Hands on practical: Online databases exploration

Roderick Slieker, Bas Heijmans

12-10-2018

Investigate age-related changes across multiple tissues

Over the past years, an enormous amount of omics data has become publically available. This data is freely available and can be used by researchers across the world. In this practical we will download genome-wide DNA methylation data from the Gene Expression Omnibus (GEO) for two tissues. We will use this data to investigate the relation between DNA methylation and age. CpG sites which change their DNA methylation with age are called age-related differentially methylated positions or *aDMPs*.

Load packages

```
library(Biobase)
library(GEOquery)
library(ggplot2)
```

Datasets

We use two datasets. The first dataset is from Tsaprouni *et al.*. Note that this is not the largest dataset available on GEO, but for for this pratical large enough. The accession number of this dataset is *GSE50660*. The accession number is a number that refers to a specific dataset on GEO and can be referred to in scientific publications. The second dataset is from Berko *et al.* with accession number *GSE50759*.

There are different ways to download data from GEO. Go to https://www.ncbi.nlm.nih.gov/geo/. On the right you can search for datasets. Search for *GSE50660*. This will take you to the page of the dataset. You can see the summary of the data/study, the authors, the paper the data is from, contact details and the number of samples.

1. How many samples are there? What is the tissue the samples were taken from?

There are different ways to get the data. One is GEO2R which allows direct parsing of data in R. You can also manually download the data at the bottom of the page.

2. In what ways is the data available? Which one would you use?

Now we download the data from GEO. We can do this for both accession numbers. Note that this will take a few minutes to download and parse.

NOTE: If this step is taking extremely long or R just crashes, you can download the ready to use data from here (expires 19/11/2018): https://filesender.surf.nl/?s=download&token=28d91b5d-6525-7f38-749f-b8ad09f47306

```
gset.blood <- getGEO("GSE50660", GSEMatrix =TRUE, getGPL=FALSE)
gset.buccal <- getGEO("GSE50759", GSEMatrix =TRUE, getGPL=FALSE)
```

This is gives back a list with one slot. This contains a ExpressionSet. This is a way to store multiple datatypes/phenotypes in one object.

```
length(gset.blood) #Length of an object
class(gset.blood[[1]]) #Class of an object
```

```
eSet.blood <- gset.blood[[1]] #Select first object from list
eSet.buccal <- gset.buccal[[1]] #Select first object from list
```

We want to extract the DNA methylation levels of this dataset. That can be achieved with `exprs()`. The phenotypes can be extracted with `pheno`, which returns an AnnotatedDataFrame from which the phenotypes can be extracted using @data. See code below.

```
exprs.blood <- exprs(eSet.blood) #Extract expression for blood
exprs.buccal <- exprs(eSet.buccal) #Extract expression for buccal

pheno.blood <- phenoData(eSet.blood)@data #Extract phenotype data for blood
pheno.buccal <- phenoData(eSet.buccal)@data #Extract phenotype data for buccal
```

3. The DNA methylation data contains quite some loci. How many (code below)? Are the number of loci measured equal? If not, why do you think not?

```
dim(exprs.blood) #Dimensions
dim(exprs.buccal)
```

For the sake of time, only a subset of the data is investigated for a relation with age.

```
aDMPs <- unique(
  read.table("https://raw.githubusercontent.com/roderickslieker/FOS18/master/aDMPs.txt",
             stringsAsFactors = F)[,1])

# Make an interesting subset of CpGs
aDMPs <- Reduce(intersect, list(aDMPs,rownames(exprs.blood),rownames(exprs.buccal))) #Find the overlap

exprs.blood <- exprs.blood[match(aDMPs, rownames(exprs.blood)),] #Match the data to the new list of aDM
exprs.buccal <- exprs.buccal[match(aDMPs, rownames(exprs.buccal)),]
exprs.buccal <- (2^exprs.buccal)/(2^exprs.buccal + 1)
```

Note that in the chunk above, for buccal the values are still in M-values, which are transformed beta-values. M-values are on a continuous scale, while beta values are 0/1. Beta values can be interpreted as methylation fraction. We use beta values only here.

4. What are the dimensions of the new datasets?

```
dim(exprs.buccal)
dim(exprs.blood)
```

Run the analysis

We are now ready for the actual analysis. See the function below, we can use this function to test one CpG, but also for many. We start with one CpG.

```
get.aDMP <- function(CpG, data.in, samplesheet, phenotype)
{
  #Just to be sure make the phenotype numeric
  age <- as.numeric(as.character(samplesheet[,phenotype]))
  fit <- lm(data.in[CpG,]~age) # Run the LM
  fit.anova <- anova(fit) #Run the anova

  coef <- fit$coefficients["age"] #Extract coef
  p.val <- fit.anova["age",] #Extract pvalue

  out <- data.frame(CpG, coef, p.val) #Create a dataframe to output
  return(out) #Return the output
}
```

