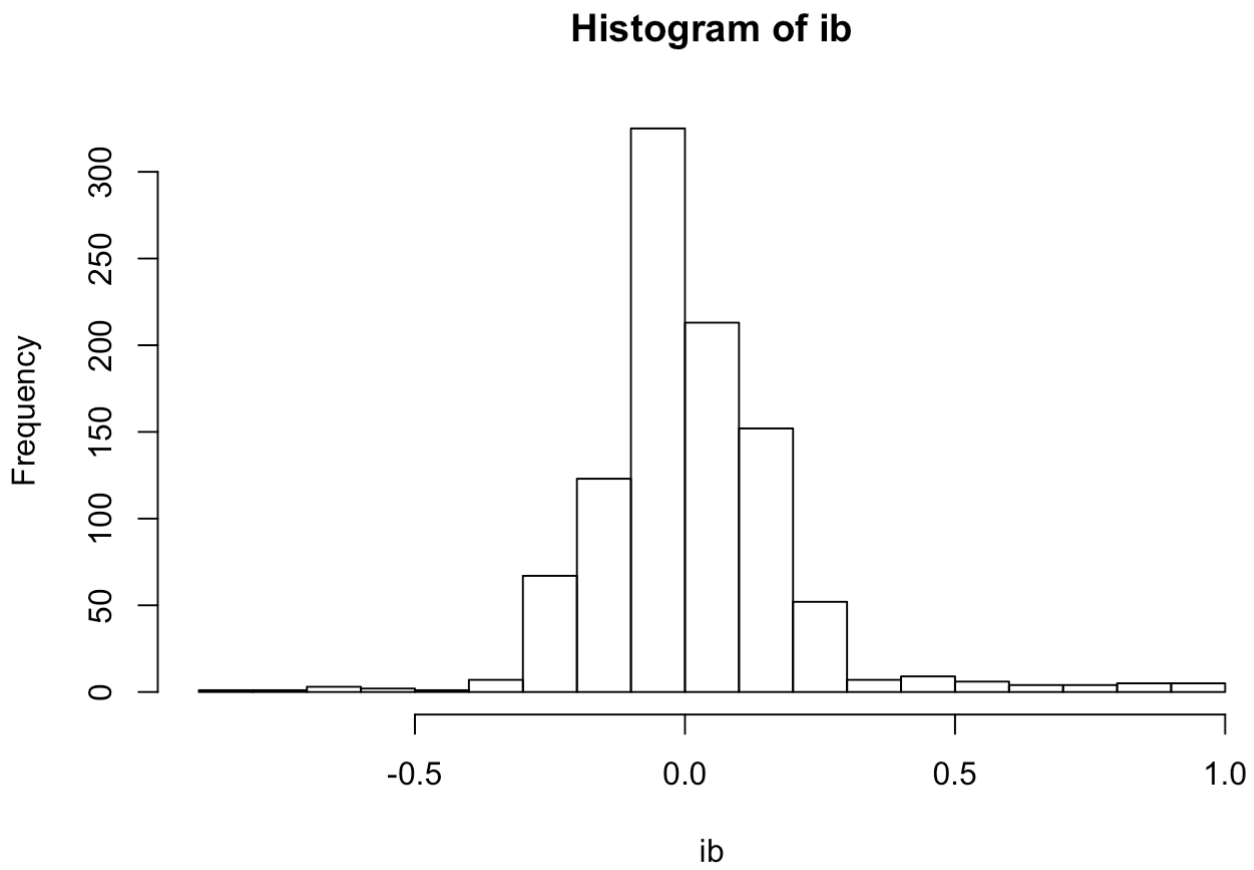
```
## [1] TRUE
```

According to the Chi-Square test conducted above and comparing the p-values we see that the statistical test will not change if the homozygotes are interchanged.

Question 11

Compute the inbreeding coefficient ($\hat{f}$) for each SNP, and make a histogram of $\hat{f}$. You can use function HWf for this purpose. Give descriptive statistics (mean, standard deviation, etc) of $\hat{f}$ calculated over the set of SNPs. What distribution do you think $\hat{f}$ follows? Use a probability plot to confirm your idea.

