Microarray data analysis

COMMANDO 1

In [2]:

```
library(affy)
```

```
Loading required package: BiocGenerics
Loading required package: parallel

Attaching package: ‘BiocGenerics’

The following objects are masked from ‘package:parallel’:

    clusterApply, clusterApplyLB, clusterCall, clusterEvalQ,
    clusterExport, clusterMap, parApply, parCapply, parLapply,
    parLapplyLB, parRapply, parSapply, parSapplyLB

The following object is masked from ‘package:stats’:

    xtabs

The following objects are masked from ‘package:base’:

    anyDuplicated, append, as.data.frame, as.vector, cbind, colnames
,
    do.call, duplicated, eval, evalq, Filter, Find, get, intersect,
    is.unsorted, lapply, Map, mapply, match, mget, order, paste, pma
x,
    pmax.int, pmin, pmin.int, Position, rank, rbind, Reduce, rep.int
,
    rownames, sapply, setdiff, sort, table, tapply, union, unique,
    unlist, unsplit

Loading required package: Biobase
Welcome to Bioconductor

    Vignettes contain introductory material; view with
    'browseVignettes()'. To cite Bioconductor, see
    'citation("Biobase")', and for packages 'citation("pkgname")'.
```

COMMANDO 2

In [3]:

```
celfiles = Sys.glob("/mnt/storage/r0384602/jupyternotebooks/groepswerk/ARRAY_GSE1
```

In [4]:

```
celfiles
```

'/mnt/storage/r0384602/jupyternotebooks/groepswerk/ARRAY_GSE18632_RAW/GSM4
'/mnt/storage/r0384602/jupyternotebooks/groepswerk/ARRAY_GSE18632_RAW/GSM4
'/mnt/storage/r0384602/jupyternotebooks/groepswerk/ARRAY_GSE18632_RAW/GSM4
'/mnt/storage/r0384602/jupyternotebooks/groepswerk/ARRAY_GSE18632_RAW/GSM4
'/mnt/storage/r0384602/jupyternotebooks/groepswerk/ARRAY_GSE18632_RAW/GSM4
'/mnt/storage/r0384602/jupyternotebooks/groepswerk/ARRAY_GSE18632_RAW/GSM4
'/mnt/storage/r0384602/jupyternotebooks/groepswerk/ARRAY_GSE18632_RAW/GSM4
'/mnt/storage/r0384602/jupyternotebooks/groepswerk/ARRAY_GSE18632_RAW/GSM4

COMMANDO 3

In [5]:

```
rawcel <- ReadAffy(filenames=celfiles)
```

COMMANDO 4

In [6]:

```
sampleNames(rawcel)
```

'GSM463015.CEL.gz' 'GSM463016.CEL.gz' 'GSM463017.CEL.gz'
'GSM463018.CEL.gz' 'GSM463019.CEL.gz' 'GSM463020.CEL.gz'
'GSM463021.CEL.gz' 'GSM463022.CEL.gz'

COMMANDO 5

In [7]:

```
transfection = c("scr_siRNA1","scr_siRNA2","scr_siRNA3","scr_siRNA4","TDP-43_siRN
table <- data.frame(Treatment = transfection)
rownames(table) <- sampleNames(rawcel)
table
```

	Treatment
GSM463015.CEL.gz	scr_siRNA1
GSM463016.CEL.gz	scr_siRNA2
GSM463017.CEL.gz	scr_siRNA3
GSM463018.CEL.gz	scr_siRNA4
GSM463019.CEL.gz	TDP-43_siRNA1
GSM463020.CEL.gz	TDP-43_siRNA2
GSM463021.CEL.gz	TDP-43_siRNA3
GSM463022.CEL.gz	TDP-43_siRNA4

COMMANDO 6

In [8]:

```
library(Biobase)
phenoData(rawcel) <- AnnotatedDataFrame(data=table)
```

COMMANDO 7

In [9]:

```
library(frma)
eset <- frma(rawcel, verbose=TRUE)
save(eset, file="/mnt/storage/r0384602/jupyternotebooks/groepswerk/ARRAY_GSE18632
```

```
Creating a generic function for ‘nchar’ from package ‘base’ in packa
ge ‘S4Vectors’
Background Correcting ...


Normalizing ...

Summarizing ...
```

COMMANDO 8

In [10]:

```
eset.labels <- c("HEK293E_scrambled_siRNA_replicate_1","HEK293E_scrambled_siRNA_r
sampleNames(eset) <- eset.labels
sampleNames(eset)
```

'HEK293E_scrambled_siRNA_replicate_1'
'HEK293E_scrambled_siRNA_replicate_2'
'HEK293E_scrambled_siRNA_replicate_3'
'HEK293E_scrambled_siRNA_replicate_4'
'HEK293E_TDP-43_siRNA_replicate_1' 'HEK293E_TDP-43_siRNA_replicate_2'
'HEK293E_TDP-43_siRNA_replicate_3' 'HEK293E_TDP-43_siRNA_replicate_4'

COMMANDO 9

In [11]:

```
probe.counts <- exprs(eset)
probe.counts [1:5,1:8]
```

	HEK293E_scrambled_siRNA_replicate_1	**HEK293E_scrambled_siRNA_repli**
1007_s_at	7.993843	8.131543
1053_at	9.412263	9.214187
117_at	5.161575	5.141424
121_at	7.442329	7.352903
1255_g_at	3.410106	3.514217

