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?

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.

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

```
# Insert your code here
```

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

- 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?

```
# Insert your code here, print only the relevant outcome
```

- Does the data support the hypothesis?

```
# Insert your code here, print only the relevant outcome
```

- What does the data suggest about the dumpy mutation?

(Type your answer here)

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`

```
# Insert your code here, print only the relevant outcome
```

4.2) Allele Frequencies 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)

- HWEPval (the p-value from the HWE test)

```
 # Insert your code to create the three data frame here, do not print any results
```

For each of the populations present:

- A histogram of minor-allele frequency,
- A plot with -log10(pValue HWE) on the y axis, and SNP order in the x-axis,

```
# Insert your code here, print the three histograms and the three plots
# Add titles to the plots  eg, plot(x,...,main='HWE p-value, Chinese')
```