

Assignment 1

Guillaume Buisson-Chavot

21 novembre 2018

Introduction:

Le but est d'étudier l'impact du double Knock-Out des gènes DNMT1 et DNMT3A sur le taux de méthylation d'un génome humain. Nous allons donc analyser le profil de méthylation de l'ADN de trois réplicats biologiques dans les conditions sauvage (contrôle) et de double KO réalisé à l'aide du la plate-forme de méthylation Illumina Infinium. J'ai utilisé les librairies GGplot2 et dplyr afin de réaliser ce devoir.

Data acquisition

Dans un premier temps, j'importe les données, je les transforme en classe numérique afin de pouvoir les utiliser et je supprime les NA s'il y en a.

```
library(ggplot2)
library(dplyr)

## 
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
## 
##     filter, lag

## The following objects are masked from 'package:base':
## 
##     intersect, setdiff, setequal, union

setwd("/home/guiom/ULB/2/génomics_proteomics_evolution/assignment_1/")
Infiniumdf <- read.table("Signal_AB.txt.gz", header = TRUE, stringsAsFactors = FALSE)
Infiniumdf[, 1:12] <- apply(Infiniumdf[,1:12], 2, as.numeric) #passage en "numérique"
Infiniumdf <- Infiniumdf[complete.cases(Infiniumdf),] # supprime les NA
Infiniumdf$names <- row.names(Infiniumdf) # rajoute une colonne names pour les sondes
```

β-value distribution

1. Compute the β-value using the Signal_A and Signal_B data (Signal_B corresponds to the Methylated signal (M) and Signal_A to the Unmethylated signal (U)).

```
B_value_WT1 <- Infiniumdf$WT1.Signal_B/(Infiniumdf$WT1.Signal_B + Infiniumdf$WT1.Signal_A)
B_value_WT2 <- Infiniumdf$WT2.Signal_B/(Infiniumdf$WT2.Signal_B + Infiniumdf$WT2.Signal_A)
B_value_WT3 <- Infiniumdf$WT3.Signal_B/(Infiniumdf$WT3.Signal_B + Infiniumdf$WT3.Signal_A)
B_value_KO1 <- Infiniumdf$KO1.Signal_B/(Infiniumdf$KO1.Signal_B + Infiniumdf$KO1.Signal_A)
B_value_KO2 <- Infiniumdf$KO2.Signal_B/(Infiniumdf$KO2.Signal_B + Infiniumdf$KO2.Signal_A)
B_value_KO3 <- Infiniumdf$KO3.Signal_B/(Infiniumdf$KO3.Signal_B + Infiniumdf$KO3.Signal_A)
```

```

IlmnID <- Infiniumdf$names
B_values <- data.frame(IlmnID, B_value_WT1, B_value_WT2,
                        B_value_WT3, B_value_KO1, B_value_KO2, B_value_KO3)

head(B_values)

##      IlmnID B_value_WT1 B_value_WT2 B_value_WT3 B_value_KO1 B_value_KO2
## 1 cg00000029    0.8472843   0.8666327   0.9092991   0.4136873   0.4188734
## 2 cg00000108    0.8627078   0.9013848   0.9359841   0.6680526   0.6466701
## 3 cg00000109    0.8479977   0.8815632   0.8872244   0.8041511   0.7533346
## 4 cg00000165    0.8415437   0.8827264   0.8928415   0.7240071   0.7251262
## 5 cg00000236    0.8413939   0.8668285   0.8910567   0.6137658   0.5407618
## 6 cg00000289    0.6420095   0.6942097   0.7533822   0.3238636   0.2677165
##      B_value_KO3
## 1    0.4729088
## 2    0.7305517
## 3    0.8226051
## 4    0.7922252
## 5    0.6566892
## 6    0.3603798

```

2. Compute and plot the β -value distribution for each sample independently. Describe the obtained results.

Avant d'effectuer les graphes, je vérifie la classe des données, je les passe en numérique afin de pouvoir continuer à travailler dessus et je supprime les NA

```

B_values[, 2:7] <- apply(B_values[,2:7], 2, as.numeric)
B_values <- na.omit(B_values)

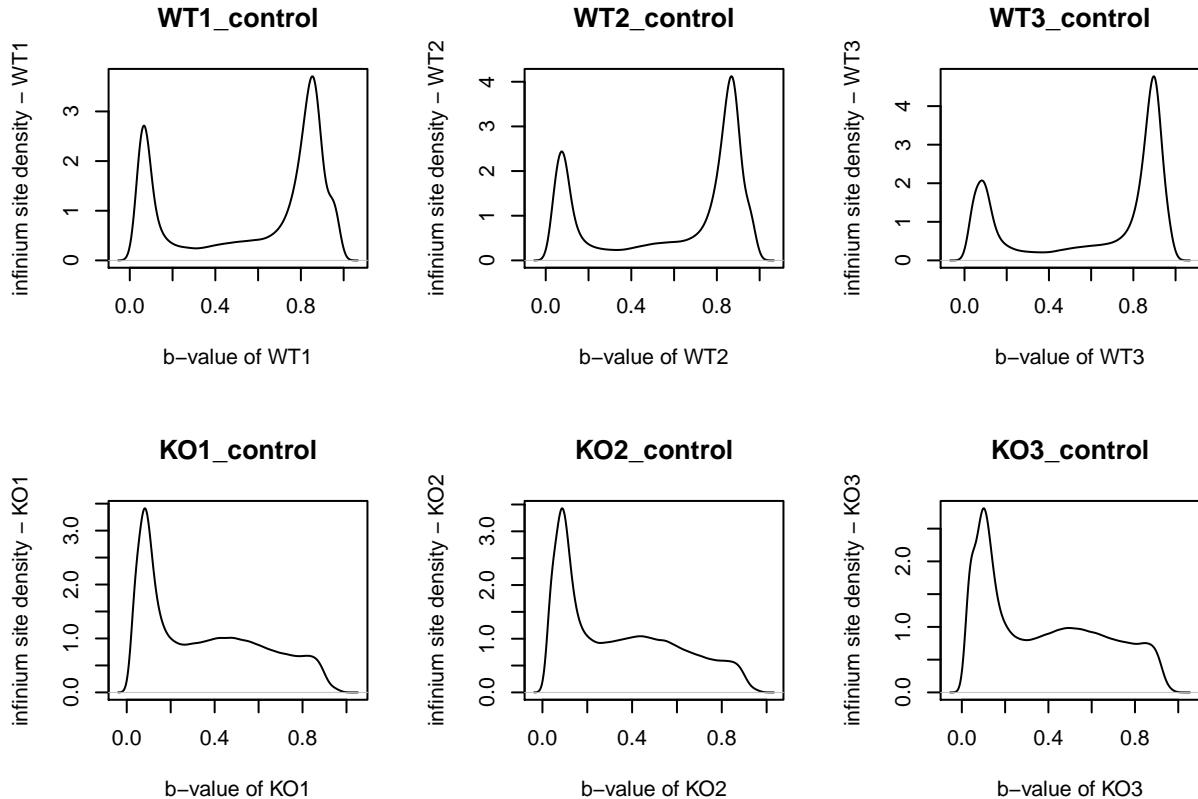
```

Je peux donc maintenant faire un plot des β -values:

```

par(mfrow=c(2,3))
plot(density(B_values$B_value_WT1), main= 'WT1_control',
      xlab = "b-value of WT1" ,ylab = 'infinium site density - WT1')
plot(density(B_values$B_value_WT2), main= 'WT2_control',
      xlab = "b-value of WT2" ,ylab = 'infinium site density - WT2')
plot(density(B_values$B_value_WT3), main= 'WT3_control',
      xlab = "b-value of WT3" ,ylab = 'infinium site density - WT3')
plot(density(B_values$B_value_KO1), main= 'K01_control',
      xlab = "b-value of K01" ,ylab = 'infinium site density - K01')
plot(density(B_values$B_value_KO2), main= 'K02_control',
      xlab = "b-value of K02" ,ylab = 'infinium site density - K02')
plot(density(B_values$B_value_KO3), main= 'K03_control',
      xlab = "b-value of K03" ,ylab = 'infinium site density - K03')

```



discussion: les beta values représentent une estimation du taux de méthylation obtenu à partir du ratio d'intensité entre les allèles méthylées et non méthylées. Elles ont donc des valeurs entre 0 et 1 avec 0 représentant les allèles non méthylées et 1 celles complètement méthylées. Ici, on voit que les représentations des 3 controles sont sensiblement identiques entre elles de même pour les 3 représentations des KO globalement identiques entre eux. En ce qui concerne les controles, on observe deux pics, donc un grand nombre de sondes méthylées et un grand nombre aussi (mais plus petit) de sondes non méthylées. Concernant les KO, les sondes méthylées ont disparues, elles ont été presque complètement démethylées, le pic des démethylées ayant augmenté à la fois en nombre (hauteur) et en largeur. Toutes ces observations paraissent normal après un double KO.

Global overview

3. Using the annotation provided in annotation.csv.gz , compute the methylation level distribution for chromosome 6, 7 and 21 in both condition and plot it using a boxplot.

```
annotations <- read.csv("annotation.csv.gz", header = TRUE, sep = ',', stringsAsFactors = FALSE)
annotations <- annotations[complete.cases(annotations), ]
chr6 <- annotations[grep("6", annotations$CHR), ]
chr7 <- annotations[grep("7", annotations$CHR), ]
chr21 <- annotations[grep("21", annotations$CHR), ]

chr6 <- suppressWarnings(merge(B_values, chr6, by="IlmnID"))
chr7 <- suppressWarnings(merge(B_values, chr7, by="IlmnID"))
chr21 <- suppressWarnings(merge(B_values, chr21, by="IlmnID"))
```

```

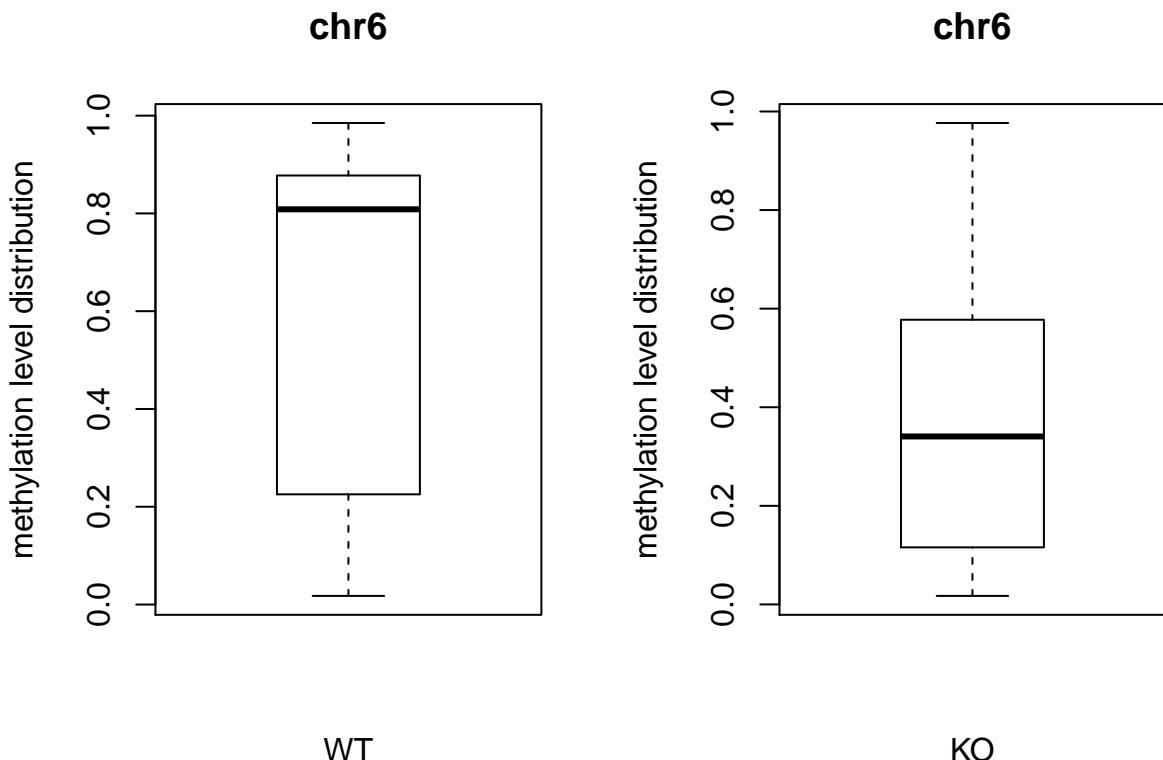
chr6$mean_WT <- (chr6$B_value_WT1 + chr6$B_value_WT2 + chr6$B_value_WT3)/3
chr6$mean_KO <- (chr6$B_value_KO1 + chr6$B_value_KO2 + chr6$B_value_KO3)/3

chr7$mean_WT <- (chr7$B_value_WT1 + chr7$B_value_WT2 + chr7$B_value_WT3)/3
chr7$mean_KO <- (chr7$B_value_KO1 + chr7$B_value_KO2 + chr7$B_value_KO3)/3

chr21$mean_WT <- (chr21$B_value_WT1 + chr21$B_value_WT2 + chr21$B_value_WT3)/3
chr21$mean_KO <- (chr21$B_value_KO1 + chr21$B_value_KO2 + chr21$B_value_KO3)/3

par(mfrow=c(1,2))
boxplot(chr6$mean_WT, xlab= "WT", ylab= "methylation level distribution" , main = "chr6")
boxplot(chr6$mean_KO, xlab= "KO", ylab= "methylation level distribution" , main = "chr6")

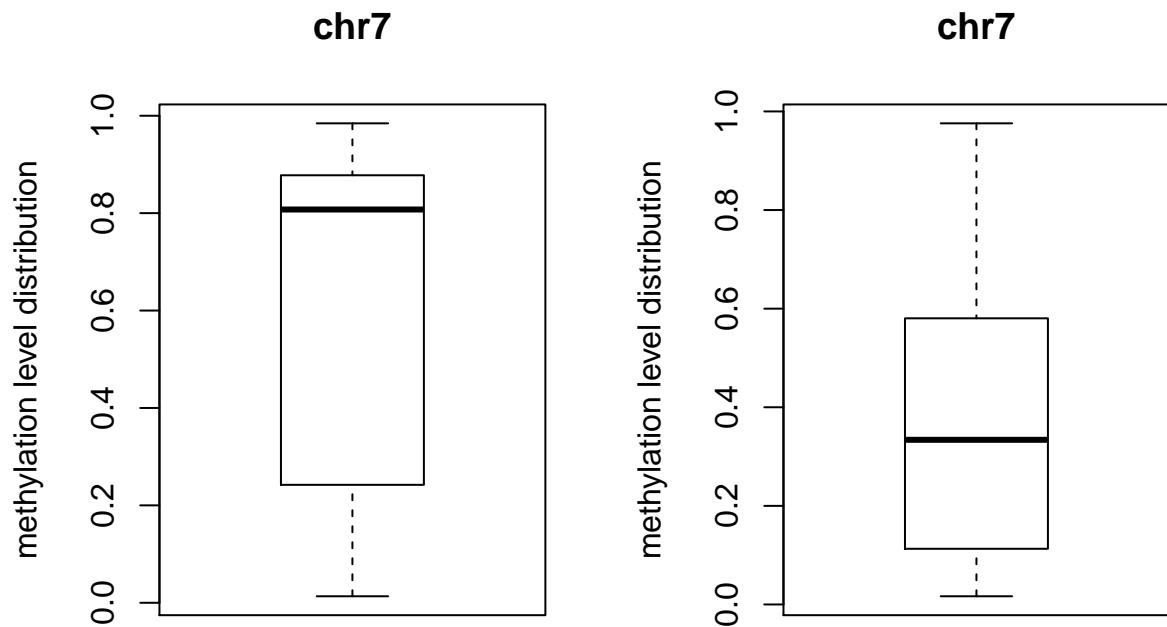
```



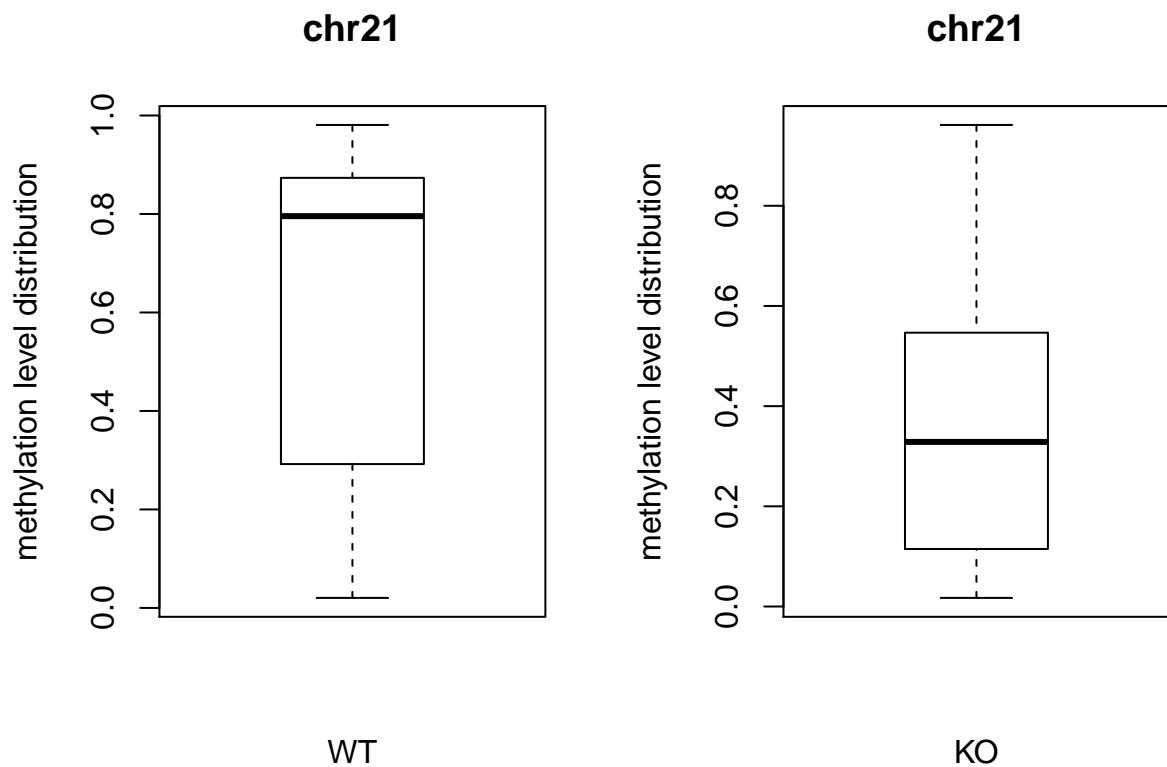
```

par(mfrow=c(1,2))
boxplot(chr7$mean_WT, xlab= "WT", ylab= "methylation level distribution", main = "chr7")
boxplot(chr7$mean_KO, xlab= "KO", ylab= "methylation level distribution", main = "chr7")

```



```
par(mfrow=c(1,2))
boxplot(chr21$mean_WT, xlab= "WT", ylab= "methylation level distribution", main = "chr21")
boxplot(chr21$mean_KO, xlab= "KO", ylab= "methylation level distribution", main = "chr21")
```



discussion: On voit que quel que soit le chromosome, si on compare les résultats aux WT, la distribution du taux de méthylation des KO s'est fortement déplacée vers le bas, l'ensemble des valeurs a diminué et

évidemment, la médiane avec.

4. Compute the average methylation for the control and the case for the following genes: MED14, CLK2 and TLR5 and plot it using a barplot. How many probes are related to gene CLK2?

```

liste <- strsplit(annotations$UCSC_RefGene_Name, split = ";")
liste2 <- lapply(liste, unique)
clean_annot <- data.frame(IlmnID = rep(annotations$IlmnID,
                                         sapply(liste2, length)),
                           UCSC_RefGene_Name = unlist(liste2))
CLK2 <- clean_annot[grep("^CLK2$", clean_annot$UCSC_RefGene_Name),]
probe_number_CLK2 <- count(CLK2)
TLR5 <- clean_annot[grep("^TLR5$", clean_annot$UCSC_RefGene_Name),]
probe_number_TLR5 <- count(TLR5)
MED14 <- clean_annot[grep("^MED14$", clean_annot$UCSC_RefGene_Name),]
probe_number_MED14 <- count(MED14)

CLK2_and_B_values <- suppressWarnings(merge(B_values, CLK2, by="IlmnID"))
TLR5_and_B_values <- suppressWarnings(merge(B_values, TLR5, by="IlmnID"))
MED14_and_B_values <- suppressWarnings(merge(B_values, MED14, by="IlmnID"))

CLK2_and_B_values$mean_WT <-(CLK2_and_B_values$B_value_WT1+CLK2_and_B_values$B_value_WT2
                                +CLK2_and_B_values$B_value_WT3)/3
CLK2_and_B_values$mean_KO <-(CLK2_and_B_values$B_value_KO1+CLK2_and_B_values$B_value_KO2
                                +CLK2_and_B_values$B_value_KO3)/3
TLR5_and_B_values$mean_WT <-(TLR5_and_B_values$B_value_WT1+TLR5_and_B_values$B_value_WT2
                                +TLR5_and_B_values$B_value_WT3)/3
TLR5_and_B_values$mean_KO <-(TLR5_and_B_values$B_value_KO1+TLR5_and_B_values$B_value_KO2
                                +TLR5_and_B_values$B_value_KO3)/3
MED14_and_B_values$mean_WT <-(MED14_and_B_values$B_value_WT1+MED14_and_B_values$B_value_WT2
                                +MED14_and_B_values$B_value_WT3)/3
MED14_and_B_values$mean_KO <-(MED14_and_B_values$B_value_KO1+MED14_and_B_values$B_value_KO2
                                +MED14_and_B_values$B_value_KO3)/3

mean_WT_CLK2 <-mean(CLK2_and_B_values$mean_WT)
mean_KO_CLK2 <-mean(CLK2_and_B_values$mean_KO)

mean_WT_TLR5 <-mean(TLR5_and_B_values$mean_WT)
mean_KO_TLR5 <-mean(TLR5_and_B_values$mean_KO)

mean_WT_MED14 <-mean(MED14_and_B_values$mean_WT)
mean_KO_MED14 <-mean(MED14_and_B_values$mean_KO)

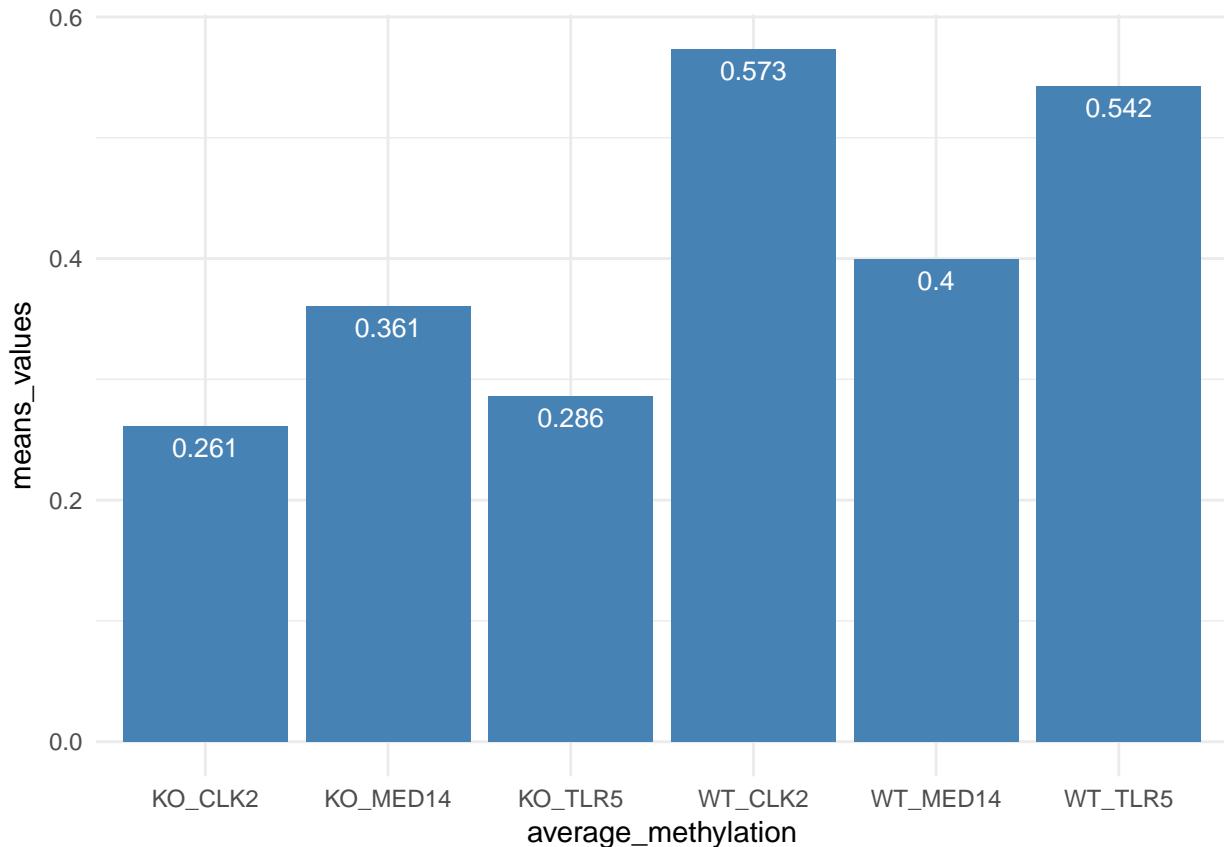
dfgenes <-data.frame(average_methylation= c("WT_CLK2", "KO_CLK2",
                                              "WT_TLR5", "KO_TLR5",
                                              "WT_MED14", "KO_MED14"),
                      means_values=c(mean_WT_CLK2, mean_KO_CLK2,
                                    mean_WT_TLR5, mean_KO_TLR5,
                                    mean_WT_MED14, mean_KO_MED14))
plot <- ggplot(data=dfgenes, aes(x=average_methylation,y=means_values))+
```

```

geom_bar(stat="identity", fill="steelblue")+
geom_text(aes(label=round(means_values,3)),
vjust=1.6, color="white", size=3.5)+ theme_minimal()

```

plot



discussion: même chose ici pour les gènes, la perte de méthylation n'est pas seulement visible au niveau global = chromosome, elle est également visible au niveau des gènes. On peut toutefois noter pour le gène MED14 qu'il y a une perte de méthylation moins importante que pour les deux autres gènes.

probe_number_CLK2

```

## # A tibble: 1 x 1
##       n
##   <int>
## 1     27

```

Le nombre de sondes relatives à CLK2 est donc de 27.

Differential analysis

5. Compute the delta- β (β -value difference between case mean and control mean) for all the probes. Using the t-test (t.test) compute the Pvalue of the differential methylation (case vs control). Plot and discuss the distribution of the delta- β .

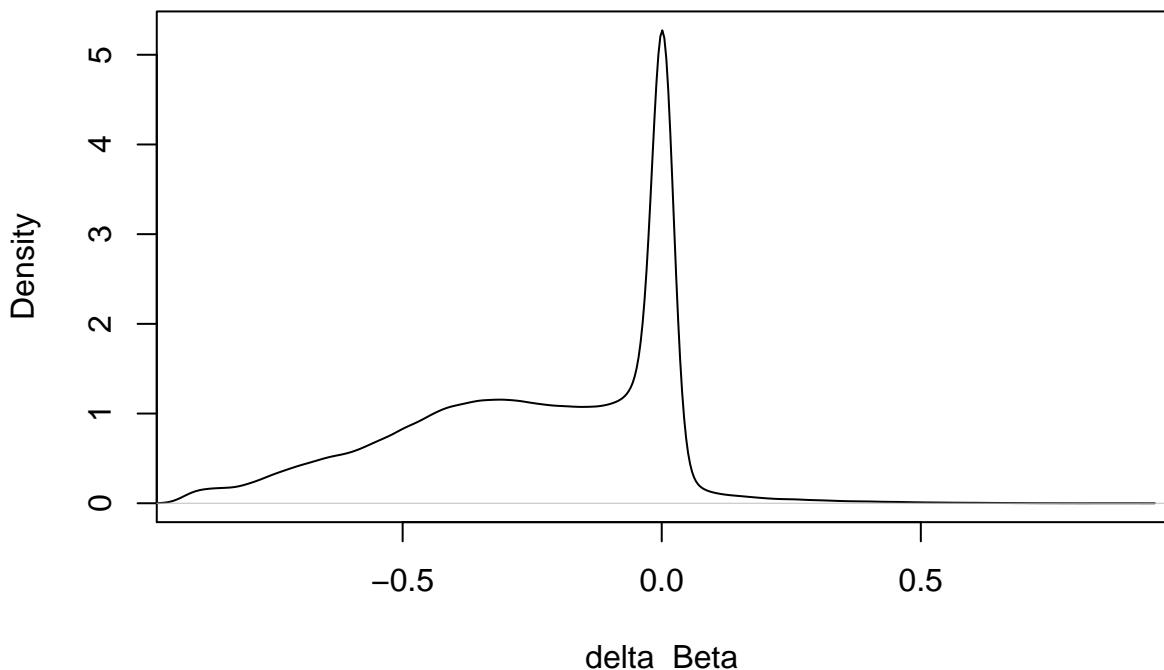
```
B_values$mean_WT <- (B_values$B_value_WT1+B_values$B_value_WT2+B_values$B_value_WT3)/3
B_values$mean_K0 <- (B_values$B_value_KO1+B_values$B_value_KO2+B_values$B_value_KO3)/3
B_values$delta_B <-B_values$mean_K0-B_values$mean_WT

v <- c(1,2,3,4,5,6)
B_values$pval = apply(B_values[,2:7], 1,
                      function(v) t.test(x=v[1:3],y=v[4:6])$p.value)
head(B_values)

##      I1mnID B_value_WT1 B_value_WT2 B_value_WT3 B_value_KO1 B_value_KO2
## 1 cg00000029  0.8472843  0.8666327  0.9092991  0.4136873  0.4188734
## 2 cg00000108  0.8627078  0.9013848  0.9359841  0.6680526  0.6466701
## 3 cg00000109  0.8479977  0.8815632  0.8872244  0.8041511  0.7533346
## 4 cg00000165  0.8415437  0.8827264  0.8928415  0.7240071  0.7251262
## 5 cg00000236  0.8413939  0.8668285  0.8910567  0.6137658  0.5407618
## 6 cg00000289  0.6420095  0.6942097  0.7533822  0.3238636  0.2677165
##      B_value_KO3   mean_WT   mean_K0   delta_B      pval
## 1    0.4729088 0.8744054 0.4351565 -0.43924888 7.641318e-05
## 2    0.7305517 0.9000256 0.6817581 -0.21826747 2.960095e-03
## 3    0.8226051 0.8722618 0.7933636 -0.07889815 4.145026e-02
## 4    0.7922252 0.8723705 0.7471195 -0.12525103 1.346769e-02
## 5    0.6566892 0.8664264 0.6037389 -0.26268744 7.967230e-03
## 6    0.3603798 0.6965338 0.3173200 -0.37921380 9.552477e-04

par(mfrow=c(1,1))
plot(density(B_values$delta_B), xlab= "delta_Beta" ,
     main= "distribution of the delta-beta" , xlim =c(-max(B_values$delta_B),
                                                   max(B_values$delta_B)))
```

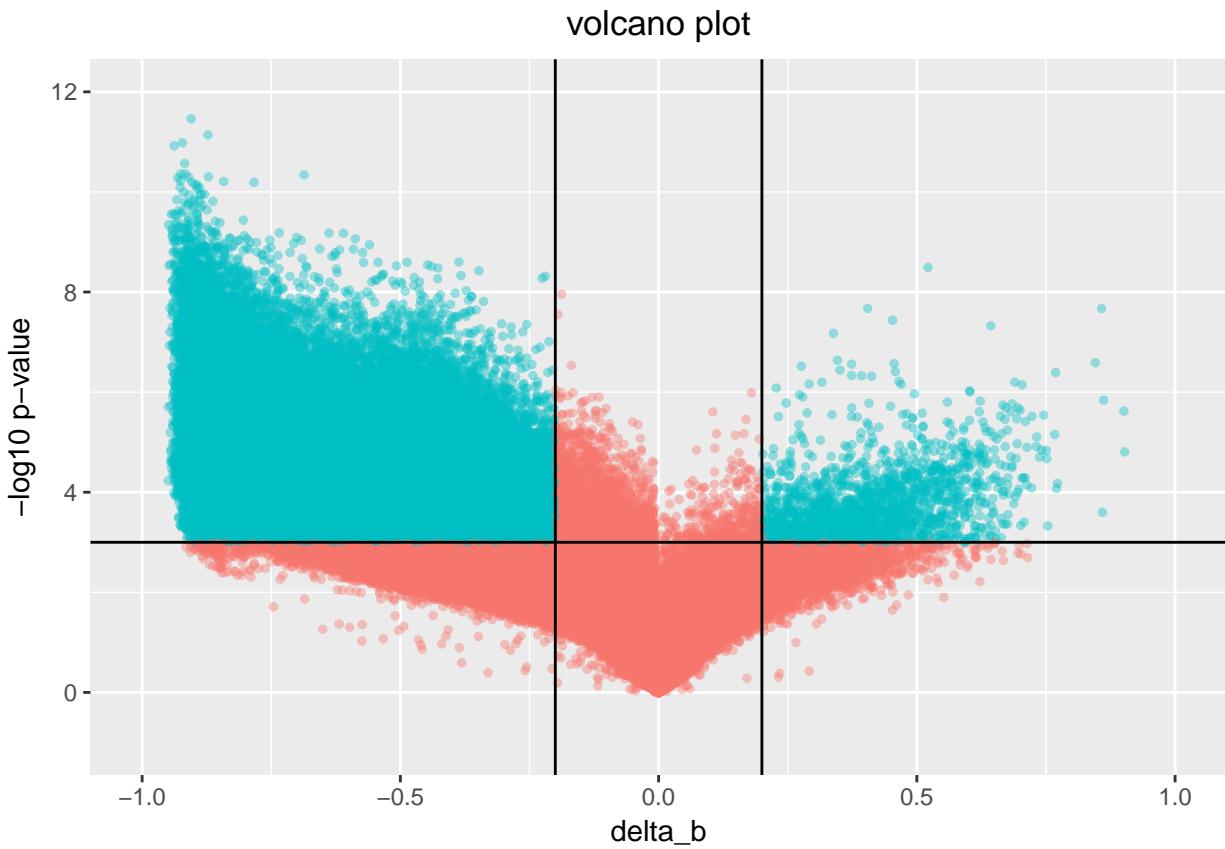
distribution of the delta-beta



discussion: Le pic se situe proche de delta_b = 0, il y a donc le plus grand nombre de sondes pour lesquelles il n'y a pas de différence de méthylation entre contrôle et KO ce qui paraît normal. Les delta_b positifs correspondent à des réplicats qui malgré le double KO seraient plus méthylés que les contrôles. On en retrouve donc très peu, c'est négligeable. Puis entre -0.6 et -0.1 se trouve la majorité de la distribution qui correspond aux réplicats moins méthylés que les contrôles suite aux KO. Le double KO a donc démethylé un grand nombre de sondes.