COMMANDO 10

In [12]:

```
library(annotate)
library(hgu133plus2.db)
```

```
Loading required package: AnnotationDbi
Loading required package: stats4
Loading required package: GenomeInfoDb
Loading required package: S4Vectors
Loading required package: IRanges
Loading required package: XML
Loading required package: org.Hs.eg.db
```

COMMANDO 11

In [13]:

```
probe.names <- rownames(probe.counts)
gene.symbols <- getSYMBOL(probe.names, "hgu133plus2")
head(gene.symbols)
```

1007_s_at
'NA'
1053_at
'RFC2'
117_at
'HSPA6'
121_at
'PAX8'
1255_g_at
'GUCA1A'
1294_at
'NA'

COMMANDO 12

In [14]:

```
probe.count.genesymbol <- as.data.frame(cbind(as.data.frame(probe.counts),gene.sy
head(probe.count.genesymbol)
```

	HEK293E_scrambled_siRNA_replicate_1	HEK293E_scrambled_siRNA_repli
1007_s_at	7.993843	8.131543
1053_at	9.412263	9.214187
117_at	5.161575	5.141424
121_at	7.442329	7.352903
1255_g_at	3.410106	3.514217
1294_at	5.697905	5.769541

COMMANDO 13

In [39]:

```
probe.count.genesymbol=probe.count.genesymbol[complete.cases(probe.count.genesymb
probe.count.genesymbol$gene.symbols <- as.character(probe.count.genesymbol$gene.s
```

COMMANDO 14

In [40]:

```
library(reshape2)
```

COMMANDO 15

In [41]:

```
m <- melt(probe.count.genesymbol, id.vars = c("gene.symbols"))
head(m)
counts <- dcast(m, gene.symbols ~ variable, fun.aggregate=median)
```

gene.symbols	variable	value
RFC2	HEK293E_scrambled_siRNA_replicate_1	9.412263
HSPA6	HEK293E_scrambled_siRNA_replicate_1	5.161575
PAX8	HEK293E_scrambled_siRNA_replicate_1	7.442329
GUCA1A	HEK293E_scrambled_siRNA_replicate_1	3.410106
THRA	HEK293E_scrambled_siRNA_replicate_1	5.159896
PTPN21	HEK293E_scrambled_siRNA_replicate_1	4.558947

In [42]:

```
counts[1:5,1:8]
```

gene.symbols	HEK293E_scrambled_siRNA_replicate_1	HEK293E_scrambled_siRNA_
A1BG	5.879163	5.708719
A1BG-AS1	5.141935	4.869427
A1CF	3.911032	3.821178
A2M	4.674781	4.668512
A2M-AS1	4.784417	4.750436

COMMANDO 16

In [43]:

```
head(counts$gene.symbols)
```

'A1BG' 'A1BG-AS1' 'A1CF' 'A2M' 'A2M-AS1' 'A2ML1'

In [44]:

```
rownames(counts) <- counts$gene.symbols
```

In [45]:

```
counts$gene.symbols <- NULL
```

In [46]:

