In [5]:
```
library("DESeq2")
library(scales)
options(repr.plot.width=4, repr.plot.height=3)
```

```
Loading required package: S4Vectors
Loading required package: stats4
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, pmax,
    pmax.int, pmin, pmin.int, Position, rank, rbind, Reduce, rep.int,
    rownames, sapply, setdiff, sort, table, tapply, union, unique,
    unlist, unsplit

Creating a generic function for 'nchar' from package 'base' in package 'S4Vecto
rs'
Loading required package: IRanges
Loading required package: GenomicRanges
Loading required package: GenomeInfoDb
Loading required package: Rcpp
Loading required package: RcppArmadillo
Warning message:
: replacing previous import 'S4Vectors::Position' by 'ggplot2::Position' when l
oading 'DESeq2'
```

In [6]:
```
setwd(path.expand("~/data/work/"))
```

In [7]:
```
matrixFile = path.expand("~/data/work/all_gene_counts_RUN.tsv")
```

In [8]:
```
countDataRUN <- as.matrix(read.csv(matrixFile, sep="\t", row.names="Geneid"))
dim(countDataRUN)
head(countDataRUN)
```

41863 4

	shNSR1.bam	shNSR2.bam	shRUNR1.bam	shRUNR2.bam
MIR1302-11	0	0	0	0
FAM138A	0	0	0	0
OR4F5	0	0	0	0
RP11-34P13.7	2	0	0	0
RP11-34P13.8	0	0	0	0
AL627309.1	13	14	6	15

In [9]:
```
colnames(countDataRUN) <- c("shNSR1", "shNSR2", "shRUNR1", "shRUNR2")
head(countDataRUN)
```

	shNSR1	shNSR2	shRUNR1	shRUNR2
MIR1302-11	0	0	0	0
FAM138A	0	0	0	0
OR4F5	0	0	0	0
RP11-34P13.7	2	0	0	0
RP11-34P13.8	0	0	0	0
AL627309.1	13	14	6	15

In [10]:
```
condition <- factor(c("NS", "NS", "S", "S"))
print(condition)
```

```
[1] NS NS S  S
Levels: NS S
```

In [11]:
```
colDataRUN <- data.frame(sampleName = colnames(countDataRUN), condition = condition
colDataRUN
```

sampleName	condition
shNSR1	NS
shNSR2	NS
shRUNR1	S
shRUNR2	S

In [12]:
```
colSums(countDataRUN)
```

shNSR1	9202975
shNSR2	13072801
shRUNR1	11202455
shRUNR2	10788451

In [13]:
```
dds <- DESeqDataSetFromMatrix(countData = countDataRUN,
                              colData = colDataRUN,
                              design = ~ condition)
#we need to do this to ensure the column names are correct.
colnames(dds) <- colnames(countDataRUN)
```

In [14]:
```
dds
dim(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): shNSR1 shNSR2 shRUNR1 shRUNR2
colData names(2): sampleName condition
```

41863 4

In [15]:
```
dds$condition
```

NS NS S S

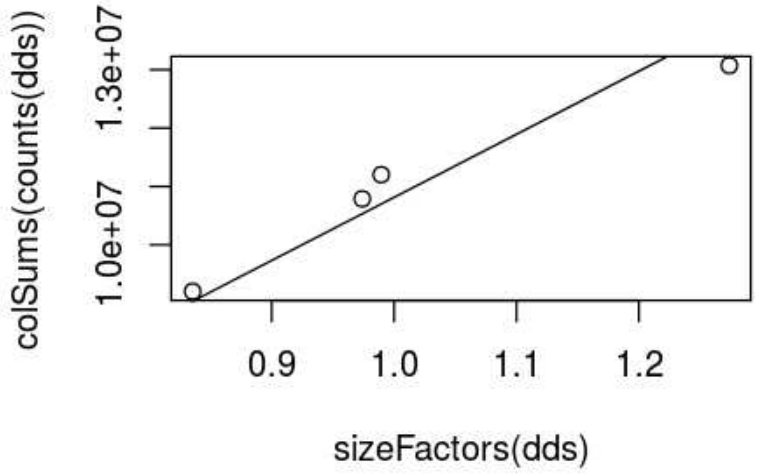
In [16]:
```
dds <- estimateSizeFactors(dds)
sizeFactors(dds)
```

shNSR1	0.835415644964619
shNSR2	1.27387288277898
shRUNR1	0.989458117216464
shRUNR2	0.974003746425297

