Performing an EWAS"

Gibran Hemani

In this practical we are running an epigenome-wide association study (EWAS).

We are using data from this paper:

> Tsaprouni et al. (2014) Cigarette smoking reduces DNA methylation levels at multiple genomic loci but the effect is partially reversible upon cessation. Epigenetics. 9(10):1382-96.

Our aim is to identify CpG sites that are associated with cigarette smoking. We are going to prepare the data for analysis, run a simple EWAS, then adjust for other important variables.

There are four sets of data:

- `betas` A matrix of methylation beta values where each row is a 450k probe and each column is a sample.
- `traits` A data frame of phenotype data for each sample. The first column (SampleID) uniquely identifies each sample.
- `cell.counts` A matrix of estimated cell counts where each row is a sample and each column is a cell type.

Loading and exploring the data

Open R on your computer. We will load the first three sets of data described above (methylation data (betas), traits and cell counts). The methylation data and cell counts are saved as .Rdata files, which can be opened in R using the `load()` function. The phenotype data is saved as a comma separated file, or .csv file. This can be opened and viewed in Excel, but also read straight into R using the `read.csv()` function.

First, set the working directory to the folder that `data/epigenetic` folder - this contains the data we'll be using

```r
setwd(".....")
```

To open the methylation data and the cell counts data in R, type:

```r
load("betas.Rdata")
load("cell.counts.Rdata")
```

To open the file containing the phenotype data, type:

```r
traits <- read.csv("traits.csv")
```

To check what data is loaded in R, type:

```r
ls()
```

This should return:

```
[1] "betas" "cell.counts" "traits"
```

We can explore these datasets using str() or dim() , for example:

```
str(traits)
```

shows us that traits contains age, smoking, gender and batch information for 464 samples.

```
dim(betas)
```

shows us that the betas dataset has 45759 probes (rows) and 464 samples (columns).

What do the methylation values look like? e.g. the 34th CpG:

```
hist(betas[34,])
```

and the 35th:

```
hist(betas[35,])
```

The values are bound between 0 and 1, and typically are not normally distributed.

It is very important to check that the ORDER of our data is the same in each data frame / matrix. If we look at the top three samples for each dataset, we can see that they are not matched up:

```
betas[1:3,1:3]
traits[1:3,]
```

We can also see this if we ask R to tell us whether the column names of betas are identical to the SampleID column of "traits":

```
identical(colnames(betas),traits$SampleID)
```

This should return FALSE. Therefore, we need to match up the phenotype data with the methylation data. To do this, we can use the match() function to match the column names of betas to the SampleIDs in traits:

```
traits <- traits[match(colnames(betas),traits$SampleID),]
```

We can check that this has happened successfully by using this line of code again:

```
identical(
    as.character(colnames(betas)),
    as.character(traits$SampleID)
)
```

This time, this should return TRUE.

Since cell counts are going to be included in our regression model as covariates in the same way as age and gender, we can make things easier by merging the cell counts and phenotype data using the merge() function.

```
traits <- merge(traits, cell.counts, by.x="SampleID", by.y="row.names")
```

You can find out the names of the variables in traits using:

```
names(traits)
```

Smoking is coded as a dummy variable where 0=non-smoker and 1=smoker. Use the `table()` function to find out how many smokers there are.

```
table(traits$smoking)
```

Running a simple EWAS

Now that we have loaded and prepared our data, we can run a simple EWAS using the `cpg.assoc()` function in the CpGassoc package. Load the package using:

```
library(CpGassoc)
```

`cpg.assoc()` performs linear regression on methylation at each CpG site on a trait of interest (independent variable). In this case, the trait of interest is smoking. We will call the results of this first EWAS "model1":

```
model1 <- cpg.assoc(
    beta.val=betas,
    indep=as.numeric(traits$smoking)
)
```

We can use `summary()` to find the number of sites that survived Bonferroni (Holm) correction for multiple testing.

```
summary(model1)
```

This shows us that 261 sites were found significant by Holm method The top three CpG sites are:

```
  CpG        P.value
1 cg21566642 1.373349e-42
2 cg05951221 2.850320e-41
3 cg05575921 5.340820e-37
```

There is lots of information stored in the object `model1` we created (see e.g. `str(model1)` for more information.

We can extract coefficients and p values from our results using:

```
model1_results <- cbind(model1$results, model1$coefficients)
```

```
model1_results$CI_lower <-
    model1_results$effect.size - (model1_results$std.error * 1.96)
model1_results$CI_upper <-
    model1_results$effect.size + (model1_results$std.error * 1.96)
```

Sort by p-value so that the "top-hits" are at the top of the data frame:

```
model1_results <-
    model1_results[order(model1_results$P.value),]
```

View the first few CpG sites using `head()`

```
head(model1_results)
```

Adjusting for confounding

As suggested by the simple EWAS, smoking is a source of variation in the data:

```
library(minfi)
mdsPlot(
    betas,
    sampGroups=as.factor(traits$smoking)
)
```

But this could be confounded by other factors, for example, batch:

```
mdsPlot(
    betas,
    sampNames=as.factor(traits$batch),
    sampGroups=as.factor(traits$batch)
)
```

There are several ways to control for batch (discussed in the lecture), but one option is to include the batch variable (e.g. the chip the sample was run on) as a covariate in your EWAS model. Batch is specified in cpg.assoc as `chip.id`:

```
model2 <- cpg.assoc(
    beta.val=betas,
    indep=as.numeric(traits$smoking),
    chip.id=as.factor(traits$batch)
)
summary(model2)
```

Methylation levels may also be influenced by other variables such as age and gender. These variables are also associated with smoking status, so they should be included in the EWAS model to adjust for confounding. A dataframe of covariates can be specified in cpg.assoc using `covariates`:

```
model3 <- cpg.assoc(
    beta.val=betas,
    indep=as.numeric(traits$smoking),
    chip.id=as.factor(traits$batch),
    covariates=traits[,c("age", "gender")]
)
summary(model3)
```

As discussed in the lecture, a significant source of variation in methylation data derived from blood comes from differences in the relative proportions of different cell types.

Add all six cell proportions to your model as covariates (as well as age and sex). Call this new model `model4`.

```
model4 <- cpg.assoc(
    beta.val=betas,
    indep=as.numeric(traits$smoking),
    chip.id=as.factor(traits$batch),
    covariates=
        traits[,c("age", "gender", "CD8T", "CD4T",
                  "NK", "Bcell", "Mono", "Gran")
              ]
)
summary(model4)
```

Create a Manhattan plot of the results

```
load("annotation.RData")
manhattan(
    model1,
    chr=annotation$CHR,
    cpgname=annotation$TargetID,
    pos=annotation$COORDINATE_37
)
```