6. Draw a volcano plot with the delta- β on the x-axis and the $-\log_{10}(P\text{value})$ on the y-axis.

```
B_values$threshold <- as.factor(abs(B_values$delta_B)>0.2  
                                & B_values$pval< 0.001)  
ggplot(data=B_values, aes(x=B_values$delta_B, y=-log10(B_values$pval),  
                           colour=threshold)) +  
  geom_point(alpha=0.4, size=1) +  
  theme(legend.position = "none") +  
  geom_hline(aes(yintercept=3))+ geom_vline(xintercept = 0.2)+  
  xlim(c(-1, 1)) + ylim(c(-1, 12)) + geom_vline(xintercept = -0.2)+  
  xlab("delta_b") + ylab("-log10 p-value") +  
  ggtitle("volcano plot") + theme(plot.title = element_text(hjust = 0.5))
```



discussion: En rouge, ce sont les sondes qui ne répondent pas aux deux seuils que j'ai fixé ($|\text{delta}_\Delta| > 0.2$ & $p\text{-value} < 1e-3$). Les sondes qui semblent statistiquement avoir été affectées par le KO sont représentées en bleu. il faut noter qu'on observe une petite quantité de sondes méthylées après KO, il faudrait donc se pencher sur celles-ci afin d'en déterminer la raison.

Random control (Bonus)

7. Perform a sequence of random controls by shuffling the data and redo the differential analysis. What are your conclusions?

```

shuffled_B_values <- B_values[,2:7]
shuffled_B_values <- shuffled_B_values[sample(ncol(shuffled_B_values),
                                              ncol(shuffled_B_values)) ]
colnames(shuffled_B_values) <- c("control_1", "control_2", "control_3",
                                  "KO_1", "KO_2", "KO_3")
shuffled_B_values$mean_WT <- (shuffled_B_values$control_1+
                                shuffled_B_values$control_2 +
                                shuffled_B_values$control_3)/3
shuffled_B_values$mean_KO <- (shuffled_B_values$KO_1+shuffled_B_values$KO_2+
                                shuffled_B_values$KO_3)/3
shuffled_B_values$delta_B <- shuffled_B_values$mean_KO-shuffled_B_values$mean_WT

head(shuffled_B_values)

##   control_1 control_2 control_3      KO_1      KO_2      KO_3  mean_WT

```

```

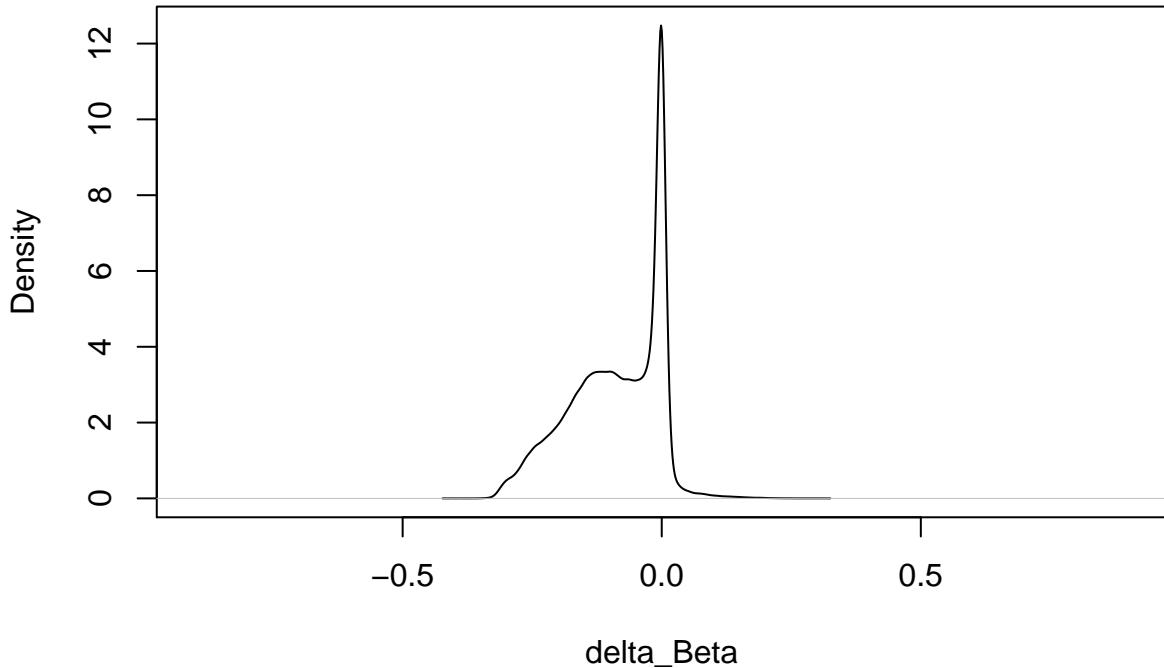
## 1 0.9092991 0.4136873 0.8666327 0.8472843 0.4729088 0.4188734 0.7298730
## 2 0.9359841 0.6680526 0.9013848 0.8627078 0.7305517 0.6466701 0.8351405
## 3 0.8872244 0.8041511 0.8815632 0.8479977 0.8226051 0.7533346 0.8576462
## 4 0.8928415 0.7240071 0.8827264 0.8415437 0.7922252 0.7251262 0.8331917
## 5 0.8910567 0.6137658 0.8668285 0.8413939 0.6566892 0.5407618 0.7905503
## 6 0.7533822 0.3238636 0.6942097 0.6420095 0.3603798 0.2677165 0.5904852
##      mean_KO      delta_B
## 1 0.5796889 -0.15018415
## 2 0.7466432 -0.08849729
## 3 0.8079791 -0.04966708
## 4 0.7862984 -0.04689332
## 5 0.6796149 -0.11093538
## 6 0.4233686 -0.16711657

v1 <- c(1,2,3,4,5,6)
shuffled_B_values$pval = apply(shuffled_B_values[,2:7], 1,
                                function(v1) t.test(x=v1[1:3],y=v1[4:6])$p.value)

par(mfrow=c(1,1))
plot(density(shuffled_B_values$delta_B), xlab= "delta_Beta" ,
     main= "distribution of the delta-beta" , xlim =c(-max(B_values$delta_B),
                                                    max(B_values$delta_B)))

