Gene Expression Project

Fenny and Daniëlle

2023-04-26

Hippocampal subfield transcriptome analysis in schizophrenia psychosis

```
library(tidyr)
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

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

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

```
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 objects are masked from 'package:dplyr':
##
##     combine, intersect, setdiff, union
```

```
## The following objects are masked from 'package:stats':
##
##     IQR, mad, sd, var, xtabs
```

```
## The following objects are masked from 'package:base':
##
##     Filter, Find, Map, Position, Reduce, anyDuplicated, append,
##     as.data.frame, basename, cbind, colnames, dirname, do.call,
##     duplicated, eval, evalq, get, grep, grepl, intersect, is.unsorted,
##     lapply, mapply, match, mget, order, paste, pmax, pmax.int, pmin,
```

```
##     pmin.int, rank, rbind, rownames, sapply, setdiff, sort, table,
##     tapply, union, unique, unsplit, which, which.max, which.min

## 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")'.
```

```
library(scales)
library(ggplot2)
library('DESeq2')
```

```
## Loading required package: S4Vectors

## Loading required package: stats4

##
## Attaching package: 'S4Vectors'

## The following objects are masked from 'package:dplyr':
##
##     first, rename

## The following object is masked from 'package:tidyr':
##
##     expand

## The following object is masked from 'package:base':
##
##     expand.grid

## Loading required package: IRanges

##
## Attaching package: 'IRanges'

## The following objects are masked from 'package:dplyr':
##
##     collapse, desc, slice

## Loading required package: GenomicRanges

## Loading required package: GenomeInfoDb

## Loading required package: SummarizedExperiment

## Loading required package: DelayedArray

## Loading required package: matrixStats

##
## Attaching package: 'matrixStats'

## The following objects are masked from 'package:Biobase':
##
##     anyMissing, rowMedians

## The following object is masked from 'package:dplyr':
##
##     count
```

```
##
## Attaching package: 'DelayedArray'

## The following objects are masked from 'package:matrixStats':
##
##     colMaxs, colMins, colRanges, rowMaxs, rowMins, rowRanges

## The following objects are masked from 'package:base':
##
##     aperm, apply, rowsum
```

The count data is loaded into R.

Table 1: Raw counts data

```
count_data <- "Sample_Subfields_Counts.txt"
counts <- read.table(file = count_data, header = TRUE)
head(counts)
```

```
##         CA1_CTL_1 CA1_CTL_2 CA1_CTL_3 CA1_CTL_4 CA1_CTL_5 CA1_CTL_6 CA1_CTL_7
## A1BG          138       117       201       183       152       223       177
## A1CF            2         5         2         1         0         2         4
## A2M         11307      4847      9954      9716      6690      7470      3251
## A2ML1         464       220       398       453       263       370       264
## A3GALT2        25        18        26        10         9        11        11
## A4GALT         79        64        84        47        44        75        36
##         CA1_CTL_8 CA1_CTL_9 CA1_CTL_10 CA1_CTL_11 CA1_CTL_12 CA1_CTL_13
## A1BG          120       106        182        139        198        138
## A1CF            1         1          9          3          0          1
## A2M          8587      1786      10807      14698      16482       5102
## A2ML1         218       229        891        641        664        724
## A3GALT2        15        82         25         30         13         32
## A4GALT         80        43        108        111         71        109
##         CA1_SZ_1_ON CA1_SZ_2_ON CA1_SZ_3_ON CA1_SZ_4_ON CA1_SZ_5_ON
## A1BG            162          91         151         174         135
## A1CF              0           0           4           1           2
## A2M            5512        6099        6575       14831        6428
## A2ML1           276         336         359         415         351
## A3GALT2          48           4          11          41           7
## A4GALT           82          53          40          62          67
##         CA1_SZ__6_ON CA1_SZ_7_OFF CA1_SZ_8_OFF CA1_SZ_9_OFF CA1_SZ_10_OFF
## A1BG             131          101          143          187           214
## A1CF               2            7            1            2             3
## A2M             5112        29178         2450         3821          8096
## A2ML1            290           38          112          269           829
## A3GALT2           17           18           15           33            25
## A4GALT            23          229           26           57            51
##         CA1_SZ_11_OFF CA1_SZ_12_OFF CA1_SZ_13_OFF CA3_CTL_1 CA3_CTL_2 CA3_CTL_3
## A1BG              237           160           158       106        99        96
## A1CF                2             1             0         2         4         4
## A2M             10538         10048          7513      7990      6763      7065
## A2ML1             423           415           370       537       494       533
## A3GALT2            65            11            10        14        26        15
## A4GALT            165            73            96        67        97        51
##         CA3_CTL_4 CA3_CTL_5 CA3_CTL_6 CA3_CTL_7 CA3_CTL_8 CA3_CTL_9 CA3_CTL_10
## A1BG          128        88        90       119       106        15         13
## A1CF            0         0         2         5         0         6          1
```