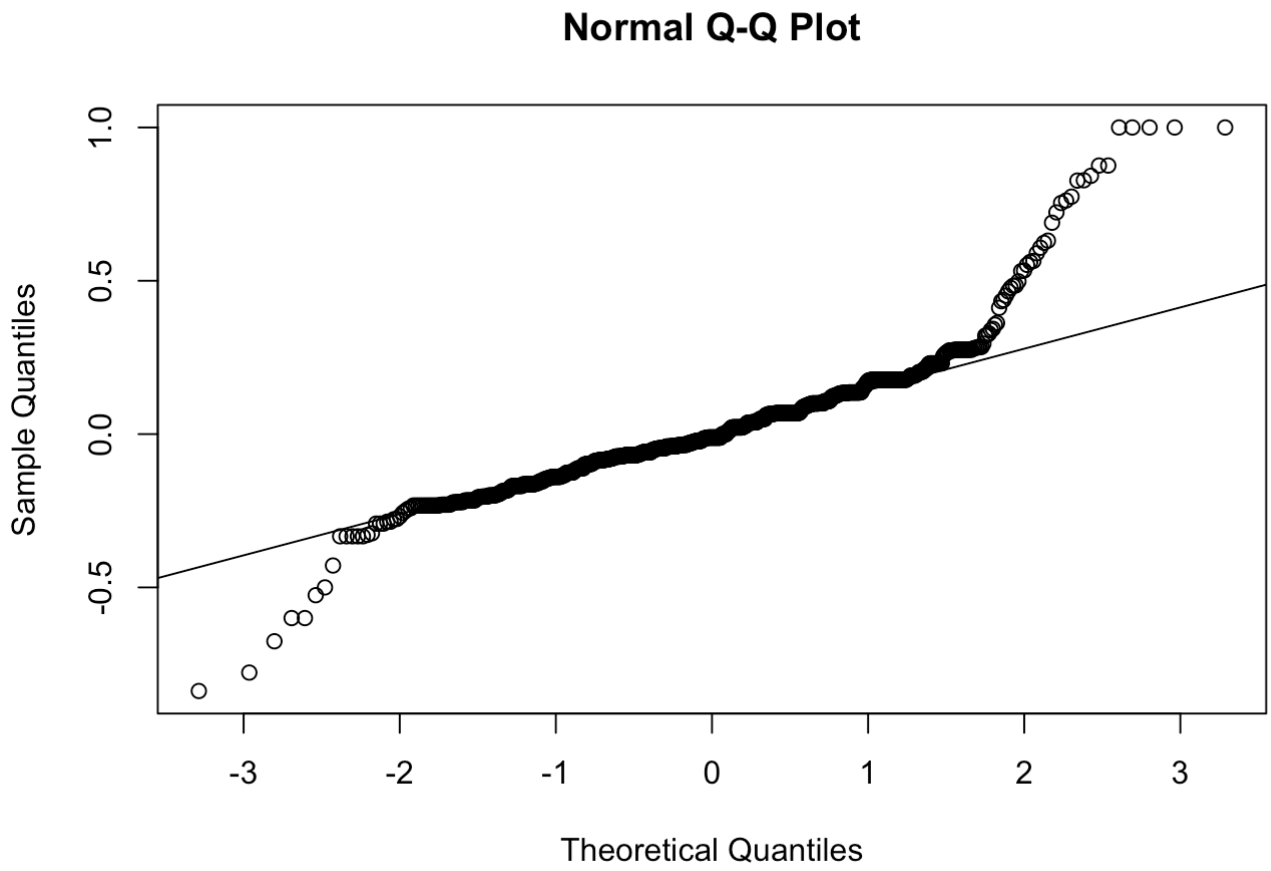
```
ib <- apply(unlist(geno_count[,1:3]),1,function(x) HWf(x))
hist(ib, breaks=20)
```




```
M = summary(ib)
paste(c("mean ","standard deviation ", "median ", "minimum ", "maximum"),
      c(M[4], round(sd(ib),2), M[3],M[1],M[6])  )
```

```
## [1] "mean  0.01612"            "standard deviation  0.19"
## [3] "median  -0.01124"         "minimum  -0.8378"
## [5] "maximum 1"
```

```
qqnorm(ib)
qqline(ib)
```