```

distribution of the delta-beta



discussion: lorsqu'on effectue un contrôle négatif(ici un mélange des colonnes), la distribution des delta_beta est moins large, plus ramassée autour de 0 car finalement, la différence de méthylation entre les nouveaux controles et les nouveaux KO est davantage diminuée que lorsqu'on utilise des données correctes ce qui paraît normal.

```

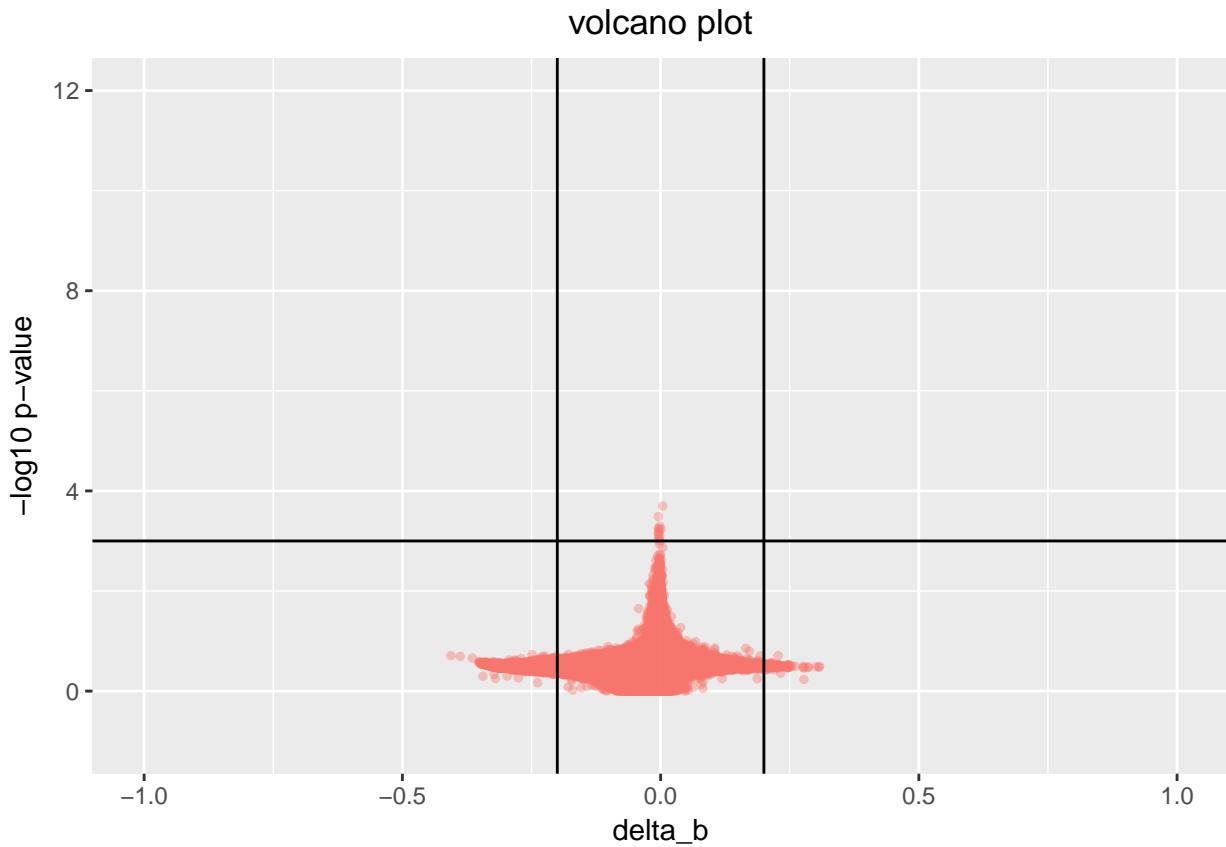
shuffled_B_values$threshold <- as.factor(abs(shuffled_B_values$delta_B)>0.2
                                         & shuffled_B_values$pval< 0.001)
ggplot(data=shuffled_B_values, aes(x=shuffled_B_values$delta_B,

```

```

y=-log10(shuffled_B_values$pval),
colour=threshold) +
geom_point(alpha=0.4, size=1) +
theme(legend.position = "none") +
geom_hline(aes(yintercept=3))+ geom_vline(xintercept = 0.2)+
xlim(c(-1, 1)) + ylim(c(-1, 12)) + geom_vline(xintercept = -0.2)+
xlab("delta_b") + ylab("-log10 p-value")+
ggtitle("volcano plot")+theme(plot.title = element_text(hjust = 0.5))

```



Discussion: Le volcano a perdu ses doubles positifs ($|\text{delta_Beta}| > \text{seuil}$ et $\text{P_value} > \text{seuil}$) c'est à dire que tout ce qui était significativement méthylé n'existe plus. En faisant le shuffle, on a donc supprimé les méthylations qui étaient significatives. On peut donc conclure que nos résultats ne sont pas dus au hasard puisque contrairement aux résultats du shuffle, ils sont cohérents et significatifs.