

## Expression quantitative trait loci mapping.

### Part 1: eQTL simulation

eQTLs are genetic loci (single nucleotide polymorphisms, SNP) whose alleles are associated with different expression levels of a specific gene. Different alleles can be associated with a decrease or increase in gene expression. Figure 1 displays how a SNP (in red) can influence gene expression.

In this example, an A allele increases gene expression in a dose-response manner, with A homozygotes displaying higher levels of expression of a gene compared to G homozygotes.

Most eQTLs are found outside of coding regions making their identifications harder due to the distance between the variant and the gene of interest.

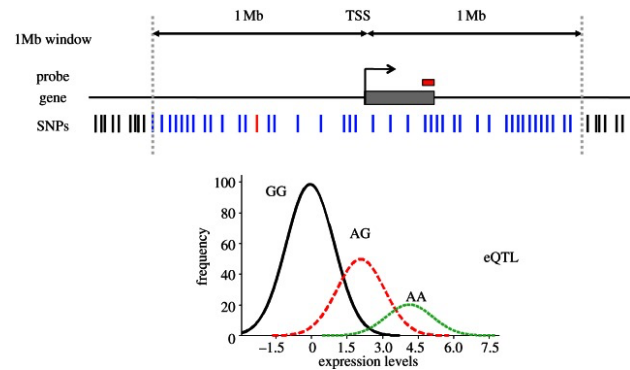


Figure 1. Representation of an eQTL effect on a gene. Figure taken from Nica, A. C. & Dermitzakis, E. T. Expression quantitative trait loci: present and future. *Philos. Trans. R. Soc. London. Ser. B, Biol. Sci.* **368**, 20120362 (2013).

In this practical, we will investigate how to identify eQTLs. To understand eQTL analysis, we will start by simulating both genotype and expression data. This simulation approach will allow us to better understand the structure of the data used for eQTL mapping. Once we finish performing the simulation, we will investigate the GTEx website and see how eQTL can be used to investigate genome-wide association study (GWAS) results.

Genetic data:

To identify eQTLs, we need to know the individual's genotype information. This genetic information will then be used to test the association between specific alleles and gene expression. Before simulating genetic data, let's look at how it is represented.

Question 1:

Think about a single genetic locus where the allele can either be A or T.

How would you represent four individuals whose genotypes are respectively:

- AA
- AT
- TA
- TT

We represent genetic information based on the number of alleles an individual carries at a specific locus. Individuals described within question 1 can therefore be represented based on

the number of A alleles each individual possesses:  $\begin{bmatrix} 2 \\ 1 \\ 1 \\ 0 \end{bmatrix}$ .

To simulate genetic data, we, therefore, have to create several vectors containing 0, 1 or 2 representing the genotype of a single individual at different loci. Those vectors can then be accumulated into a matrix with the columns representing different individuals and the row a specific genetic locus.

Connection to the cluster:

To simulate the data, we will need to connect to the High-Performance Computing (HPC) cluster set up for this class. You should have been given the credentials necessary to connect to the cluster. Follow the information presented in the introductory guide to connect to the HPC cluster.

In this practical, you can either use the command line or the interactive R session. If you are using the command line, you will have to download the plots to your local machine using the *scp* commands found within the text. If you are using an interactive R session, you can ignore those steps.

Simulation of genotypes:

First, we will set up the HPC folders to keep your analyses organised. The following bash script allow you to create three folders:

```
cd ~
mkdir eQTLPrac
cd eQTLPrac
mkdir Genotype
mkdir Expression
mkdir eQTL
```

You will now save and download files based on the folder that we created earlier. The following R code can be used to simulate genetic data. Start your R session and , copy, and paste it within the interpreter.

```
set.seed(6543456)
frequency <- 0.5
SNP <- rbinom(5000, size = 2, frequency)
SNP_number <- 10000
indv_number <- 5000
p <- runif(SNP_number, min = 0, max = 0.9) # probability of each alleles.
genotypes <- replicate(indv_number, rbinom(SNP_number, 2, p)) # Generate our genotype matrix
rownames(genotypes) <- paste0('SNP', seq(1, nrow(genotypes)))
colnames(genotypes) <- paste0('Indv', seq(1, ncol(genotypes)))
print(nrow(genotypes))
print(ncol(genotypes))
print(genotypes[1:10,1:10])
```

Note: The *set.seed* function allow the code to be reproducible, by fixing the random processes. A different seed would change the results.

