DRD report

Roberto Polverelli Monti
LM Bioinformatics
A.Y. 2018/2019
Bologna

Step 1-4

Preliminary work

- Start with a clean environment, and load the *minfi* package:

```r
rm(list=ls())
setwd("~/reportDRD")
library("minfi")
```

- Import targets, then load the manifest:

```r
targets <- read.metharray.sheet("inputData")
load("Illumina450Manifest.RData")
```

- Subset the manifest as to remove those probes that do not map to any chromosome:

```r
Illumina450Manifest <- droplevels(Illumina450Manifest[Illumina450Manifest$CHR!="",])
```

- Extract probe intensities:

```r
RGset <- read.metharray.exp(targets=targets)
save(RGset,file = "RGset.Rdata")
```

- Store the red and green fluorescence intensities:

```r
red <- data.frame(getRed(RGset))
green <- data.frame(getGreen(RGset))
```

- For convenience's sake, store the address and obtain the probe type:

```r
my_address <- 52682510
probe_type <- as.character(Illumina450Manifest[Illumina450Manifest$AddressA_ID==my_address,]$Infinium_Design_Type)
```

- Separately obtain the red and green intensities for the address:

```r
my_address_green <- green[row.names(green)==my_address,]
my_address_red <- red[row.names(red)==my_address,]
my_intensities <- cbind(t(my_address_red),t(my_address_green),probe_type)
my_intensities <- cbind(row.names(my_intensities),my_intensities)
rownames(my_intensities) <- NULL
colnames(my_intensities) <- c("Sample","Red","Green","Type")
View(my_intensities)
```

	Sample	Red	Green	Type
1	X5775278008_R01C01	941	3573	II
2	X5775278008_R02C01	1259	3820	II
3	X5775278008_R04C01	875	866	II
4	X5775278008_R05C01	745	716	II
5	X5775278035_R01C01	675	1652	II
6	X5775278035_R02C01	462	393	II
7	X5775278035_R04C01	1846	4182	II
8	X5775278035_R05C01	1331	923	II

Table 1: the red and green intensities for the given address.

- Convert red/green channel intensities to methylation signal, and save it to a file:

```r
MSet.raw <- preprocessRaw(RGset)
save(MSet.raw,file="MSet.raw.RData")
```

Step 5

Quality checks

- Obtain and plot the chipwide medians of the Meth and Unmeth channels:

```r
pdf(file="step5.1.pdf")
qc <- getQC(MSet.raw)
plotQC(qc)
dev.off()
```

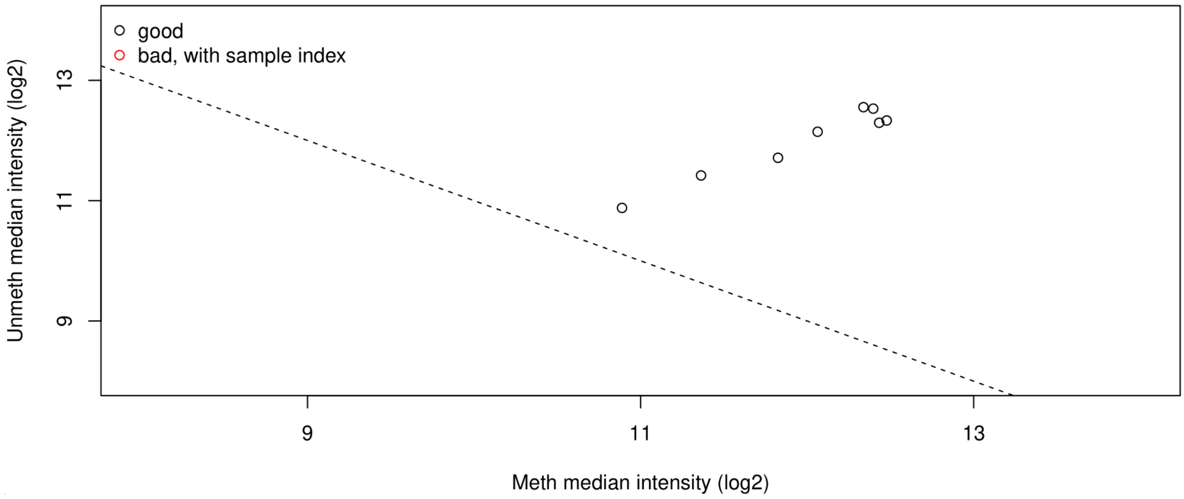

Figure 1: no sample reporting bad detection intensity.

- Produce the negative control probe signal plot component of the QC report:

```
pdf(file="step5.2.pdf")
controlStripPlot(RGset,controls="NEGATIVE")
dev.off()
```

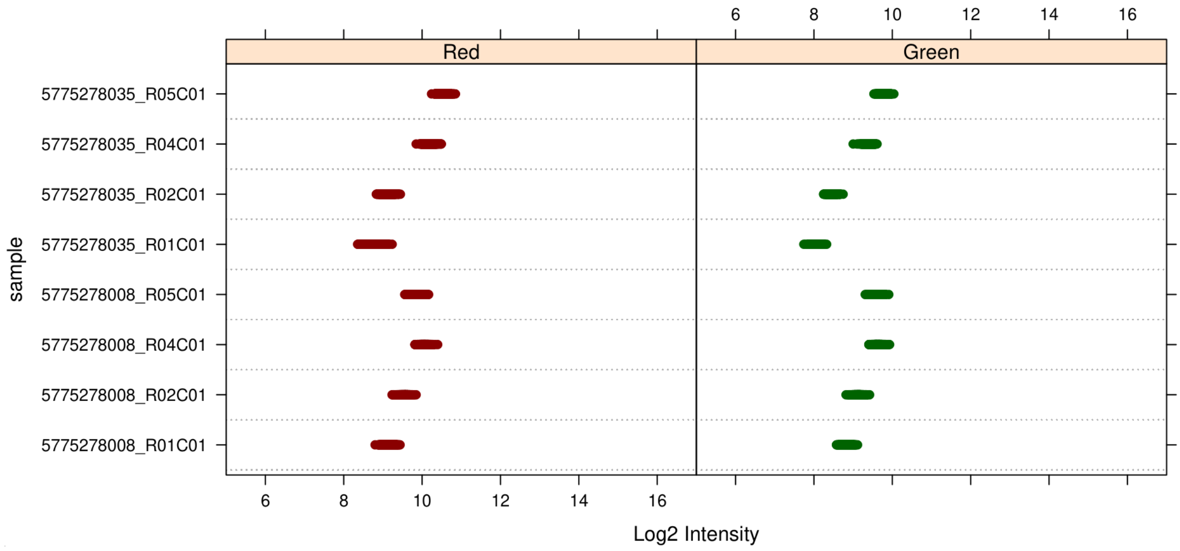

Figure 2: for all eight samples, Log2 intensities fall within detection range.