In [17]:
```
colSums(countDataRUN)
```

shNSR1	9202975
shNSR2	13072801
shRUNR1	11202455
shRUNR2	10788451

In [40]:
```
plot(sizeFactors(dds), colSums(counts(dds)))
abline(lm(colSums(counts(dds)) ~ sizeFactors(dds) + 0))
```



In [ ]:

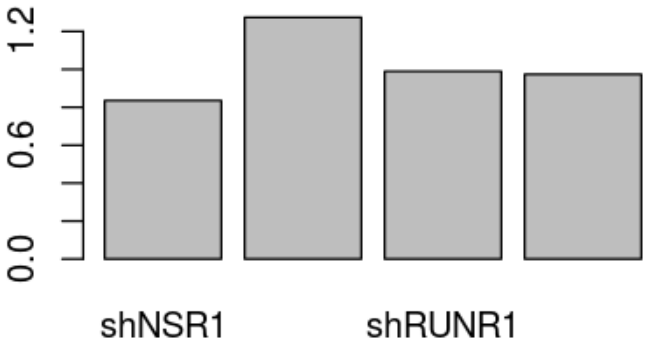
In [18]:
```
dds <- DESeq(dds)
```

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

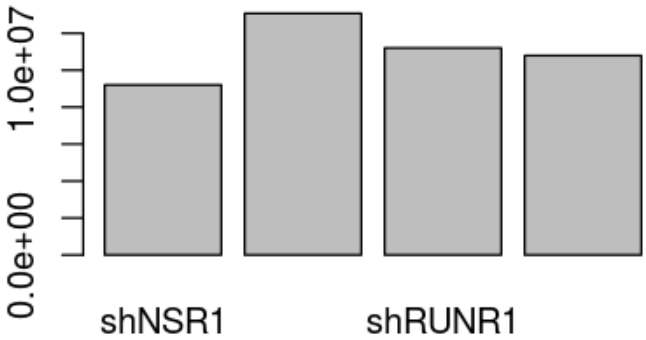
In [42]:
```
head(dds)
```

```
class: DESeqDataSet
dim: 1 4
exptData(0):
assays(3): counts mu cooks
rownames(1): MIR1302-11
rowRanges metadata column names(27): baseMean baseVar ... deviance
  maxCooks
colnames(4): shNSR1 shNSR2 shRUNR1 shRUNR2
colData names(3): sampleName condition sizeFactor
```

