RNA-seq analysis of CSC (TPC) and noncsc datasets

July 22nd, 2016

Here we describe the following steps we had taken in the RNA-seq analysis of CSC and non-csc samples :

I. ALIGNMENT to mm10 mouse genome by using STAR aligner :

The STAR aligner is available at /N/users/btanasa/Software/STAR-2.5/bin/Linux_x86_64 The alignment procedure started from the fastq files containing the clipped reads:

```
STAR --runThreadN 12  --outWigStrand Stranded --readFilesCommand zcat \
--genomeDir /N/users/btanasa/GENENTECH-ERIKA/mm10-STAR --sjdbOverhang 100 \
--readFilesIn file1.fastq file2.fastq \
--outFileNamePrefix file
```

II. READ COUNTING with HOMER on RefSeq genes of mm10 genome :

The read counting procedure followed the guidelines available on HOMER website :

http://homer.salk.edu/homer/ngs/rnaseq/index.html

```
makeTagDirectory folder/  -format sam -sspe

makeUCSCfile folder/ -fragLength given -o auto -strand separate
```

Specifically, we have used the following commands :

```
#############################################################################
#############################################################################

makeTagDirectory CSC_011515-1_L3/ CSC_011515-1_L3.sam -format sam  -sspe
makeTagDirectory CSC_060215-1_L5/ CSC_060215-1_L5.sam -format sam  -sspe
makeTagDirectory CSC_092415-1_L5/ CSC_092415-1_L5.sam -format sam  -sspe
makeTagDirectory CSC_092915-1_L5/ CSC_092915-1_L5.sam -format sam  -sspe
makeTagDirectory CSC_102215-1_L4/ CSC_102215-1_L4.sam -format sam  -sspe
makeTagDirectory CSC_102315-1_L2/ CSC_102315-1_L2.sam -format sam  -sspe


makeTagDirectory noncsc_011515-0_L5/ noncsc_011515-0_L5.sam -format sam  -sspe
makeTagDirectory noncsc_060215-0_L1/ noncsc_060215-0_L1.sam -format sam  -sspe
makeTagDirectory noncsc_092415-0_L5/ noncsc_092415-0_L5.sam -format sam  -sspe
makeTagDirectory noncsc_092915-0_L5/ noncsc_092915-0_L5.sam -format sam  -sspe
makeTagDirectory noncsc_102215-0_L4/ noncsc_102215-0_L4.sam -format sam  -sspe
makeTagDirectory noncsc_102315-0_L1/ noncsc_102315-0_L1.sam -format sam  -sspe


#############################################################################
#############################################################################


makeUCSCfile CSC_011515-1_L3 -fragLength given -o auto -strand separate
makeUCSCfile CSC_060215-1_L5 -fragLength given -o auto -strand separate
makeUCSCfile CSC_092415-1_L5 -fragLength given -o auto -strand separate
makeUCSCfile CSC_092915-1_L5 -fragLength given -o auto -strand separate
makeUCSCfile CSC_102215-1_L4 -fragLength given -o auto -strand separate
makeUCSCfile CSC_102315-1_L2 -fragLength given -o auto -strand separate

makeUCSCfile noncsc_011515-0_L5 -fragLength given -o auto -strand separate
makeUCSCfile noncsc_060215-0_L1 -fragLength given -o auto -strand separate
makeUCSCfile noncsc_092415-0_L5 -fragLength given -o auto -strand separate
makeUCSCfile noncsc_092915-0_L5 -fragLength given -o auto -strand separate
makeUCSCfile noncsc_102215-0_L4 -fragLength given -o auto -strand separate
makeUCSCfile noncsc_102315-0_L1 -fragLength given -o auto -strand separate


#############################################################################
#############################################################################
```

Computing the gene expression, both the RAW COUNTS and the FPKM values :

```
analyzeRepeats.pl rna mm10 -strand both -count exons -d the-list-of-folders -rpkm
> mm10.expression.rpkm.txt

analyzeRepeats.pl rna mm10 -strand both -count exons -d the list-of-folders -noadj
> mm10.expression.noadj.txt
```

We have worked with an isoform per gene, and we have used the option -condenseGenes :

```
analyzeRepeats.pl rna mm10 -strand both -count exons -d the-list-of-folders -rpkm -condenseGenes
> mm10.expression.rpkm.condense-genes.txt
```

```
analyzeRepeats.pl rna mm10 -strand both -count exons -d the list-of-folders -noadj -condenseGenes
> mm10.expression.noadj.condense-genes.txt
```

More precisely :

```
analyzeRepeats.pl rna mm10 -strand both -count exons -condenseGenes -d \
CSC_011515-1_L3 \
CSC_060215-1_L5 \
CSC_092415-1_L5 \
CSC_092915-1_L5 \
CSC_102215-1_L4 \
CSC_102315-1_L2 \
noncsc_011515-0_L5 \
noncsc_060215-0_L1 \
noncsc_092415-0_L5 \
noncsc_092915-0_L5 \
noncsc_102215-0_L4 \
noncsc_102315-0_L1 \
> mm10.strand.both.count.exons.condense.genes.normalizedCOUNTS.txt


analyzeRepeats.pl rna mm10 -strand both -count exons -condenseGenes -rpkm -d \
CSC_011515-1_L3 \
CSC_060215-1_L5 \
CSC_092415-1_L5 \
CSC_092915-1_L5 \
CSC_102215-1_L4 \
CSC_102315-1_L2 \
noncsc_011515-0_L5 \
noncsc_060215-0_L1 \
noncsc_092415-0_L5 \
noncsc_092915-0_L5 \
noncsc_102215-0_L4 \
noncsc_102315-0_L1 \
> mm10.strand.both.count.exons.condense.genes.FPKM.txt

analyzeRepeats.pl rna mm10 -strand both -count exons -condenseGenes -noadj -d \
CSC_011515-1_L3 \
CSC_060215-1_L5 \
CSC_092415-1_L5 \
CSC_092915-1_L5 \
CSC_102215-1_L4 \
CSC_102315-1_L2 \
noncsc_011515-0_L5 \
noncsc_060215-0_L1 \
noncsc_092415-0_L5 \
noncsc_092915-0_L5 \
noncsc_102215-0_L4 \
noncsc_102315-0_L1 \
> mm10.strand.both.count.exons.condense.genes.NOADJ.txt
```

III. DIFFERENTIAL EXPRESSION with EDGER :

(here it is the general R code that could be applied to any RNA-seq data)

Initially, in the previous versions, we have used edgeR calls directly from HOMER (in the following way) :

```
getDiffExpression.pl RAW.txt CSC CSC CSC CSC CSC noncsc noncsc noncsc noncsc noncsc
> diffExpression.output.txt
```

Or we can do the calculations in edgeR in the following way :

```
library("edgeR")

eset <- read.delim("mm10.strand.both.count.exons.condense.genes.NOADJ.with.gene.names.simple.no060215",
                   row.names="Symbol")

group <- factor(c("csc","csc","csc","csc","csc","non","non","non","non","non"))
group <- relevel(group,ref="non")

subject <- factor(c(1,2,3,4,5,1,2,3,4,5))
design <- model.matrix(~group+subject)

y <- DGEList(counts=eset,group=group)

keep <- rowSums(cpm(y) > 0.5) >= 6

y <- y[keep,,keep.lib.sizes=FALSE]
y <- calcNormFactors(y)

logCPM <- cpm(y,log=TRUE,prior.count=3)
fit <- lmFit(logCPM, design)
fit <- eBayes(fit,trend=TRUE, robust=TRUE)

pdf("display.mm10.strand.both.count.exons.condense.genes.NOADJ.with.gene.names.simple.no060215.SA_fit_e
plotSA(fit)
dev.off()

results <- topTable(fit, coef=2, adjust="fdr", number=Inf)

write.table(results, file="mm10.strand.both.count.exons.condense.genes.NOADJ.with.gene.names.simple.no0
            sep="\t", eol="\n", row.names=TRUE, col.names=TRUE)
```

Or we can perform differential expression with limma, as it is shown below

(the results are similar ~90%-95% with the results obtained with edgeR).

IV. DIFFERENTIAL EXPRESSION with LIMMA :

(here it is the general code that can be used with any RNA-seq data)

```
library("limma")
library("edgeR")

### reading the expression dataset

eset <- read.delim("mm10.strand.both.count.exons.condense.genes.NOADJ.with.gene.names.simple.no060215",
                   row.names="Symbol")
```

```
### reading the library sizes, if we wish.

### libsize <- c(28182007, 28599704, 26386209, 28069058,
### 26361374, 27108450, 27097918, 12155292,
### 27124214, 26774125, 26241068, 26296757)

##########################################

### setting up the groups and the subjects

group <- factor(c("csc","csc","csc","csc","csc","non","non","non","non","non"))
subject <- factor(c(1,2,3,4,5,1,2,3,4,5))

### setting up the design and the contrast matrix

design <- model.matrix(~0+group+subject)
contrast.matrix <- makeContrasts(groupcsc-groupnon, levels=design)

### filtering the genes based on CPM :

y <- DGEList(counts=eset,group=group)

### keep <- rowSums(cpm(y, lib.size=libsize)>1) >= 3

keep <- rowSums(cpm(y)>0.5) >= 6
y <- y[keep,]

y$samples$lib.size <- colSums(y$counts)

### computing the normalization factors :

y <- calcNormFactors(y)

### using the VOOM transformation :

v <- voom(y,design,plot=FALSE)

pdf("display.mm10.strand.both.count.exons.condense.genes.NOADJ.with.gene.names.simple.no060215.limma.mea
v <- voom(y,design,plot=TRUE)
dev.off()

### doing the LINEAR FIT in LIMMA :

fit <- lmFit(v, design)

fit2 <- contrasts.fit(fit, contrast.matrix)
fit2 <- eBayes(fit2)

### obtaining and writing the results :

results <- topTable(fit2, coef=1, adjust="fdr", number=Inf)

write.table(results, file="mm10.strand.both.count.exons.condense.genes.NOADJ.with.gene.names.simple.no0
```

```
                    sep="\t", eol="\n", row.names=TRUE, col.names=TRUE)
```