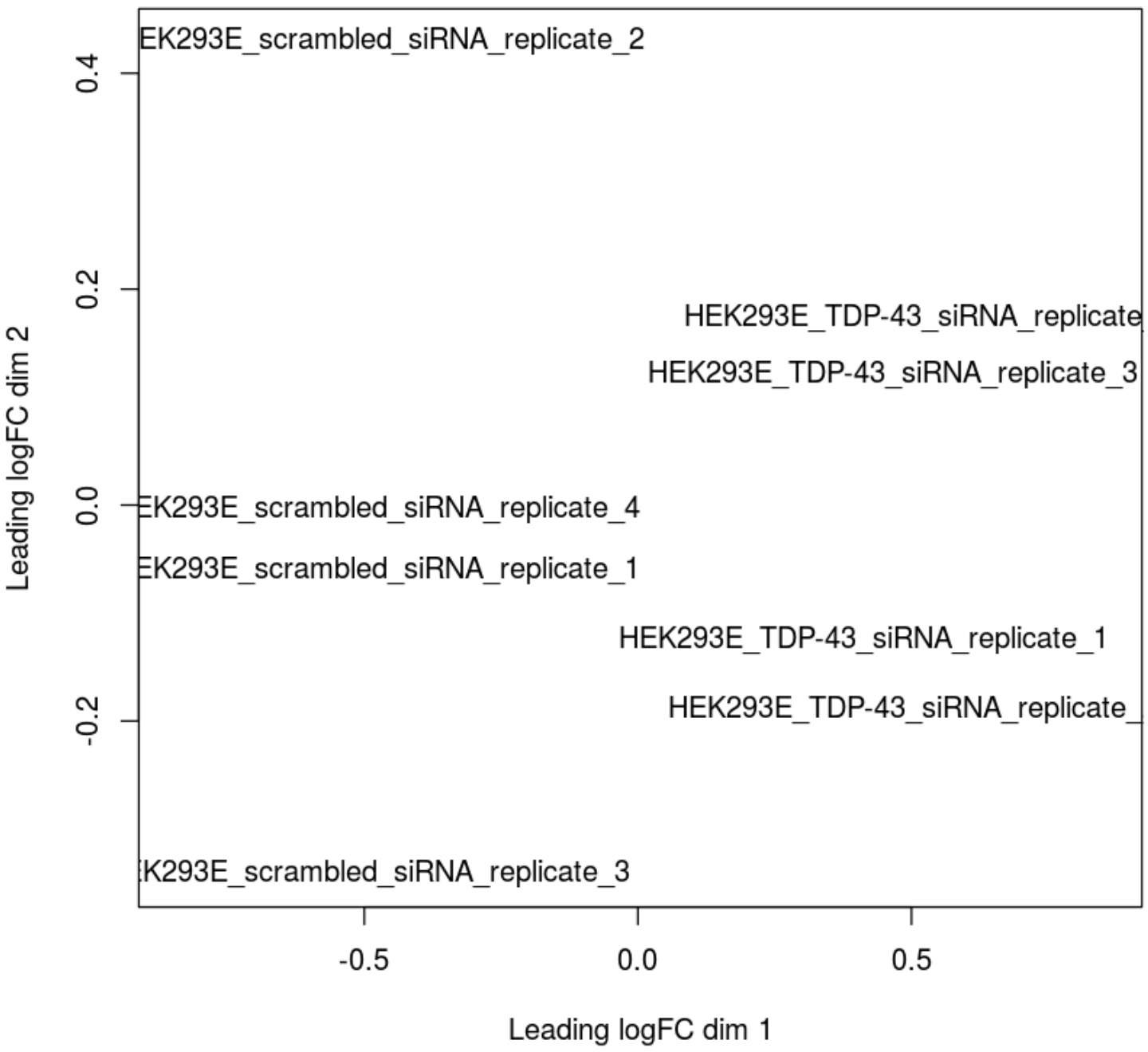
```
counts[1:7,1:8]
```

	HEK293E_scrambled_siRNA_replicate_1	HEK293E_scrambled_siRNA_replicat
A1BG	5.879163	5.708719
A1BG-AS1	5.141935	4.869427
A1CF	3.911032	3.821178
A2M	4.674781	4.668512
A2M-AS1	4.784417	4.750436
A2ML1	3.947888	3.851730
A2MP1	4.342936	4.249094

COMMANDO 17

In [48]:

```
plotMDS(eset)
```



COMMANDO 18

In [49]:

```
gene.of.interest <- c("HDAC6")
counts[gene.of.interest,]
```

	HEK293E_scrambled_siRNA_replicate_1	HEK293E_scrambled_siRNA_replica
HDAC6	6.942547	6.92049

COMMANDO 19

In [50]:

```
condition <- factor(c('scrambled','scrambled','scrambled','scrambled','TDP-43','T
levels=c('scrambled','TDP-43')
condition
```

scrambled scrambled scrambled scrambled TDP-43 TDP-43 TDP-43
TDP-43

COMMANDO 20

In [51]:

```
design <- model.matrix(~condition)
design
```

	(Intercept)	conditionTDP-43
1	1	0
2	1	0
3	1	0
4	1	0
5	1	1
6	1	1
7	1	1
8	1	1

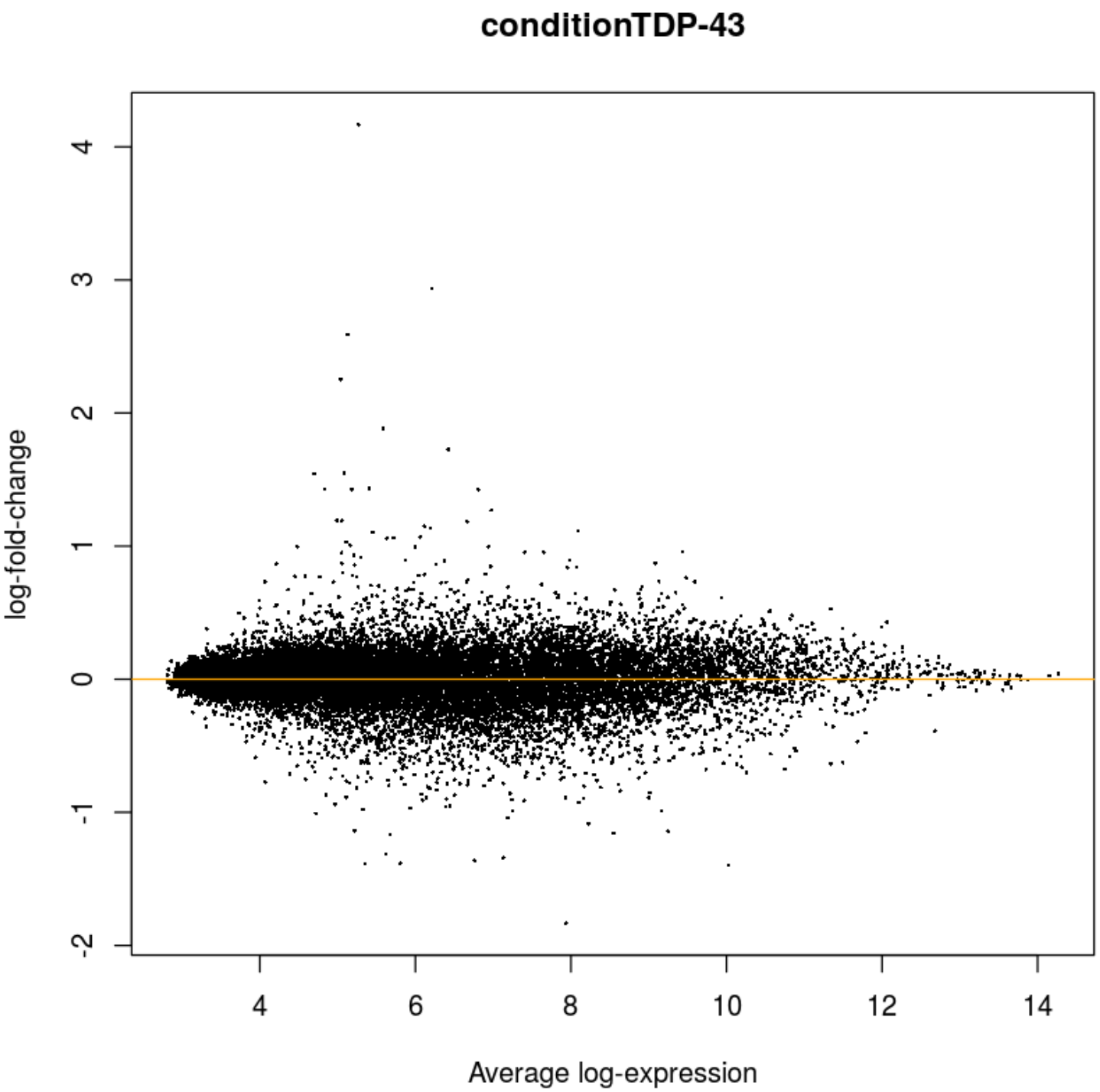
In [52]:

```
fit <- eBayes(lmFit(counts, design))
```