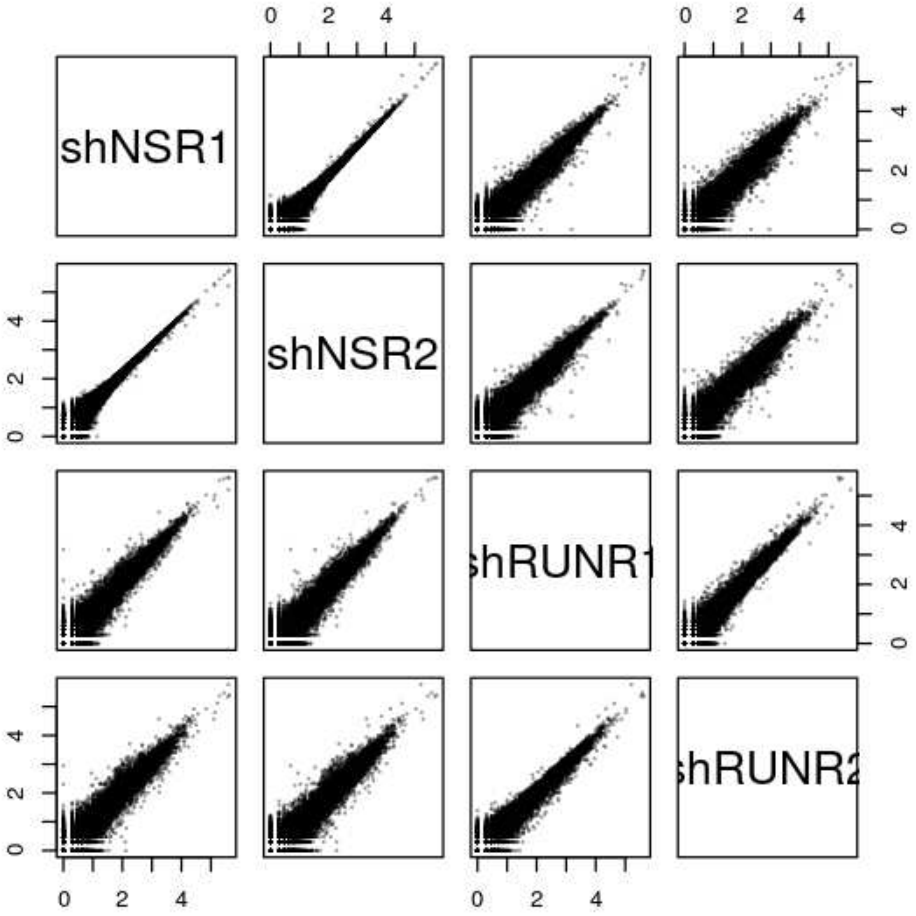
In [43]:
```
options(repr.plot.width=4, repr.plot.height=3)
barplot(sizeFactors(dds))
```



