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.

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.

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

```
## Found 1 file(s)
```

```
## GSE50660_series_matrix.txt.gz
```

```
## Parsed with column specification:
## cols(
##   .default = col_double(),
##   ID_REF = col_character()
## )
```

```
## See spec(...) for full column specifications.
```

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

```
## Found 1 file(s)
```

```
## GSE50759_series_matrix.txt.gz
```

```
## Parsed with column specification:
## cols(
##   .default = col_double(),
##   ID_REF = col_character()
## )
```

```
## See spec(...) for full column specifications.
```

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)
```

```
## [1] 1
```

```
class(gset.blood[[1]])
```

```
## [1] "ExpressionSet"
## attr(,"package")
## [1] "Biobase"
```

```
eSet.blood <- gset.blood[[1]]
eSet.buccal <- gset.buccal[[1]]
```

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)
exprs.buccal <- exprs(eSet.buccal)

pheno.blood <- phenoData(eSet.blood)@data
pheno.buccal <- phenoData(eSet.buccal)@data
```

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)
```

```
## [1] 482739    464
```

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

```
## [1] 461339     96
```

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/rodericksliekerlieker/FOS18/master/aDMPs.txt",
             stringsAsFactors = F)[,1])

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

exprs.blood <- exprs.blood[match(aDMPs, rownames(exprs.blood)),]
```

```
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 different from beta-values. M-values are on a continuous scale, while beta values are easier to interpret, but less optimal in statistical analyses. For clarity we use beta values only here.

4. What are the dimensions of the new datasets?

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

```
## [1] 4766   96
```

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

```
## [1] 4766  464
```

Run the analysis

We are now ready for the actual analysis. For this we use function to loop over the CpGs. Otherwise we would have to many tests manually while now we can run them at once. See the function below.

```
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)
}
```

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
```

```
##            CpG        coef Df    Sum.Sq   Mean.Sq  F.value       Pr..F.
## age cg16867657 0.004276927  1 0.3760846 0.3760846 411.5677 6.676537e-66
```

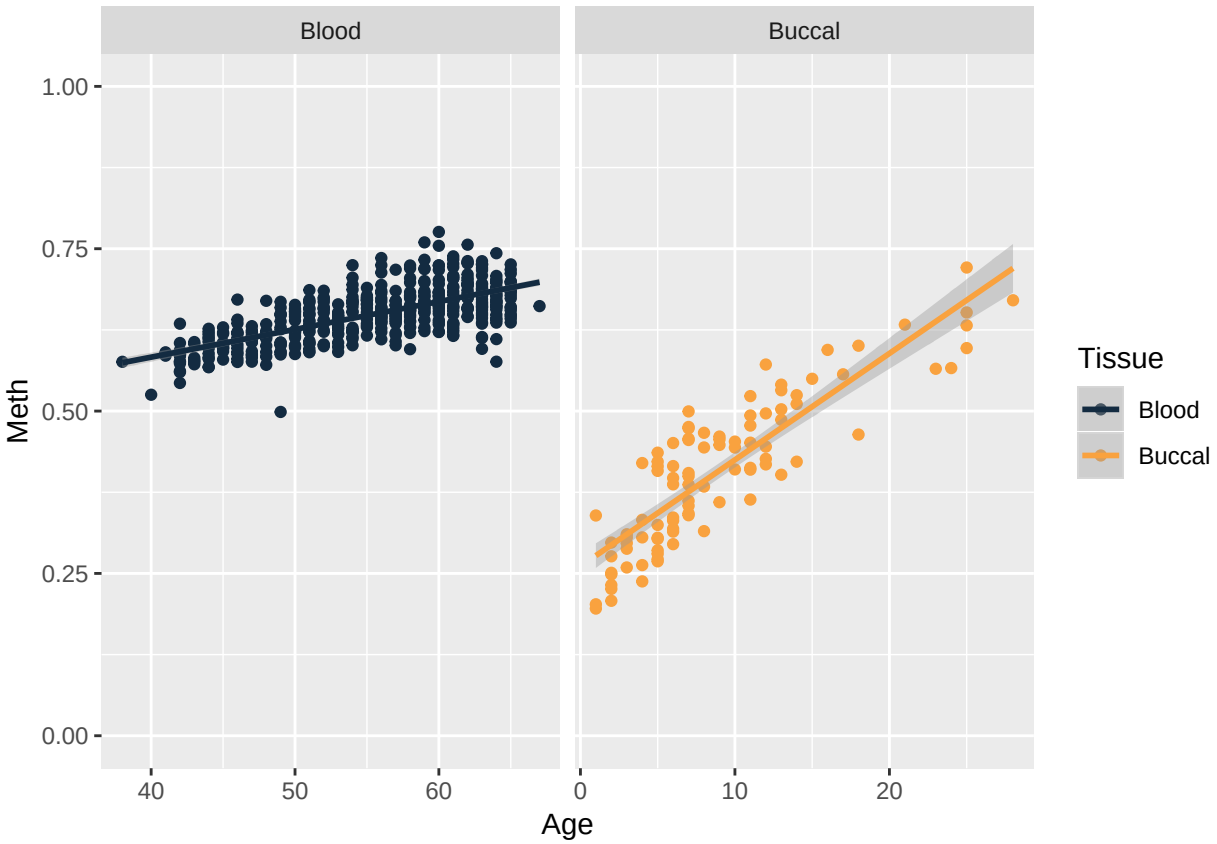
```
fit.buccal
```

```
##            CpG      coef Df    Sum.Sq   Mean.Sq  F.value       Pr..F.
## age cg16867657 0.0163895  1 0.9780046 0.9780046 300.7359 4.813671e-31
```