COMMANDO 21

In [53]:

```
library(limma)
limma::plotMA(fit,2)
abline(0,0,col='orange')
```



In [54]:

```
topTable(fit, coef=2)
```

	logFC	AveExpr	t	P.Value	adj.P.Val	B
SCG3	4.167059	5.267898	61.52144	5.887576e-17	1.200418e-12	23.63992
CPLX1	2.590921	5.132486	46.22347	2.101845e-15	2.142725e-11	22.22292
SNAP25	2.252762	5.038594	36.37211	4.174506e-14	2.837134e-10	20.63059
AP3B2	1.882561	5.585829	27.35921	1.429601e-12	7.287033e-09	18.26637
ASPHD1	1.727798	6.422915	24.45641	5.706432e-12	2.326969e-08	17.21013
TARDBP	-1.832799	7.938111	-23.82506	7.873933e-12	2.675694e-08	16.95497
USP13	-1.342251	7.130792	-23.49131	9.366001e-12	2.728049e-08	16.81604
CHGB	2.935207	6.212650	22.76635	1.376736e-11	3.337140e-08	16.50416
PFKP	-1.398034	10.027743	-22.64130	1.473062e-11	3.337140e-08	16.44893
ACTL6B	1.425856	5.183743	20.44550	5.138498e-11	1.047688e-07	15.40367

COMMANDO 22

In [55]:

```
select = p.adjust(fit[,2]$p.value, method="BH") < 0.05
```

In [56]:

```
summary(select)
```

```
   Mode   FALSE    TRUE    NA's
logical   16874    3515       0
```

COMMANDO 23

In [57]:

```
top = topTable(fit, coef=2, number=100000)
write.table(top, "GSE18632.tophits.tsv", sep="\t", quote=FALSE, col.names=NA)
head(top)
```

	logFC	AveExpr	t	P.Value	adj.P.Val	B
SCG3	4.167059	5.267898	61.52144	5.887576e-17	1.200418e-12	23.63992
CPLX1	2.590921	5.132486	46.22347	2.101845e-15	2.142725e-11	22.22292
SNAP25	2.252762	5.038594	36.37211	4.174506e-14	2.837134e-10	20.63059
AP3B2	1.882561	5.585829	27.35921	1.429601e-12	7.287033e-09	18.26637
ASPHD1	1.727798	6.422915	24.45641	5.706432e-12	2.326969e-08	17.21013
TARDBP	-1.832799	7.938111	-23.82506	7.873933e-12	2.675694e-08	16.95497

RNASeq Analysis

COMMANDO 1

In [17]:

```
library("DESeq2")
library(scales)
options(repr.plot.width=4, repr.plot.height=3)
```

```
Loading required package: GenomicRanges
Loading required package: Rcpp
Loading required package: RcppArmadillo
Warning message:
: replacing previous import ‘S4Vectors::Position’ by ‘ggplot2::Posit
ion’ when loading ‘DESeq2’
```

COMMANDO 2

In [18]:

```
setwd(path.expand("/mnt/storage/r0384602/jupyternotebooks/groepswerk/RNAseq"))
```

COMMANDO 3

In [19]:

```
matrixFile = path.expand("/mnt/storage/r0384602/jupyternotebooks/groepswerk/RNAse
countData <- as.matrix(read.csv(matrixFile, sep="\t", row.names="Geneid"))
head(countData)
```

	scr_1.bam	scr_2.bam	TDP.43_1.bam	TDP.43_2.bam
MIR1302-11	0	0	0	0
FAM138A	0	0	0	0
OR4F5	0	0	0	0
RP11-34P13.7	2	1	0	0
RP11-34P13.8	0	0	0	0
AL627309.1	6	6	0	2

COMMANDO 4

In [20]:

```
colnames(countData) <- c("scrambled_replicate1", "scrambled_replicate2","TDP-43_r
head(countData)
```

	scrambled_replicate1	scrambled_replicate2	TDP-43_replicate1	TDP-43_replicate2
MIR1302-11	0	0	0	0
FAM138A	0	0	0	0
OR4F5	0	0	0	0
RP11-34P13.7	2	1	0	0
RP11-34P13.8	0	0	0	0
AL627309.1	6	6	0	2

COMMANDO 5

In [21]:

```
condition <- factor(c("scrambled", "scrambled", "TDP-43", "TDP-43"))
print(condition)
```

```
[1] scrambled scrambled TDP-43    TDP-43
Levels: scrambled TDP-43
```