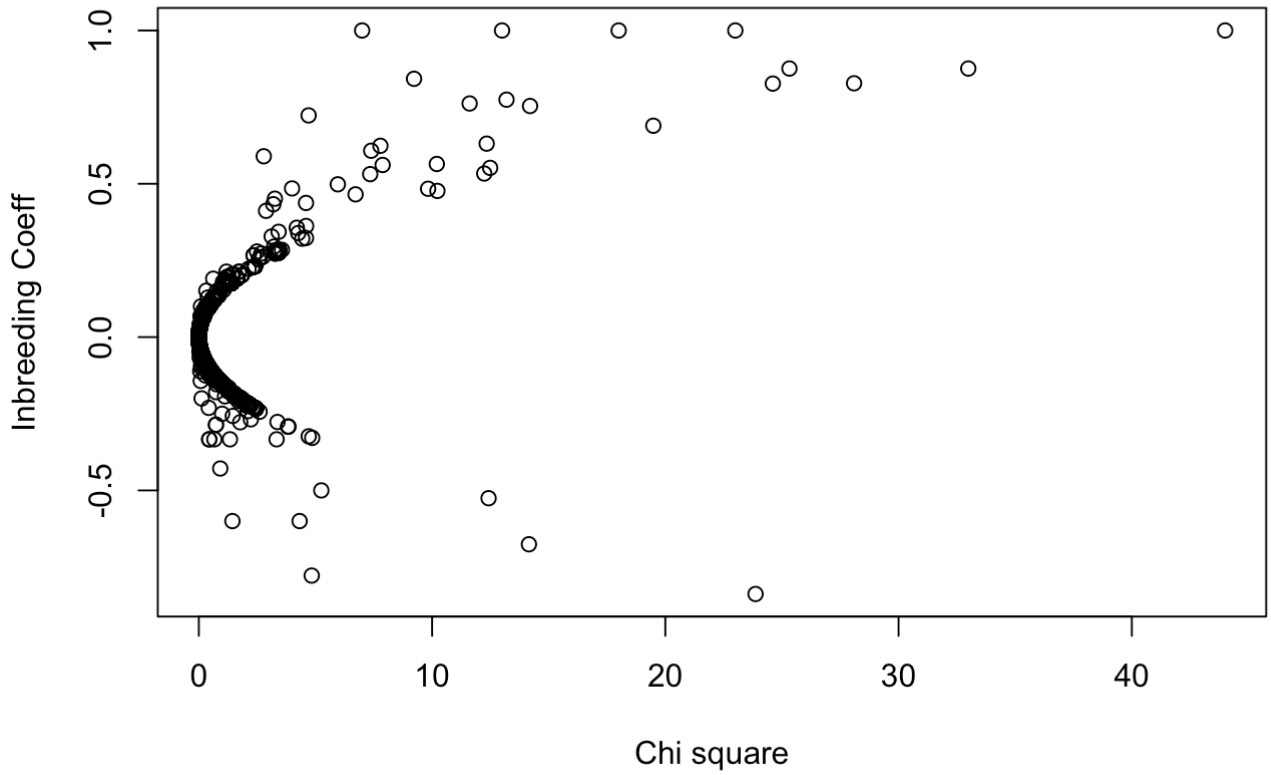
The QQplot distribution looks normal, since majority of the observations are on the diagonal line. A coefficient of inbreeding can be calculated for an individual, as a measure for the amount of pedigree collapse within that individual's genealogy.In general, the higher the level of inbreeding the closer the coefficient of relationship approaches a value of 1, expressed as a percentage,and approaches a value of 0 for individuals with arbitrarily remote common ancestors. In the histogram we observe that the inbreeding coefficient seems to be centered around 0.

Question 12

Make a plot of the observed chi-square statistics against the inbreeding coefficient ($\hat{f}$). What do you observe? Can you give an equation that relates the two statistics?

```
chi_sq <- apply(unlist(geno_count[,1:3]),1,function(x){
  HWChisq(x,cc = 0,verbose=F)$chi
  })

plot(x=chi_sq,y=ib, xlab= "Chi square", ylab="Inbreeding Coeff")
```



