eQTL Homework Assignment

Solution to tasks 1-4.

Task 1: Take a look at the sub_geno.tab, sub_expr.tab and design.tab files.

a. What do the -1,0,1,2 values represent in the sub_geno.tab file?
b. What is sored in the sub_expr.tab file and what has been done with this data?
c. What information is stored in the design.txt file?

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

#Use dim() and head() to explore the data
```

Task 2: Genotype data

a. Calculate the number of missing genotypes for each SNP across all individuals.
b. Calculate the minor allele frequency (MAF) for all SNPs across all individuals.
c. Filter our SNPs that have missing genotypes or a MAF<0.05 and use the filtered snps for the rest of the exercise.

```
#For all individuals
missing <- rowCounts(as.matrix(geno),value=-1)
maf <- apply(geno,1,function(x) mean(x[x>=0]))/2
maf <- pmin(maf,1-maf)


filt_geno <- geno[maf>=0.05&missing==0,]

dim(geno)
```

```
## [1]  39 462
```

```
dim(filt_geno)
```

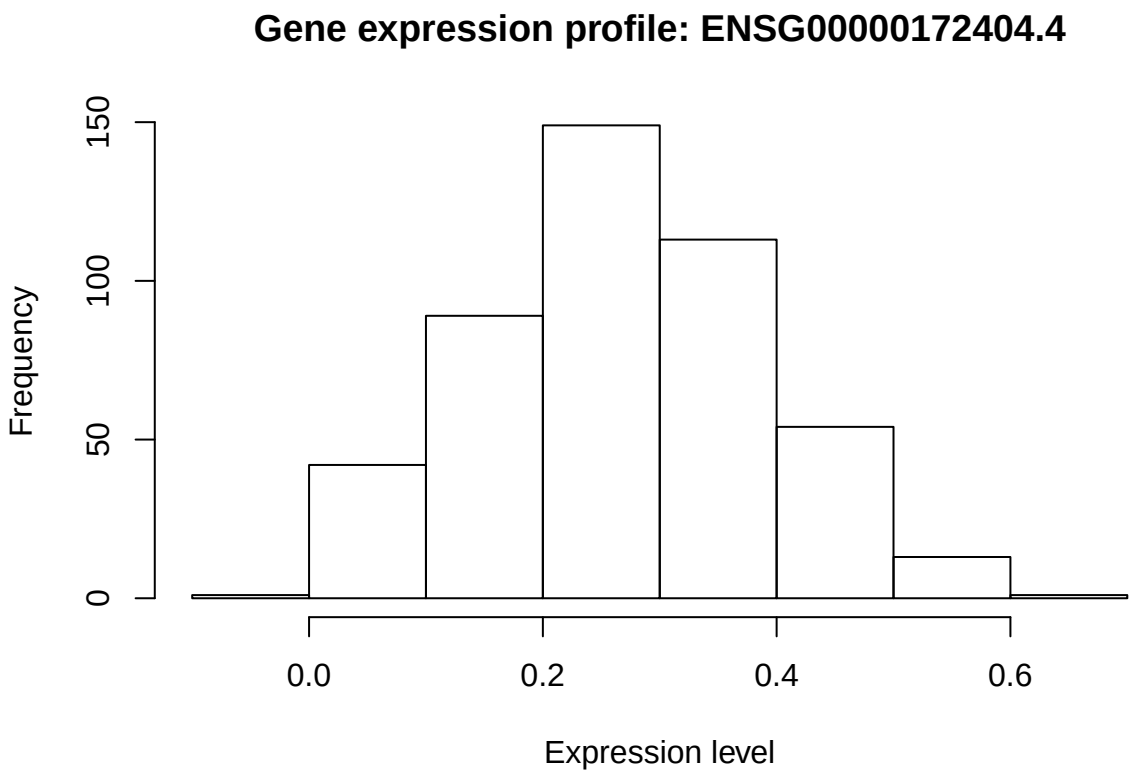
```
## [1]  32 462
```

Task 3: Gene expression profiles

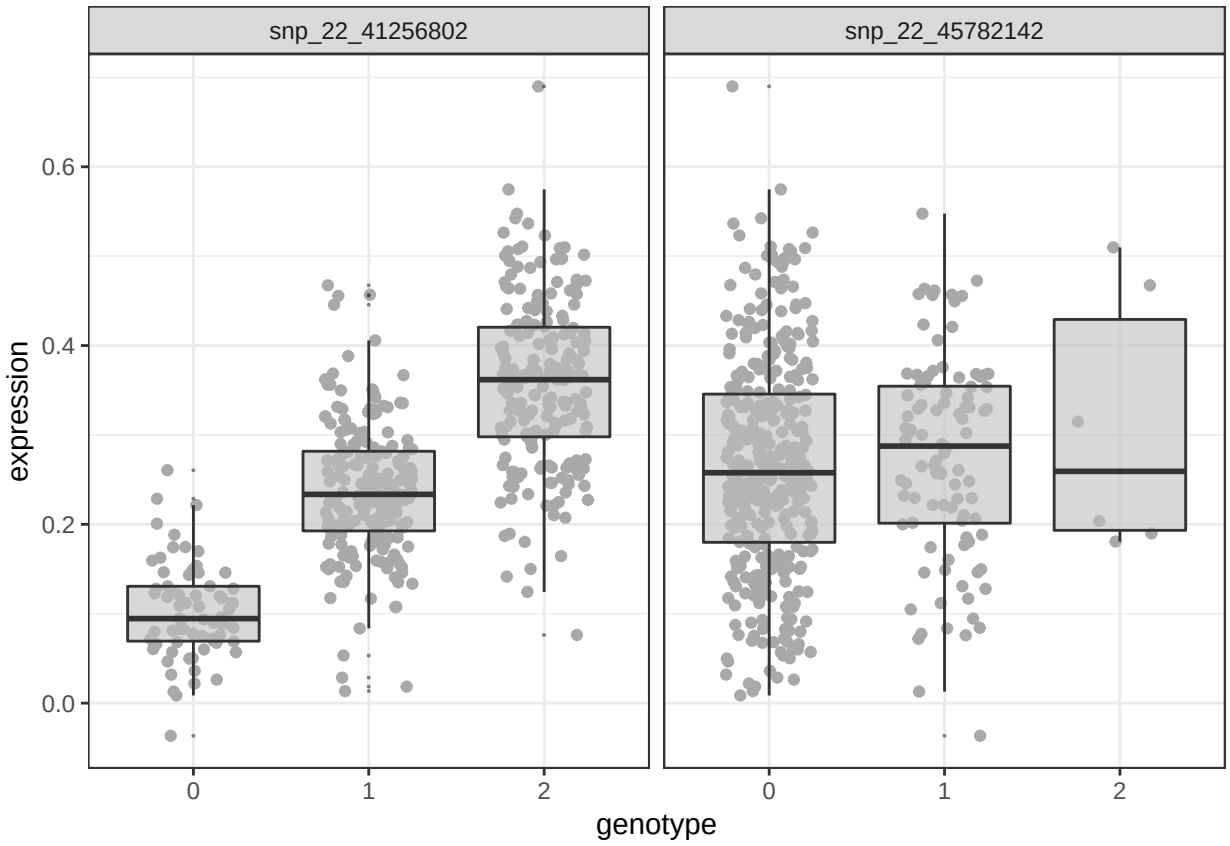
a. Plot the distribution of expression levels across all samples for the ENSG00000172404.4 gene
b. Plot the expression levels of ENSG00000172404.4 against the genotypes of snp_22_41256802 and snp_22_45782142