```
## A2M          4806      3394      4893      6740      5196       645       3177
## A2ML1         353       209       182       355       241        90        186
## A3GALT2        22        15         9        11        14        23          4
## A4GALT         63        32        53        73        51        25         23
##         CA3_CTL_11 CA3_CTL_12 CA3_CTL_13 CA3_SZ_1_ON CA3_SZ_2_ON CA3_SZ_3_OFF
## A1BG            26         19         11          57          74           74
## A1CF             0         84         13           0           3            1
## A2M           3844       1273       2139        6029        5428         9681
## A2ML1          249        184        143         278         414          188
## A3GALT2          5         23         24          10          12           23
## A4GALT          32         31         30          68         123          129
##         CA3_SZ_4_OFF CA3_SZ_5_ON CA3_SZ_6_ON CA3_SZ_7_ON CA3_SZ_8_ON
## A1BG             135         182         141         133          94
## A1CF               1           0           2           1           3
## A2M             3715        4949        5609       10541        5222
## A2ML1            256         406         350         405         251
## A3GALT2           15          31          29          18          29
## A4GALT            78          58          60         103          28
##         CA3_SZ_9_OFF CA3_SZ_10_OFF CA3_SZ_11_OFF CA3_SZ_12_OFF CA3_SZ_13_OFF
## A1BG              18            17            13            10            17
## A1CF               1             1             2             2             1
## A2M              535          2937          1603          1289          1800
## A2ML1             59            84           134           125           240
## A3GALT2           16             7             8             5            31
## A4GALT            21            26            27            10            40
##         DG_CTL_1 DG_CTL_2 DG_CTL_3 DG_CTL_4 DG_CTL_5 DG_CTL_6 DG_CTL_7 DG_CTL_8
## A1BG          60       81      118       72      113      248       64      109
## A1CF           4        3        2        3        0        0        3        1
## A2M         7138     6770     7227     4404     4948     8029     3901     6646
## A2ML1        522      444      430      265      266      413      128      239
## A3GALT2       17       15       30        3       17       28        5       15
## A4GALT        37       88       75       41       39       97       41       48
##         DG_CTL_9 DG_CTL_10 DG_CTL_11 DG_CTL_12 DG_CTL_13 DG_SZ_1_ON DG_SZ_2_ON
## A1BG         140       142        77        87        94         83        113
## A1CF           2         1         1         0         1          1          3
## A2M         3181     10757     10016      8633      7695       4269       9368
## A2ML1        339       359       532       406       467        194        488
## A3GALT2       72        15        25        15        21         49         24
## A4GALT        48        53        90        47       132         49        107
##         DG_SZ_3_ON DG_SZ_4_ON DG_SZ_5_OFF DG_SZ_6_OFF DG_SZ_7_ON DG_SZ_8_ON
## A1BG            84         69          68          58        102         16
## A1CF             1          0           3           2          3          4
## A2M           5985       5565        9162        2308       5571        942
## A2ML1          245        314          86         107        349         32
## A3GALT2         22          5          13           8         46          2
## A4GALT          35        102         109          25         28          7
##         DG_SZ_9_OFF DG_SZ_10_OFF DG_SZ_11_OFF DG_SZ_12_OFF DG_SZ_13_OFF
## A1BG             76           81          166          214          164
## A1CF              1            3            1            0            0
## A2M            2482         6924         8688         9412         7300
## A2ML1           265          423          369          478          249
## A3GALT2          26           15           20           12            2
## A4GALT           44           50           78           27           52
```

```
dim(counts)
```

```
## [1] 20268    78
```

```
#str(counts)

control_CA1 <- c(1:13)
case_CA1 <- c(14:26)
control_CA3 <- c(27:39)
case_CA3 <- c(40:52)
control_DG <- c(53:65)
case_DG <- c(66:78)


counts_CA1 <- counts[, 1:26]
#counts_CA1
counts_CA3 <- counts[, 27:52]
#counts_CA3
counts_DG <- counts[, 53:78]
#counts_DG
```

The data of the counts was split based on the hippocampal subfield and tidied using tidyr.

Table 2: Tidy CA1 data

```
counts_CA1_tidy <- pivot_longer(data = counts_CA1,
             cols=1:26,
             #names_pattern="(CA1_CTL|CA1_SZ).",
             names_to ="sample",
             values_to = "value")
head(counts_CA1_tidy)
```