In [44]:
```
barplot(colSums(countDataRUN))
```

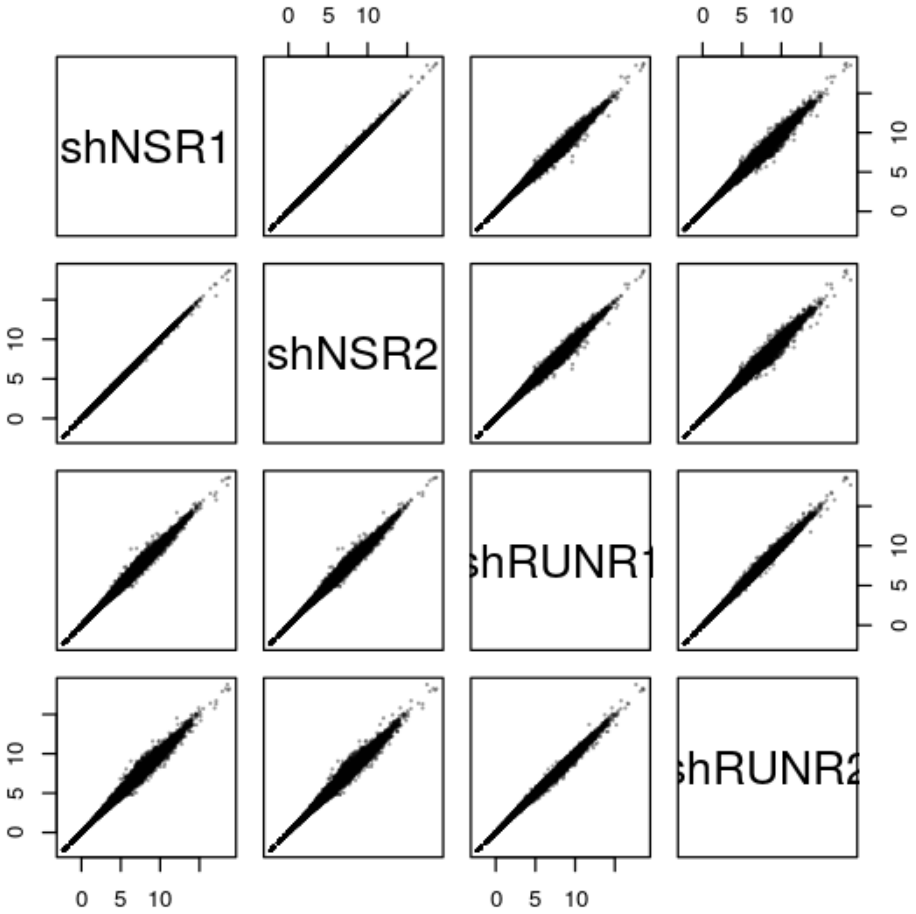


In [45]:
```
options(repr.plot.width=5, repr.plot.height=5)
plot(data.frame(log10(counts(dds))), cex=0.1, col=alpha("black", 0.4))
```

