Topics in advanced bioinformatics Homework Part II

Part II

In this part of the exercise, we explore cis and trans eQTLs for gene located on chromosome 20. The data used for this homework is generated by the GEUVADIS project. The expression data (RPKM unit) is from from LCL cell lines. To run the eQTL analysis, we will use `MatrixEQTL`

```
install.packages("MatrixEQTL")
```

Data files

Below are listed all the files needed for this exercise. This data can be downloaded from `ricco` path `/home/hjolli/ATB19/homework/`

```
# Expression data for chromosome 20
"expr_ceu_chr20.tab"
# Gene positions for genes on chromosome 20
"expr_chr20.pos"
# Genotype data for chromosome 20
"geno_ceu_chr20_strict.tab"
# Position of genotype data for chromosome 20
"geno_ceu_chr20_strict.pos"
# Genotype data for chromosome 22
"geno_ceu_chr22_strict.tab"
# Position of genotype data for chromosome 22
"geno_ceu_chr22_strict.pos"
```

Task 1

First, we will explore the data.

1. How many samples are included in this dataset?
2. How many variants are present on chromosome 20?
3. Generate a histogram of allele frequencies for chromosome 20.
4. What is the lowest allele frequency observed?
5. How many genes are included?
6. What gene shows the highest mean expression?

Task 2 - cis-eQTL

In this task, we will calculate all the cis-eQTLs on chromosome 20. A SNP must be at most $1e6$ bases from a gene to be considered for this analysis. Run the following code to calculate the statistics for all cis SNP-Gene pairs. The output will be saved to a data frame named `cis_eqtls`.

```
#Matrix eQTL
library(MatrixEQTL)

# Genotype file names
SNP_file_name = "geno_ceu_chr20_strict.tab" ; #Genotype file path
snps_location_file_name = "geno_ceu_chr20_strict.pos" ; #snp position file path

# Gene expression file names
```

```
expression_file_name = "expr_ceu_chr20.tab" ;#Expression file path
gene_location_file_name = "expr_chr20.pos" ;#gene position file path

# Only associations significant at this level will be saved
pvOutputThreshold_cis = 1; #p.value threshold for cis eqtls
pvOutputThreshold_tra = 0; #p.value threshold for trans eqtls

#Covariates file names
covariates_file_name = character();# Set to character() for no covariates

# Distance for local gene-SNP pairs
cisDist = 1e6; #Define cis distance

## Load genotype data
snps = SlicedData$new();
snps$fileDelimiter = "\t";      # the TAB character
snps$fileOmitCharacters = "NA"; # denote missing values;
snps$fileSkipRows = 1;          # one row of column labels
snps$fileSkipColumns = 1;       # one column of row labels
snps$fileSliceSize = 2000;      # read file in slices of 2,000 rows
snps$LoadFile(SNP_file_name);

## Load gene expression data
gene = SlicedData$new();
gene$fileDelimiter = "\t";      # the TAB character
gene$fileOmitCharacters = "NA"; # denote missing values;
gene$fileSkipRows = 1;          # one row of column labels
gene$fileSkipColumns = 1;       # one column of row labels
gene$fileSliceSize = 2000;      # read file in slices of 2,000 rows
gene$LoadFile(expression_file_name);

#Load position files
snpspos = read.table(snps_location_file_name, header = TRUE, stringsAsFactors = FALSE);
genepos = read.table(gene_location_file_name, header = TRUE, stringsAsFactors = FALSE);

## Run the analysis
me = Matrix_eQTL_main(
  snps = snps,
  gene = gene,
  output_file_name=NULL,
  pvOutputThreshold = pvOutputThreshold_tra,
  useModel = modelLINEAR,
  errorCovariance =numeric(),
  verbose = TRUE,
  output_file_name.cis  = NULL, #Do not write out cis results
  pvOutputThreshold.cis = pvOutputThreshold_cis,
  snpspos = snpspos,
  genepos = genepos,
  cisDist = cisDist,
  min.pv.by.genesnp = FALSE,
  noFDRsaveMemory = FALSE,
  pvalue.hist = FALSE)
```

```
cis_eqtls = me$cis$eqtls[,-c(5)]
cis_eqtls["beta_se"] = cis_eqtls["beta"]/cis_eqtls["statistic"]
rm(me)
```

Questions

1. How many tests were conducted?
2. Using a bonferroni correction ($\alpha = 0.05$), how many genes are significant?
3. What gene-snp pair show the lowest pvalue? What is the effect size of this snp-gene pair?
4. What is the biotype of this gene?