```
## # A tibble: 6 x 2
##   sample    value
##   <chr>     <int>
## 1 CA1_CTL_1   138
## 2 CA1_CTL_2   117
## 3 CA1_CTL_3   201
## 4 CA1_CTL_4   183
## 5 CA1_CTL_5   152
## 6 CA1_CTL_6   223
```

Table 3: Tidy CA3 data

```
counts_CA3_tidy <- pivot_longer(data = counts_CA3,
             cols=1:26,
             #names_pattern="(CA1_CTL|CA1_SZ).",
             names_to ="sample",
             values_to = "value")
head(counts_CA3_tidy)
```

```
## # A tibble: 6 x 2
##   sample    value
##   <chr>     <int>
## 1 CA3_CTL_1   106
## 2 CA3_CTL_2    99
## 3 CA3_CTL_3    96
```

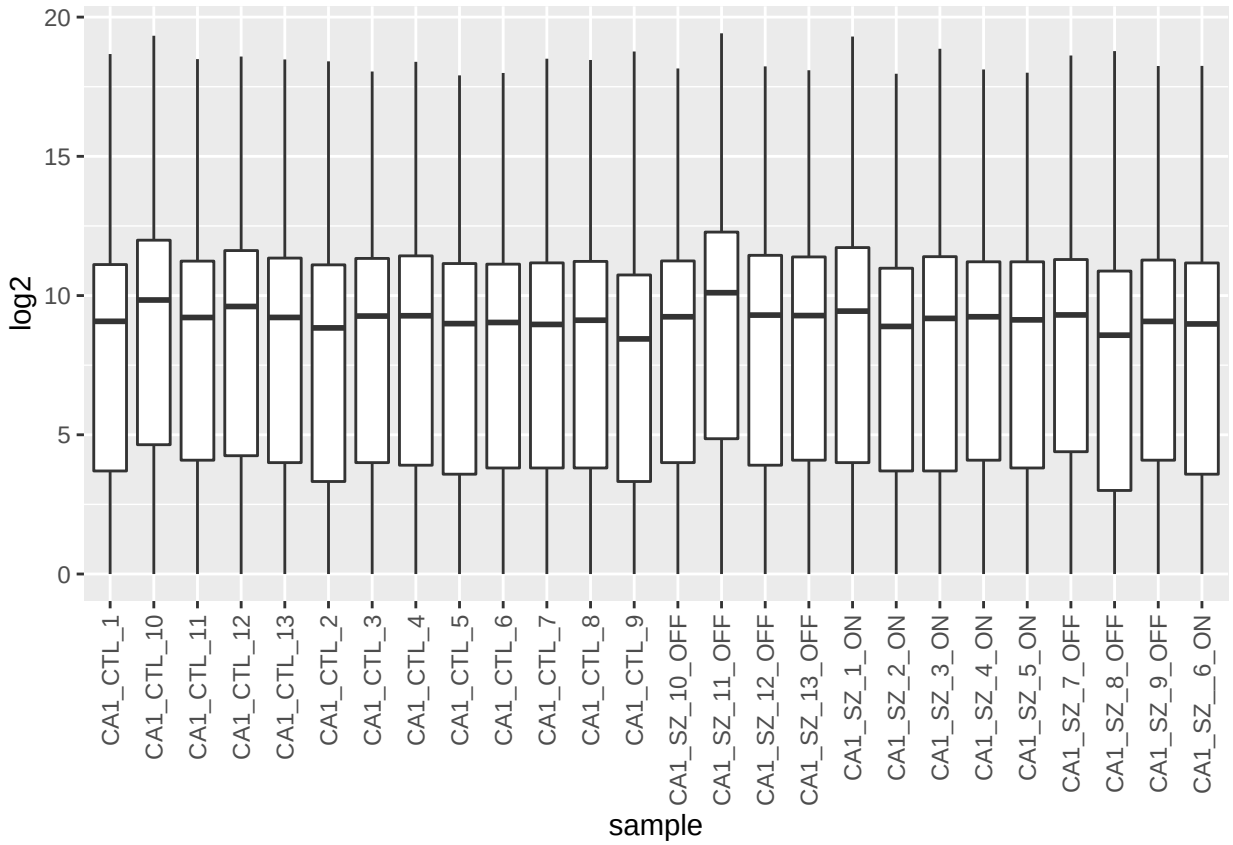
```
## 4 CA3_CTL_4   128
## 5 CA3_CTL_5    88
## 6 CA3_CTL_6    90
```

Table 4: Tidy DG data

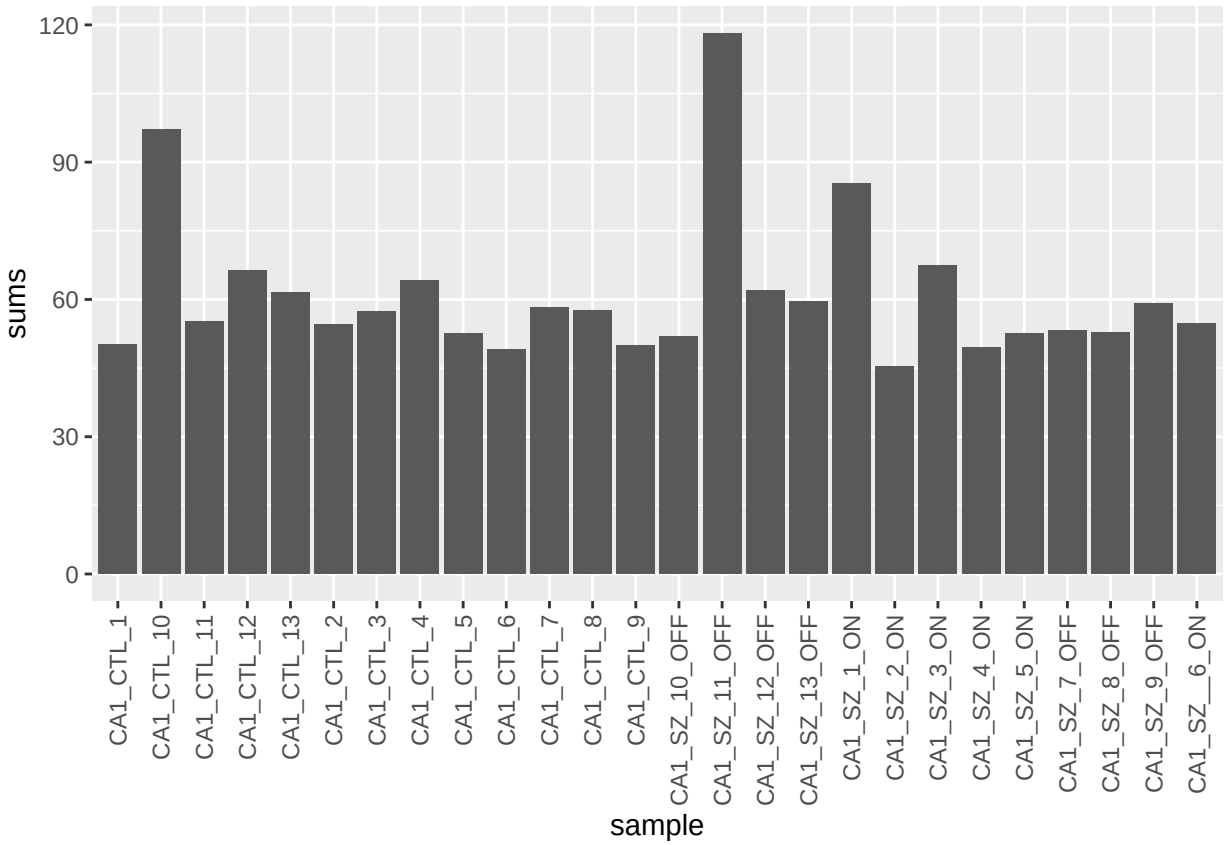
```
counts_DG_tidy <- pivot_longer(data = counts_DG,
             cols=1:26,
             #names_pattern="(CA1_CTL|CA1_SZ).",
             names_to ="sample",
             values_to = "value")
head(counts_DG_tidy)
```