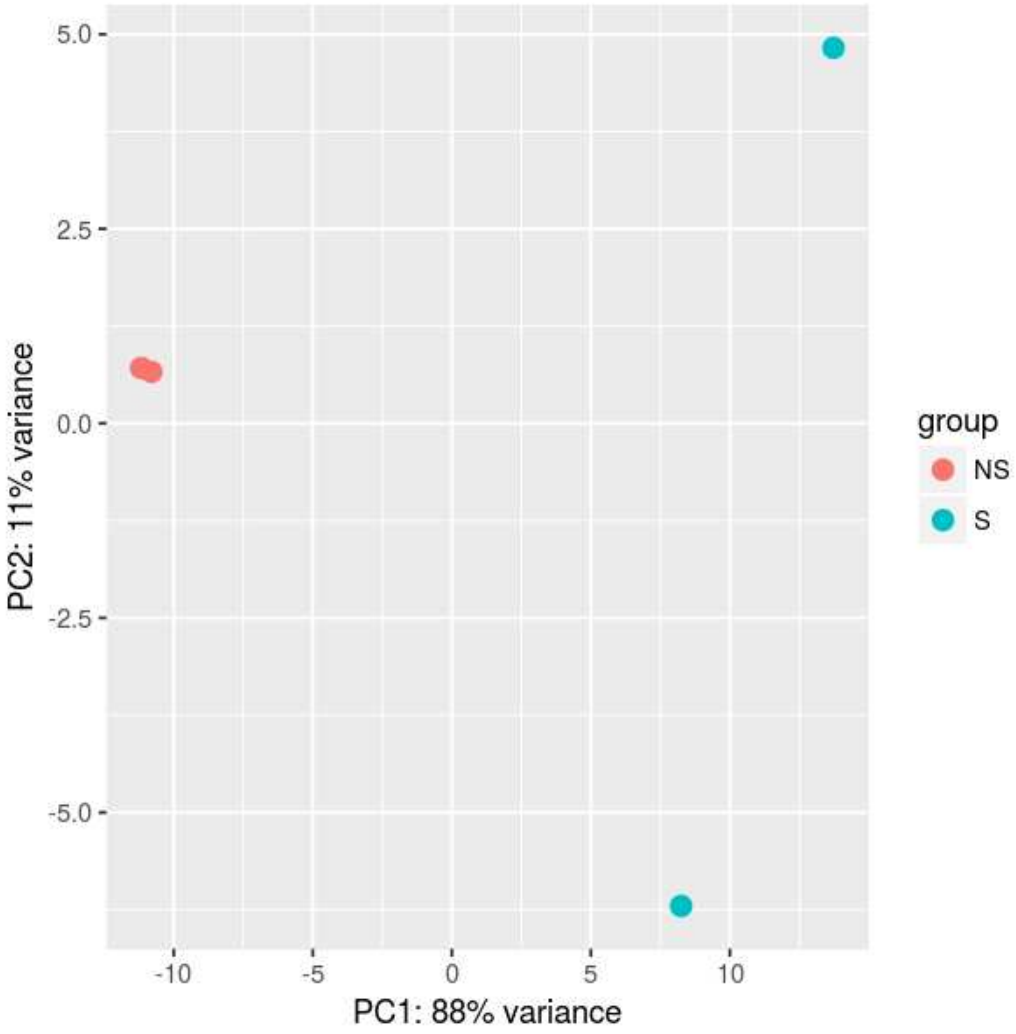


In [19]:
```
rld <- rlogTransformation(dds, blind=TRUE)
```

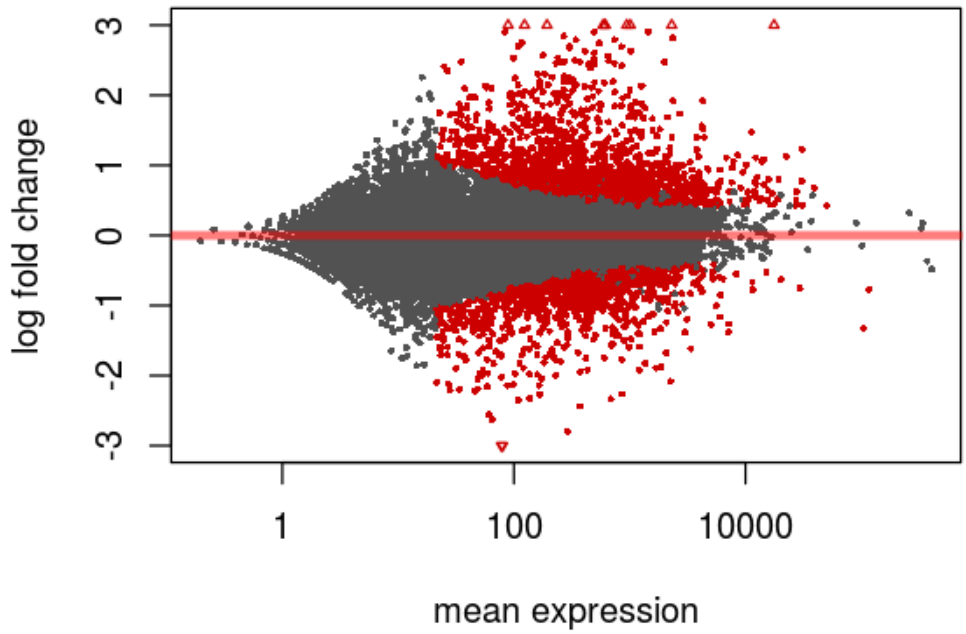
In [47]:
```
options(repr.plot.width=5, repr.plot.height=5)
plot(data.frame(assay(rld)), cex=0.1, col=alpha("black", 0.4))
```



In [48]:


5.0
2.5
0.0
-2.5
-5.0
PC2: 11% variance
-10
-5
0
5
10
PC1: 88% variance
group
NS
S

In [49]:
```
options(repr.plot.width=5, repr.plot.height=4)
plotMA(dds,ylim=c(-3,3))
```



In [20]:
```
res <- results(dds)
res <- res[order(res$padj),]
head(res)
dim(res)
```

```
log2 fold change (MAP): condition S vs NS
Wald test p-value: condition S vs NS
DataFrame with 6 rows and 6 columns
            baseMean log2FoldChange     lfcSE      stat       pvalue
           <numeric>      <numeric> <numeric> <numeric>    <numeric>
MIR663A   17657.6449       3.950579 0.2065394  19.12749 1.490956e-81
MIR30A      613.7356       6.664756 0.3489097  19.10167 2.445557e-81
ANKRD30BL  2306.0849       4.197999 0.2224707  18.86990 2.016852e-79
RRM2       1012.7201       3.216305 0.2037762  15.78352 4.040307e-56
MIR3687     589.4134       4.501850 0.2928766  15.37115 2.555944e-53
HIST1H3B   1444.3334       2.907871 0.2014228  14.43666 3.042535e-47
                 padj
            <numeric>
MIR663A   1.517957e-77
MIR30A    1.517957e-77
ANKRD30BL 8.345733e-76
RRM2      1.253909e-52
MIR3687   6.345898e-50
HIST1H3B  6.295005e-44
```