- Check the number of bad probes for each sample, with detection P-value 0.01 as threshold:

```
det_p <- detectionP(RGset)
failed_positions <- det_p > 0.01
failed_count <- c()
for (i in colnames(failed_positions)){
  failed_count.newrow <- t(as.integer(summary(failed_positions[,i])
[-1]))
  failed_count.newrow <- cbind(i,failed_count.newrow)
  failed_count <- rbind(failed_count,failed_count.newrow)
}
rm(i,failed_count.newrow)
colnames(failed_count) <- c("Sample","Passed","Failed")
View(failed_count)
```

	Sample	Passed	Failed
1	5775278008_R01C01	484946	566
2	5775278008_R02C01	485061	451
3	5775278008_R04C01	484665	847
4	5775278008_R05C01	484631	881
5	5775278035_R01C01	485038	474
6	5775278035_R02C01	484845	667
7	5775278035_R04C01	485182	330
8	5775278035_R05C01	484481	1031

Table 2: number of positions reporting a detection P-value lower or higher than the threshold.

- Store in a vector all those probes with a reported detection P-value higher than 0.01 in more than 1% of the samples, and check whether this number (i.e. 2191) is in agreement with the number of good probes and the total number of probes:

```
failed_proportion <- rowMeans(failed_positions)
bad_probes <- names(failed_proportion[failed_proportion > 0.01])
good_probes <- names(failed_proportion[failed_proportion <= 0.01])

length(good_probes) + length(bad_probes) == length(failed_proportion)
# [1] TRUE
length(bad_probes)
# [1] 2191
```

Step 6

β-values and M-values

- Obtain the raw β-values and M-values from the MethylSet object:

```
beta <- getBeta(MSet.raw)
M <- getM(MSet.raw)
```

- Subset the data into *WT* and *DS* subjects:

```r
wt_names <- targets[targets$Group=="WT","Basename"]
wt_names <- sub(".+/","",wt_names)
wt_beta <- beta[,wt_names]
wt_M <- M[,wt_names]

ds_names <- targets[targets$Group=="DS","Basename"]
ds_names <- sub(".+/","",ds_names)
ds_beta <- beta[,ds_names]
ds_M <- M[,ds_names]
```

- Obtain the mean methylation values for each subset, and plot their densities:

```r
mean_wt_beta <- apply(wt_beta,1,mean)
mean_ds_beta <- apply(ds_beta,1,mean)

mean_wt_M <- apply(wt_M,1,mean)
mean_ds_M <- apply(ds_M,1,mean)

pdf(file="step6.1.pdf")
plot(density(mean_wt_beta,na.rm=TRUE),col=4,main="Mean Beta-values, density",xlab="Beta")
lines(density(mean_ds_beta,na.rm=TRUE),col=3)
legend('topright',pch=c(16,16),col=c(4,3),legend=c("WT","DS"))
dev.off()
```

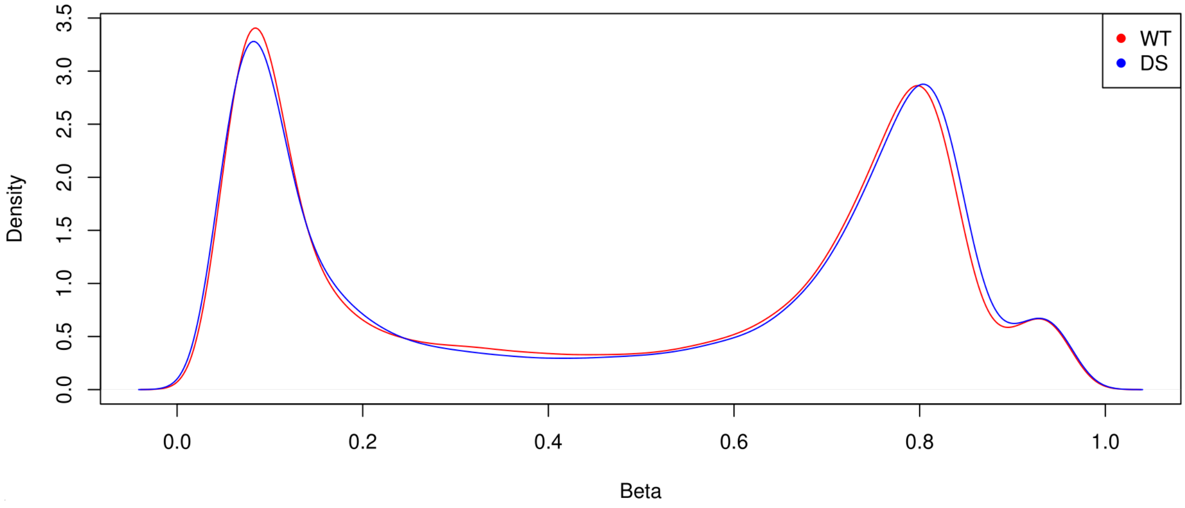

Figure 3: density of mean β-values.

```r
pdf(file="step6.2.pdf")
plot(density(mean_wt_M,na.rm=TRUE),col=4,main="Mean M-values, density",xlab="M")
lines(density(mean_ds_M,na.rm=TRUE),col=3)
legend('topright',pch=c(16,16),col=c(4,3),legend=c("WT","DS"))
dev.off()
```

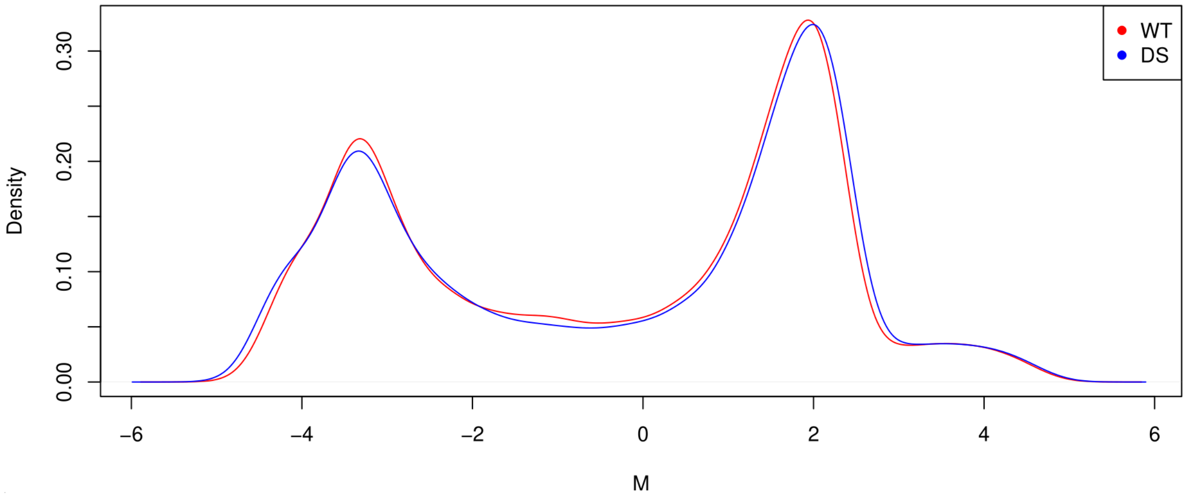

Figure 4: density of mean M-values.

Step 7

Subsetting

- Subset the manifest according to probe chemistry:

```
type_I_manifest <-
Illumina450Manifest[Illumina450Manifest$Infinium_Design_Type=="I",]
type_II_manifest <-
Illumina450Manifest[Illumina450Manifest$Infinium_Design_Type=="II",]
```

- Subset beta according to type of probe, and compute both the mean and S.D.:

```
beta_I <- beta[rownames(beta) %in% type_I_manifest$IlmnID,]
mean_beta_I <- apply(beta_I,1,mean)
sd_beta_I <- apply(beta_I,1,sd)

beta_II <- beta[rownames(beta) %in% type_II_manifest$IlmnID,]
mean_beta_II <- apply(beta_II,1,mean)
sd_beta_II <- apply(beta_II,1,sd)
```

Normalization

- Normalize fluorescence intensities, subset them according to type of probe, and compute the mean and the S.D.:

```r
norm_RGset <- preprocessFunnorm(RGset)
norm_beta <- getBeta(norm_RGset)

norm_beta_I <- norm_beta[rownames(norm_beta) %in% type_I_manifest$IlmnID,]
mean_norm_beta_I <- apply(norm_beta_I,1,mean)
sd_norm_beta_I <- apply(norm_beta_I,1,sd)

norm_beta_II <- norm_beta[rownames(norm_beta) %in% type_II_manifest$IlmnID,]
mean_norm_beta_II <- apply(norm_beta_II,1,mean)
sd_norm_beta_II <- apply(norm_beta_II,1,sd)
```

Comparison

- Produce a six panel plot comparing raw and normalized data:

```r
pdf(file="step7.pdf",width=10,height=5)
par(mfrow=c(2,3))

plot(density(mean_beta_I,na.rm=T),main="Mean of Raw Beta-values",xlab="mean(Beta)",col=4)
lines(density(mean_beta_II,na.rm=T),col=3)
legend('top',pch=c(16,16),col=c(4,3),legend=c("I","II"))

plot(density(sd_beta_I,na.rm=T),main="S.D. of Raw Beta-values",xlab="sd(Beta)",col=4)
lines(density(sd_beta_II,na.rm=T),col=3)
legend('top',pch=c(16,16),col=c(4,3),legend=c("I","II"))

colnames(beta) <- sub("57752780","",colnames(beta))
boxplot(beta,main="Raw Beta-values",las=2,cex.axis=0.7)
colnames(beta) <- sub("","57752780",colnames(beta))

plot(density(mean_norm_beta_I,na.rm=T),main="Mean of Norm. Beta-values",xlab="mean(Beta)",col=4)
lines(density(mean_norm_beta_II,na.rm=T),col=3)
legend('top',pch=c(16,16),col=c(4,3),legend=c("I","II"))

plot(density(sd_norm_beta_I,na.rm=T),main="S.D. of Norm. Beta-values",xlab="sd(Beta)",col=4)
lines(density(sd_norm_beta_II,na.rm=T),col=3)
legend('top',pch=c(16,16),col=c(4,3),legend=c("I","II"))

colnames(norm_beta) <- sub("57752780","",colnames(norm_beta))
boxplot(norm_beta,main="Normalized Beta-values",las=2,cex.axis=0.7)
colnames(norm_beta) <- sub("","57752780",colnames(norm_beta))
dev.off()
```

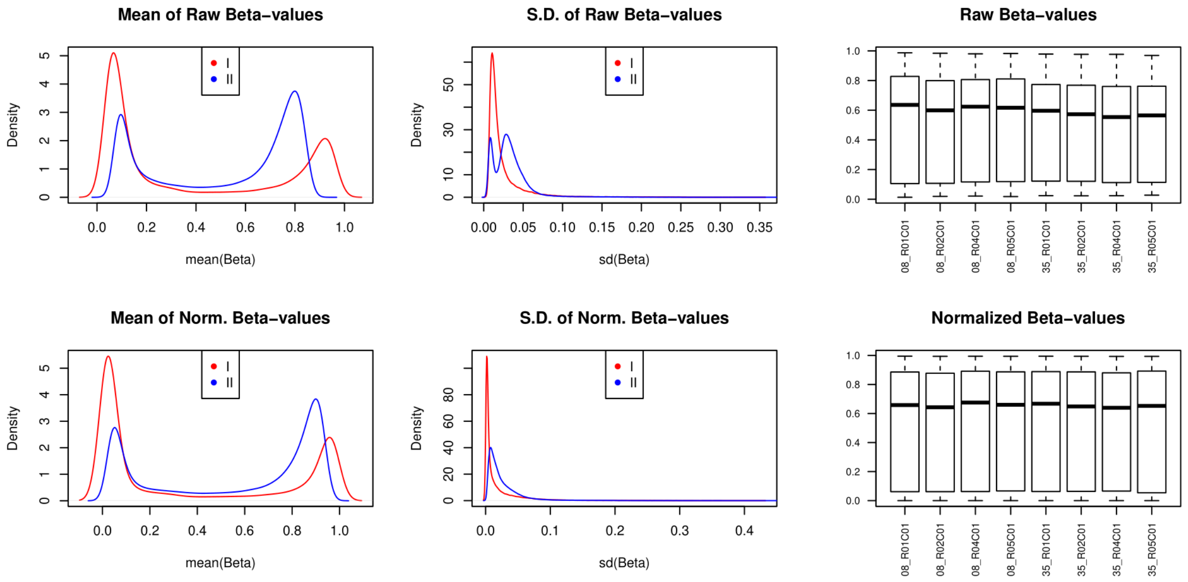

Figure 5: after normalization, shifing of strongly methylated sites allows for type I and II probes distribution to better coincide; type I probes are much more represented in weakly methylated sited, as would be expected giving that they preferentially target CpG islands. Normalized data also shows shorter variance which makes type I and II more easily comparable. Experimental variation across samples is also reduced in normalized data.

Step 8

Filtering

- Extract the M-values from the normalized RGset, then filter the normalized data retaining only the good probes:

```
norm_M <- getM(norm_RGset)

filt_norm_beta <- norm_beta[!row.names(norm_beta) %in% bad_probes,]
filt_norm_M <- norm_M[!row.names(norm_M) %in% bad_probes,]
```

- Checking sizes just to be sure:

```
nrow(norm_beta) - nrow(filt_norm_beta) == length(bad_probes)
# [1] TRUE
nrow(norm_M) - nrow(filt_norm_M) == length(bad_probes)
# [1] TRUE
nrow(filt_norm_beta) == length(good_probes)
# [1] TRUE
nrow(filt_norm_M) == length(good_probes)
# [1] TRUE
```

Step 9

Homo/heteroscedasticity

- Compute mean and S.D. for the now filtered data:

```
mean_filt_norm_beta <- apply(filt_norm_beta,1,mean)
sd_filt_norm_beta <- apply(filt_norm_beta,1,sd)

mean_filt_norm_M <- apply(filt_norm_M,1,mean)
sd_filt_norm_M <- apply(filt_norm_M,1,sd)
```

- We can assess the homo/heteroscedasticity of the data by plotting the mean against the S.D.:

```
pdf(file="step9.pdf",width=10,height=5)
par(mfrow=c(1,2))
smoothScatter(mean_filt_norm_beta,sd_filt_norm_beta,main="Beta-
Values",xlab="Mean(Beta-Value)",ylab="SD(Beta-Value)")
lines(lowess(mean_filt_norm_beta,sd_filt_norm_beta),col="red")
smoothScatter(mean_filt_norm_M,sd_filt_norm_M,main="M-
Values",xlab="Mean(M-Value)",ylab="SD(M-Value)")
lines(lowess(mean_filt_norm_beta,sd_filt_norm_beta),col="red")
dev.off()
```

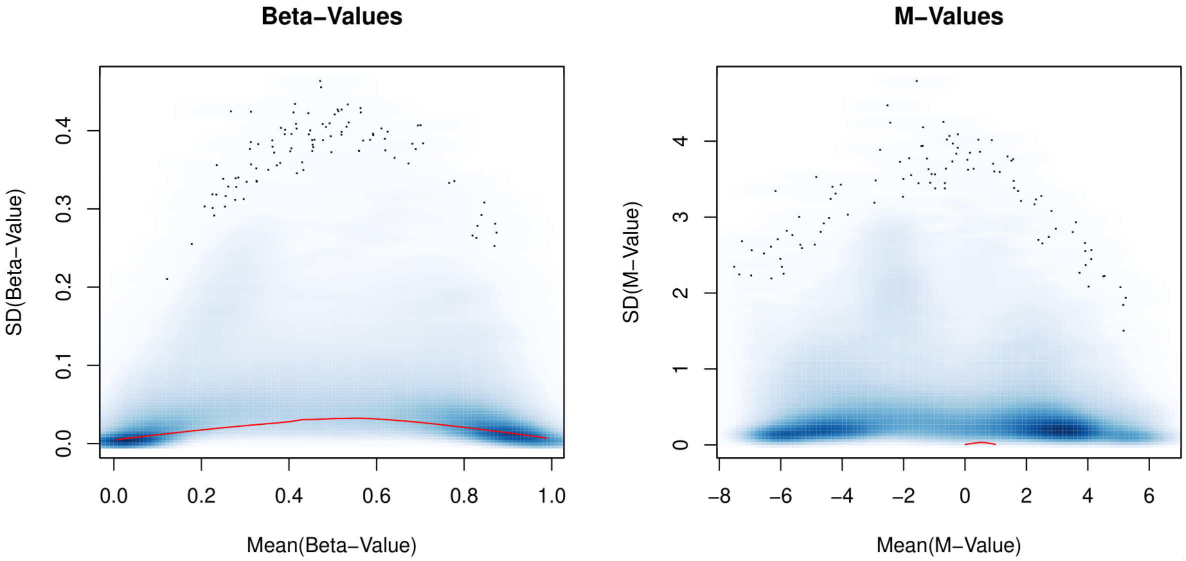

Figure 6: β-values appear to be more heavily heteroscedastic, showing lower variance within the two groups coinciding with near-zero and near-one methylation signal; M-values are still heteroscedastic in nature, only less prominently.

Step 10

Principal Component Analysis

- Perform PCA on the normalized β-values:

```
pca <- prcomp(t(filt_norm_beta),scale=TRUE)
summary(pca)

# Importance of components:
#                PC1      PC2      PC3      PC4      PC5      PC6      PC7       PC8
# (1)     322.4629 287.7958 276.2468 256.3168 238.4123 224.9492 216.93110 4.614e-12
# (2)       0.2151   0.1714   0.1579   0.1359   0.1176   0.1047   0.09737 0.000e+00
# (3)       0.2151   0.3865   0.5444   0.6803   0.7979   0.9026   1.00000 1.000e+00
#
## (1):  Standard deviation
## (2):  Proportion of Variance
## (3):  Cumulative Proportion
```

- Plot PC1 and PC2:

```
pdf(file="step10.pdf",width=7,height=5)
targets$Group <- as.factor(targets$Group)
plot(pca$x[,1],pca$x[,2],main="PCA of Norm. Beta Values",xlab="PC1",ylab="PC2",xlim=c(-800,800),ylim=c(-600,600),cex=2,col=targets$Group)
text(pca$x[,1],pca$x[,2],labels=rownames(pca$x),cex=0.5,pos=1)
legend("topright",legend=levels(targets$Group),col=c(1:nlevels(targets$Group)),pch=1)
dev.off()
```

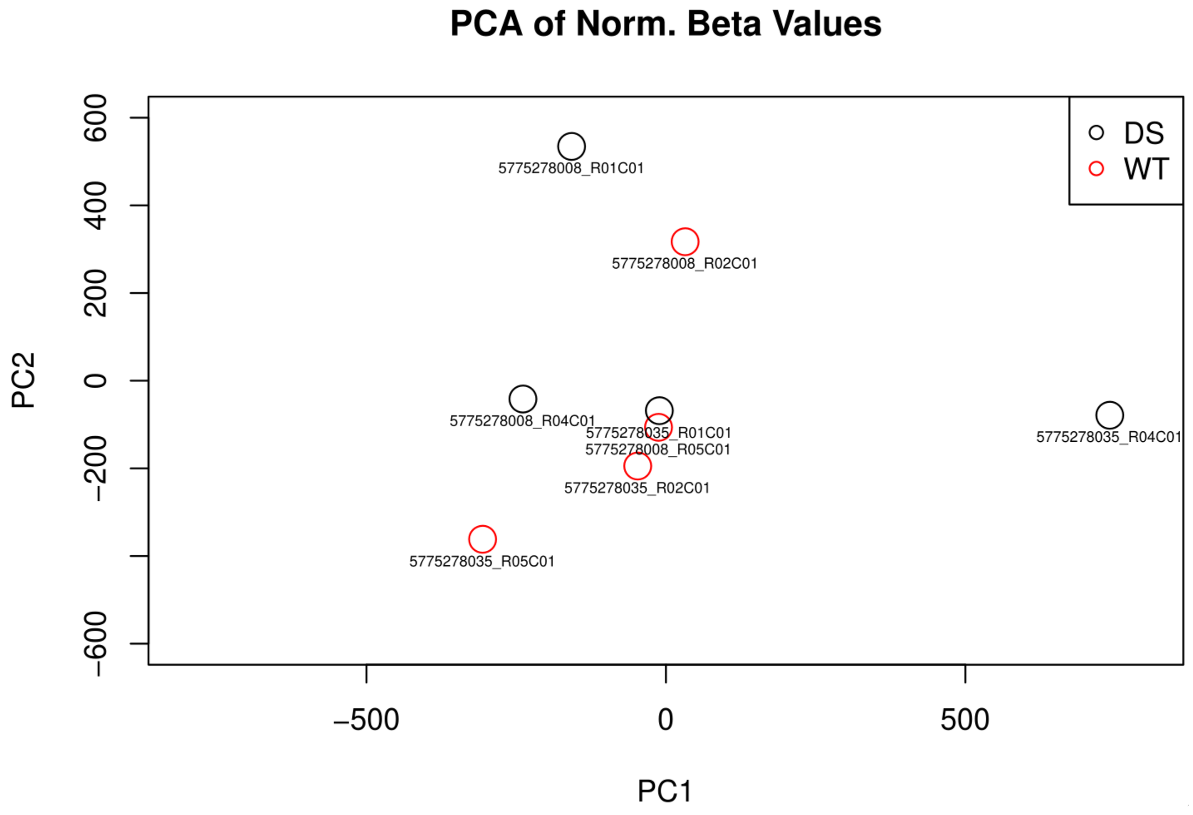

Figure 7: *5775278008_R02C01* appears to be an outlier in the PCA plot.

Step 11

ANOVA

- Perform ANOVA in order to identify differentially methylated probes between groups *WT* and *DS*, using sex as a covariate:

```
my_anova_fun <- function(x) {
  aov_test <- aov(x~ targets$Group+as.factor(targets$Female))
  return(summary(aov_test)[[1]][[5]][1])
}

p_values_anova <- apply(filt_norm_beta,1,my_anova_fun)
```

Step 12

Correction

- Count the number of differentially methylated probes considering a significance threshold of 0.05:

```
length(p_values_anova[p_values_anova<=0.05])

# [1] 30493
```

- Same thing but after Bonferroni correction:

```
bonf_p_values_anova <- p.adjust(p_values_anova,"bonferroni")
length(bonf_p_values_anova[bonf_p_values_anova<=0.05])

# [1] 1
```

- After Benjamini-Hochberg:

```
benj_p_values_anova <- p.adjust(p_values_anova,"BH")
length(benj_p_values_anova[benj_p_values_anova<=0.05])

# [1] 10
```

- A table showing the probes just before mentioned:

```
final_anova <- data.frame(filt_norm_beta,p_values_anova,bonf_p_values_anova,benj_p_values_anova)
final_anova <- final_anova[order(benj_p_values_anova),]
final_anova_benj_0.05 <- final_anova[final_anova$benj_p_values_anova<=0.05,]
View(final_anova_benj_0.05[,9:11])
```

	P-values	Bonferroni	BH
cg19246080	3.841172e-09	0.001856519	0.001856519
cg03639185	1.606309e-07	0.077636264	0.031566401
cg23260026	1.959344e-07	0.094699204	0.031566401
cg07867687	3.359398e-07	0.162366767	0.040591692
cg16044109	6.457099e-07	0.312085155	0.044583594
cg25542438	5.535240e-07	0.267529754	0.044583594
cg01086462	4.848629e-07	0.234344410	0.044583594
cg08967584	7.640929e-07	0.369302139	0.046162767
cg06621691	9.697929e-07	0.468721281	0.048031239
cg20348344	9.937751e-07	0.480312394	0.048031239

Table 3: showing uncorrected and corrected P-values for the ten probes that have a reported FDR adjusted P-value lower than 0.05 with Benjamini-Hochberg.