```
## # A tibble: 6 x 2
##   sample   value
##   <chr>    <int>
## 1 DG_CTL_1    60
## 2 DG_CTL_2    81
## 3 DG_CTL_3   118
## 4 DG_CTL_4    72
## 5 DG_CTL_5   113
## 6 DG_CTL_6   248
```

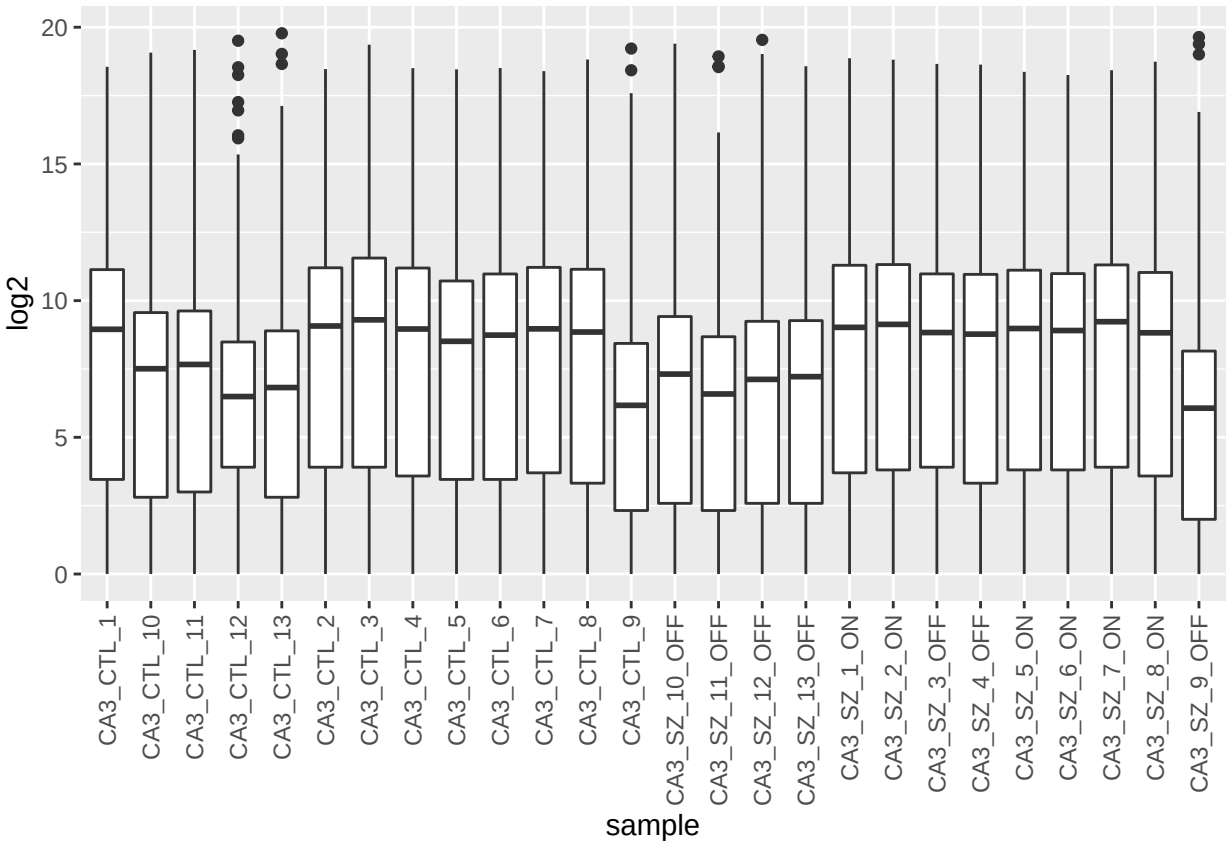
```
counts_CA1_tidy$log2 <- log2(counts_CA1_tidy$value + 1)
counts_CA3_tidy$log2 <- log2(counts_CA3_tidy$value + 1)
counts_DG_tidy$log2 <- log2(counts_DG_tidy$value + 1)