5. Run the model for CpG *cg16867657*. What do you observe?

```
fit.blood <- get.aDMP(CpG = "cg16867657", data.in = exprs.blood,
                      samplesheet = pheno.blood, phenotype = "age:ch1")
fit.buccal <- get.aDMP(CpG = "cg16867657", data.in = exprs.buccal,
                       samplesheet = pheno.buccal, phenotype = "age at draw (years):ch1")


#Look at the results
fit.blood
fit.buccal
```

6. Plot this one CpG (see code below). What do you observe? Is this locus really associated with age? Compare the slope of the line, what do you observe?

```
getPlot <- function(CpG)
{
  plotdata <- data.frame(Age = as.numeric(c(pheno.blood$`age:ch1`,  #Make a new data.frame
                                  pheno.buccal$`age at draw (years):ch1`)),
            Meth = c(exprs.blood[CpG,], exprs.buccal[CpG,]),
            Tissue = c(rep("Blood",464), rep("Buccal",96)))
  ggplot2::ggplot(plotdata, aes(x=Age, y=Meth, col=Tissue))+ #Plot
    geom_point()+
    geom_smooth(method=lm)+
    facet_grid(~Tissue, scale="free_x")+ #Set the x-axis free
    ylim(0,1)+ #Limit y axis
    scale_color_manual(values = c("#132B41","#F9A23F")) #CUstom colors
}

getPlot("cg16867657") #Plot for one
```

7. Now we loop over all CpGs; otherwise we would have to many tests manually. Run the model for all CpGs. Look at the top results. What is the top locus for blood? Is it the same for buccal?

```
res.blood <- lapply(aDMPs, get.aDMP, data.in=exprs.blood,  #Loop over all CpGs
                    samplesheet=pheno.blood, phenotype="age:ch1")
res.buccal <- lapply(aDMPs, get.aDMP, data.in=exprs.buccal,
```

```
                    samplesheet=pheno.buccal, phenotype="age at draw (years):ch1")

#Combine data to dataframe
results.blood <- as.data.frame(do.call(rbind, res.blood)) #Combine results in table
results.buccal <- as.data.frame(do.call(rbind, res.buccal))

#Sort on P-value
results.blood <- results.blood[order(results.blood$Pr..F., decreasing=F),] #Order on P-value
results.buccal <- results.buccal[order(results.buccal$Pr..F., decreasing=F),]

#Look at the top
head(results.blood) #View
head(results.buccal)
```

8. How many CpGs are significant in each of the tissues? Adjust for the number of tests performed!

```
alpha <- 0.05/nrow(results.blood) #Set alpha

results.blood.sign <- results.blood[results.blood$Pr..F. <= alpha,] #Subset results on alpha
results.buccal.sign <- results.buccal[results.buccal$Pr..F. <= alpha,]

nrow(results.blood.sign) #Check how many rows
nrow(results.buccal.sign)
```

9. Plot the coefficients against each other. What do observe in this subset?

```
results.blood <- results.blood[match(results.buccal$CpG, results.blood$CpG),] #Make a new df
coefs <- data.frame(Blood  = results.blood$coef, Buccal = results.buccal$coef)
rownames(coefs) <- results.blood$CpG


#Check which are significant
#Uses ifelse: when a condition is met, it will output the first argument,otherwise the second
coefs$Sign <- NA
coefs$Sign <- ifelse(rownames(coefs) %in% results.blood.sign$CpG, "Blood",coefs$Sign)
coefs$Sign <- ifelse(rownames(coefs) %in% results.buccal.sign$CpG, "Buccal",coefs$Sign)
coefs$Sign <- ifelse(rownames(coefs) %in% results.buccal.sign$CpG &
                       rownames(coefs) %in% results.blood.sign$CpG, "Both",coefs$Sign)


ggplot(coefs, aes(x=Blood, y=Buccal, col=Sign))+
  geom_point(size=0.5)+
  scale_colour_manual(values = c("#009AC7","#F9A23F","#8B1A4F"))
```

10. How many are shared between blood and buccal?

```
table(results.buccal.sign$CpG %in% results.blood.sign$CpG)
```

11. What do you notice in the figure of Q9 in terms of effect size? Look at the axis range. Do you think that aDMPs are tissue specific?

12. One could also add an effect size threshold to have biological relevant differences.Set the threshold to 2%/10years. How many are significant and how many overlap?

```
results.blood.sign.eff <- results.blood.sign[results.blood.sign$coef >= 0.002,] #Subset on effect size
results.buccal.sign.eff <- results.buccal.sign[results.buccal.sign$coef >= 0.002,]
```

```r
nrow(results.blood.sign.eff)
nrow(results.buccal.sign.eff)

table(results.blood.sign.eff$CpG %in% results.buccal.sign.eff$CpG)
```

13. One of the loci is *cg16867657* which we looked at before. Go to UCSC (https://genome.ucsc.edu), Genomes Buid hg19. Look-up *cg16867657*. Zoom out, what protein coding gene(s) are close to this CpG? Is the CpG in a CpG island?

14. Go to GeneCards (https://www.genecards.org), search for the gene found in 11.. What is the function of the gene? Does that make sense?