Step 13

Manhattan plot

- Produce a Manhattan plot of the uncorrected P-values:

```
library("gap")
final_anova_with_IlmnID <- data.frame("IlmnID"=rownames(final_anova),final_anova)
final_anova_annotated <- merge(final_anova_with_IlmnID,Illumina450Manifest,by="IlmnID")

chrs <- c(as.character(1:22),"X","Y")
db <- data.frame(final_anova_annotated$CHR, final_anova_annotated$MAPINFO, final_anova_annotated$p_values_anova)
db$final_anova_annotated.CHR <- factor(db$final_anova_annotated.CHR,levels=chrs)

pdf(file="step13.1.pdf",width=8,height=4)
mhtplot(db,xlab="",ylab="",control=mht.control(colors=rep(c("darkgrey","blue"),12)))
```

```
axis(2,cex.axis=0.8)
abline(a=-log10(0.05),b=0)
title(xlab="Chromosome",line=0.5)
title(ylab="-Log10(P)",line=2.5)
dev.off()
```

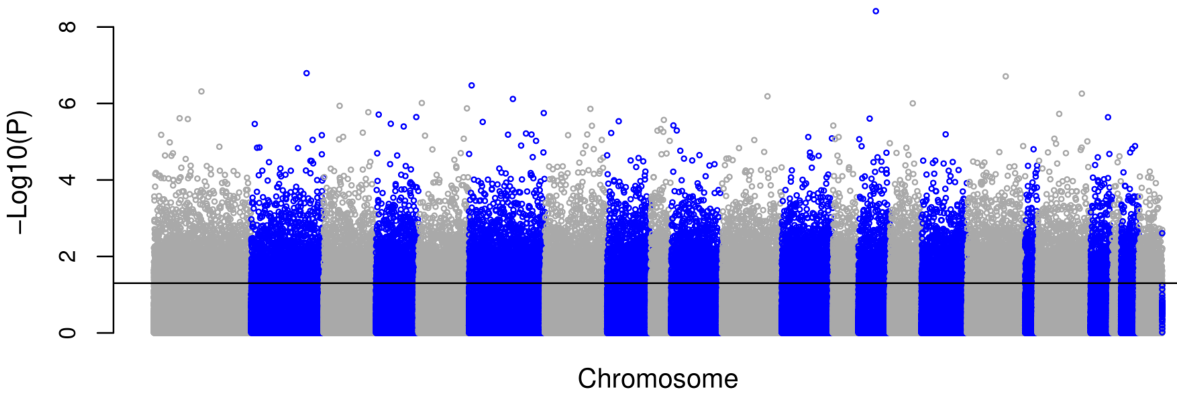

Figure 8: the distribution of P-values is mostly uniform across all chromosomes.

Volcano plot

- Produce a Volcano plot of the uncorrected P-values:

```
to_volc <- data.frame(mean_ds_beta - mean_wt_beta)
sites <- rownames(to_volc)
to_volc <- data.frame(sites,to_volc)

colog_p <- data.frame(-log10(p_values_anova))
sites <- rownames(colog_p)
colog_p <- data.frame(sites,colog_p)

to_volc <- merge(to_volc,colog_p,by="sites")
names(to_volc) <- c("sites","delta","colog_p")

pdf(file="step13.2.pdf",width=8,height=6)
plot(to_volc$delta, to_volc$colog_p, main="", xlab="mean(beta DS) - mean(beta WT)", ylab="-log10(P)",cex=0.5,col="darkgrey")
dev.off()
```

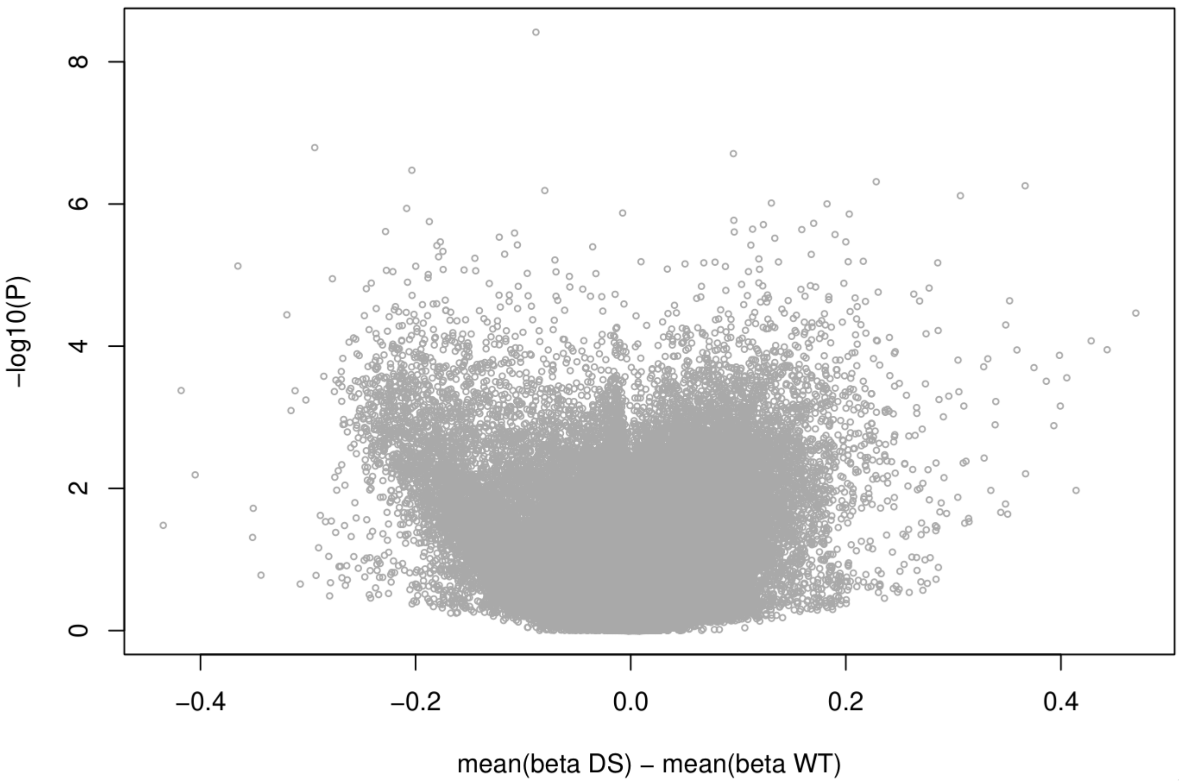

Figure 9: the difference in methylation level doesn't seem to be varying significantly at different reported P-values.

Step 14

Heatmap

- Produce a heatmap of the top one hundred methylated probes:

```
library(gplots)

final_anova_annotated <- final_anova_annotated[order(final_anova_annotated$p_values_anova),]

matrix=as.matrix(final_anova_annotated[1:100,2:9])
targets$Group
# [1] DS WT DS WT DS WT DS WT
# Levels: DS WT
colorbar <- rep(c("orange","green"),4)

pdf("step14.complete.pdf")
par(mar=c(2,2,2,2))
heatmap.2(matrix,col=terrain.colors(100),Rowv=T,Colv=T,
          dendrogram="both",key=T,ColSideColors=colorbar,
          density.info="none",trace="none",scale="none",symm=F,
          cexRow=1,cexCol=1,margins=c(10,8),srtCol=45)
dev.off()
```

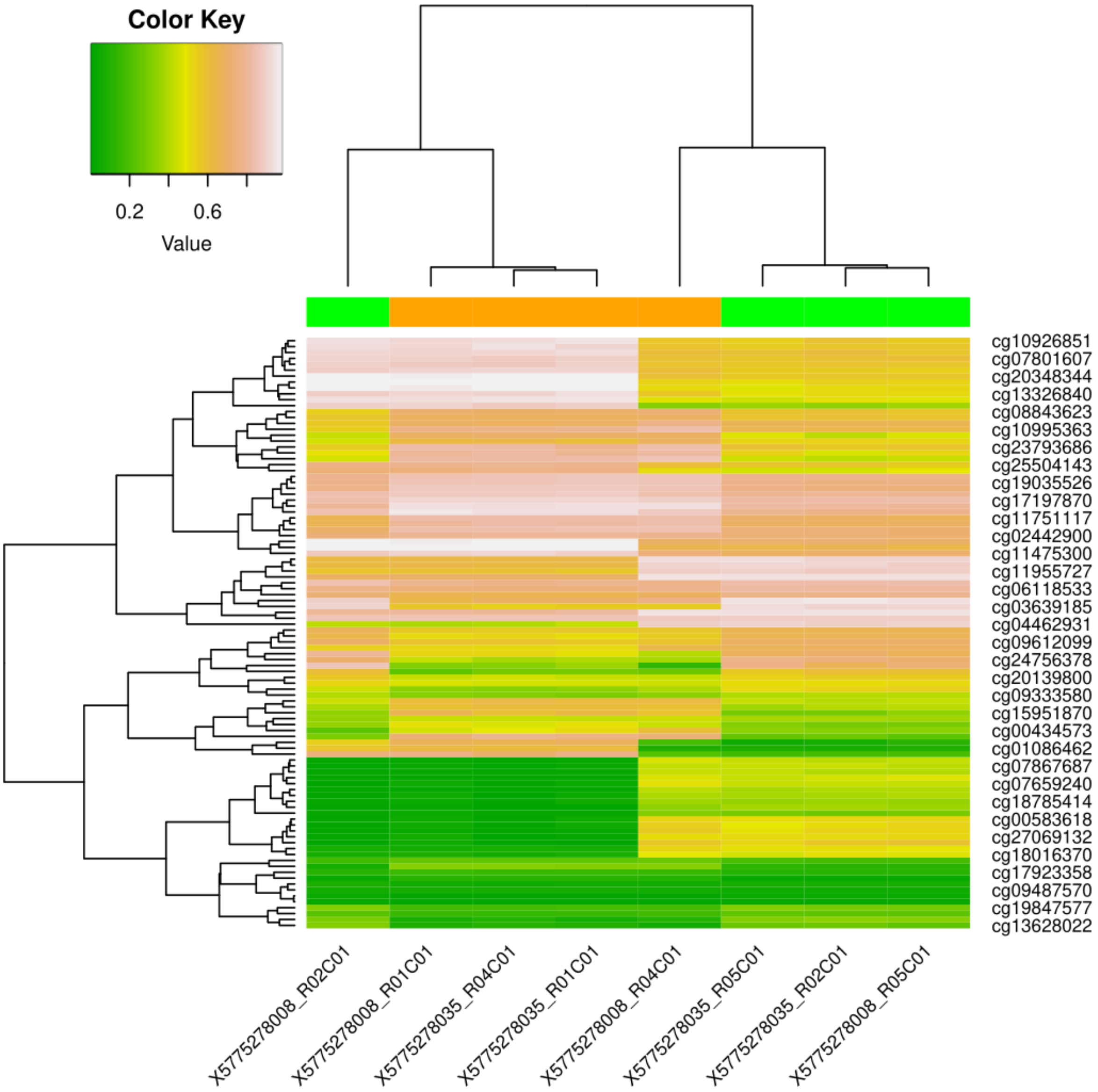

Figure 10: the clustering here doesn't seem to be discriminating very well between *WT* and *DS* individuals. Different agglomeration methods were tried other than complete linkage, namely single linkage and unweighted average linkage clustering, but there was no noticeable difference.
It seems that, for this small sample size, correcting for sex has affected more the output than *DS* and *WT* groups, as in fact the algorithm has correctly clusterized the targets on the basis of sex.

```r
targets$Female
# [1] 0 0 1 1 0 1 0 1
colorbar <- c("cyan","cyan","pink","pink","cyan","pink","cyan","pink")

pdf("step14.complete.sex.pdf")
par(mar=c(15,2,2,2))
heatmap.2(matrix,col=terrain.colors(100),Rowv=T,Colv=T,
     dendrogram="both",key=T,ColSideColors=colorbar,
     density.info="none",trace="none",scale="none",symm=F,
     cexRow=1,cexCol=1,margins=c(10,8),srtCol=45)
dev.off()
```

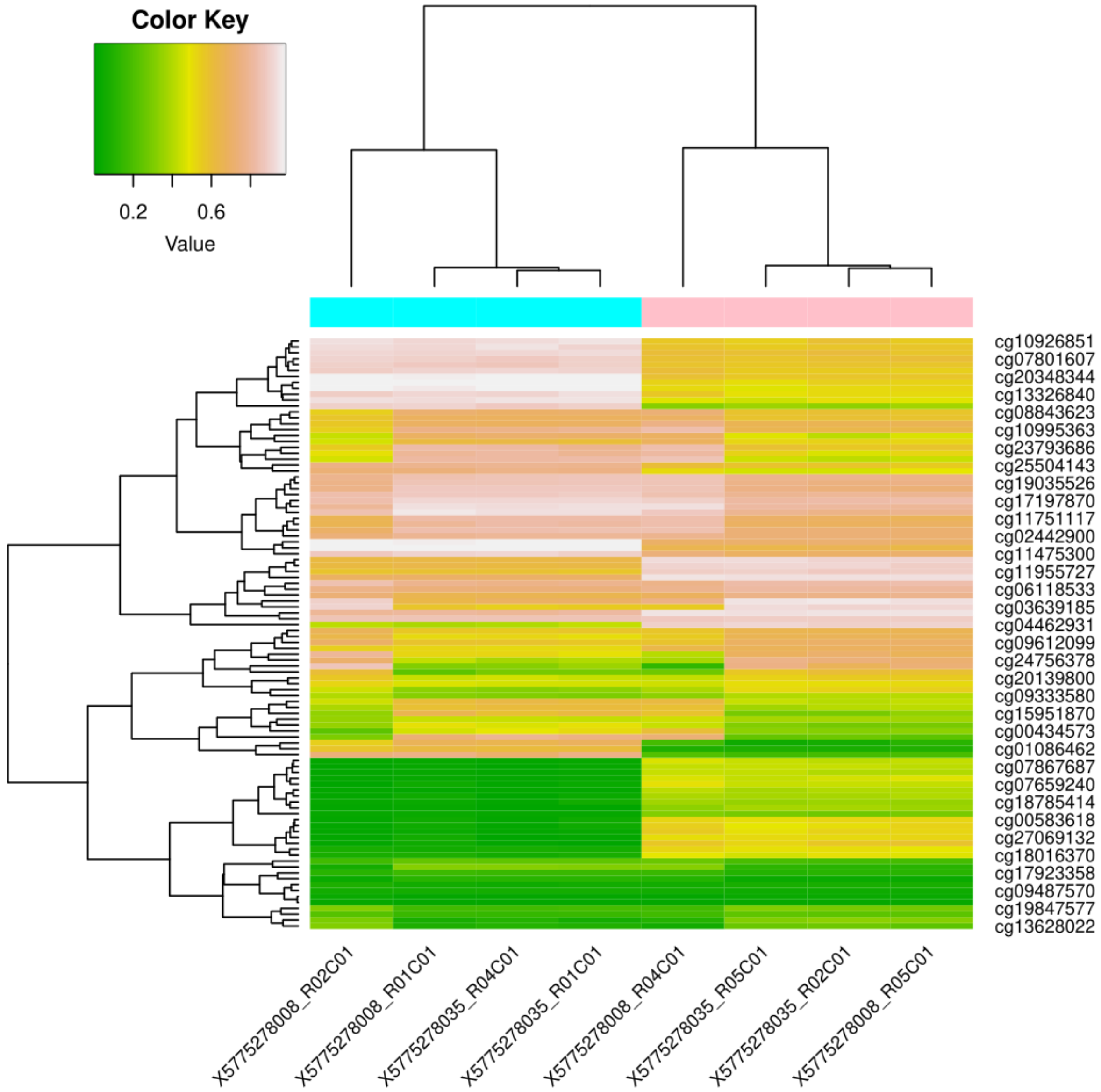

Figure 11: same heatmap but with sex labelled targets.

Step 15

Chromosome 21

- Plot the mean beta density only of those probes that map to chromosome 21 (of which there are 4243):

```
data <- data.frame(rownames(beta),beta)
colnames(data)[1] <- "IlmnID"
data <- merge(data,Illumina450Manifest[,c(1,12)],by="IlmnID")
rownames(data) <- data[,1]
data <- data[,-1]
data <- subset(data,data$CHR=="21")
beta_c21 <- data[,1:8]
rm(data)

dim(beta_c21)
# [1] 4243    8

wt_names <- targets[targets$Group=="WT","Basename"]
wt_names <- sub(".+/","X",wt_names)
wt_beta_c21 <- beta_c21[,wt_names]

ds_names <- targets[targets$Group=="DS","Basename"]
ds_names <- sub(".+/","X",ds_names)
ds_beta_c21 <- beta_c21[,ds_names]

mean_wt_beta_c21 <- apply(wt_beta_c21,1,mean)
mean_ds_beta_c21 <- apply(ds_beta_c21,1,mean)

pdf(file="step15.pdf",width=10,height=5)
plot(density(mean_wt_beta_c21,na.rm=TRUE),col=2,main="Mean Beta-values, density (Chr21)",xlab="Beta")
lines(density(mean_ds_beta_c21,na.rm=TRUE),col=4)
legend('topright',pch=c(16,16),col=c(2,4),legend=c("WT","DS"))
dev.off()
```

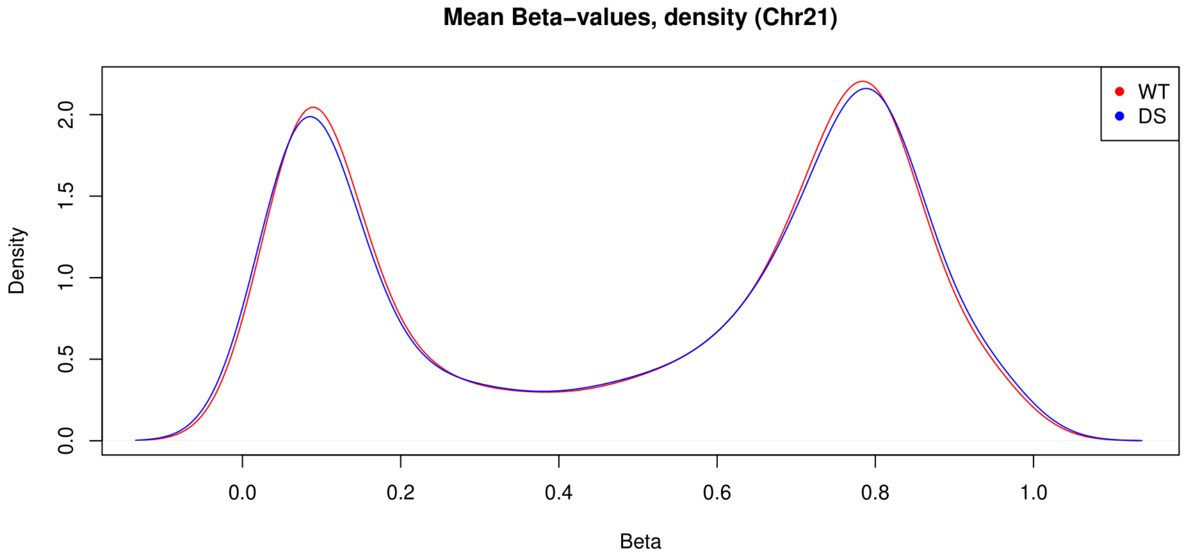

Figure 12: the two lines are so similar here that it doesn't seem there is a significant difference in methylation level between *WT* and *DS* on chromosome 21.