COMMANDO 6

In [22]:

```
colData <- data.frame(sampleName = colnames(countData),
                      condition = condition)
colData
```

sampleName	condition
scrambled_replicate1	scrambled
scrambled_replicate2	scrambled
TDP-43_replicate1	TDP-43
TDP-43_replicate2	TDP-43

COMMANDO 7

In [23]:

```
dds <- DESeqDataSetFromMatrix(countData = countData,
                              colData = colData,
                              design = ~ condition)
colnames(dds) <- colnames(countData)
dds
```

```
class: DESeqDataSet
dim: 41863 4
exptData(0):
assays(1): counts
rownames(41863): MIR1302-11 FAM138A ... AC007965.1 RNU6-1314P
rowRanges metadata column names(0):
colnames(4): scrambled_replicate1 scrambled_replicate2
  TDP-43_replicate1 TDP-43_replicate2
colData names(2): sampleName condition
```

COMMANDO 8

In [24]:

```
dds$condition
```

```
scrambled scrambled TDP-43 TDP-43
```

COMMANDO 9

In [25]:

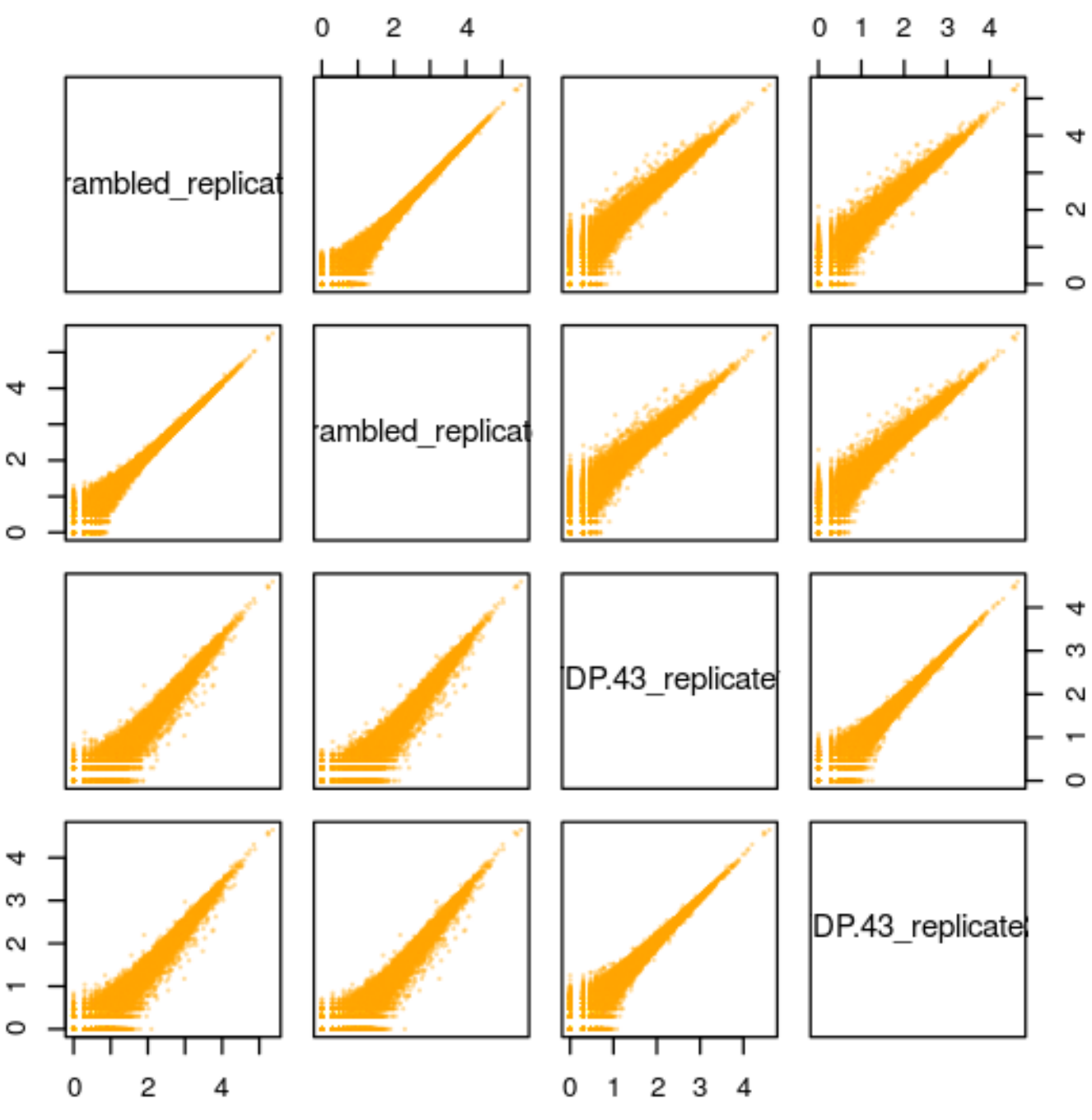
```
dds <- DESeq(dds)
```

```
estimating size factors
estimating dispersions
gene-wise dispersion estimates
mean-dispersion relationship
final dispersion estimates
fitting model and testing
```

COMMANDO 10

In [26]:

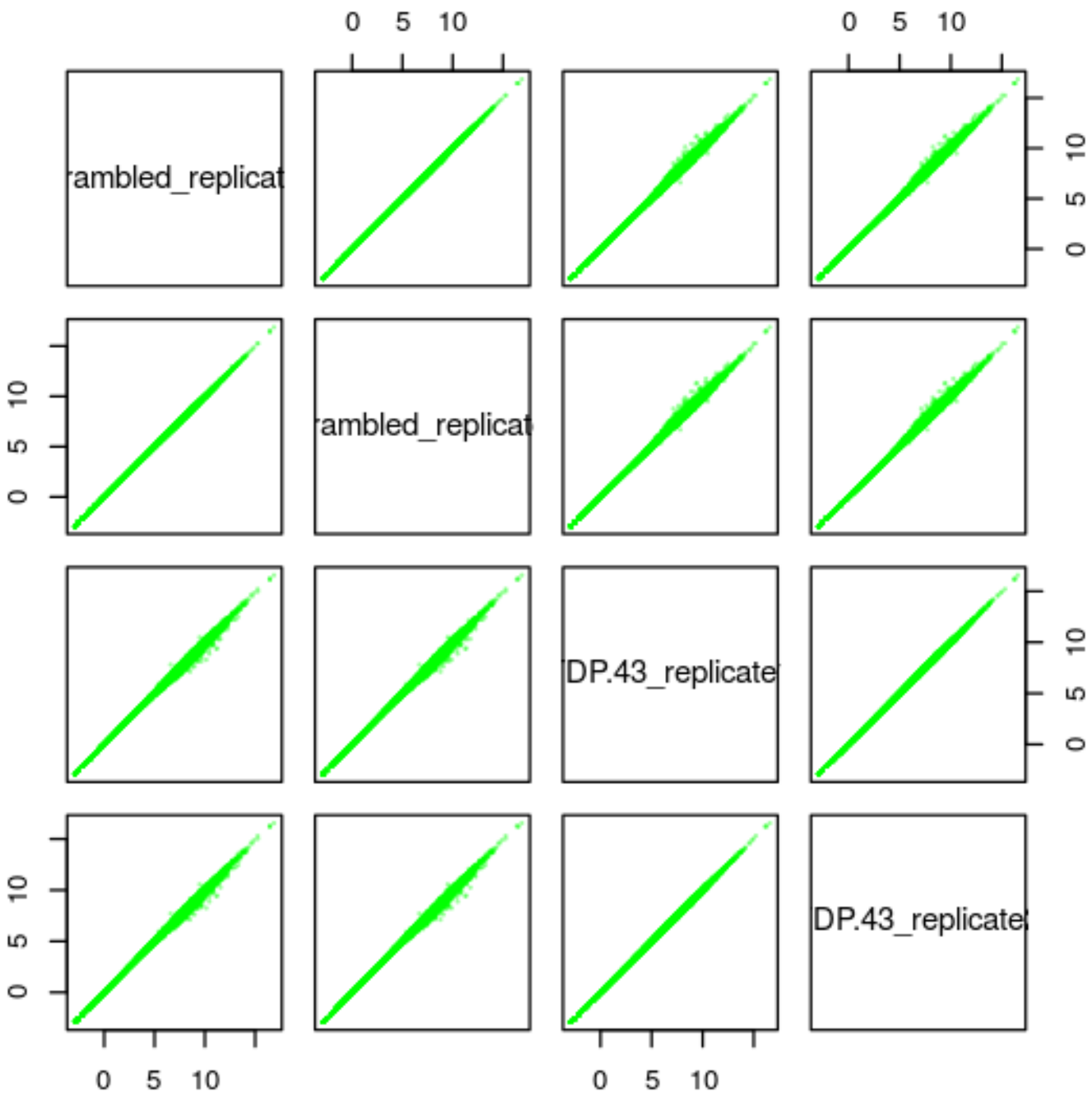
```
options(repr.plot.width=5, repr.plot.height=5)
plot(data.frame(log10(counts(dds))), cex=0.1, col=alpha("orange", 0.4))
```



COMMANDO 11

In [27]:

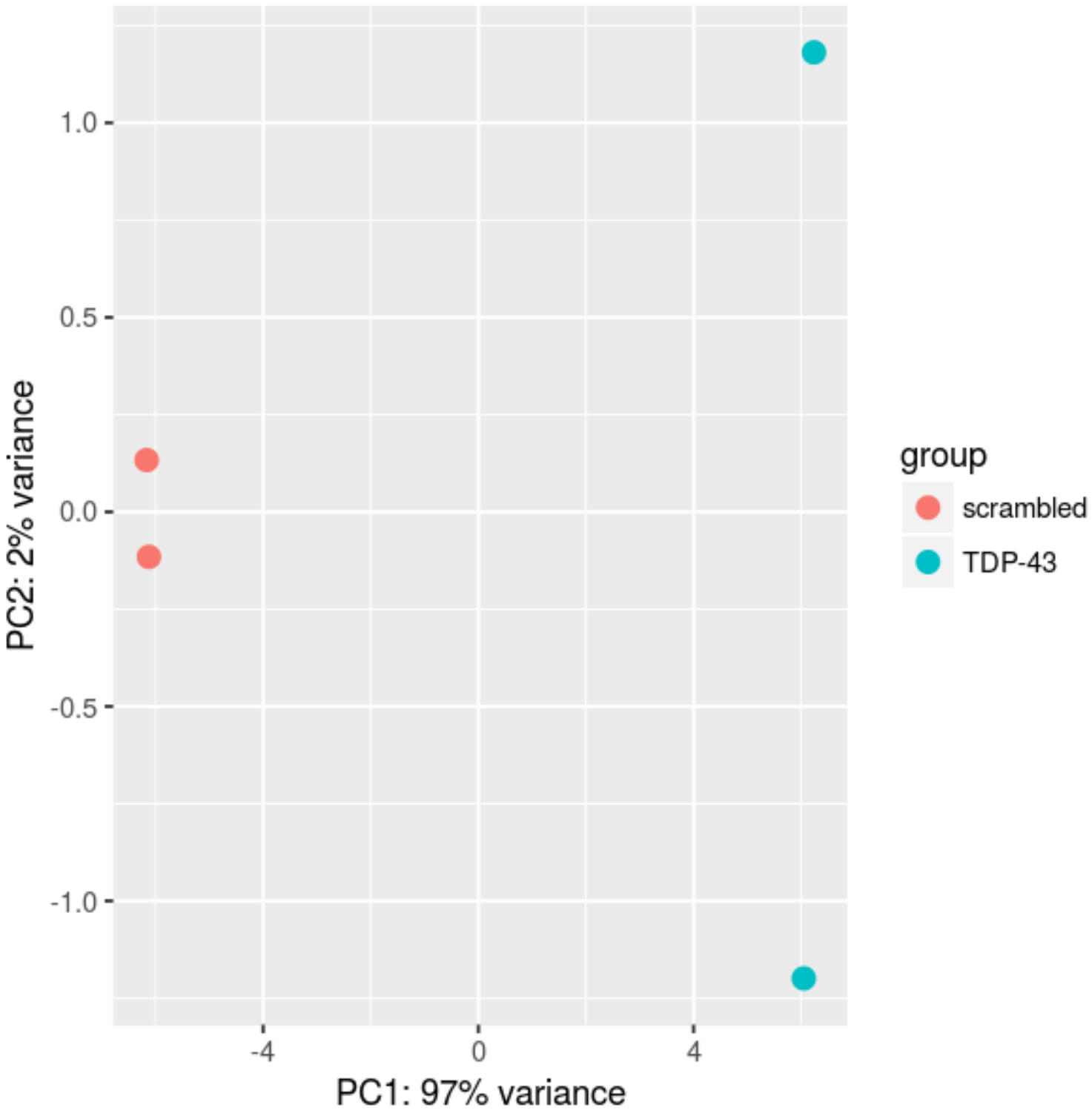
```
rld <- rlogTransformation(dds, blind=TRUE)
options(repr.plot.width=5, repr.plot.height=5)
plot(data.frame(assay(rld)), cex=0.1, col=alpha("green", 0.4))
```



COMMANDO 12

In [28]:

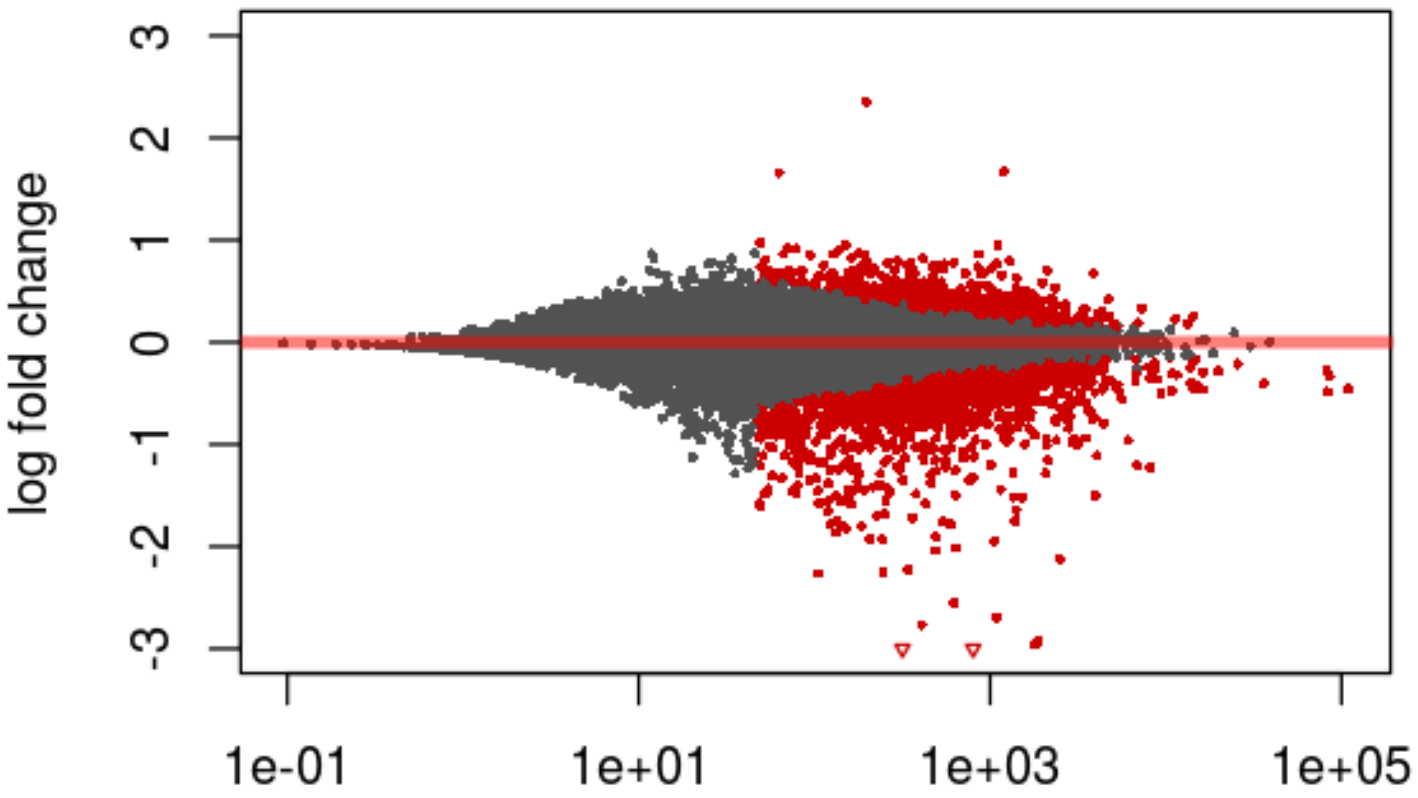
```
plotPCA(rld, intgroup=c("condition"))
```