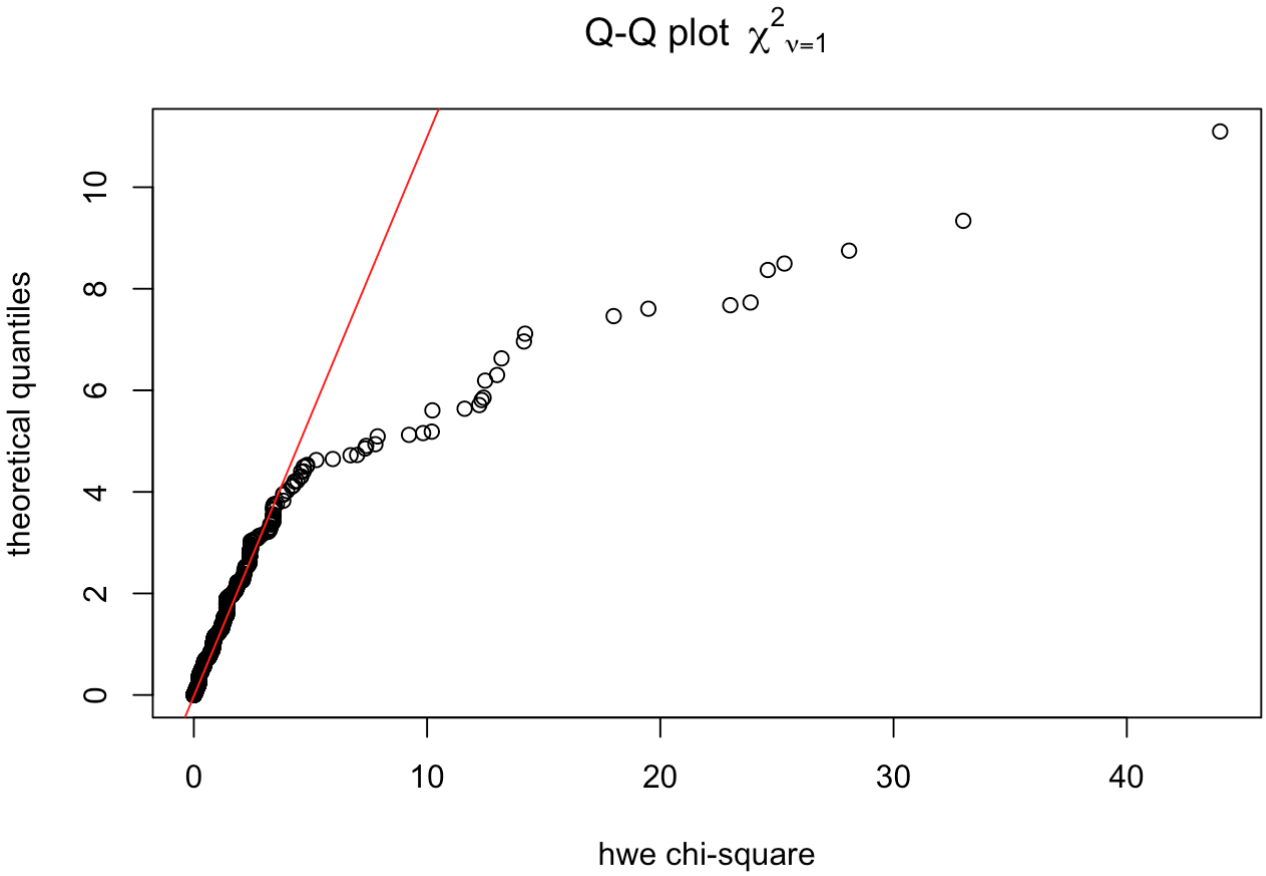
We observe an eliptical trend for the above plot. It seems a squared function for y. This being similar to $x = y^2$. We also observe that number of observations decrease as Chi-Squre increases.

Question 13

Make a chi-square probability plot of the observed chi-square statistics against their theoretical quantiles. Does the sample statistic follow a chi-square distribution?

```
y <- rchisq(length(chi_sq), df = 1)

qqplot(chi_sq, y,main = expression("Q-Q plot" ~~ {chi^2}[nu == 1]),
       ylab = "theoretical quantiles",xlab = "hwe chi-square")
qqline(y, distribution = function(p) qchisq(p, df = 1),col=2)
```