V. Here it is the R code that we have used for the following analyses :

1. Differential expression analysis in LIMMA

4. PCA analysis

5. MDS analysis

6. Heatmaps of the DE genes

7. MA display

8. Scatter plot

9. Volcano plot

1. Differential expression analysis in LIMMA

```
library("limma")
library("edgeR")
library("ggplot2")
library("gplots")
```

```
##
## Attaching package: 'gplots'

## The following object is masked from 'package:stats':
##
##     lowess
```

```
library("pheatmap")
library("RColorBrewer")

### reading the expression dataset (the raw counts)
### the samples 060215 were not included here as they did have lower read counts (only 12 mil reads)

eset <- read.delim("mm10.strand.both.count.exons.condense.genes.NOADJ.with.gene.names.simple.no060215",
                   row.names="Symbol")

### reading the library sizes

### libsize <- c(28182007, 28599704, 26386209, 28069058, 26361374, 27108450,
###              27097918, 12155292, 27124214, 26774125, 26241068, 26296757)

### setting up the groups and the subjects

group <- factor(c("csc","csc","csc","csc","csc","non","non","non","non","non"))
subject <- factor(c(1,2,3,4,5,1,2,3,4,5))

### setting up the design and the contrast matrix

design <- model.matrix(~0+group+subject)
contrast.matrix <- makeContrasts(groupcsc-groupnon, levels=design)
```

```
### filtering the genes based on CPM :

y <- DGEList(counts=eset,group=group)

### keep <- rowSums(cpm(y, lib.size=libsize)>1) >= 3

keep <- rowSums(cpm(y)>0.5) >= 6
y <- y[keep,]

y$samples$lib.size <- colSums(y$counts)
########## we can use y$counts for PCA analysis

### computing the normalization factors :

y <- calcNormFactors(y)

### using the VOOM transformation :

v <- voom(y,design,plot=FALSE)

### the LINEAR FIT in LIMMA :

fit <- lmFit(v, design)

fit2 <- contrasts.fit(fit, contrast.matrix)
fit2 <- eBayes(fit2)

### obtaining and writing the results :

results_limma <- topTable(fit2, coef=1, adjust="fdr", number=Inf)

write.table(results_limma, file="mm10.strand.both.count.exons.condense.genes.NOADJ.with.gene.names.simpl
                           sep="\t", eol="\n", row.names=TRUE, col.names=TRUE)
```

2. TABLES and DATAFRAMES containing DEG genes according to FDR, FC and FPKM :

Here we combine the following tables (these have been post-processed in order to have a matrix of GENE NAMES and EXPRESSION FPKM values :

- **the TABLE with the FPKM values of gene expression**
- **the TABLE of DEG (from limma, as described above)**

```
results_limma$Symbol <- rownames(results_limma)

fpkm_results <- read.delim("mm10.strand.both.count.exons.condense.genes.FPKM.with.gene.names.complex.no
                           row.names="Symbol")
fpkm_results_name <- basename("mm10.strand.both.count.exons.condense.genes.FPKM.with.gene.names.complex

fpkm_results$Symbol <- rownames(fpkm_results)

fpkm_and_limma <- merge(fpkm_results, results_limma, by="Symbol", all=TRUE)
fpkm_and_limma_file <- paste(fpkm_results_name,".with.limma.results.txt", sep="")
```

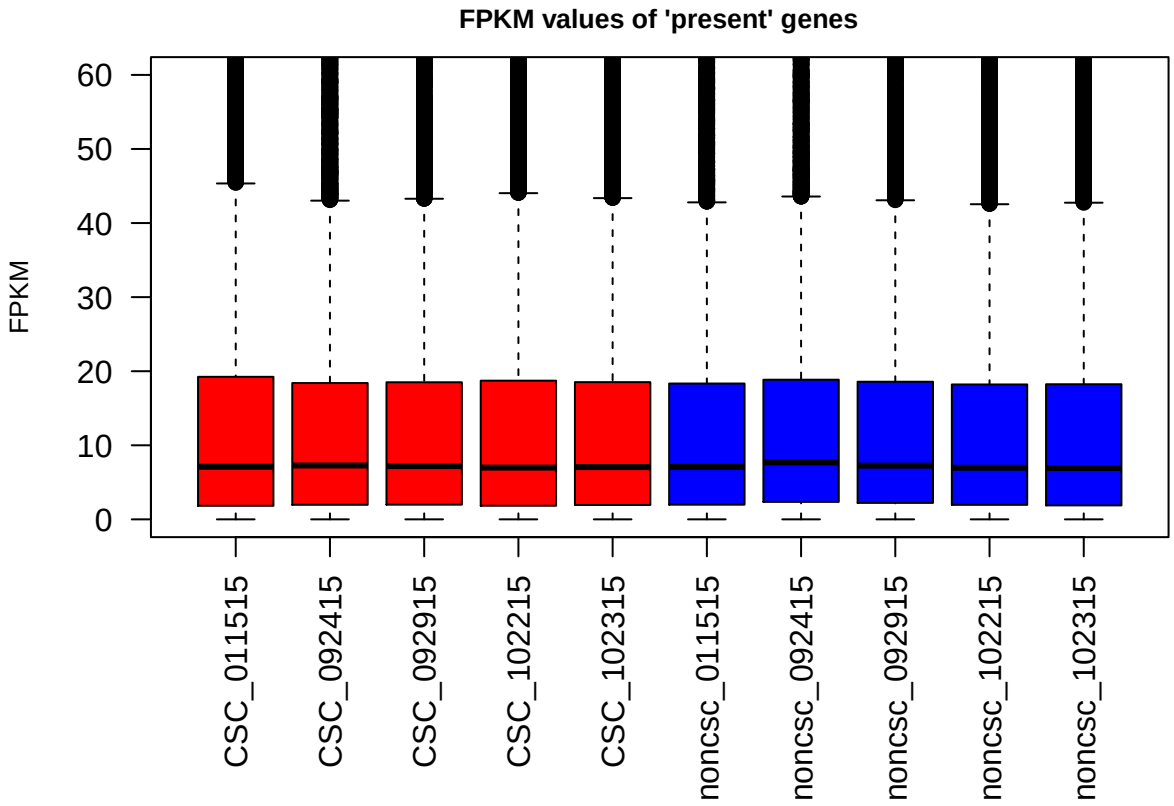
```
write.table(fpkm_and_limma, file=fpkm_and_limma_file, sep="\t", eol="\n",
                                          row.names=FALSE, col.names=TRUE)
```

```
################################################################################
################## here we do add the average of FPKM values :
################################################################################

fpkm_and_limma$average_FPKM_CSC <- rowMeans(subset(fpkm_and_limma,
                                          select = c(CSC_011515,
                                                     CSC_092415,
                                                     CSC_092915,
                                                     CSC_102215,
                                                     CSC_102315)),
                                          na.rm = TRUE)


fpkm_and_limma$average_FPKM_noncsc <- rowMeans(subset(fpkm_and_limma,
                                             select = c(noncsc_011515,
                                                        noncsc_092415,
                                                        noncsc_092915,
                                                        noncsc_102215,
                                                        noncsc_102315)),
                                          na.rm = TRUE)


fpkm_and_limma$average_FPKM_all_samples <- rowMeans(subset(fpkm_and_limma,
                                                  select = c(CSC_011515,
                                                             CSC_092415,
                                                             CSC_092915,
                                                             CSC_102215,
                                                             noncsc_011515,
                                                             noncsc_092415,
                                                             noncsc_092915,
                                                             noncsc_102215,
                                                             noncsc_102315)),
                                          na.rm = TRUE)


fpkm_and_limma$average_FPKM_all_samples_log2 <- log2(fpkm_and_limma$average_FPKM_all_samples)
fpkm_and_limma$average_FPKM_all_samples_log10 <- log10(fpkm_and_limma$average_FPKM_all_samples)

### head(fpkm_and_limma)

################################################################################
################## computing the set of regulated genes for a FDR < 0.1 :
################################################################################

fpkm_and_limma_FDR01 <- subset(fpkm_and_limma, (adj.P.Val < 0.1) )


################################################################################
################## computing the UP- and DOWN- regulated genes for a FDR < 0.1;
################## in order to extract the genes with FDR < 0.1 and FC > 1.2 :
```

```
#############################################################################
################### selecting the genes with FDR < 0.1, FC > 1.2, and FPKM > 1
#############################################################################


fpkm_and_limma_upreg_FDR01_FC12_FPKM1 <- subset(fpkm_and_limma, (adj.P.Val < 0.1) &
                                                                (logFC > 0.264) &
                                                                (average_FPKM_CSC > 1) )

fpkm_and_limma_downreg_FDR01_FC12_FPKM1 <- subset(fpkm_and_limma, (adj.P.Val < 0.1) &
                                                                  (logFC < -0.264) &
                                                                  (average_FPKM_noncsc > 1) )


#########################################################################################################
################### selecting the genes with FDR < 0.1, FC > 1.5, and FPKM > 1 :
#########################################################################################################


fpkm_and_limma_upreg_FDR01_FC15_FPKM1 <- subset(fpkm_and_limma, (adj.P.Val < 0.1) &
                                                               (logFC > 0.585) &
                                                          (average_FPKM_CSC > 1) )

fpkm_and_limma_downreg_FDR01_FC15_FPKM1 <- subset(fpkm_and_limma, (adj.P.Val < 0.1) &
                                                                (logFC < -0.585) &
                                                          (average_FPKM_noncsc > 1) )
```

3. BOXPLOTS of all present genes and DE genes :

```
#### Here we have started from the dataframe "fpkm_and_limma" and select only the genes
#### with no NA columns (ie these genes were used in limma)

fpkm_and_limma_expression <- subset(fpkm_and_limma, logFC!="NA" ,
                                    select = c(CSC_011515,
                                               CSC_092415,
                                               CSC_092915,
                                               CSC_102215,
                                               CSC_102315,
                                               noncsc_011515,
                                               noncsc_092415,
                                               noncsc_092915,
                                               noncsc_102215,
                                               noncsc_102315), na.rm = TRUE)


### making the BOXPLOTS for the genes that have been kept for LIMMA based on CPM values.

par(las=2)
par(mar=c(8,4,2,2))

boxplot(fpkm_and_limma_expression, ylim=c(0,60), col=c(rep("red",5), rep("blue",5)),
                                   ylab="FPKM", main="FPKM values of 'present' genes",
                                   cex.main=0.8, cex.lab=0.8)
```