#### Question 2:

- How many SNPs were simulated?
- How many individuals were simulated?
- Given that the SNP3 reference allele is G and the alternate allele C, what is the genotype of individual 5?

Now that we simulated genotype data, we can calculate the frequency of the alleles simulated with the following code:

```
library(MASS)
maf = rowMeans(genotypes)/2
maf <- pmin(maf, 1-maf)

jpeg('~/.eQTLPrac/Genotype/HistogramMAFsimulated.jpeg',width = 21, height = 12, res = 300, units =
'cm')
truehist(maf, main = "Histogram of minor allele frequency", col = "light grey")
lines(density(maf), lty = 2, col = "dark red", lwd = 3)
dev.off()
```

You can download the plot that you created by using the following command on your local machine:

```
scp <username>@203.101.xxx.xxx: ~/.eQTLPrac/Genotype/HistogramMAFsimulated.jpeg .
```

#### Question 3:

Look at the allele frequency of the genotype data you simulated.

- What is the allele frequency occurring more often?
- Why is allele frequency important for eQTL analysis?

Allele frequency is important during eQTL analysis due to the lack of representation of some genotype.

For eQTL analysis to be possible, it needs to include individuals with all possible genotypes (0, 1 or 2 alleles). If we consider a genetic locus with a minor allele frequency (MAF) of 1%, we will have the following proportions of individuals:

*Table 1. Proportion of the possible genotypes for a genetic loci with a minor allele frequency of 0.01 (Calculation based on the Hardy–Weinberg principle)*

	Proportion	1,000 individuals	10,000 individuals	100,000 individuals
AA	0.9801	980.1	9801	98010
AT	0.0198	19.8	198	1980
TT	0.0001	0.1	1	10

To identify eQTL with a minor allele frequency of 1%, we would therefore need a population larger than 100,000 individuals. Current large genetic studies such as the eQTLgen consortium<sup>1</sup> contain 31,684 individuals, we therefore need to pay attention to the allele frequency before performing eQTL mapping.

Simulation of gene expression data:

Now that we simulated genetic data, we need to create matching gene expression data. While gene expression is not normally distributed, most analyses will start by normalising the data. As such, simulating gene expression data can be performed either at the discrete level or at the normalised level.

```
set.seed(58944)
genesTotal <- 5000
geneswithQTL <- 2000
geneswithoutQTL <- genesTotal - geneswithQTL
# Select the SNPs associated with each of the gene:
SNPs <- rownames(genotypes)
SNPswithQTL <- sample(SNPs, size = geneswithQTL)
SNPswithoutQTL <- SNPs[-which(SNPs %in% SNPswithQTL)]
# Create the expression matrix for associated SNPs
expMatrixNotAssociated <- do.call(cbind, lapply(SNPswithoutQTL, function(x) {
  meanForAlleles <- c(rnorm(1,10))
  yWithQTL <- rnorm(indv_number, meanForAlleles)
  return(yWithQTL)}))
#Associated genes:
expMatrixAssociated <- do.call(cbind, lapply(SNPswithQTL, function(i) {
  meanForAlleles <- c(rnorm(1,5), rnorm(1,8), rnorm(1,10))
  yWithQTL <- rnorm(indv_number, meanForAlleles[factor(genotypes[i,])])
  return(yWithQTL)
}))
colnames(expMatrixAssociated) <- paste0('Gene', 1:ncol(expMatrixAssociated))

print(geneswithQTL)
print(geneswithoutQTL)
print(ncol(t(expMatrixAssociated)))
```

Question 5:

- How many genes were simulated?
  - How many of those were associated with SNPs?