41863 6

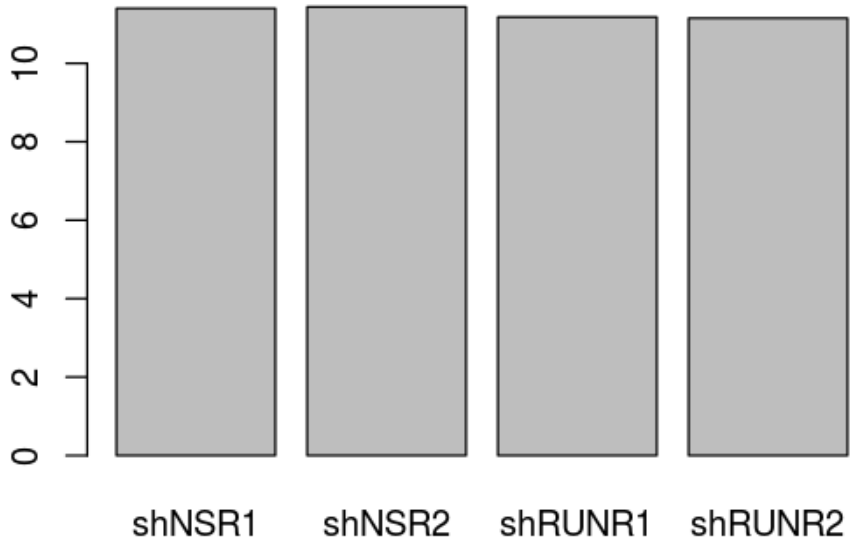
In [21]:
```
res.sign <- res[(!is.na(res$padj)) & (res$padj < 0.05),]
head(res.sign)
```

```
log2 fold change (MAP): condition S vs NS
Wald test p-value: condition S vs NS
DataFrame with 6 rows and 6 columns
           baseMean log2FoldChange     lfcSE      stat       pvalue
          <numeric>      <numeric> <numeric> <numeric>    <numeric>
MIR663A   17657.6449       3.950579 0.2065394  19.12749 1.490956e-81
MIR30A      613.7356       6.664756 0.3489097  19.10167 2.445557e-81
ANKRD30BL  2306.0849       4.197999 0.2224707  18.86990 2.016852e-79
RRM2       1012.7201       3.216305 0.2037762  15.78352 4.040307e-56
MIR3687     589.4134       4.501850 0.2928766  15.37115 2.555944e-53
HIST1H3B   1444.3334       2.907871 0.2014228  14.43666 3.042535e-47
                 padj
            <numeric>
MIR663A   1.517957e-77
MIR30A    1.517957e-77
ANKRD30BL 8.345733e-76
RRM2      1.253909e-52
MIR3687   6.345898e-50
HIST1H3B  6.295005e-44
```

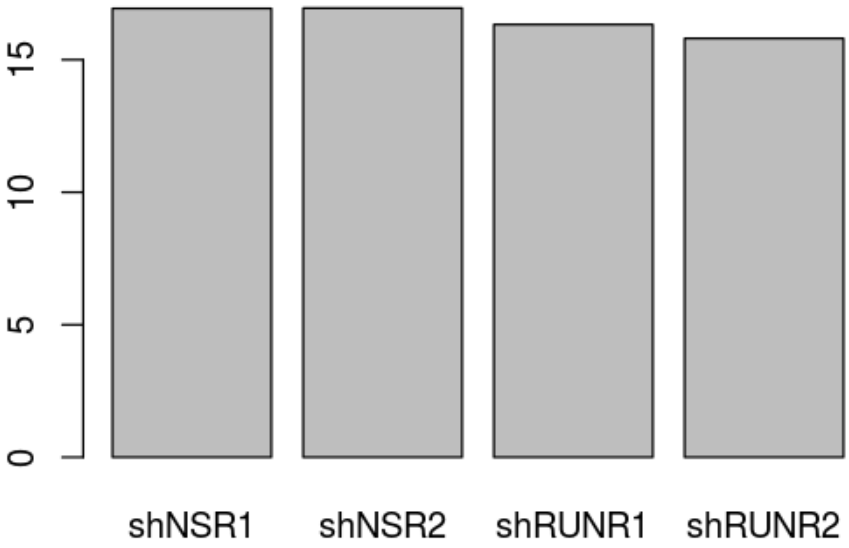
```
2072  6
```

In [57]:
```
write.table(res, "deseq_results_RUN.tsv", sep="\t", col.names=NA, quote=FALSE)
```