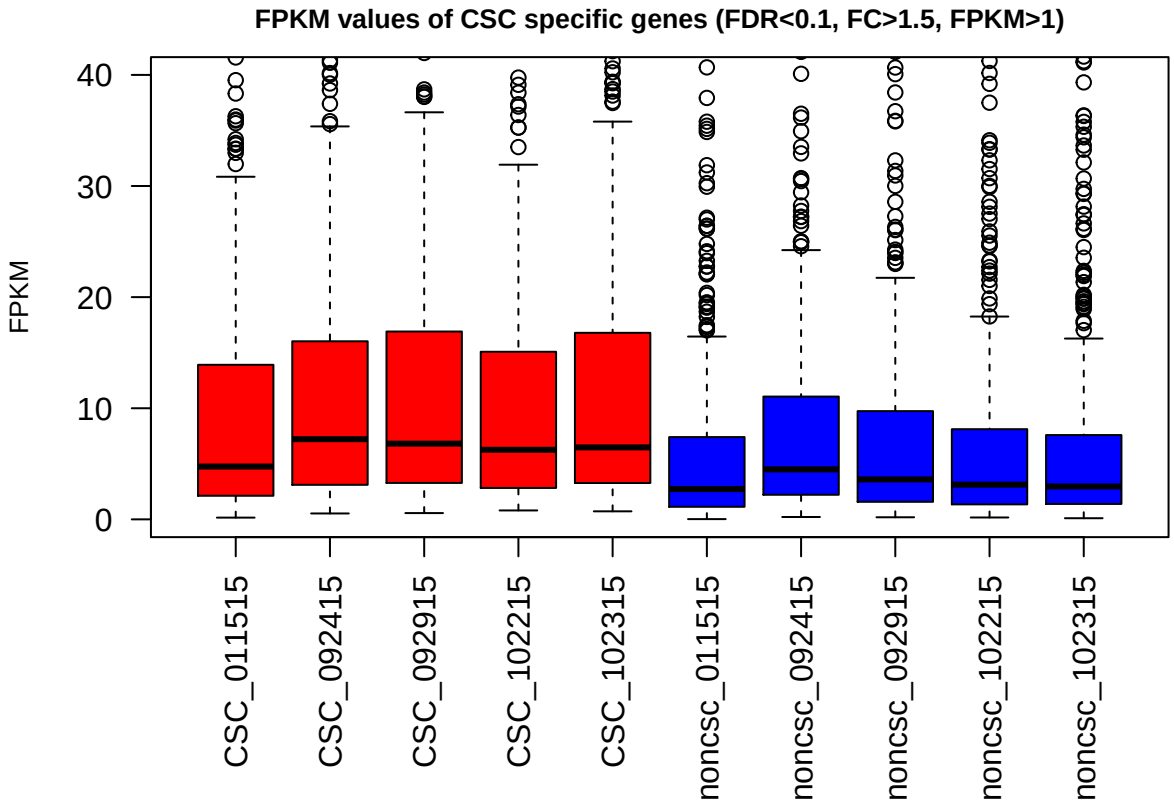



```
pdf("display.boxplots.FPKM values of the genes for LIMMA.pdf")
par(las=2)
par(mar=c(8,4,2,2))
boxplot(fpkm_and_limma_expression, ylim=c(0,60), col=c(rep("red",5), rep("blue",5)),
                                   ylab="FPKM", main="FPKM values of 'present' genes",
                                   cex.main=0.8, cex.lab=0.8)
dev.off()
```

```
## pdf
##   2
```

```
# pca <- prcomp(t(fpkm_and_limma_expression))
# plot(pca$x[,1], pca$x[,2], xlab="PCA1", ylab="PCA2")
# text(pca$x[,1], pca$x[,2], labels=row.names(pca$x), pos=1)


# pca <- prcomp(t(fpkm_and_limma_expression))
# plot(pca$x[,1], pca$x[,2], xlab="PCA1", ylab="PCA2")
# text(pca$x[,1], pca$x[,2], labels=row.names(pca$x), pos=1)

#### Here we have started from the dataframe "fpkm_and_limma_FDR01" and select the genes with FDR < 0.1

fpkm_and_limma_FDR01_expression <- subset(fpkm_and_limma_FDR01,
                                          select = c(CSC_011515,
                                                     CSC_092415,
                                                     CSC_092915,
                                                     CSC_102215,
```

```
                                                    CSC_102315,
                                                    noncsc_011515,
                                                    noncsc_092415,
                                                    noncsc_092915,
                                                    noncsc_102215,
                                                    noncsc_102315), na.rm = TRUE)



# boxplot(fpkm_and_limma_FDR01_expression,ylim=c(0,100),
#         col=c(rep("red",5), rep("blue",5)), ylab="FPKM",
#         main="FPKM values of DEG (FDR<0.1)")

# pdf("display.boxplots.FPKM values of DEG.pdf")
# boxplot(fpkm_and_limma_FDR01_expression,ylim=c(0,100),
#         col=c(rep("red",5), rep("blue",5)), ylab="FPKM",
#         main="FPKM values of DEG (FDR<0.1)")
# dev.off()

#### Or we can plot only the DEG genes :

fpkm_and_limma_upreg_FDR01_FC15_FPKM1_expression <- subset(fpkm_and_limma_upreg_FDR01_FC15_FPKM1,
                                                           select = c(CSC_011515,
                                                                      CSC_092415,
                                                                      CSC_092915,
                                                                      CSC_102215,
                                                                      CSC_102315,
                                                                      noncsc_011515,
                                                                      noncsc_092415,
                                                                      noncsc_092915,
                                                                      noncsc_102215,
                                                                      noncsc_102315), na.rm = TRUE)

par(las=2)
par(mar=c(8,4,2,2))
boxplot(fpkm_and_limma_upreg_FDR01_FC15_FPKM1_expression , ylim=c(0,40), col=c(rep("red",5), rep("blue"
                                    ylab="FPKM", main="FPKM values of CSC specific genes (FDR<0.1, FC>1.5
                                    cex.main=0.8, cex.lab=0.8)
```

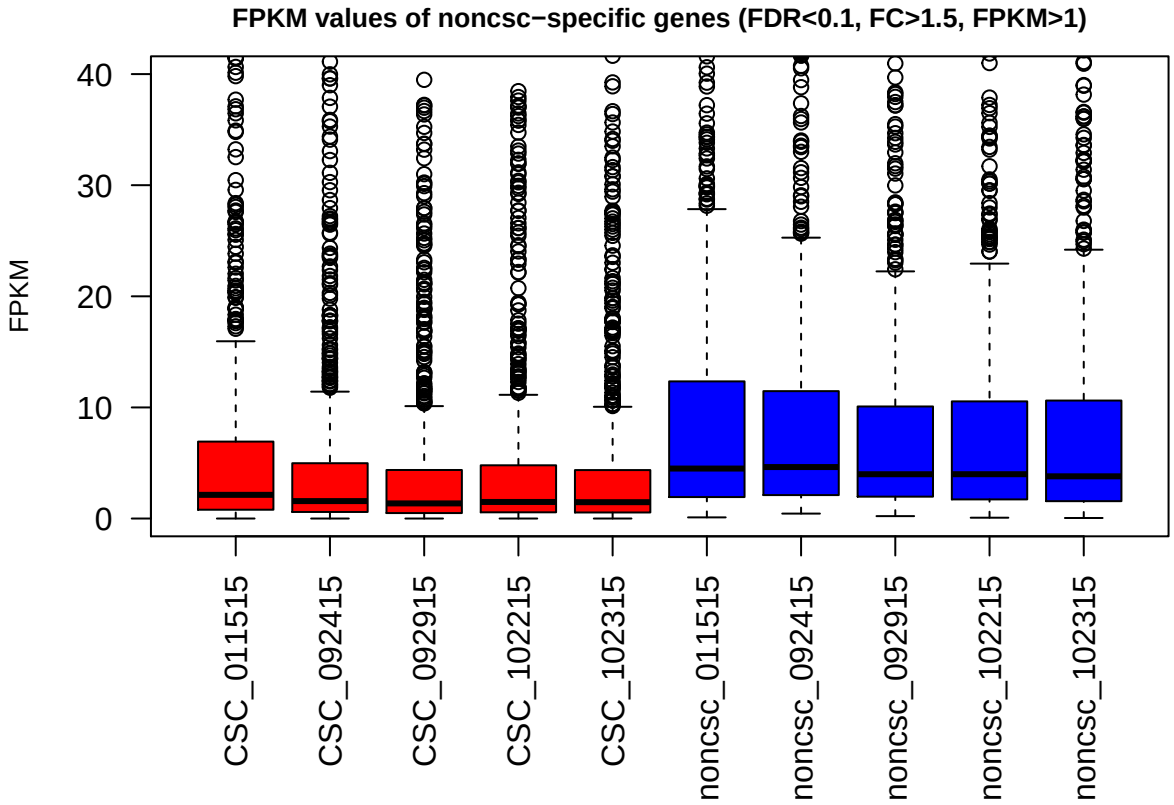



```
pdf("display.boxplots.FPKM values of CSC-specific genes (FDR<0.1, FC>1.5, FPKM>1).pdf")
par(las=2)
par(mar=c(8,4,2,2))
boxplot(fpkm_and_limma_upreg_FDR01_FC15_FPKM1_expression , ylim=c(0,40), col=c(rep("red",5), rep("blue"
                                  ylab="FPKM", main="FPKM values of CSC-specific genes (FDR<0.1, FC>1.5
                                  cex.main=0.8, cex.lab=0.8)
dev.off()
```

```
## pdf
##   2
```

```
fpkm_and_limma_downreg_FDR01_FC15_FPKM1_expression <- subset(fpkm_and_limma_downreg_FDR01_FC15_FPKM1,
                                                             select = c(CSC_011515,
                                                                        CSC_092415,
                                                                        CSC_092915,
                                                                        CSC_102215,
                                                                        CSC_102315,
                                                                        noncsc_011515,
                                                                        noncsc_092415,
                                                                        noncsc_092915,
                                                                        noncsc_102215,
                                                                        noncsc_102315), na.rm = TRUE)



par(las=2)
```

```
par(mar=c(8,4,2,2))
boxplot(fpkm_and_limma_downreg_FDR01_FC15_FPKM1_expression , ylim=c(0,40), col=c(rep("red",5), rep("blue
                                    ylab="FPKM", main="FPKM values of noncsc-specific genes (FDR<0.1, FC
                                    cex.main=0.8, cex.lab=0.8)
```




```
pdf("display.boxplots.FPKM values of noncsc-specific genes (FDR<0.1, FC>1.5, FPKM>1).pdf")
par(las=2)
par(mar=c(8,4,2,2))
boxplot(fpkm_and_limma_downreg_FDR01_FC15_FPKM1_expression , ylim=c(0,40), col=c(rep("red",5), rep("blue
                                    ylab="FPKM", main="FPKM values of noncsc-specific genes (FDR<0.1, FC
                                    cex.main=0.8, cex.lab=0.8)
dev.off()
```

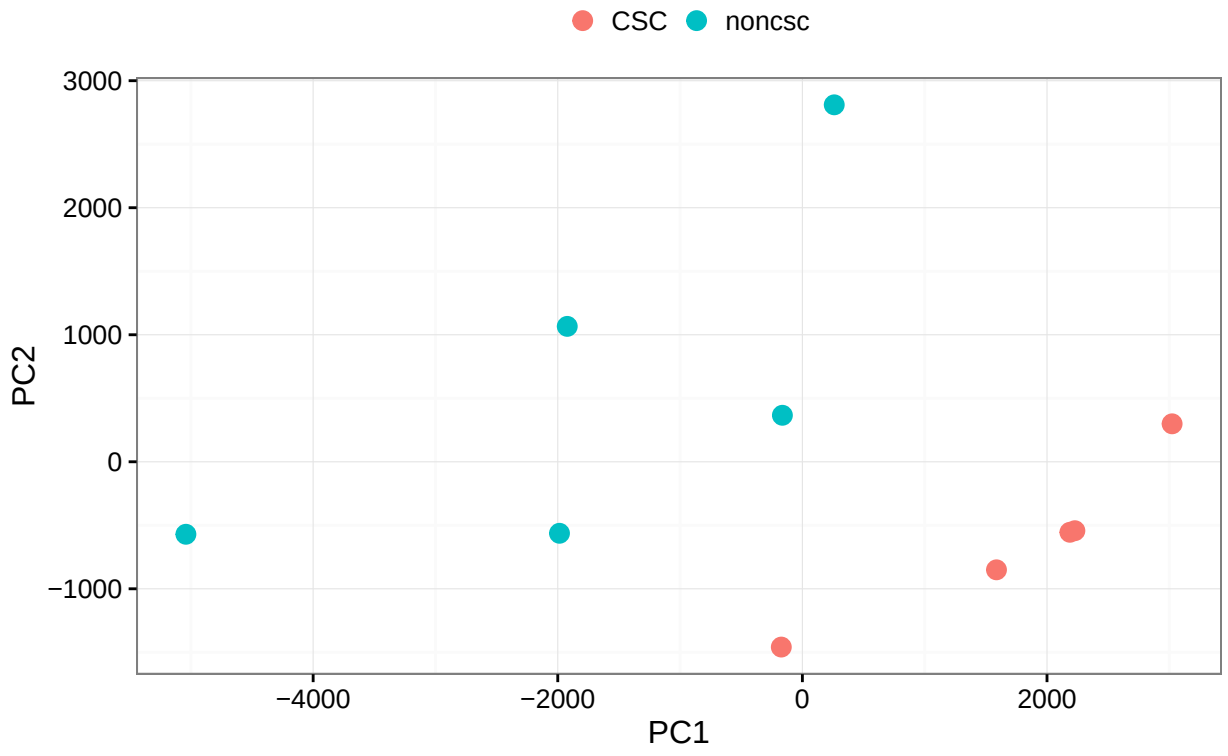
```
## pdf
##   2
```

```
#################################################################
################################################################# here, showing the principal components :
```

4. PCA analysis :

```
## In order to set up the color palette :
## col.rainbow <- rainbow(12)
## col.topo <- topo.colors(12)
## col.terrain <- terrain.colors(12)
## palette(col.rainbow)
```

```
# plot(pca$x[,1], pca$x[,2], xlab="PCA1", ylab="PCA2", pch=20, col=c(rep("red",5),rep("blue",5)))
# legend('topleft', legend = c("CSC","noncsc"), col=c("red","blue"))
# text(pca$x[,1], pca$x[,2], labels=row.names(pca$x), pos=1)

#########################################################################################################
## Here we are plotting the genes with FDR < 0.1, starting from the dataframe :
## fpkm_and_limma_FDR01_expression :

group <- factor(c(rep("CSC",5),rep("noncsc",5)))

pca <- prcomp(t(fpkm_and_limma_FDR01_expression))

pca.df <- data.frame(PCA1=pca$x[,1], PCA2=pca$x[,2], PCA3=pca$x[,3], group=group)

## Here we are plotting PCA1 vs PCA2

ggplot(pca.df, aes(x=PCA1, y=PCA2, color=group, label=rownames(pca.df))) +
       geom_point(size=3) +
     # geom_text(col='black', size=4) +
       theme_bw() +
       theme(legend.position="top", legend.title=element_blank(), legend.key = element_blank()) +
       labs(x="PC1", y="PC2") + ggtitle("PCA analysis of DEG : PC1 vs PC2")
```

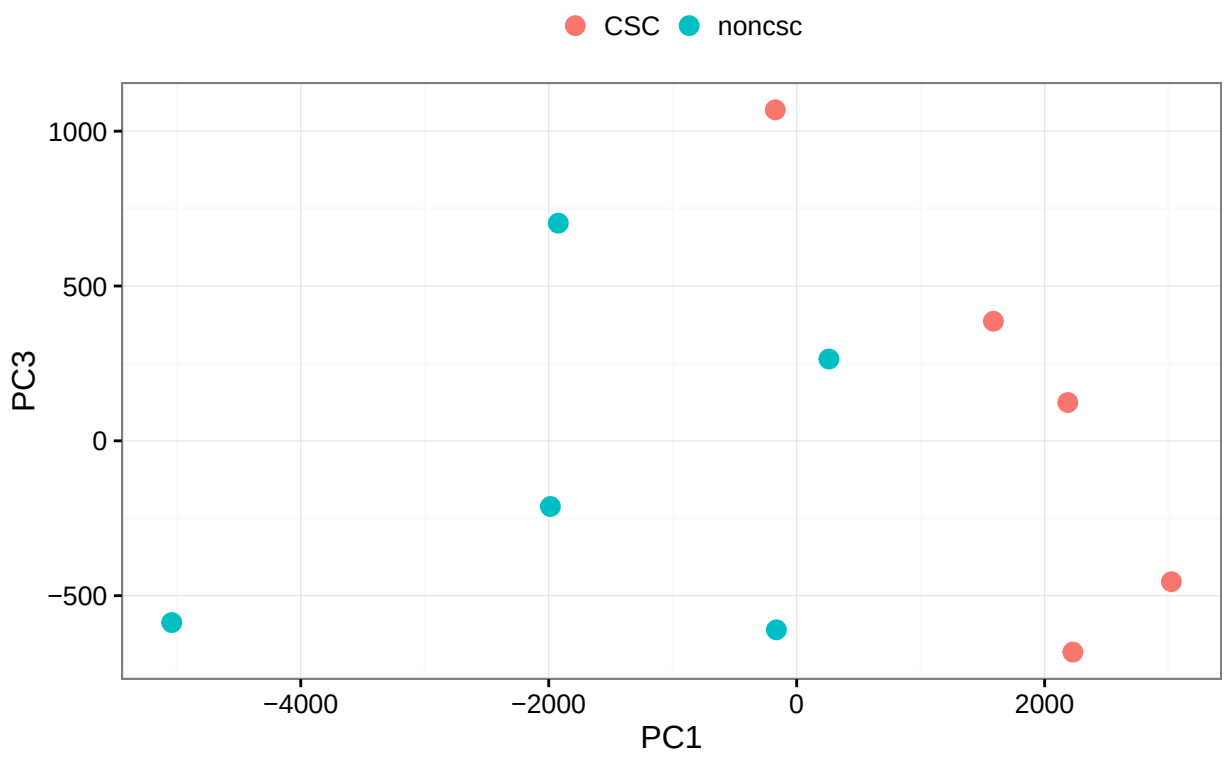



```
## in order to save it in a pdf file :

pdf("display.PCA.PC1-vs-PC2.pdf")
```

```
ggplot(pca.df, aes(x=PCA1, y=PCA2, color=group, label=rownames(pca.df))) +
       geom_point(size=3) +
     # geom_text(col='black', size=4) +
       theme_bw() +
       theme(legend.position="top", legend.title=element_blank(), legend.key = element_blank()) +
       labs(x="PC1", y="PC2") + ggtitle("PCA analysis of DEG : PC1 vs PC2")
dev.off()
```

```
## pdf
##   2
```

```
## Here we are plotting PCA1 vs PCA3

ggplot(pca.df, aes(x=PCA1, y=PCA3, color=group, label=rownames(pca.df))) +
       geom_point(size=3) +
     # geom_text(col='black', size=4) +
       theme_bw() +
       theme(legend.position="top", legend.title=element_blank(), legend.key = element_blank()) +
       labs(x="PC1", y="PC3") + ggtitle("PCA analysis of DEG : PCA1 vs PCA3")
```

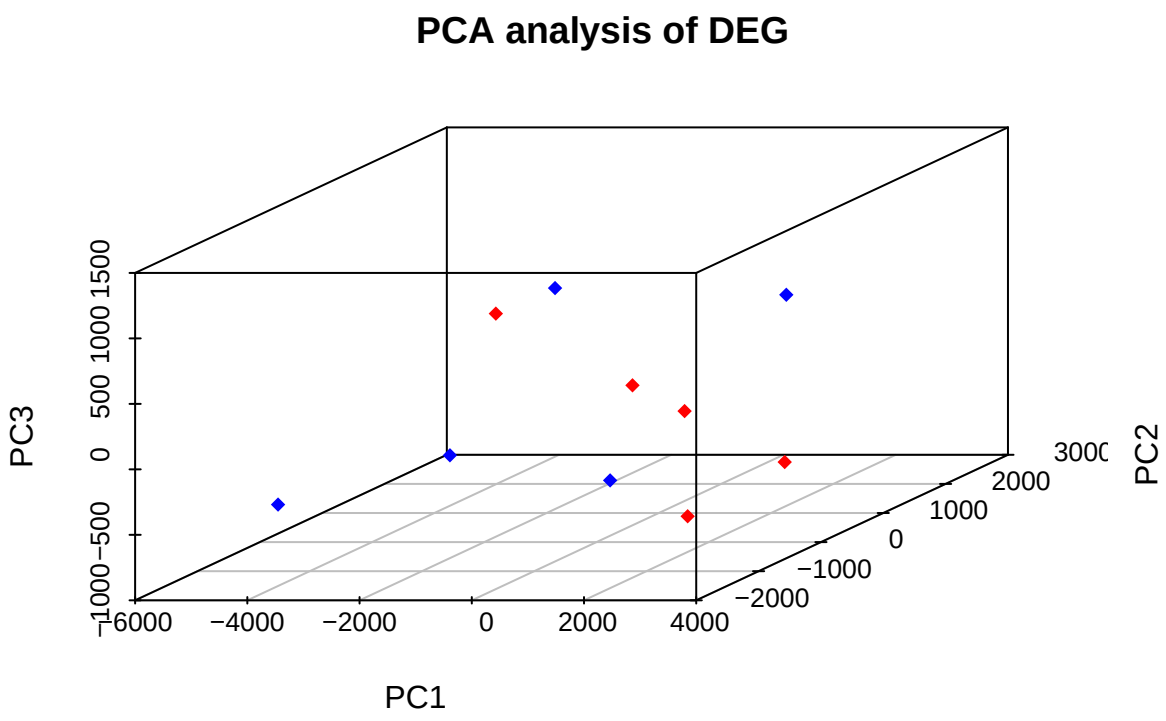



```
## in order to save it in a pdf file :

pdf("display.PCA.PC1-vs-PC3.pdf")
ggplot(pca.df, aes(x=PCA1, y=PCA3, color=group, label=rownames(pca.df))) +
       geom_point(size=3) +
```

```
     # geom_text(col='black', size=4) +
       theme_bw() +
       theme(legend.position="top", legend.title=element_blank(), legend.key = element_blank()) +
       labs(x="PC1", y="PC3") + ggtitle("PCA analysis of DEG : PCA1 vs PCA3")
dev.off()
```

```
## pdf
##   2
```

```
######################################################################
### We display here a 3D plot :
######################################################################

library(scatterplot3d)

scatterplot3d(pca$x[,1:3],
              color=c(rep("red",5),rep("blue",5)),
              pch=18,
              main="PCA analysis of DEG")
```

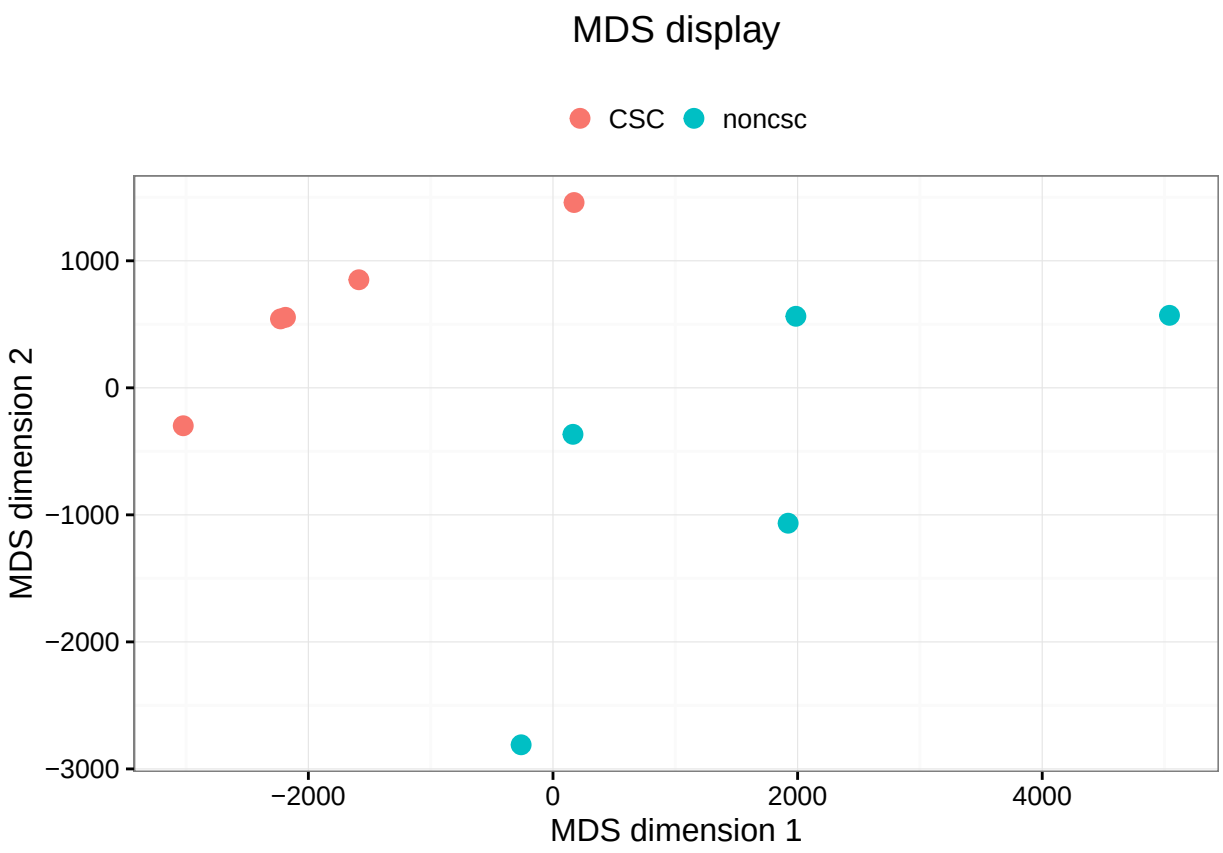



```
pdf("display.PCA.in.3D.pdf")
scatterplot3d(pca$x[,1:3],
              color=c(rep("red",5),rep("blue",5)),
              pch=18,
              main="PCA analysis of DEG")
dev.off()
```

```
## pdf
##   2
```

5. MDS analysis :

```
#############################################################################################################
## Here we are plotting the genes with FDR < 0.1, starting from fpkm_and_limma_FDR01_expression :

group <- factor(c(rep("CSC",5),rep("noncsc",5)))

### We can use the function plotMDS from LIMMA or we can use the function cmdscale :

### mds <- plotMDS(fpkm_and_limma_FDR01_expression)
### mds.df <- data.frame(MDSx=mds$x, MDSy=mds$y, group=group)

mds <- cmdscale(dist(t(fpkm_and_limma_FDR01_expression)))
mds.df <- data.frame(MDSx=mds[,1], MDSy=mds[,2], group=group)

### plot(cmdscale(dist(t(fpkm_and_limma_FDR01_expression))))
### text(cmdscale(dist(t(fpkm_and_limma_FDR01_expression))), labels=colnames(fpkm_and_limma_FDR01_expres

ggplot(mds.df, aes(x=MDSx, y=MDSy, color=group, label=rownames(mds.df))) +
        geom_point(size=3) +
        # geom_text(col='black', size=4) +
        theme_bw() +
        theme(legend.position="top", legend.title=element_blank(), legend.key = element_blank()) +
        labs(x="MDS dimension 1", y="MDS dimension 2") +
        ggtitle("MDS display")
```



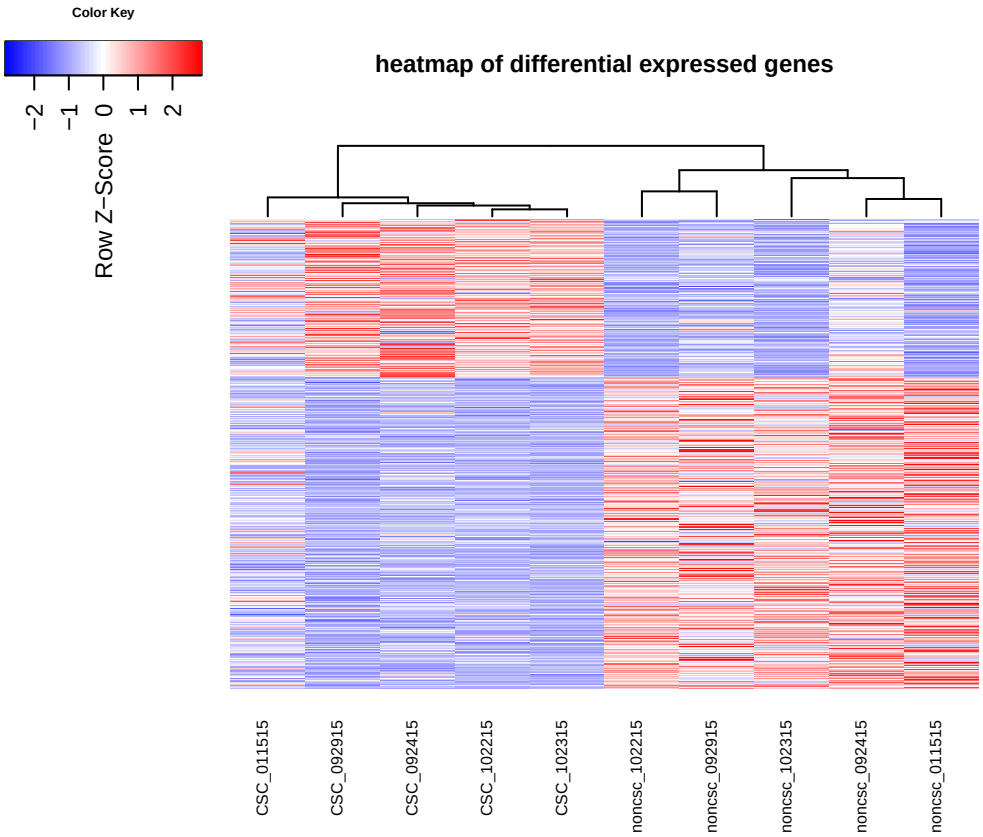
```
pdf("display.MDS.pdf")
ggplot(mds.df, aes(x=MDSx, y=MDSy, color=group, label=rownames(mds.df))) +
        geom_point(size=3) +
        # geom_text(col='black', size=4) +
        theme_bw() +
        theme(legend.position="top", legend.title=element_blank(), legend.key = element_blank()) +
        labs(x="MDS dimension1", y="MDS dimension2") +
        ggtitle("MDS plot")
dev.off()
```

```
## pdf
##   2
```

6. Displaying the heatmaps of the DE genes :

```
##############################################################################################
####### putting together the dataframes for UP- and DOWN-regulated genes (for HEATMAPS)
##############################################################################################

fpkm_and_limma_reg_FDR01_FC15_FPKM1_expression <- rbind(fpkm_and_limma_upreg_FDR01_FC15_FPKM1_expressio
                                                        fpkm_and_limma_downreg_FDR01_FC15_FPKM1_expressi


par(las=2)
par(mar=c(8,4,2,2))
par(cex.main=0.6)
heatmap.2(as.matrix(fpkm_and_limma_reg_FDR01_FC15_FPKM1_expression), col=bluered(149),
          scale="row",trace="none",
          cexRow=0.6, cexCol=0.6, cex.main=0.6,
          Rowv=FALSE, symkey=FALSE, labRow=NA,
          key=T, keysize=1.5, density.info="none",
          main="heatmap of differential expressed genes")
```

```
## Warning in heatmap.
## 2(as.matrix(fpkm_and_limma_reg_FDR01_FC15_FPKM1_expression), : Discrepancy:
## Rowv is FALSE, while dendrogram is `both'. Omitting row dendogram.
```




```
pdf("display.heatmap.regulated.genes.FDR01.FC15.FPKM1.pdf")
par(las=2)
par(mar=c(8,4,2,2))
par(cex.main=0.6)
heatmap.2(as.matrix(fpkm_and_limma_reg_FDR01_FC15_FPKM1_expression), col=bluered(149),
          scale="row",trace="none",
          cexRow=0.6, cexCol=0.6, cex.main=0.6,
          Rowv=FALSE, symkey=FALSE, labRow=NA,
          key=T, keysize=1.5, density.info="none",
          main="heatmap of differential expressed genes")
```

```
## Warning in heatmap.
## 2(as.matrix(fpkm_and_limma_reg_FDR01_FC15_FPKM1_expression), : Discrepancy:
## Rowv is FALSE, while dendrogram is `both'. Omitting row dendogram.
```

```
dev.off()
```