ggplot(data = counts_CA1_tidy, mapping = aes(x = sample, y = log2))+
geom_boxplot()+
  scale_x_discrete(guide = guide_axis(angle = 90))
```



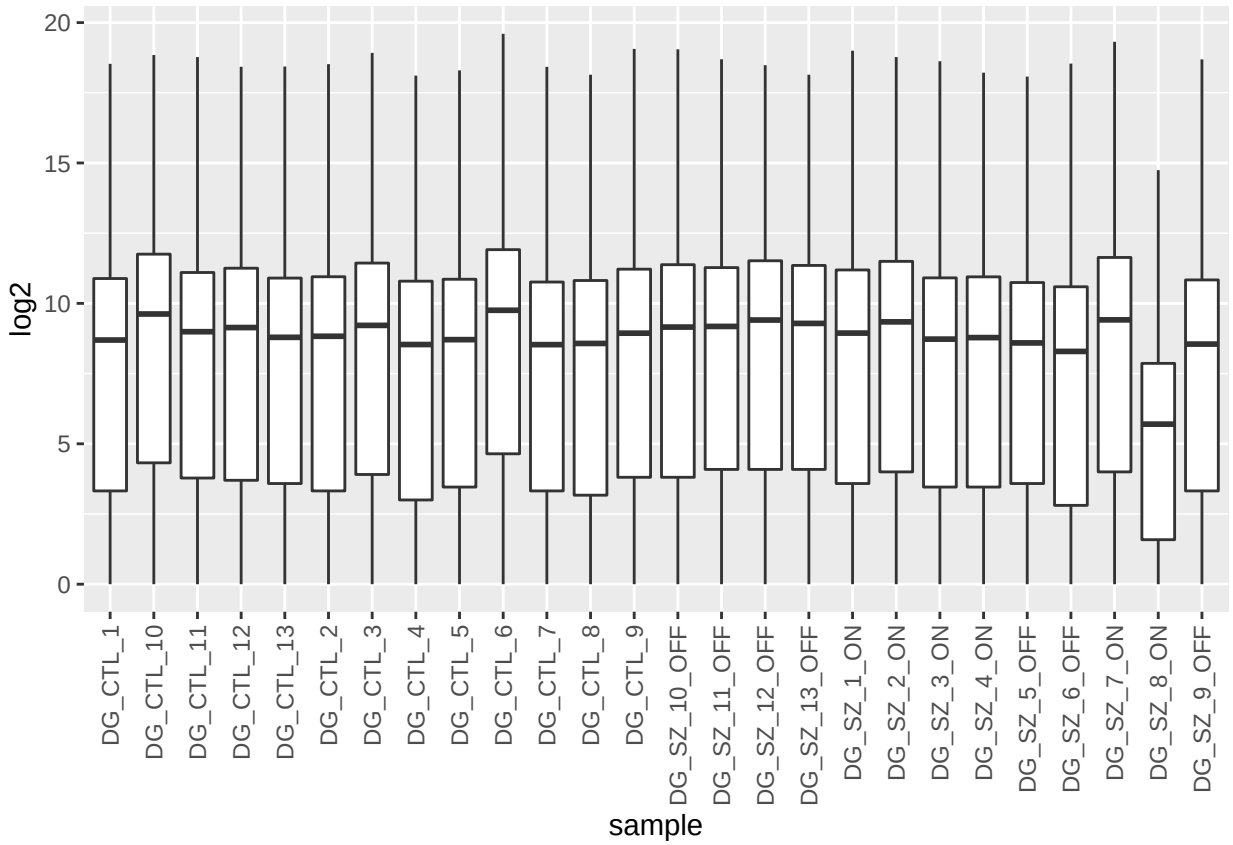