The following code will generate plots showing the association between SNPs and genotypes:

```

library(ggplot2);library(cowplot)
SNPassociationPlot <- function(SNPID, GeneID) {
  ggplot(data.frame(snp=genotypes[SNPID,], y=expMatrixAssociated[,GeneID]),
    aes(x = factor(snp), y = y)) +
  ggtitle(paste0('Association between Gene 2 and ', SNPID)) +
  geom_boxplot(fill='dark red') + geom_point(col='dark grey') + xlab("Reference allele count") +
  theme_minimal()+ theme(plot.title = element_text(hjust = 0.5))
}
p1 <- SNPassociationPlot(SNPID = 'SNP8621', GeneID = 'Gene2')
p2 <- SNPassociationPlot(SNPID = 'SNP2044', GeneID = 'Gene2')
p3 <- SNPassociationPlot(SNPID = 'SNP9521', GeneID = 'Gene2')
p4 <- SNPassociationPlot(SNPID = 'SNP5564', GeneID = 'Gene2')
p <- plot_grid(p1,p2,p3,p4)
ggsave(p, filename = '~/eQTLPrac/Expression/AssociationPlot.jpeg', width = 14, height=14, dpi =
300)

```

Download the plot created using the following command:

```
scp <username>@203.101.229.143:~/eQTLPrac/Expression/AssociationPlot.jpeg .
```

#### Question 6:

Based on the plot that you generated answer the following questions:

- What is the mean expression of the gene you simulated?
- Which SNP (if any) is associated with the expression of Gene2 ?
  - How would you identify SNP statistically associated with gene expression?

You can change the previous code (by change the name of the SNPs in red) to visually inspect the association between different SNPs and genes.

While identifying SNP and gene expression pair visually is already time-consuming, the human genome is composed of 3.2 billion base pairs and roughly 20,000 genes rendering it impossible.

eQTL mapping, simple linear regression:

The most common way of testing the association between a SNP and gene expression is to perform a linear regression. This linear regression of the form:

$$y = \beta_0 + \beta_1 G_i + \varepsilon,$$

With  $y$  being a vector containing the expression of a gene for all individuals,  $\beta_1$  the effect of an allele on gene expression,  $G_i$  a vector containing the genotype of each individual for a specific genetic loci and  $\varepsilon$  being an error term.

To find the association between a specific gene and genetic loci we need to estimate  $\beta_1$ , the effect of each SNP on gene expression. The following code will perform a linear regression on for all SNPs and gene2:

```

library(tidyverse)
GeneID='Gene2'

# Set the first test:
Association <- summary(lm(expMatrixAssociated[,GeneID]~genotypes['SNP1',]))
Association <- as.data.frame(Association$coefficients)[2,]
rownames(Association) <- SNPID

for(SNPID in rownames(genotypes)){
  test <- summary(lm(expMatrixAssociated[,GeneID]~genotypes[SNPID,]))
  test <- as.data.frame(test$coefficients)[2,]
  rownames(test) <- SNPID
  Association <- rbind(Association, test)
}
colnames(Association) <- c("Estimate", "Std.Error", "t_value", "P")
Association %>% arrange(P) %>% head()

```

#### Question 7:

- What SNP is significantly associated with Gene 2?
  - Let's assume that the reference allele of that SNP is A and alternate allele is T
    - What will be the expected gene expression of an individual with a genotype of AA?
    - With a genotype of TT?
- Use the code used previously to plot the association between the significant SNP and Gene 2 and save it in the eQTL folder

The association test between Gene 2 and 10,000 SNPs that we just performed took a few minutes. We can see how eQTL analyses quickly result in an exponential computation time as we increase the number of SNPs and individuals tested. Software such as `matrixeQTL`<sup>2</sup> and `fastQTL`<sup>3</sup> have been developed to decrease the computational resources and time necessary for eQTL analyses. While we will not go into details on their working here, the underlying mechanisms of that software remain similar to the analysis performed within this practical. Methodology used to improve computational efficiency range from limiting the SNPs tested for a gene to the closest SNPs to developing mathematical approximations to computationally heavy calculations.

## Part 2: Real world eQTL:

Genotype-Tissue Expression (GTEx):

We will now investigate real-world eQTLs data. For this, we will go to the GTEx website. You can access it through this [link \(https://gtexportal.org/home/\)](https://gtexportal.org/home/).

The GTEx consortium collected post-mortem samples for 948 donors. We know that eQTLs are dynamic and evolve over time and with exposure to the environment. Characteristics such as sex, age or disease status can influence eQTL association and are therefore important.