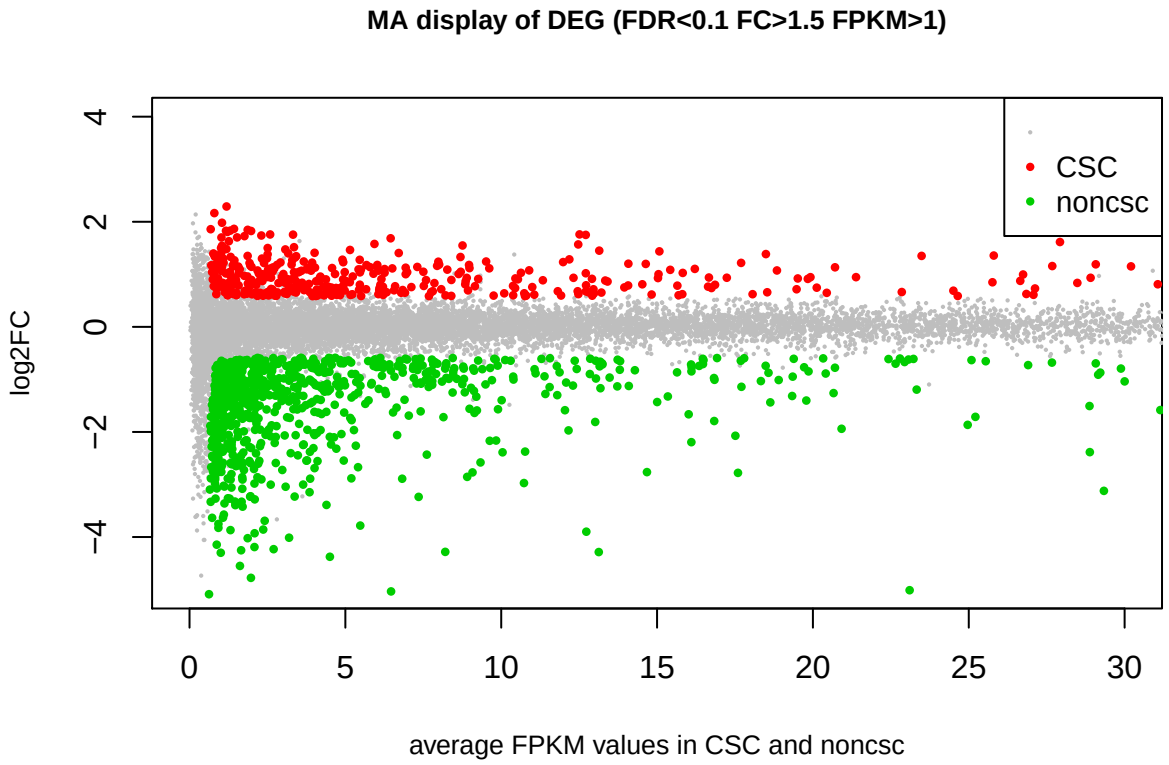
```
## pdf
##   2
```

An alternative way to visualize the heatmaps is by using MeV : **http://www.tm4.org/mev.html**

7. Displaying the MA plots :

```
#######################################################################
## Using the dataframe in order to assign "U" and "D" genes ----
## starting from the dataframe where we have computed the average
## in order to assign up- and down-regulation based on a FC of 1.2
#######################################################################

fpkm_and_limma$regulation_FDR01_FC12_FPKM1 <- ''

fpkm_and_limma$regulation_FDR01_FC12_FPKM1[ ( (fpkm_and_limma$adj.P.Val < 0.1) &
                                              (fpkm_and_limma$logFC >  0.264) &
                                              (fpkm_and_limma$average_FPKM_CSC > 1) ) ] <- "CSC"

fpkm_and_limma$regulation_FDR01_FC12_FPKM1[ ( (fpkm_and_limma$adj.P.Val < 0.1)
                                               & (fpkm_and_limma$logFC < -0.264)
                                               & (fpkm_and_limma$average_FPKM_noncsc > 1) ) ] <- "noncs


### for verification :
### write.table(fpkm_and_limma, file="results_fpkm_and_limma_regulation_FDR01_logFC12_FPKM1_with_U_and_
###             sep="\t", eol="\n", row.names=FALSE, col.names=TRUE)


#############################################################################
####### Using the dataframe where we have computed the average
####### in order to assign up- and down-regulation based on a FC of 1.5 :
#############################################################################


fpkm_and_limma$regulation_FDR01_FC15_FPKM1 <- ''

fpkm_and_limma$regulation_FDR01_FC15_FPKM1[ ( (fpkm_and_limma$adj.P.Val < 0.1) &
                                              (fpkm_and_limma$logFC >  0.585) &
                                              (fpkm_and_limma$average_FPKM_CSC > 1) ) ] <- "CSC"

fpkm_and_limma$regulation_FDR01_FC15_FPKM1[ ( (fpkm_and_limma$adj.P.Val < 0.1) &
                                              (fpkm_and_limma$logFC < -0.585) &
                                              (fpkm_and_limma$average_FPKM_noncsc > 1) ) ] <- "noncsc"

### for verification :
### write.table(fpkm_and_limma, file="results_fpkm_and_limma_regulation_FDR01_logFC15_FPKM1_with_U_and_
###                             sep="\t", eol="\n", row.names=FALSE, col.names=TRUE)
```

```
###  here the MA PLOTS : the FOLD CHANGE is on log2 scale :

plotWithHighlights(fpkm_and_limma$average_FPKM_all_samples, fpkm_and_limma$logFC,
                   status=fpkm_and_limma$regulation_FDR01_FC15_FPKM1, values=c("CSC","noncsc"),
                   bg.col="grey", xlim=c(0,30), ylim=c(-5,4), hl.cex=0.6, cex.main=0.8,  cex.lab =0.8,
                   xlab="average FPKM values in CSC and noncsc", ylab="log2FC", legend= "topright",
                   main="MA display of DEG (FDR<0.1 FC>1.5 FPKM>1)" )
```



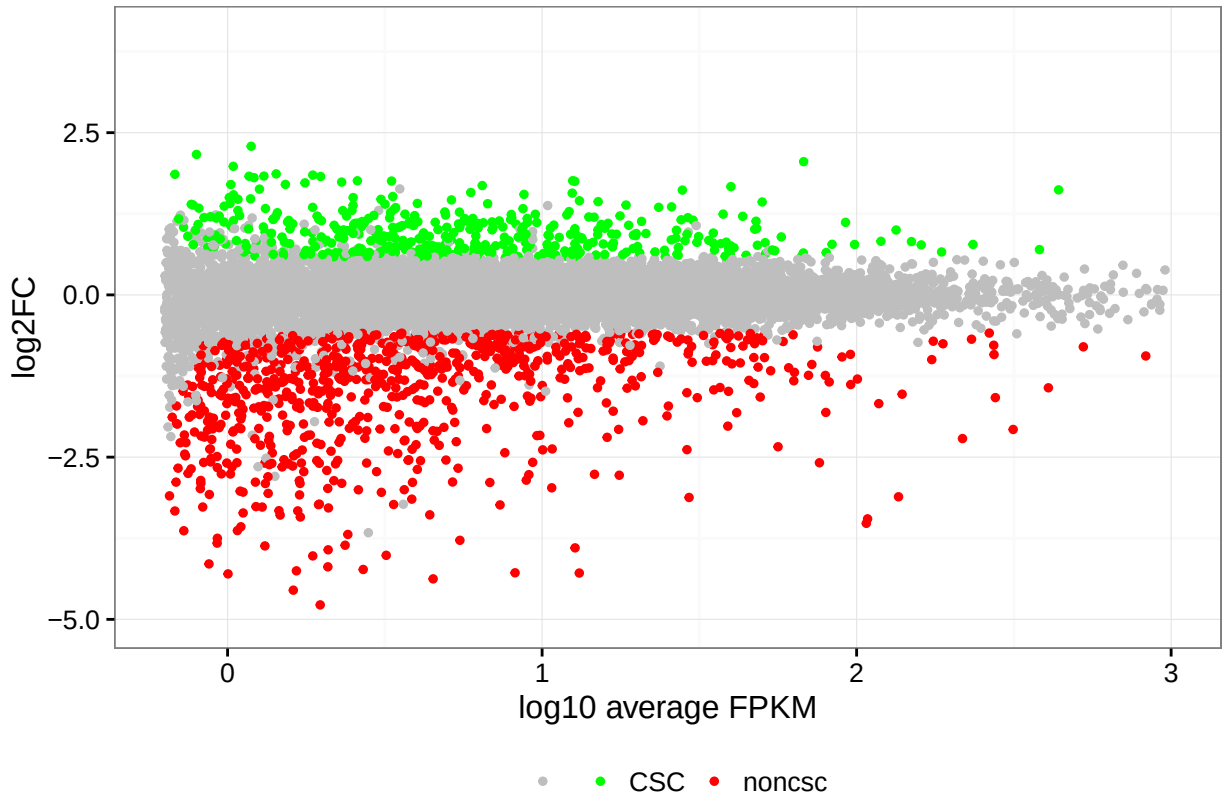

```
pdf("display.MA plot of the DEG (FDR<0.1 FC>1.5 FPKM>1).with.limma.pdf")
plotWithHighlights(fpkm_and_limma$average_FPKM_all_samples, fpkm_and_limma$logFC,
                   status=fpkm_and_limma$regulation_FDR01_FC15_FPKM1, values=c("CSC","noncsc"),
                   bg.col="grey", xlim=c(0,30), ylim=c(-5,4), hl.cex=0.6, cex.main=0.8,  cex.lab =0.8,
                   xlab="average FPKM values in CSC and noncsc", ylab="log2FC", legend= "topright",
                   main="MA display of DEG (FDR<0.1 FC>1.5 FPKM>1)" )
dev.off()
```

```
## pdf
##   2
```

```
#################################################################
#### here, doing the MA PLOTS : the FOLD CHANGE is on LOG2 scale but also a LOG10 AVERAGE FPKM :
#################################################################

#plotWithHighlights(fpkm_and_limma$average_FPKM_all_samples_log10, fpkm_and_limma$logFC,
#                   status=fpkm_and_limma$regulation_FDR01_FC15_FPKM1, c("CSC","noncsc"),
#                   bg.col="grey", xlim=c(-0.2,3), ylim=c(-5,4), hl.cex=0.6, cex.main=0.8,  cex.lab =0.
#                   xlab="log10 average FPKM values in CSC and noncsc", ylab="log2FC", legend= "toprigh
#                   main="MA plot of the DEG (FDR<0.1 FC>1.5 FPKM>1)" )

#pdf("display.MA plot of the DEG (FDR<0.1 FC>1.5 FPKM>1).log10.pdf")
#plotWithHighlights(fpkm_and_limma$average_FPKM_all_samples_log10, fpkm_and_limma$logFC,
#                   status=fpkm_and_limma$regulation_FDR01_FC15_FPKM1, c("CSC","noncsc"),
#                   bg.col="grey", xlim=c(-0.2,3), ylim=c(-5,4), hl.cex=0.6, cex.main=0.8,  cex.lab =0.
#                   xlab="log10 average FPKM values in CSC and noncsc", ylab="log2FC", legend= "toprigh
#                   main="MA plot of the DEG (FDR<0.1 FC>1.5 FPKM>1)" )
#dev.off()
```

```
##########################################################################################################
######################### here, doing the same MA plots : the FOLD CHANGE is on log2 scale : but using G
##########################################################################################################

ggplot(fpkm_and_limma, aes(x=average_FPKM_all_samples_log10, y=logFC, col=factor(regulation_FDR01_FC15_
       geom_point(size=1) +
       theme_bw() +
       xlim(-0.2, 3) +
       ylim(-5, 4) +
       theme(legend.position="bottom", legend.title=element_blank(), legend.key = element_blank()) +
       scale_colour_manual(values = c("grey","CSC"="green", "noncsc"="red")) +
       labs(x="log10 average FPKM", y="log2FC")
```

```
## Warning: Removed 11277 rows containing missing values (geom_point).
```




```
pdf("display.MA plot of DEG (FDR<0.1 FC>1.5 FPKM>1).with.ggplot2.pdf")
ggplot(fpkm_and_limma, aes(x=average_FPKM_all_samples_log10, y=logFC, col=factor(regulation_FDR01_FC15_
       geom_point(size=1) +
       theme_bw() +
       xlim(-0.2, 3) +
       ylim(-5, 4) +
       theme(legend.position="bottom", legend.title=element_blank(), legend.key = element_blank()) +
       scale_colour_manual(values = c("grey","CSC"="green", "noncsc"="red")) +
       labs(x="log10 average FPKM", y="log2FC")
```