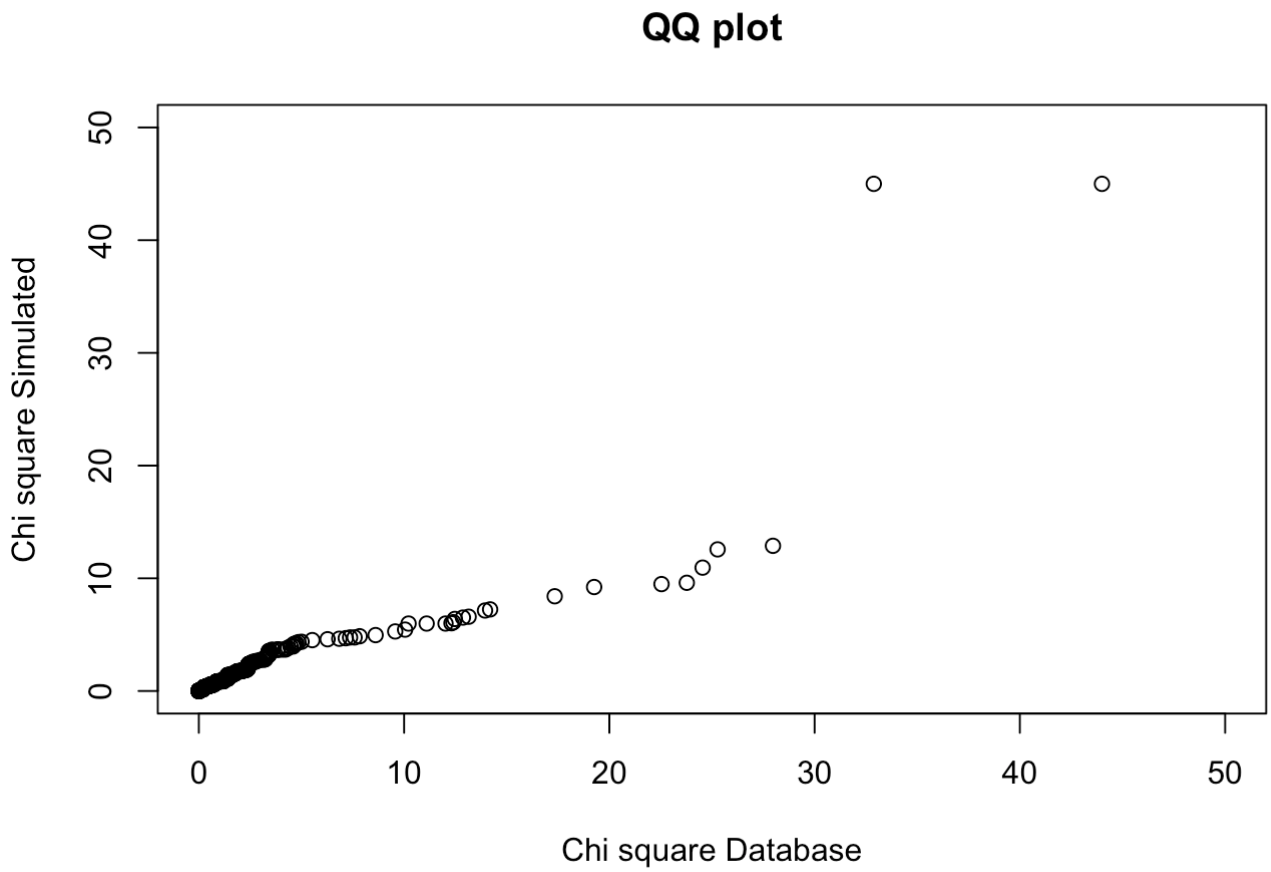
It appears that most of the data of the SNPs, do follow a chi-square distribution, since by plotting the chi square values of the HWE (without continuity) against the computed theoretical quantiles for chi-square, it appears to follow a linear trend. Some other values seem to differ, but they are not the mayority. In any case, a proof of hypothesis would have to be computed before being certain of this conclusion.

Question 14

Simulate SNPs under the assumption of Hardy-Weinberg equilibrium. Simulate the SNPs of this database, and take care to match each of the SNPs in your database with a simulated SNP that has the same sample size and allele frequency. You can use function HWData of the HardyWeinberg package for this purpose. Compare the distribution of the observed chi-square statistics with the distribution of the chi-square statistics of the simulated SNPs by making a Q-Q plot. What do you observe? State your conclusions.

```
geno <- unlist(geno_count[,1:3])
n = unname(rowSums(unlist(geno_count[,1:3])))
p = unname(apply(geno,1,maf))
simul <-c()
for(i in 1:length(n)){
  s <- HWData(1,n=n[i],p=p[i])
  simul <- rbind(simul,s)
}
chisq_simul <- apply(t(simul),2,function(x){
  HWChisq(x,cc = 0,verbose=F)$chi
  })

qqplot(x=chi_sq,y=chisq_simul,xlab = "Chi square Database",
       ylab="Chi square Simulated", main="QQ plot",
       xlim=c(0,50),ylim=c(0,50))
```



The QQ plot suggests the the distribution is normal, the Chisquare values observed seem to be correleted to the actual values in the database. Since density of the observation is less for higher chisqares we see sparse observations.

Question 15

Do you think genotyping error is a problem for the database you just studied? Explain your opinion.

Genotyping errors are present in almost all genetic data and can affect biological conclusions of a study, particularly for studies based on individual identification and parentage. We can use statistical approach to estimate these error. I used the internet where they mentioned that

- repeat genotyping 5% of samples
- comparing unintentionally recaptured individuals
- Mendelian inheritance error checking

However in our dataset we might have genotyping error. We ould estimate it by using a statistical appoach.