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$`age:ch1`,
                                pheno.buccal$`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))+
    geom_point()+
    geom_smooth(method=lm)+
    facet_grid(~Tissue, scale="free_x")+
    ylim(0,1)+
    scale_color_manual(values = c("#132B41","#F9A23F"))
}

getPlot("cg16867657")
```



7. 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,
                    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))
results.buccal <- as.data.frame(do.call(rbind, res.buccal))

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

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

```
##                CpG         coef Df    Sum.Sq   Mean.Sq  F.value
## age2346 cg16867657  0.004276927  1 0.3760846 0.3760846 411.5677
## age1004 cg06639320  0.002833567  1 0.1650779 0.1650779 177.1369
## age751  cg04875128  0.004563590  1 0.4281886 0.4281886 142.6360
## age1146 cg07553761  0.002625216  1 0.1416942 0.1416942 115.6333
## age1143 cg07547549  0.002429714  1 0.1213759 0.1213759 103.6875
## age791  cg05093315 -0.002227195  1 0.1019855 0.1019855 100.1526
##                Pr..F.
## age2346 6.676537e-66
```

```
## age1004 1.930728e-34
## age751  7.712853e-29
## age1146 3.203591e-24
## age1143 4.176716e-22
## age791  1.801928e-21
```

```
head(results.buccal)
```

```
##                CpG        coef Df    Sum.Sq   Mean.Sq  F.value
## age1146 cg07553761 0.009550805  1 0.3321157 0.3321157 377.9884
## age1660 cg11705975 0.012481875  1 0.5672430 0.5672430 317.4373
## age2346 cg16867657 0.016389500  1 0.9780046 0.9780046 300.7359
## age2542 cg18473521 0.010724208  1 0.4187357 0.4187357 270.2991
## age1142 cg07544187 0.007633596  1 0.2121621 0.2121621 266.0294
## age2257 cg16181396 0.006763067  1 0.1665317 0.1665317 241.5071
##               Pr..F.
## age1146 1.055759e-34
## age1660 6.822646e-32
## age2346 4.813671e-31
## age2542 2.118108e-29
## age1142 3.693736e-29
## age2257 1.030124e-27
```

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

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

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

nrow(results.blood.sign)
```

```
## [1] 205
```

```
nrow(results.buccal.sign)
```

```
## [1] 1842
```