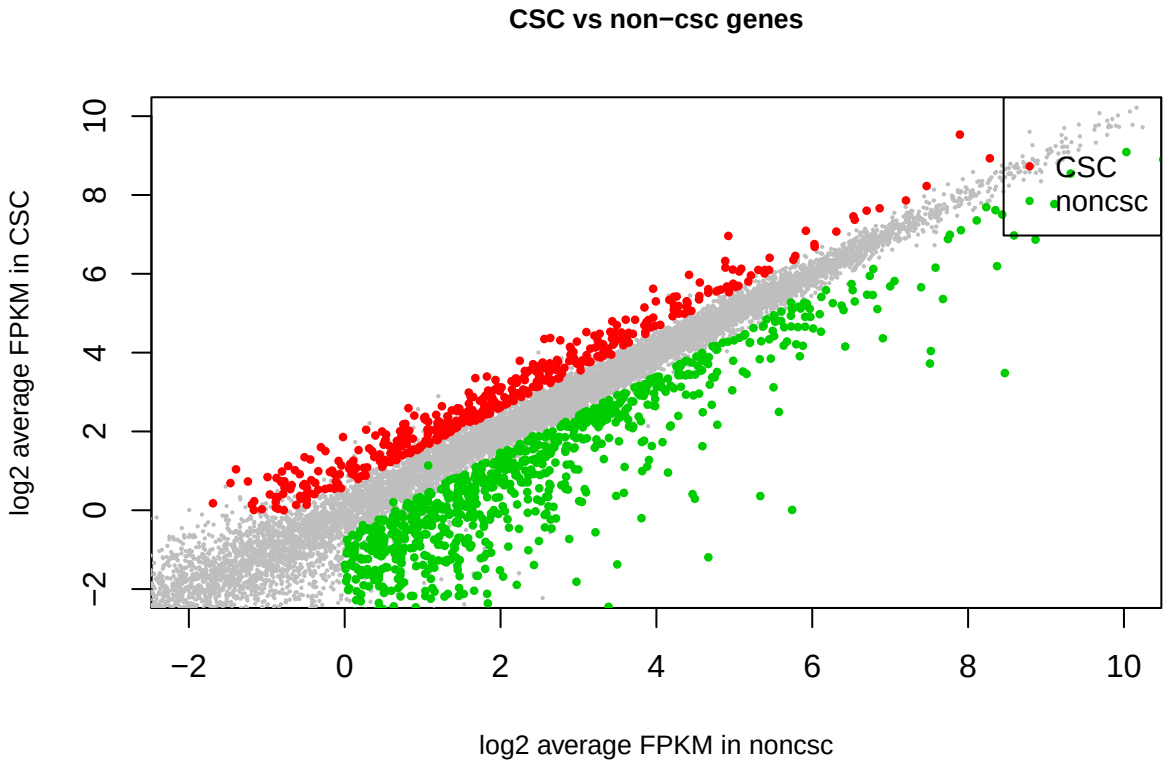
```
## Warning: Removed 11277 rows containing missing values (geom_point).
```

```
dev.off()
```

```
## pdf
##   2
```

8. Displaying the scatter plots :

```
#################################################################
#################### here, doing the SCATTER PLOTS of "U" and "D" genes :
#################################################################


plotWithHighlights(log2(fpkm_and_limma$average_FPKM_noncsc), log2(fpkm_and_limma$average_FPKM_CSC),
                   status=fpkm_and_limma$regulation_FDR01_FC15_FPKM1,
                   values=c("CSC","noncsc"),
                   bg.col="grey",
                   xlim=c(-2,10), ylim=c(-2,10),
                   hl.cex=0.6, cex.main=0.8,  cex.lab =0.8,
                   xlab="log2 average FPKM in noncsc", ylab="log2 average FPKM in CSC", legend= "toprigh
                   main="CSC vs non-csc genes" )
```



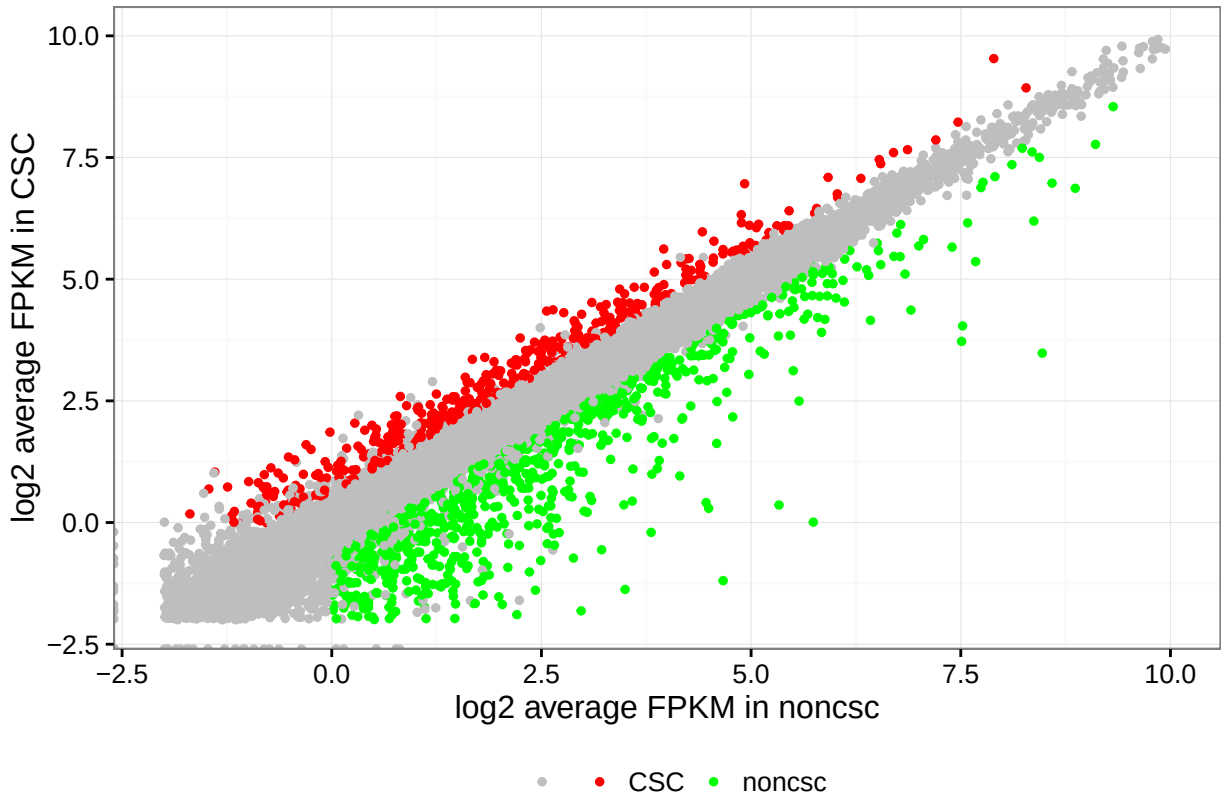

```
pdf("display.scatter plot of the FPKM values on the log2 scale.with.limma.pdf")
plotWithHighlights(log2(fpkm_and_limma$average_FPKM_noncsc), log2(fpkm_and_limma$average_FPKM_CSC),
                   status=fpkm_and_limma$regulation_FDR01_FC15_FPKM1,
                   values=c("CSC","noncsc"),
                   bg.col="grey",
                   xlim=c(-2,10), ylim=c(-2,10),
                   hl.cex=0.6, cex.main=0.8,  cex.lab =0.8,
```

```
                  xlab="log2 average FPKM in noncsc", ylab="log2 average FPKM in CSC", legend= "toprigh
                  main="CSC vs non-csc genes" )
dev.off()
```

```
## pdf
##   2
```

```
ggplot(fpkm_and_limma, aes(x=log2(fpkm_and_limma$average_FPKM_noncsc), y=log2(fpkm_and_limma$average_FPK
       geom_point(size=1) +
       theme_bw() +
       xlim(-2, 10) +
       ylim(-2, 10) +
       theme(legend.position="bottom", legend.title=element_blank(), legend.key = element_blank()) +
       scale_colour_manual(values = c("grey","noncsc"="green", "CSC"="red")) +
       labs(x="log2 average FPKM in noncsc", y="log2 average FPKM in CSC")
```

```
## Warning: Removed 6054 rows containing missing values (geom_point).
```




```
pdf("display.scatter plot of the FPKM values on the log2 scale.with.ggplot2.pdf")
ggplot(fpkm_and_limma, aes(x=log2(fpkm_and_limma$average_FPKM_noncsc), y=log2(fpkm_and_limma$average_FPK
       geom_point(size=1) +
       theme_bw() +
       xlim(-2, 10) +
       ylim(-2, 10) +
       theme(legend.position="bottom", legend.title=element_blank(), legend.key = element_blank()) +
       scale_colour_manual(values = c("grey","noncsc"="green", "CSC"="red")) +
       labs(x="log2 average FPKM in noncsc", y="log2 average FPKM in CSC")
```

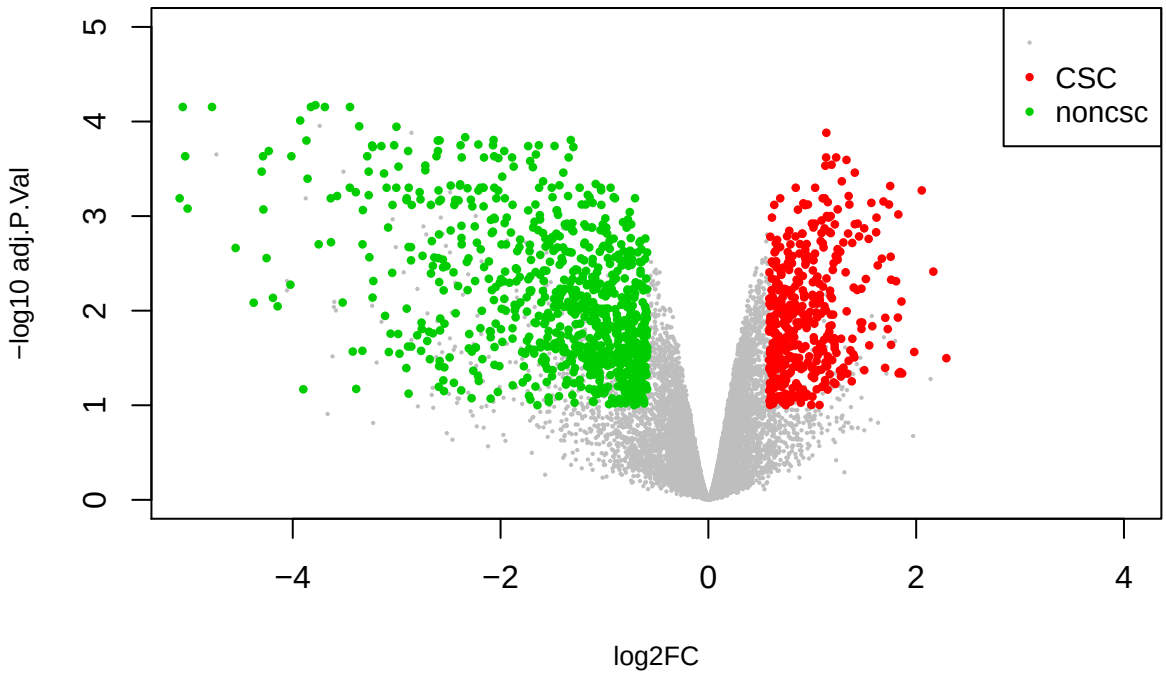
```
## Warning: Removed 6054 rows containing missing values (geom_point).
```

```
dev.off()
```

```
## pdf
##   2
```

