# qQTL exercise

Mahdi

September 25, 2020

## Understanding the data

### Task 1

1. In the `sub_geno.tab` file, 0, 1 and 2 most likely represent the two homozygous and hgeteroguzous genotypes. -1 probably means missing data.
2. In the `sub_expr.tab` file, the rows are genes/transcripts and the columns are different samples and their gene expression values.
3. The `design.tab` file contains information about each column of the `sub_expr.tab` file. It says which population they belong to and other characteristics.

### Task 2

1. Calculate the number of missing genotypes for each SNP across all individuals.

```
snps <- read.table("sub_geno.tab")
missing_count <- apply((snps == -1), 1, sum)
head(missing_count)
```

```
## snp_22_30772686 snp_22_34965577 snp_22_49436707 snp_22_30631851 snp_22_46215888
##               0               0               0               0               0
## snp_22_34153853
##               0
```

2. Calculate the minor allele frequency (MAF) for all SNPs across all individuals.

```
get_geno_frequency <- function(geno, snp_mat){
  boolean_matrix <- snp_mat == geno
  count <- apply(boolean_matrix, 1, sum)
  freq <- count/dim(snps)[2]
  return (freq)
}

get_maf <- function(snp_mat){
  genotypes <- c(0,1,2)
  #get matrix of genotype frequencies
  geno_freq <- sapply(genotypes, get_geno_frequency, snp_mat)
  colnames(geno_freq) <- genotypes
  allele_freq <- geno_freq + geno_freq[,2]/2
  allele_freq <- allele_freq[,-2]
  maf <- apply(allele_freq, 1, min)
  return (maf)
}
```

```
maf <- get_maf(snps)
head(maf)
```

```
## snp_22_30772686 snp_22_34965577 snp_22_49436707 snp_22_30631851 snp_22_46215888
##     0.097402597     0.135281385     0.083333333     0.000000000     0.001082251
## snp_22_34153853
##     0.199134199
```

3. Filter our SNPs that have missing genotypes or a MAF<0.05 and use the filtered snps for the rest of the exercise.

```
keep <- maf > 0.05 & missing_count == 0
snps_filtered <- snps[keep,]
dim(snps_filtered)
```

```
## [1]  32 462
```

4. Calculate the MAF for africans and non-africans separately. Is there a difference?

```
design <- read.table("design.tab", header = T, sep = "\t")

filter_snp_population <- function(pop, snp_mat, design_mat, inv=F){
  if(inv){
    cols <- design_mat$Source.Name[design_mat$Characteristics.population. != pop]
  } else{
    cols <- design_mat$Source.Name[design_mat$Characteristics.population. == pop]
  }
  snp_mat <- (snp_mat[, cols])
  return(snp_mat)
}

african_snps <- filter_snp_population("YRI", snps_filtered, design)
non_african_snps <- filter_snp_population("YRI", snps_filtered, design, inv = T)

african_maf <- get_maf(african_snps)
non_african_maf <- get_maf(non_african_snps)
print("African")
```

```
## [1] "African"
```

```
print(head(african_maf))
```

```
## snp_22_30772686 snp_22_34965577 snp_22_49436707 snp_22_34153853 snp_22_21970216
##     0.002164502     0.049783550     0.004329004     0.076839827     0.067099567
## snp_22_48286671
##     0.000000000
```

```
print("Non-African")
```

```
## [1] "Non-African"
```

```
print(head(non_african_maf))
```

```
## snp_22_30772686 snp_22_34965577 snp_22_49436707 snp_22_34153853 snp_22_21970216
##      0.09523810      0.08549784      0.07900433      0.12229437      0.00000000
## snp_22_48286671
##      0.05735931
```

Yes, there is a difference between the two groups.