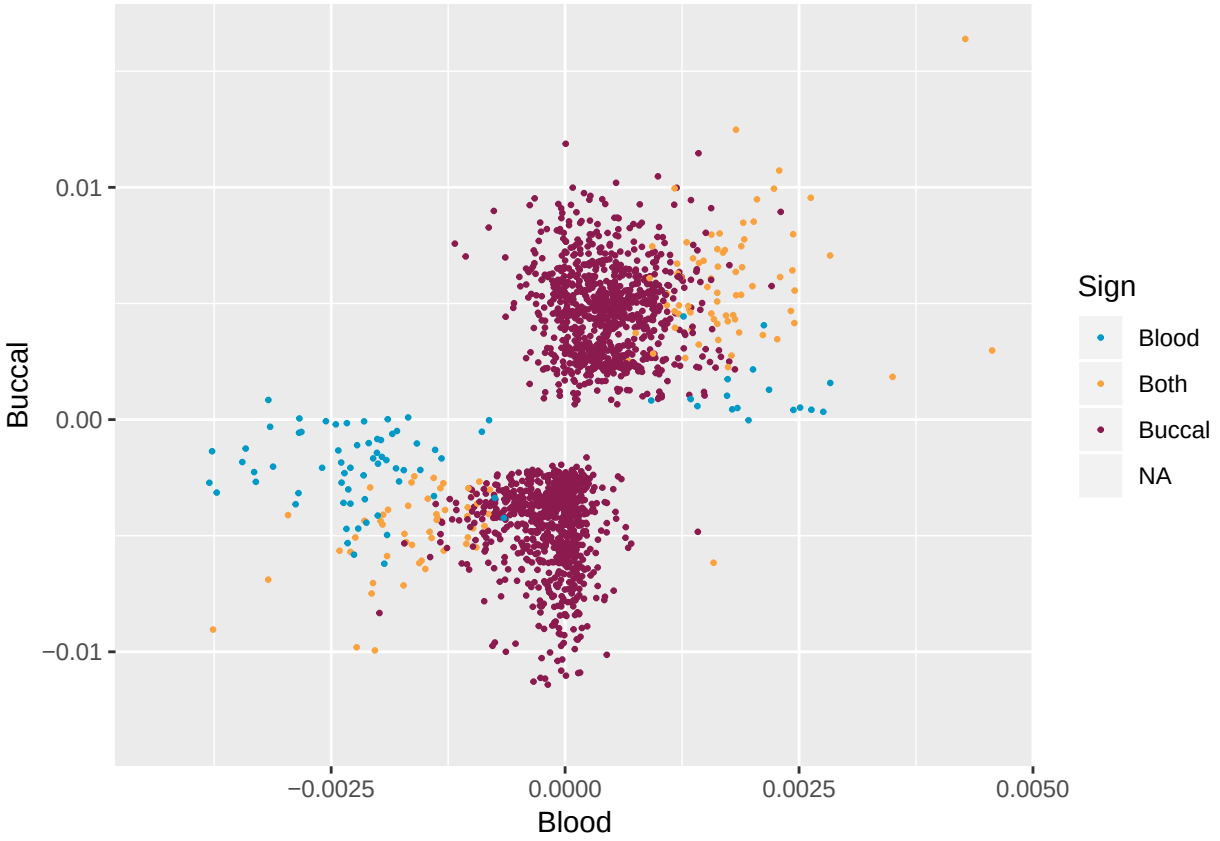
9. Plot the coefficients against each other. What do observe in this subset?

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


#Check which are significant
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"))
```

```
## Warning: Removed 2844 rows containing missing values (geom_point).
```



10. How many are shared between blood and buccal?

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

```
##
## FALSE  TRUE
##  1717   125
```

11. What do you notice in the figure of Q9 in terms of effect size? Look at the axis range. 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$coef >= 0.002,]
results.buccal.sign.eff <- results.buccal.sign[results.buccal.sign$coef >= 0.002,]

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

```
## [1] 26
```

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

```
## [1] 924
```

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

```
##
## FALSE  TRUE
##     9    17
```

12. 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?

13. 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?