9. Displaying the volcano plots :

```
################################################################
######### here, doing the VOLCANO PLOTS (on log2 and on log10) by using plotWithHighlights (in limma) :
################################################################

plotWithHighlights(fpkm_and_limma$logFC, -log10(fpkm_and_limma$adj.P.Val), status=fpkm_and_limma$regula
                   values=c("CSC","noncsc"),
                   bg.col="grey",
                   xlim=c(-5,4), ylim=c(0,5),
                   hl.cex=0.6, cex.main=0.8,  cex.lab =0.8,
                   xlab="log2FC", ylab="-log10 adj.P.Val", legend= "topright",
                   main="" )
```

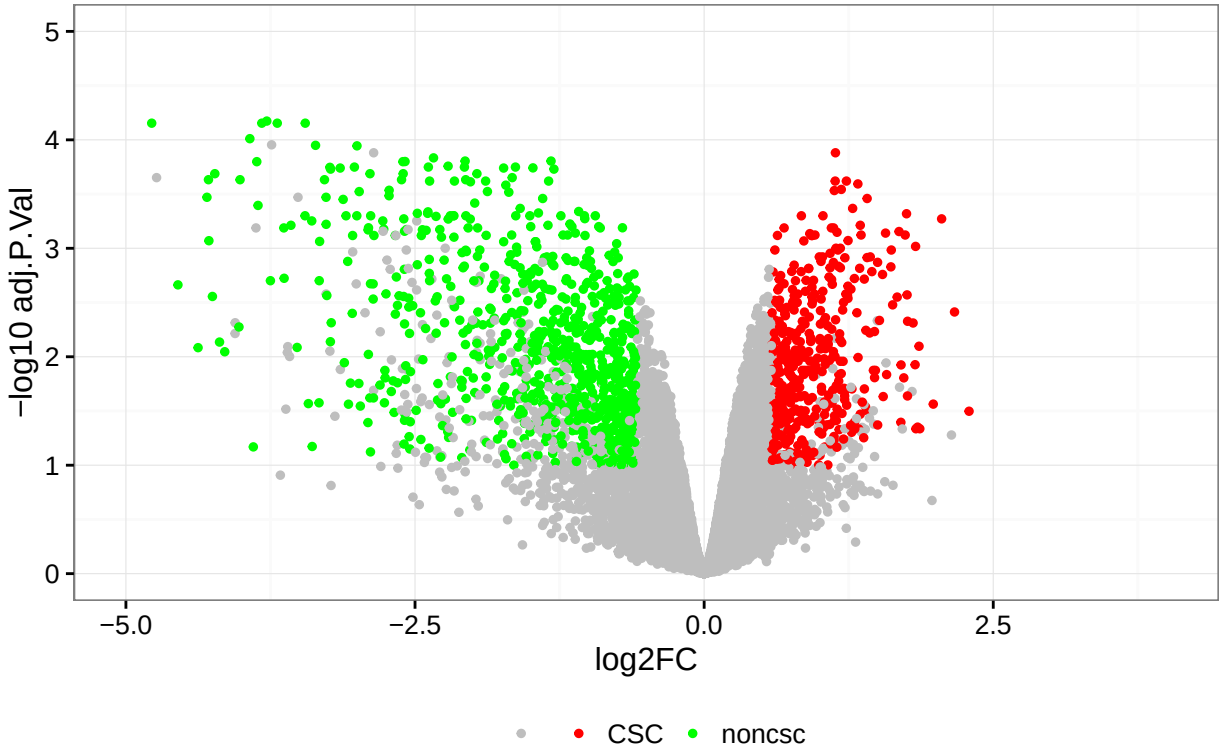



```
pdf("display.Volcano plot of the DEG (FDR<0.1 FC>1.5 FPKM>1) with limma.pdf")
plotWithHighlights(fpkm_and_limma$logFC, -log10(fpkm_and_limma$adj.P.Val), status=fpkm_and_limma$regula
                   values=c("CSC","noncsc"),
                   bg.col="grey",
                   xlim=c(-5,4), ylim=c(0,5),
                   hl.cex=0.6, cex.main=0.8,  cex.lab =0.8,
                   xlab="log2FC", ylab="-log10 adj.P.Val", legend= "topright",
                   main="" )
dev.off()
```

```
## pdf
##   2
```

```
################################################################
######### here, doing the VOLCANO PLOTS (on log2 and on log10) by using gpplot2 :
################################################################

ggplot(fpkm_and_limma, aes(x=logFC, y=-log10(adj.P.Val), col=factor(regulation_FDR01_FC15_FPKM1))) +
       geom_point(size=1) +
       theme_bw() +
       xlim(-5, 4) +
       ylim(0, 5) +
       theme(legend.position="bottom", legend.title=element_blank(), legend.key = element_blank()) +
       scale_colour_manual(values = c("grey","noncsc"="green", "CSC"="red")) +
       labs(x="log2FC", y="-log10 adj.P.Val") + ggtitle("")
```

```
## Warning: Removed 9786 rows containing missing values (geom_point).
```




```
pdf("display.Volcano plot of the DEG (FDR<0.1 FC>1.5 FPKM>1) with ggplot2.pdf")
ggplot(fpkm_and_limma, aes(x=logFC, y=-log10(adj.P.Val), col=factor(regulation_FDR01_FC15_FPKM1))) +
       geom_point(size=1) +
       theme_bw() +
       xlim(-5, 4) +
       ylim(0, 5) +
       theme(legend.position="bottom", legend.title=element_blank(), legend.key = element_blank()) +
       scale_colour_manual(values = c("grey","noncsc"="green", "CSC"="red")) +
       labs(x="log2FC", y="-log10 adj.P.Val") + ggtitle("")
```

```
## Warning: Removed 9786 rows containing missing values (geom_point).
```

```
dev.off()
```

```
## pdf
##   2
```

VI. MOTIF ENRICHMENT on the promoters of the DEG genes

For the motif enrichment, we have used HOMER and the sequences encompassing a region -150,+50 bp around the TSS of the regulated genes. As background for motif enrichment, we have selected :

a. the promoters in the opposite dataset
b. the promoters of the expressed genes
c. the genome regions that have been randomly selected from the mouse mm10 genome

The files describing the results could be found in the DROPBOX.

```
findMotifsGenome.pl file.promoters mm10 file.promoters.MOTIFS -background file.background
```

VII. ADDITIONAL ANALYSES

For GO/pathway analysis, we have used enrichR, ToppFun, consensusPathwayDB and DAVID/EASE, that can be accessed via their webserver page. We may include additional tools such as CAMERA, ROAST, and ROMER (the code is available via the limma package and other R packages.)

VIII. Printing the tables with the sets of genes

```
#########################################################################################
################### we print the sets of genes in separate files, starting with FDR < 0.1
#########################################################################################

fpkm_and_limma_FDR01_file <- paste(fpkm_results_name,".with_limma_results.and.FDR0p1.txt", sep="")
write.table(fpkm_and_limma_FDR01, file=fpkm_and_limma_FDR01_file,
                                  sep="\t", eol="\n", row.names=FALSE, col.names=TRUE)

#########################################################################################
################### we print the sets of genes in separate files, starting with FDR < 0.1
################### and FC > 1.2, and FPKM > 1
#########################################################################################

fpkm_and_limma_FDR01_FC12_FPKM1_up_file <- paste(fpkm_results_name,".with_limma_results.and.FDR0p1.FC1p1
fpkm_and_limma_FDR01_FC12_FPKM1_down_file <- paste(fpkm_results_name,".with_limma_results.and.FDR0p1.FC

write.table(fpkm_and_limma_upreg_FDR01_FC12_FPKM1, file=fpkm_and_limma_FDR01_FC12_FPKM1_up_file,
                                          sep="\t", eol="\n", row.names=FALSE, col.names=TRUE)

write.table(fpkm_and_limma_downreg_FDR01_FC12_FPKM1, file=fpkm_and_limma_FDR01_FC12_FPKM1_down_file,
                                          sep="\t", eol="\n", row.names=FALSE, col.names=TRUE)

#########################################################################################
################### we print the sets of genes in separate files, starting with FDR < 0.1
################### and FC > 1.5, and FPKM > 1
#########################################################################################
```

```
fpkm_and_limma_FDR01_FC15_FPKM1_up_file <- paste(fpkm_results_name,".with_limma_results.and.FDR0p1.FC1p5
fpkm_and_limma_FDR01_FC15_FPKM1_down_file <- paste(fpkm_results_name,".with_limma_results.and.FDR0p1.FC


write.table(fpkm_and_limma_upreg_FDR01_FC15_FPKM1, file=fpkm_and_limma_FDR01_FC15_FPKM1_up_file,
                                                    sep="\t", eol="\n", row.names=FALSE, col.names=TRUE)

write.table(fpkm_and_limma_downreg_FDR01_FC15_FPKM1, file=fpkm_and_limma_FDR01_FC15_FPKM1_down_file,
                                                   sep="\t", eol="\n", row.names=FALSE, col.names=TRUE)
```

```
# Here printing the number of genes  in each category :

# 1.  nr. genes that are CSC-specific (FDR < 0.1, FC > 1.2, FPKM > 1):

    dim(fpkm_and_limma_upreg_FDR01_FC12_FPKM1)[1]
```

```
## [1] 1187
```

```
# 2. nr. genes that are noncsc-specific (FDR < 0.1, FC > 1.2, FPKM > 1):

   dim(fpkm_and_limma_downreg_FDR01_FC12_FPKM1)[1]
```

```
## [1] 1338
```

```
# 3.  nr. genes that are CSC-specific (FDR < 0.1, FC > 1.5, FPKM > 1):

    dim(fpkm_and_limma_upreg_FDR01_FC15_FPKM1)[1]
```

```
## [1] 431
```

```
# 4. nr. genes that are noncsc-specific (FDR < 0.1, FC > 1.5, FPKM > 1):

   dim(fpkm_and_limma_downreg_FDR01_FC15_FPKM1)[1]
```

```
## [1] 843
```