```
sums <- colSums(counts_CA1)/1e6
CA1_sequence_depth = data.frame(sample=names(sums), depth=sums)
CA1_sequence_depth %>% ggplot(mapping = aes(x = sample, y=sums)) + geom_col() +
   scale_x_discrete(guide = guide_axis(angle = 90))
```



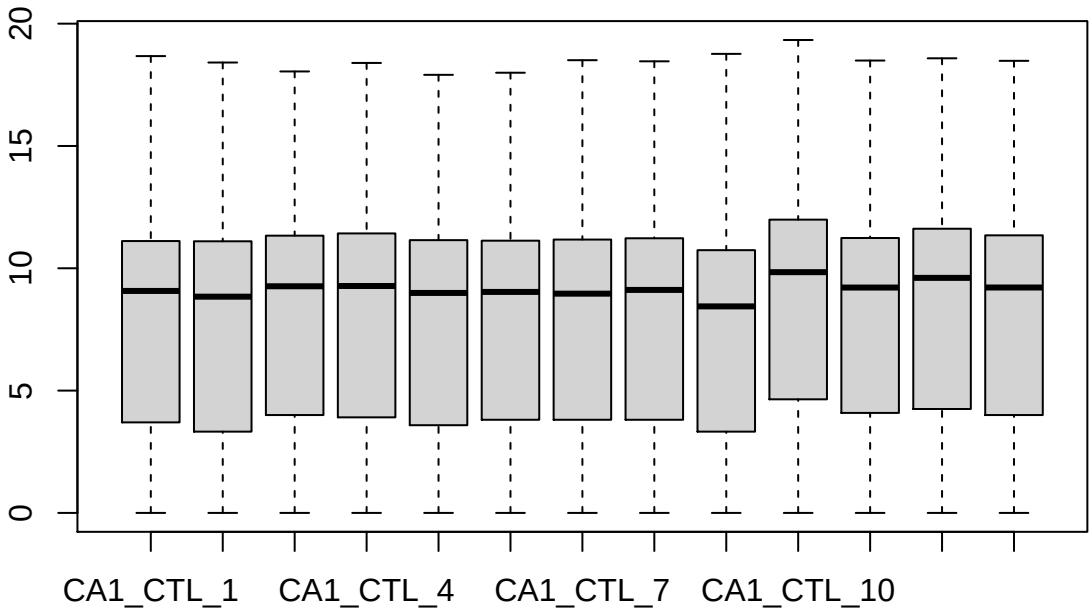