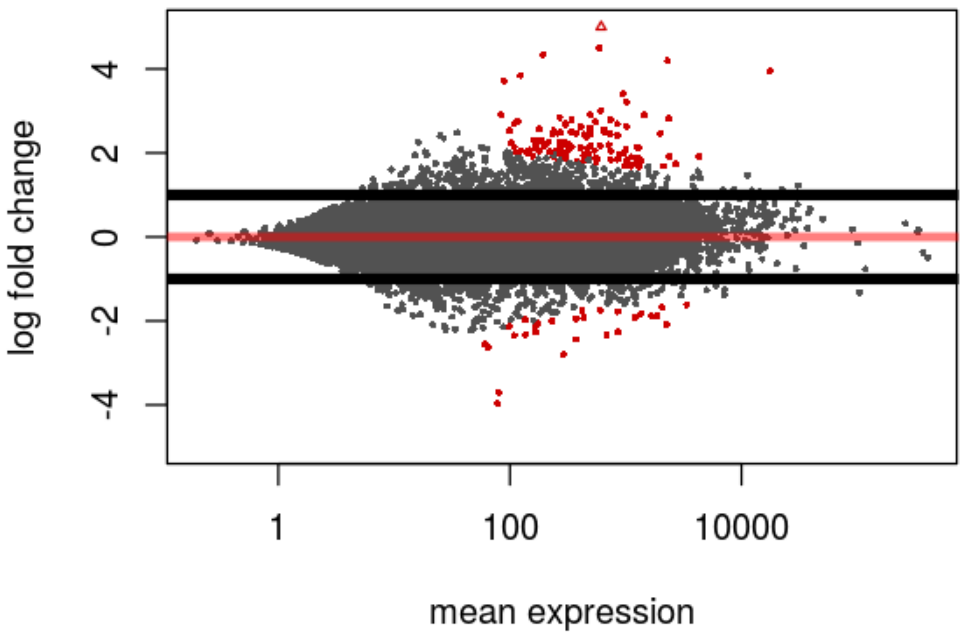
In [54]:
```
barplot(assay(rld)["CDKN1A",])
```



In [55]:
```
barplot(assay(rld)["MALAT1",])
```



In [58]:
```
resBigFC <- results(dds, lfcThreshold=1, altHypothesis="greaterAbs")
plotMA(resBigFC, ylim=c(-5,5))
abline(h=c(-1,1), lwd=5)
```



In [60]:
```
resSort <- res[order(res$padj),]
head(resSort)
```

```
log2 fold change (MAP): condition S vs NS
Wald test p-value: condition S vs NS
DataFrame with 6 rows and 6 columns
            baseMean log2FoldChange     lfcSE      stat       pvalue
           <numeric>      <numeric> <numeric> <numeric>    <numeric>
MIR663A   17657.6449       3.950579 0.2065394  19.12749 1.490956e-81
MIR30A      613.7356       6.664756 0.3489097  19.10167 2.445557e-81
ANKRD30BL  2306.0849       4.197999 0.2224707  18.86990 2.016852e-79
RRM2       1012.7201       3.216305 0.2037762  15.78352 4.040307e-56
MIR3687     589.4134       4.501850 0.2928766  15.37115 2.555944e-53
HIST1H3B   1444.3334       2.907871 0.2014228  14.43666 3.042535e-47
                 padj
            <numeric>
MIR663A   1.517957e-77
MIR30A    1.517957e-77
ANKRD30BL 8.345733e-76
RRM2      1.253909e-52
MIR3687   6.345898e-50
HIST1H3B  6.295005e-44
```

In [61]:
```
write.table(res, "deseq_results_RUN_Sorted.tsv", sep="\t", col.names=NA, quote=FALS
```

In [ ]:

In [ ]:

In [ ]: