For our lab sessions, we will be analyzing public dataset from GEO website (Serial No. GSE59592). Genome wide DNA methylation profiling in infant's blood from a mother/child cohort in The Gambia. The main variables of the analyses were the intra-uterine exposure to aflatoxin B1 (AFB1) and the season of conception. The Illumina Infinium HumanMethylation 450k Beadchip was used to obtain DNA methylation profiles across approximately 450,000 CpGs in whole peripheral blood obtained at 3-6 months of age. A total of 124 samples were analysed, including 3 technical replicates.

1. There are three files that you will need: (1) The data file, (2) annotation file, (3) covariate file. Note that before you read the files in, you will need to change your working directory to an appropriate folder where the files are located. On my linux machine, the appropriate folder is located at '/home/asalim/methyl'

```
#data
data <- read.csv('GSE59592_GEO_betas_raw.csv', header=T,
            colClasses=c('character',rep('numeric',124)))
# covariates
info <- read.table('GSE59592_info.txt',header=T,skip=7,sep='\t')
# probe annotation
anno <- read.csv('GPL13534_HumanMethylation450_anno.csv', header=T, skip=7)
```

   Questions: Do you know we use read.table command for some files and read.csv for others? What does the header=T and skip=7 argument do? If you are unsure about your answers, open the read.table and read.csv help files to learn

2. There are 124 samples in this study, however for simplicity we are going to remove a number of samples with missing covariate values and also probes with missing methylation values. This can be achieved using the following command:

```
# don't use sample with missing AFB1 levels
info <- info[!is.na(info$AFB1),]
# don't use probes with missing values
data <- data[apply(is.na(data[,-1]),1,sum)==0,]
```

   Now, the methylation dataset should have 480,050 probes [you can check this by typing nrow(data), and 118 samples [ncol(info)]

3. The data is now almost ready to be explored and analyzed. But before we start, we need to perform variance-stabilizing transformation to the data. This is because our methylation data is in the form of proportion and proportion data has variance that depends on its mean. Further downstream, we will be analyzing the data using methods that assume equality of variance, so it is important to perform the transformation beforehand.

```
# variance stabilizing transformation using arc-sine sqrt transformation
y <- asin(sqrt(data[,-1]))
# re-arrange y vector to follow the probe order in annotation file
y <- y[na.omit(match(anno$IlmnID,data$X)),]
row.id <- na.omit(match(anno$IlmnID,data$X))

# we also need to re-arrange columns of data so that it corresponds to row of info
new.colnames <- unlist(strsplit(colnames(y),"X"))
new.colnames <- new.colnames[new.colnames!=""]
y <- y[,na.omit(match(info$sample,new.colnames))]
```

Questions: Why do we need to leave out the first column of the data when performing variance-stabilizing transformation? Why do we need to ensure the probes order in the annotation and data matrices are the same?

4. We can start exploring the data now. First generate boxplots that will compare the distribution of the transformed data across the different samples [`boxplot(y)`]. If the data is well-normalized, the distribution across different samples should be similar. Are they similar now?

5. If the distribution of the data is not similar, we need to perform normalization step. We will use the quantile-normalization step and visualize the boxplots of the normalized data. Are they similar now?

```
#normalize data
require('preprocessCore')
y.norm <- normalize.quantiles( as.matrix(y))

# boxplot normalized data, by sample
boxplot(y.norm)
```

6. You can also use the heatmap function to visualize the data using classical heatmap. Note: we will be only producing heatmap based on 1000 randomly-selected probes here. A heatmap based on the full set of probes ($\sim$ 480 K) will take too much memory and space! What does the red/green colors tell you about the methylation levels? Why there are blocks of green and red?

```
# show heatmap
require(gplots)
heatmap.2(y.norm[sample(nrow(y.norm),1000),],col=redgreen(50),key=TRUE,
            symkey=FALSE, trace="none", cexRow=0.5)
```

7. We will perform a linear regression with the transformed (normalized) data for the each probe as outcome and Season of Birth (SoB) and AFB1 levels as covariates. For this, we will be using the `RUV2` command from the `ruv` package. However, we will not be performing any batch-effect adjustment as yet and will set the number of unwanted factors to adjust (k) to zero

```
require(ruv)
Xmat <- model.matrix(~as.factor(info$SoB)+info$AFB1)
# define arbitrary (fake) control
control <- sample(c(TRUE,FALSE),size=nrow(y.norm),prob=c(0.01,0.99),replace=T)
# no adjustment
ruv2.adj0 <- RUV2(Y=t(y.norm),X=Xmat[,-1],ctl=control,k=0)
```

8. **Challenge:** The output of `RUV2` is a list with one of its element (called `p`) contains the p-values associated with factor of interests. Plot $-\log_{10}$ of these p-values using different colors for probes belonging to different chromosome. How many probes have -$\log_{10}$ p-value greater than 7? These probes are the one that more likely to be statistically-significant. Do you know why?

9. When you finished don't forget to save the commands you have been executing and the workspace containing all your works, but first you need to save the three datasets `data,anno,info` for further use later. The following commands can be used:

```
# remove big datasets
save(data,anno,info, file= paste('Datasets_','YOURNAME','.Rdata',sep=""))
rm(data);rm(anno);rm(info)
# save commands
filaneme1 <- paste('Lab2_','YOURNAME','.R',sep="")
```

```
savehistory(filename1)
# save workspace
filaneme2 <- paste('Lab2_','YOURNAME','.RData',sep="")
save.image(filename2)
```