Task 3 - trans-eQTL

In this task, we will calculate all the trans-eQTLs for genes on chromosome 20. We will use a set of SNPs on chromosome 22. Run the following code to calculate the statistics for all trans SNP-Gene pairs. The output will be saved to a data frame named `trans_eqtls`.

```
#Matrix eQTL
library(MatrixEQTL)

# Genotype file names
SNP_file_name = "geno_ceu_chr22_strict.tab" ; #Genotype file path
snps_location_file_name = "geno_ceu_chr22_strict.pos" ; #snp position file path

# Gene expression file names
expression_file_name = "expr_ceu_chr20.tab" ;#Expression file path
gene_location_file_name = "expr_chr20.pos" ;#gene position file path

# Only associations significant at this level will be saved
pvOutputThreshold_cis = 0; #p.value threshold for cis eqtls
pvOutputThreshold_tra = 1; #p.value threshold for trans eqtls

#Covariates file names
covariates_file_name = character();# Set to character() for no covariates

# Distance for local gene-SNP pairs
cisDist = 1e6; #Define cis distance

## Load genotype data
snps = SlicedData$new();
snps$fileDelimiter = "\t";      # the TAB character
snps$fileOmitCharacters = "NA"; # denote missing values;
snps$fileSkipRows = 1;          # one row of column labels
snps$fileSkipColumns = 1;       # one column of row labels
snps$fileSliceSize = 2000;      # read file in slices of 2,000 rows
snps$LoadFile(SNP_file_name);

## Load gene expression data
gene = SlicedData$new();
gene$fileDelimiter = "\t";      # the TAB character
gene$fileOmitCharacters = "NA"; # denote missing values;
gene$fileSkipRows = 1;          # one row of column labels
gene$fileSkipColumns = 1;       # one column of row labels
gene$fileSliceSize = 2000;      # read file in slices of 2,000 rows
gene$LoadFile(expression_file_name);
```

```
#Load position files
snpspos = read.table(snps_location_file_name, header = TRUE, stringsAsFactors = FALSE);
genepos = read.table(gene_location_file_name, header = TRUE, stringsAsFactors = FALSE);

## Run the analysis
me = Matrix_eQTL_main(
  snps = snps,
  gene = gene,
  output_file_name=NULL,
  pvOutputThreshold = pvOutputThreshold_tra,
  useModel = modelLINEAR,
  errorCovariance =numeric(),
  verbose = TRUE,
  output_file_name.cis  = NULL, #Do not write out cis results
  pvOutputThreshold.cis = pvOutputThreshold_cis,
  snpspos = snpspos,
  genepos = genepos,
  cisDist = cisDist,
  min.pv.by.genesnp = FALSE,
  noFDRsaveMemory = FALSE,
  pvalue.hist = FALSE)

trans_eqtls = me$all$eqtls[,-c(5)]
trans_eqtls["beta_se"] = trans_eqtls["beta"]/trans_eqtls["statistic"]
rm(me)
```

Questions

1. How many tests were conducted?
2. Using a bonferroni correction ($\alpha = 0.05$), how many genes are significant?

Task 4 - QQ-plot

In this section we will explore QQ-plots (Quantile-Quantile plots) for both the cis-eQTLS and the trans-eQTLS.

The following function can be used to generate a QQ-plot of p-values:

```
qqp<-function(x, maxLogP=30,...){
    x<-x[!is.na(x)]
    if(!missing(maxLogP)){
        x[x<10^-maxLogP]<-10^-maxLogP
    }
    N<-length(x)
    chi1<-qchisq(1-x,1)
    x<-sort(x)
    e<- -log((1:N-0.5)/N,10)
    plot(e,-log(x,10),ylab="Observed log10(p-value)",xlab="Expected log10(p-value)",...)
    abline(0,1,col=2,lwd=2)

    c95<-qbeta(0.95,1:N,N-(1:N)+1)
    c05<-qbeta(0.05,1:N,N-(1:N)+1)
    lines(e,-log(c95,10))
    lines(e,-log(c05,10))
}
```

Questions

1. Briefly explain what a QQ-plot can be used for (2-3 sentences)
2. Compute the QQ-plot for both the cis and trans eqtl separately
3. Explain the plots
4. What is the main difference between these two QQ-plots?
5. Explain what drives this?

Task 5 - PVE

In the last exercise, we will calculate how much of the variance in gene expression can be explained by a SNP. This is called proporation of variance explained (PVE).

The formula for calculating PVE is shown below:

$$\text{PVE} = \frac{2\beta^2 \text{MAF}(1-\text{MAF})}{2\beta^2 \text{MAF}(1-\text{MAF}) + se(\beta)^2 2N\text{MAF}(1-\text{MAF})} \tag{1}$$

where N = number of samples, β = effect size, $se(\beta)$ = standard error of the effect size, and MAF = Minor allele frequency.

Questions

1. Calculate the PVE for all cis SNP-gene pairs and make a histogram of them
2. What gene has the highest PVE
3. what other factors can explain the remaining variance (mention 2)?