COMMANDO 13

In [29]:

```
options(repr.plot.width=5, repr.plot.height=4)
plotMA(dds,ylim=c(-3,3))
```



COMMANDO 14

In [30]:

```
res <- results(dds)
res <- res[order(res$padj),]
head(res)
```

```
log2 fold change (MAP): condition TDP-43 vs scrambled
Wald test p-value: condition TDP-43 vs scrambled
DataFrame with 6 rows and 6 columns
        baseMean log2FoldChange      lfcSE      stat        pvalue
       <numeric>      <numeric>  <numeric> <numeric>     <numeric>
SPARC  1793.2785      -2.963376 0.10715856 -27.65413 2.489449e-168
COL4A1 1877.0533      -2.924304 0.10784369 -27.11614 6.353914e-162
COL5A2 1084.9239      -2.692911 0.11913465 -22.60393 3.964745e-113
PEG10  3952.0176      -1.501478 0.06933588 -21.65514 5.436272e-104
FN1    2493.4552      -2.120828 0.10055558 -21.09111  9.599254e-99
COL1A1  801.0936      -3.081836 0.15512669 -19.86657  7.922809e-88
                padj
           <numeric>
SPARC  2.838221e-164
COL4A1 3.622049e-158
COL5A2 1.506735e-109
PEG10  1.549474e-100
FN1     2.188822e-95
COL1A1  1.505466e-84
```

COMMANDO 15

In [31]:

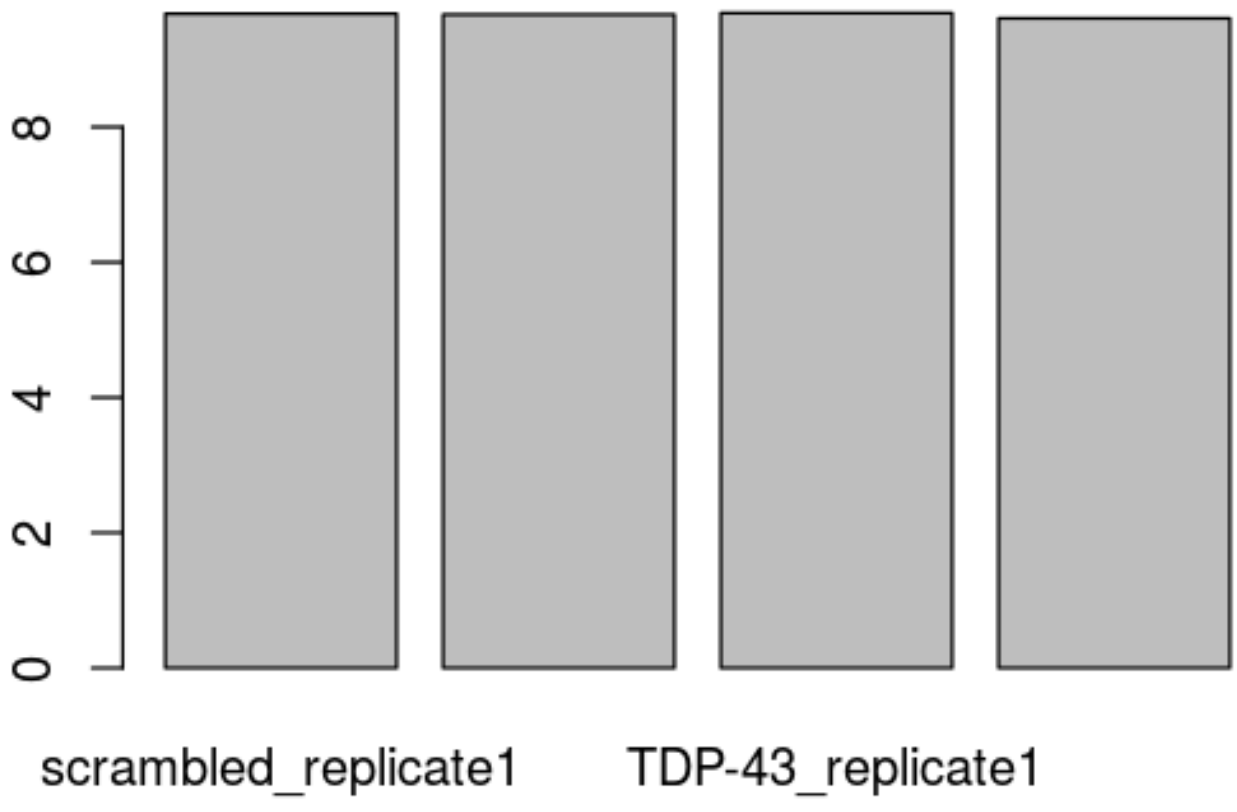
```
res.sign <- res[(!is.na(res$padj)) & (res$padj < 0.05),]
dim(res.sign)
```

```
1579  6
```

COMMANDO 16

In [32]:

```
barplot(assay(rld)["HDAC6",])
```



COMMANDO 17

In [ ]:

```
write.table(res, "deseq.results.tsv", sep="\t", col.names=NA, quote=FALSE)
```

In [ ]:

```

```