```
ggplot(data = counts_CA3_tidy, mapping = aes(x = sample, y = log2))+
geom_boxplot()+
  scale_x_discrete(guide = guide_axis(angle = 90))
```



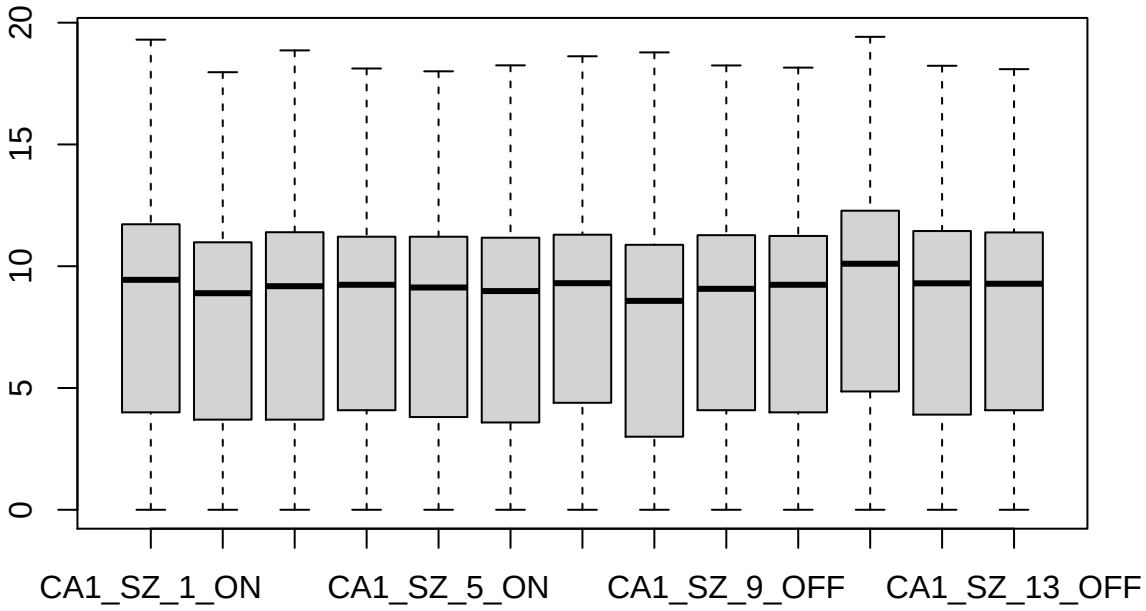

```
ggplot(data = counts_DG_tidy, mapping = aes(x = sample, y = log2))+
geom_boxplot()+
scale_x_discrete(guide = guide_axis(angle = 90))
```



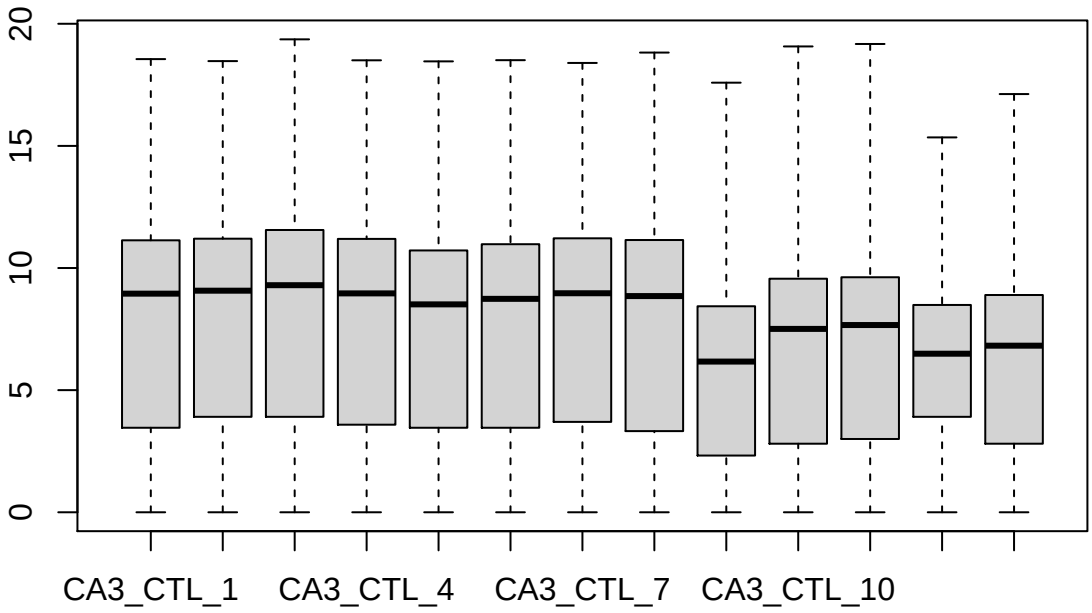

