STAT-GEN HW1

Posted: Thursday, Sept. 4

Due: Friday, Sept. 12 at 10:00pm in D2L

Question 1

The prevalence of sickle cell anemia in a population is 0.01.

Assuming an autosomal recessive disease model and HWE, What do you estimate is the frequency of sickle cell mutation at the hemoglobin locus in this population?

```
q=sqrt(0.01)
print(q)
```

```
## [1] 0.1
```

Question 2

The following numbers of the human M-N blood groups were recorded in a sample of European Americans

M	MN	N
1787	3039	1303

Source: Wiener, A.S. (1943), published by Stern C., 1973. Principles of Human Genetics, Freeman SF, USA.

- (2.1) What are the genotypic frequencies observed in this sample? Use at least 3 decimal points.
- (2.2) What are the gene frequencies?
- (2.3) What are the genotypic frequencies expected under Hardy-Weinberg equilibrium?
- (2.4) Test if the gene is in HWE.

```
O=c('M'=1787,'MN'=3039,'N'=1303)

n=sum(O)
nAlleles=2*n

OFreq=O/n

pN=(O[2]+2*O[3])/(2*n)
pM=1-pN

EFreq=c(pM^2,2*pN*pM,pN^2)
E=n*EFreq

message('(2.1) => Observed and Expected Genotype Frequencies:')
```

```
## (2.1) => Observed and Expected Genotype Frequencies:
```

```
knitr:::kable(rbind(O,E)/n,digits=3)
```

	M	MN	N
O	0.292	0.496	0.213
E	0.291	0.497	0.212

```
message('(2.2) => Gene Frequencies:')
```

```
## (2.2) => Gene Frequencies:
```

```
geneFreq=c(pM,pN)
names(geneFreq)=c('pM','pN')
knitr:::kable(t(as.matrix(geneFreq)))
```

pM	pN
0.5394844	0.4605156

```
Chisq=sum(((O-E)^2)/E)

pVal=pchisq(Chisq,df=1,lower.tail=FALSE)

message('(2.3) & (2.4) => Chi-square test:')
```

```
## (2.3) & (2.4) => Chi-square test:
```

```
print(c('DF'=1,'Chisq'=round(Chisq,3),'pVal'=round(pVal,4)))
```

```
##     DF  Chisq   pVal
## 1.0000 0.0270 0.8695
```

Question 3

The dumpy mutation is a single autosomal gene variant that produces truncated wings, as a variation to the non-mutated flies that have normal long wings.

A parental cross of inbred fly lines one having normal long wings and the other one having dumpy wings produces an F1 generation with all flies having long wings.

Mating F1 individuals at random produced an F2 generation with the following phenotypic counts:

Long	Dumpy
792	208

- (3.1) What proportion of individuals with long and dumpy wings would you expect in the F2 generation under the hypothesis that the dumpy wing trait is recessive and the two lines were fully homozygous for opposite alleles of the dumpy gene?

```
O=c('Long'=792,'Dumpy'=208)
n=sum(O)

## The allele frequencies in the base population are 0.5 because both lines are inbred.
pL=0.5
pD=0.5
```

```
E=c('Long'=1-pD^2,'Dumpy'=pD^2)
ECount=E*n
ANS=rbind(E,ECount)
rownames(ANS)=c('Expected Prop.','Expected Count')
knitr:::kable(ANS,digits=3,caption='=> (3.1) Expected Proportion of Long and Dumpy Flies in the F2 ge
```

Table 5: => (3.1) Expected Proportion of Long and Dumpy Flies in the F2 generation

	Long	Dumpy
Expected Prop.	0.75	0.25
Expected Count	750.00	250.00

- (3.2) Does the data support the hypothesis?

```
Chisq=sum(((O-ECount)^2)/ECount)
pVal=pchisq(df=1,q=Chisq,lower.tail=FALSE)

ANS=c('Chisq'=Chisq,'df'=1,'pValue'=pVal)
knitr:::kable(ANS,digits=5,caption='=> (3.2) Chi-square test')
```

Table 6: => (3.2) Chi-square test

	x
Chisq	9.40800
df	1.00000
pValue	0.00216

- (3.3) What does the data suggest about the dumpy mutation?

Table 7: Observed and Expected Counts

	Long	Dumpy
O	792	208
ECount	750	250

```
##  (3.3) Some things we leanred from the data
##    - The chi-square test rejects the null.
##    - The dumpy mutation seems to be recessive because all the F1 had long wings.
##    - One possible explanation of the discrepancy between predicted and observed proportions
##      is that the survival rate of Long and Dumpy wing flies is different.
##    - Another possible explanation is that the Long-wing line was not fully inbred,
##      but for this to hold the proportion of heterozygous must be very small to produce,
##      by chance, an F1 of only long wing flies.
```

Question 4

In this question we will use data from the 1000 Genomes (1KG) project. This project sampled individuals from diverse ancestry groups and fully sequenced their genomes.

You can load the data set in an R environment using the following script.

```
root="https://www.dropbox.com/scl/fi/"
genoFile="e0fhhoyscig34oo2amcwl/genos.csv?rlkey=66qgdanwo45egad24hc9khkkh&dl=1"
GENOS=read.csv(paste0(root,genoFile),row.names=1)
GENOS=as.matrix(GENOS)

famFile="p6hw5zv8q00bnnpeo11y3/FAM.csv?rlkey=f1p8zmeakelej8rx357p1vkb8&dl=1"
FAM=read.csv(paste0(root,famFile))

dim(GENOS)
```

```
## [1]   298 10642
```

```
head(FAM)
```

```
##       FID     IID Sex Population
## 1 HG00171 HG00171   2        FIN
## 2 HG00173 HG00173   2        FIN
## 3 HG00174 HG00174   2        FIN
## 4 HG00176 HG00176   2        FIN
## 5 HG00177 HG00177   2        FIN
## 6 HG00178 HG00178   2        FIN
```

The `GENO` matrix has genotypes of 298 individuals at 10642 SNPs.

Thw `FAM` file has the family and individual ID, Sex, and Population (ancestry) of the individual.

The `rownames` of `GENOS` are based on the family and individual IDs.

```
all(rownames(GENOS)==paste0(FAM$FID,'_',FAM$IID))
```

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

(4.1) Provide a table with the number of individuals per country of origin.

- CHB=Chinese
- LWK=Kenya
- FIN=Finland

```
ANS=table(FAM$Population)
knitr:::kable(ANS,digits=3,caption='=> (4.1) Number of individuals per country of origin.')
```

Table 8: => (4.1) Number of individuals per country of origin.

Var1	Freq
CHB	103
FIN	99
LWK	96

(4.2) Allele Frequencies, Inbreeding, and HWE test within population Create three data frames (one for each population) each with the following columns:

- snp (the snp ID),
- a1 (the allele being counted),
- a1_freq (the frequency of the a1 allele),
- maf (or minor-allele-frequencies, i.e., the frequency of the least frequent allele)
- F, the inbreeding coefficient of the SNP
- HWEPval (the p-value from the HWE test)

```
snp=substr(colnames(GENOS),start=1,stop=nchar(colnames(GENOS))-1)
a1=substr(colnames(GENOS),start=nchar(colnames(GENOS)),stop=nchar(colnames(GENOS)))
MAP=data.frame(snp=snp,a1=a1,a1_freq=NA,maf=NA,F=NA,HWEPval=NA)
```

```
# A function to get the HWE p-value
getHWEPval=function(x){

 p=mean(x,na.rm=TRUE)/2
 O=table(factor(x,levels=0:2))

 n=sum(O)
 nAlleles=2*n
 E=n*c((1-p)^2,2*p*(1-p),p^2)

 Chisq=sum(((O-E)^2)/E)
 pVal=pchisq(Chisq,df=1,lower.tail=FALSE)
 return(pVal)
}

getF=function(x){
  pA=mean(x,na.rm=TRUE)/2
  O=mean(x==1,na.rm=TRUE)
  E=2*pA*(1-pA)
  PI=O/E
  F=1-PI
  return(F)
}

# China
RES=MAP
DATA=GENOS[FAM$Population=='CHB',]
RES$a1_freq=colMeans(DATA,na.rm=TRUE)/2
RES$maf=ifelse(RES$a1_freq<.5,RES$a1_freq,1-RES$a1_freq)
RES$HWEPval=apply(FUN=getHWEPval,X=DATA,MARGIN=2)
RES$F=apply(FUN=getF,X=DATA,MARGIN=2)
RES_CHB=RES

# Kenya
RES=MAP
DATA=GENOS[FAM$Population=='LWK',]
RES$a1_freq=colMeans(DATA,na.rm=TRUE)/2
RES$maf=ifelse(RES$a1_freq<.5,RES$a1_freq,1-RES$a1_freq)
RES$HWEPval=apply(FUN=getHWEPval,X=DATA,MARGIN=2)
RES$F=apply(FUN=getF,X=DATA,MARGIN=2)
```

```
RES_LWK=RES

# Finland
RES=MAP
DATA=GENOS[FAM$Population=='FIN',]
RES$a1_freq=colMeans(DATA,na.rm=TRUE)/2
RES$maf=ifelse(RES$a1_freq<.5,RES$a1_freq,1-RES$a1_freq)
RES$HWEPval=apply(FUN=getHWEPval,X=DATA,MARGIN=2)
RES$F=apply(FUN=getF,X=DATA,MARGIN=2)
RES_FIN=RES

print(head(RES_CHB))
```

```
##           snp a1     a1_freq         maf            F   HWEPval
## 1 rs62053745_  C 0.834951456 0.165048544 -0.127222982 0.1966440
## 2  rs9747082_  A 0.834951456 0.165048544 -0.127222982 0.1966440
## 3 rs62053747_  A 0.019417476 0.019417476 -0.019801980 0.8407235
## 4 rs34151105_  C 0.000000000 0.000000000          NaN       NaN
## 5 rs77383171_  T 0.004854369 0.004854369 -0.004878049 0.9605154
## 6 rs75157665_  A 0.000000000 0.000000000          NaN       NaN
```

```
print(head(RES_LWK))
```

```
##           snp a1    a1_freq        maf           F   HWEPval
## 1 rs62053745_  C 0.56250000 0.43750000 -0.05820106 0.5685071
## 2  rs9747082_  A 0.71875000 0.28125000  0.12399356 0.2244104
## 3 rs62053747_  A 0.01041667 0.01041667 -0.01052632 0.9178547
## 4 rs34151105_  C 0.00000000 0.00000000         NaN       NaN
## 5 rs77383171_  T 0.37500000 0.37500000  0.11111111 0.2763029
## 6 rs75157665_  A 0.00000000 0.00000000         NaN       NaN
```

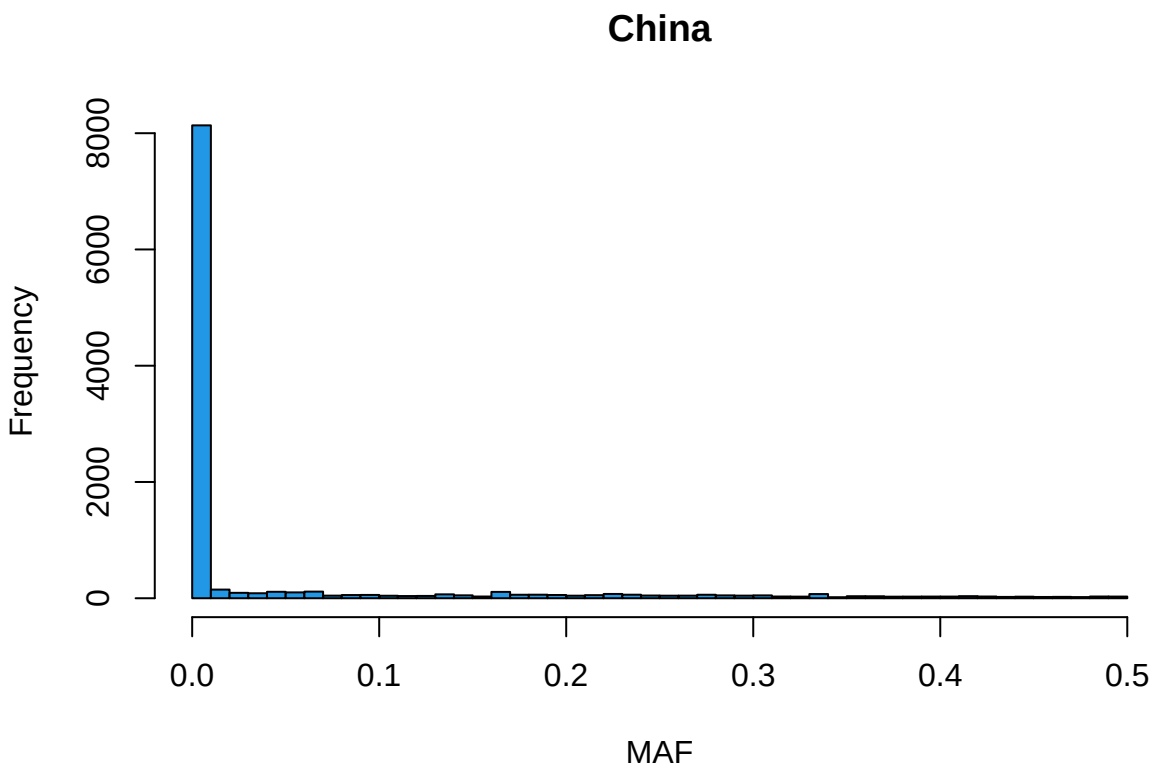
```
print(head(RES_FIN))
```

```
##           snp a1   a1_freq       maf           F   HWEPval
## 1 rs62053745_  C 0.7979798 0.2020202  0.06012658 0.5496718
## 2  rs9747082_  A 0.8282828 0.1717172 -0.06527977 0.5159988
## 3 rs62053747_  A 0.1313131 0.1313131  0.02593918 0.7963363
## 4 rs34151105_  C 0.1262626 0.1262626 -0.05294798 0.5983146
## 5 rs77383171_  T 0.1111111 0.1111111  0.07954545 0.4286714
## 6 rs75157665_  A 0.1262626 0.1262626 -0.05294798 0.5983146
```

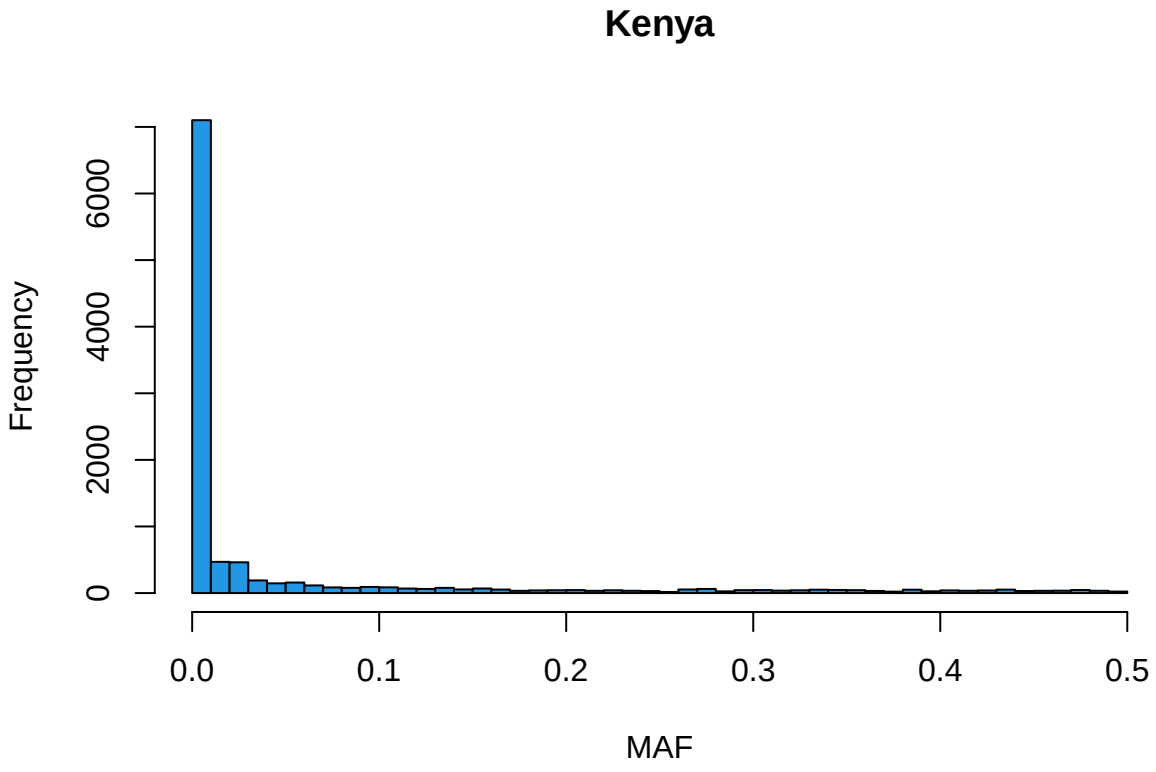
For each of the populations present:

- (4.2.1) A histogram of minor-allele frequency

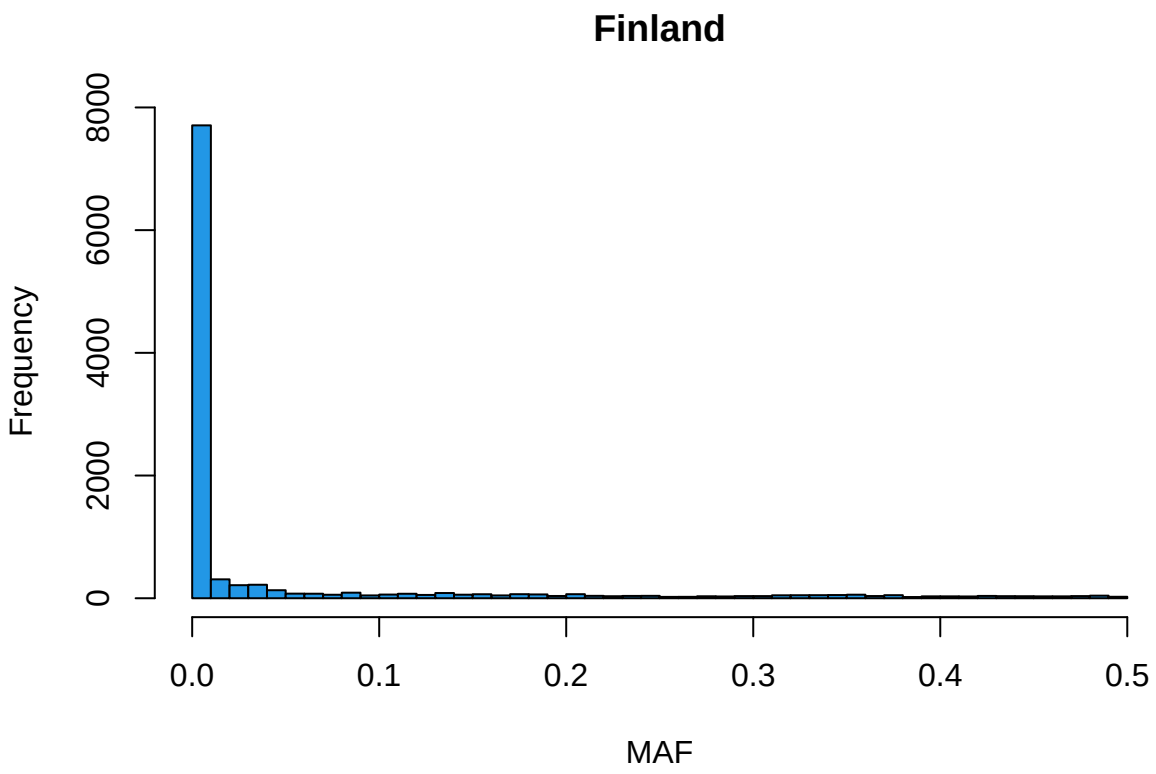
```
hist(RES_CHB$maf,50,main='China',xlab='MAF',col=4)
```




```
hist(RES_LWK$maf,50,main='Kenya',xlab='MAF',col=4)
```

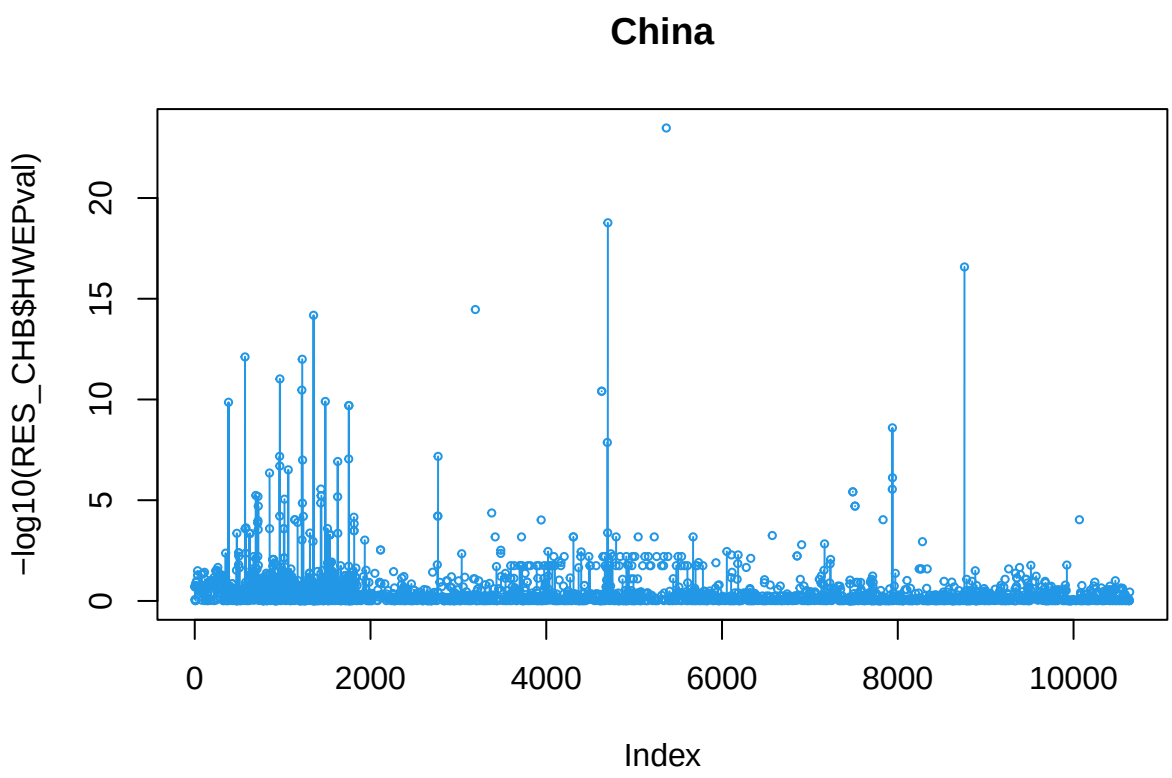



```
hist(RES_FIN$maf,50,main='Finland',xlab='MAF',col=4)
```

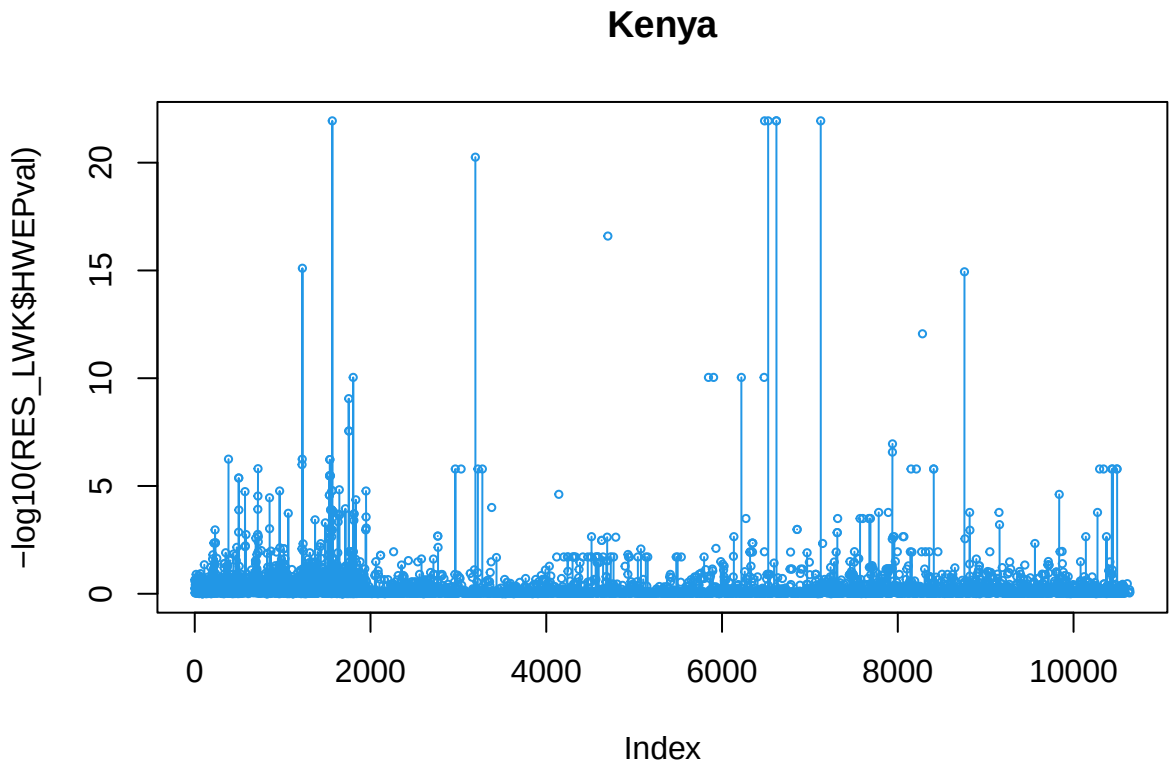


- (4.2.2) A plot with -log10(pValue HWE) on the y axis, and SNP order in the x-axis

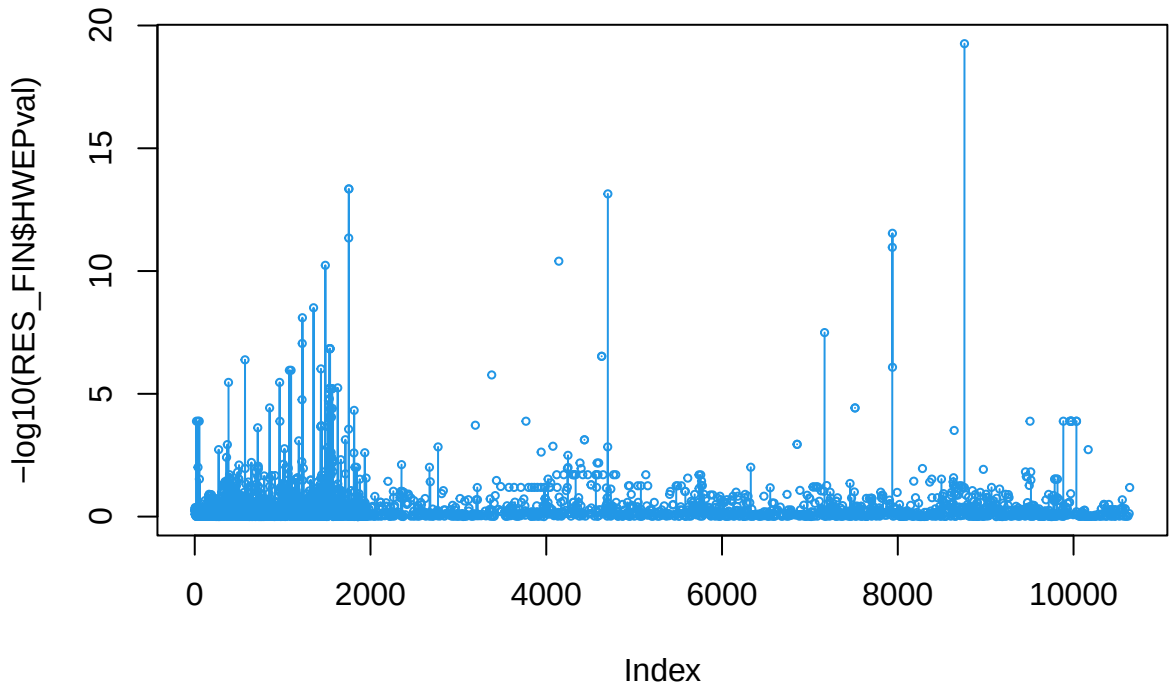
```
plot(-log10(RES_CHB$HWEPval),main='China',type='o',cex=.5,col=4)
```



```
plot(-log10(RES_LWK$HWEPval),main='Kenya',type='o',cex=.5,col=4)
```

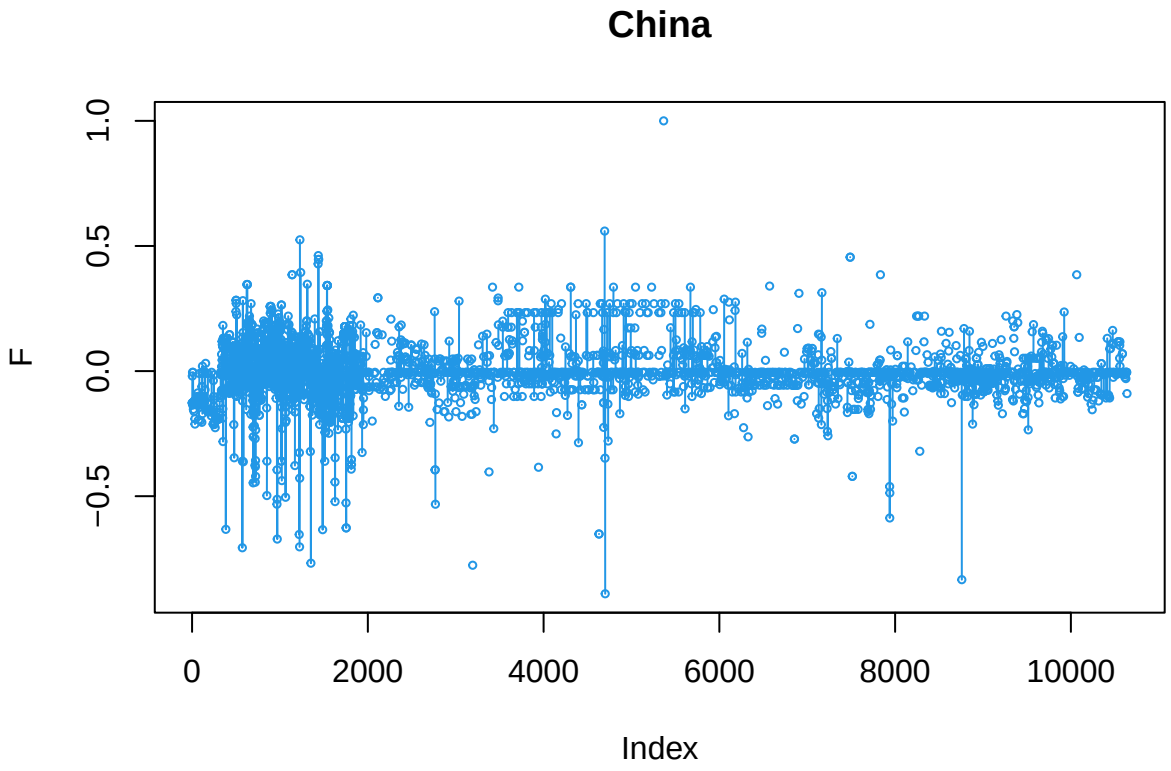



```
plot(-log10(RES_FIN$HWEPval),main='Finland',type='o',cex=.5,col=4)
```

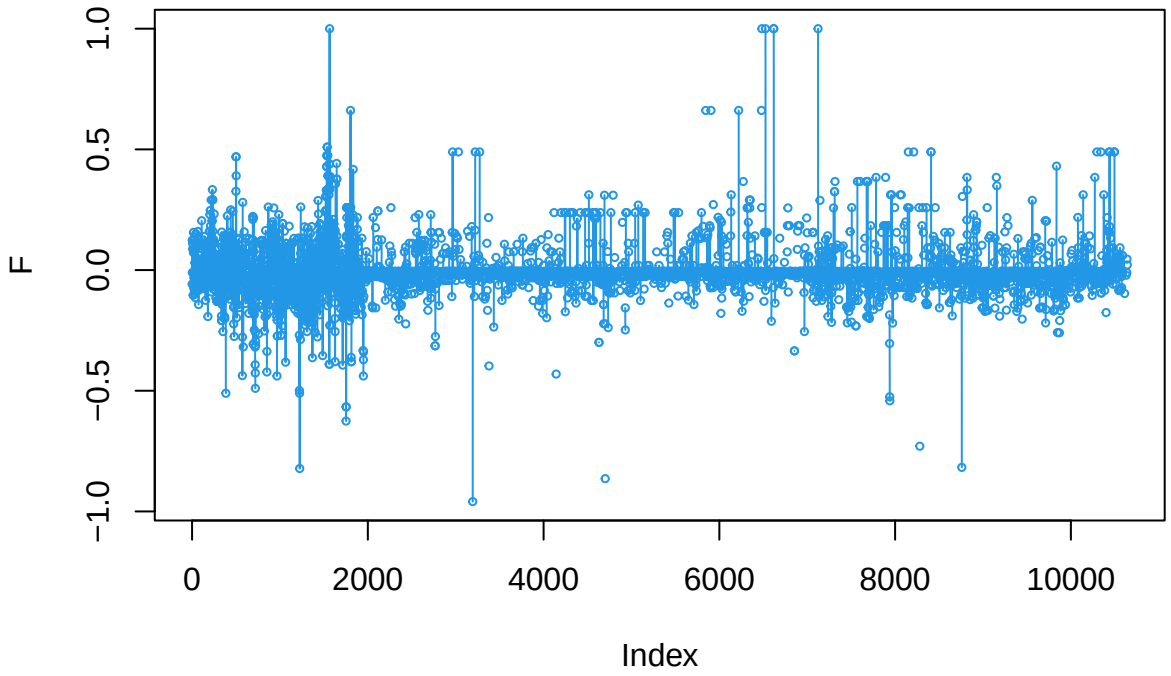


- (4.2.3) A plot with the inbreeding coefficient on the y axis and SNP order in the x-axis.

```
plot(RES_CHB$F,main='China',type='o',cex=.5,col=4,ylab='F')
```




```
plot(RES_LWK$F,main='Kenya',type='o',cex=.5,col=4,ylab='F')
```



```r
plot(RES_FIN$F,main='Finland',type='o',cex=.5,col=4,ylab='F')
```