Question 8:

- On the GTEx website, look for the sample characteristics that could influence eQTL association study.
  - *Hint: Navigate to the Tissue & Sample statistics page*

eQTL are influenced by both age<sup>4</sup>, sex<sup>5</sup> and ancestry<sup>6</sup>; the observed unbalanced number of males and females, as well as a largely white and aging (84.6% white, 68.1% of samples older than 50) cohort, therefore, need to be taken into account when performing eQTL analysis. Additionally, the cohort can be split in half with younger donors succumbing to traumatic injury while older donors displaying non-traumatic pathologies.

Sample characteristics, therefore, need to be considered when performing QTL mapping. You can read the landmark GTEx publication in 2020<sup>7</sup> to observe which sample characteristics were corrected for when testing for QTL associations.

Investigation of GWAS signal:

We will now investigate a real example of an eQTL association. For this, we will start by looking at a genome-wide association study of lipids published in 2013<sup>8</sup>:

*Discovery and refinement of loci associated with lipid levels*  
(<https://www.nature.com/articles/ng.2797>)

Question 9:

- Read the abstract of the GWAS paper, what is the goals of this paper?

This paper aimed to identify the genetic control of blood lipid levels. As such, they identified associations between SNP and blood lipid levels. They then mapped those SNPs to the closest genes, concluding on their role on blood lipid levels.

We will investigate how eQTL can give us more information regarding the genetic control of blood low-density lipoprotein (LDL) cholesterol.



Open the Supplementary figures from the paper and go to the supplementary table 3.

Question 10:

- Finds the gene with the strongest negative effect on LDL blood levels.
  - What is the impact of each alternate allele?
  - If the average person has an LDL blood level of 209.7mg/dL, what would be the expected LDL level of an individual with a genotype of GG at locus rs6511720?

Let's investigate the effect of rs6511720, the genetic loci associated with the highest decrease in LDL blood levels. Search the GTEx website for rs6511720 and answer the following question:

Question 11:

- With which genes is rs6511720 associated?
- In which tissues are those association located?
- Do you think that a change in gene expression is responsible for the association observed between LDL levels and rs6511720?

We will now look at genetic loci associated with LDL cholesterol levels. rs12916 is associated with HMGCR, a gene coding for HMG-CoA reductase an enzyme playing a central role in cholesterol synthesis. Let's investigate eQTL associated with rs12916, search the GTEx website for rs12916.

Question 12:

- In which tissue is rs12916 associated with HMGCR?
- Where does the SNP fall? (*hint: open the IGV browser*)
- Do you think that a change in gene expression is responsible for the association observed between LDL levels and rs12916?

In conclusion, eQTL can help interpreting the functional significance of GWAS signals. They can provide biological interpretation of non-coding variants helping to hint at the mechanisms underlying complex traits and diseases.

## References:

1. Võsa, U. *et al.* Large-scale cis- and trans-eQTL analyses identify thousands of genetic loci and polygenic scores that regulate blood gene expression. *Nat. Genet.* **53**, 1300–1310 (2021).
2. Shabalin, A. A. Matrix eQTL: ultra fast eQTL analysis via large matrix operations. *Bioinformatics* **28**, 1353–1358 (2012).
3. Ongen, H., Buil, A., Brown, A. A., Dermitzakis, E. T. & Delaneau, O. Fast and efficient QTL mapper for thousands of molecular phenotypes. *Bioinformatics* **32**, 1479–1485 (2016).
4. Yamamoto, R. *et al.* Tissue-specific impacts of aging and genetics on gene expression patterns in humans. *Nat. Commun.* **13**, 5803 (2022).
5. Oliva, M. *et al.* The impact of sex on gene expression across human tissues. *Science* **369**, (2020).
6. Zeng, B. *et al.* Multi-ancestry eQTL meta-analysis of human brain identifies candidate causal variants for brain-related traits. *Nat. Genet.* **54**, 161–169 (2022).
7. The GTEx Consortium atlas of genetic regulatory effects across human tissues. *Science* **369**, 1318–1330 (2020).
8. Willer, C. J. *et al.* Discovery and refinement of loci associated with lipid levels. *Nat. Genet.* **45**, 1274–1283 (2013).