```
snps = c("snp_22_41256802", "snp_22_45782142")
genes = ("ENSG00000172404.4")

hist(as.matrix(expr[genes,]), main=paste("Gene expression profile:",genes),xlab="Expression level")
```




```
geneLong <- melt(expr[genes,])
snpLong <- melt(t(filt_geno[snps,]))
dataLong <- data.frame(cbind(snpLong[,2:3]),rbind(geneLong,geneLong))
colnames(dataLong) <- c("snp","genotype","sample","expression")
dataLong$genotype <- as.factor(dataLong$genotype)
ggplot(dataLong, aes(genotype, expression)) +
  geom_jitter(colour="darkgrey", position=position_jitter(width=0.25)) +
  geom_boxplot(outlier.size=0, alpha=0.6, fill="grey") +
  facet_wrap(~snp) + theme_bw()
```



Task 4: Do a linear regression of all sample genotypes on sample gene expression:

a. For snp_22_41256802 on ENSG00000172404.4
b. For snp_22_45782142 on ENSG00000172404.4

```
texpr <- t(expr)
tgeno <- t(filt_geno)
lm_a = lm(texpr[,"ENSG00000172404.4"] ~ tgeno[,"snp_22_41256802"])
summary(lm_a)
```

```
##
## Call:
## lm(formula = texpr[, "ENSG00000172404.4"] ~ tgeno[, "snp_22_41256802"])
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.28348 -0.04934 -0.00143  0.04950  0.33024
##
## Coefficients:
##                              Estimate Std. Error t value Pr(>|t|)
## (Intercept)                  0.109393   0.007824   13.98   <2e-16 ***
## tgeno[, "snp_22_41256802"] 0.125135   0.005391   23.21   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.08216 on 460 degrees of freedom
```

```
## Multiple R-squared:  0.5394, Adjusted R-squared:  0.5384
## F-statistic: 538.7 on 1 and 460 DF,  p-value: < 2.2e-16
```

```
lm_b = lm(texpr[,"ENSG00000172404.4"] ~ tgeno[,"snp_22_45782142"])
summary(lm_b)
```

```
##
## Call:
## lm(formula = texpr[, "ENSG00000172404.4"] ~ tgeno[, "snp_22_45782142"])
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.31360 -0.08515 -0.00588  0.07998  0.42486
##
## Coefficients:
##                             Estimate Std. Error t value Pr(>|t|)
## (Intercept)                 0.265040   0.006332  41.856   <2e-16 ***
## tgeno[, "snp_22_45782142"] 0.011987   0.012425   0.965    0.335
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1209 on 460 degrees of freedom
## Multiple R-squared:  0.002019,   Adjusted R-squared:  -0.0001502
## F-statistic: 0.9308 on 1 and 460 DF,  p-value: 0.3352
```

Part 1 Understanding the basics

In part 1 you are working with selected snps and genes from the Geuvadis consortium. You are supposed to work on this individually (not in groups)! Copying results/answers from others will result in you failing the assignment.

To pass this part of the homework you are required to:
1. Answer the following questions. Keep your answers short and to the point.
2. Solve the following tasks. Include the code you used for solving the tasks.

Questions 1-4:

1. What do the -1,0,1,2 values represent in the sub_geno.tab file?
2. What information is stored in the design.txt file?
3. Explain the results from the linear model in Task 4. What are the important values to look at and what do they tell you?

Task 5: Do a linear regression for snp_22_43336231 on ENSG00000100266.11

a. Without covariates
b. Using the genotype PCs from pc_cvrt.tab as covariates
c. Separately for african and non-africans without covariates. Hint: Use the information in the design.tab
d. Make a dotplot of PC1 vs PC2 and color the dots by population.

Questions 5:

1. Is there a difference in your results in a and b? If so explain why.
2. Is there a difference between african and non-africans? If so explain why.

3. What is it we are including in our model with the pc_cvrt.tab?

Task 6: Do a linear regression on 1st snp on 1st gene, 2nd snp on 2nd gene etc.

a. Create a matrix containing the gene_id, snp_id, effect size, t.value and p.value.
b. Do a multiple testing correction on the resulting p.values using fdr.
c. Do the same but now include the genotype PCs from pc_cvrt.tab as covariates.
d. Plot the most significant hit.

Questions 6:

1. How many tests did you perform in a? and c?
2. What are you correcting for with the fdr? Why is this important for eQTL analysis?
3. Is there a difference in number of significant hits (FDR<0.05) in the two models?

Task 7: Use this Matrix_eQTL_main function to do eQTL analysis on the data.

```
#Run if you have not installed MatrixEQTL
#install.packages("MatrixEQTL")
library(MatrixEQTL)


snps <- SlicedData$new()
snps$CreateFromMatrix(as.matrix(filt_geno)) #filt_geno is your filtered genotype matrix
genes <- SlicedData$new()
genes$CreateFromMatrix(as.matrix(expr)) #expr is the unchanged expression matrix

snp_pos <- read.table("sample_geno.pos",sep="\t",header=T)
snp_pos <- snp_pos[snp_pos$snp %in% row.names(filt_geno),]
gene_pos <- read.table("sample_expr.pos",sep="\t",header=T)

all(colnames(snps) == colnames(genes))

eQTL <- Matrix_eQTL_main(snps, genes, output_file_name=NULL,
  output_file_name.cis=NULL,
  pvOutputThreshold.cis=1, pvOutputThreshold=1,
  snpspos=snp_pos, genepos=gene_pos,
  cisDist = 0)
```

Questions 7:

1. How many tests were performed in the eQTL analysis?
2. Compare the results from MatrixeQTL to your results from Task 6 a and b. Explain any similarities and/or differences that you see.