```
control_ca1 <- c(1:13)
case_ca1 <- c(14:26)
control_ca3 <- c(27:39)
case_ca3 <- c(40:52)
control_dg <- c(53:65)
case_dg <- c(66:78)

boxplot(log2(counts[control_ca1] + 1), outline = F, cex.names = 0.2)
```



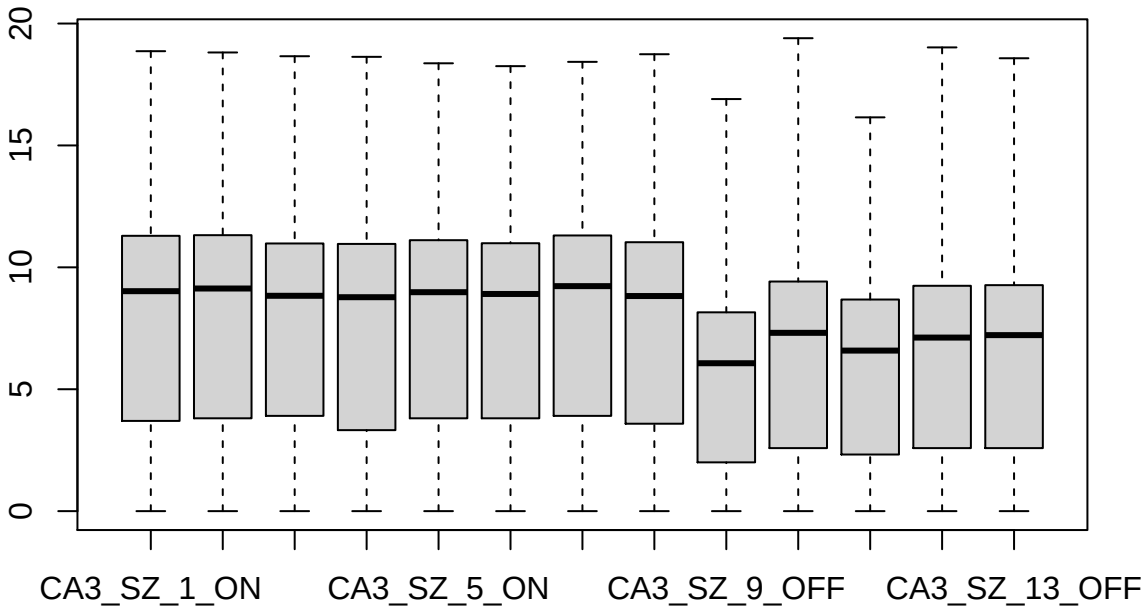

```
boxplot(log2(counts[case_ca1] + 1), outline = F, cex.names = 0.2)
```



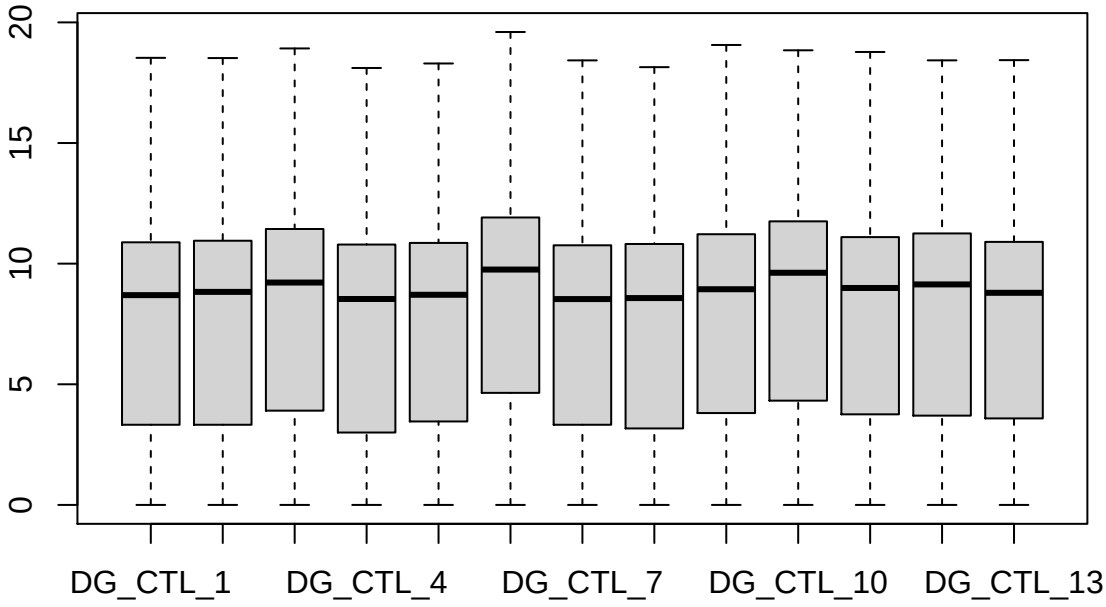

```
boxplot(log2(counts[control_ca3] + 1), outline = F, cex.names = 0.2)
```



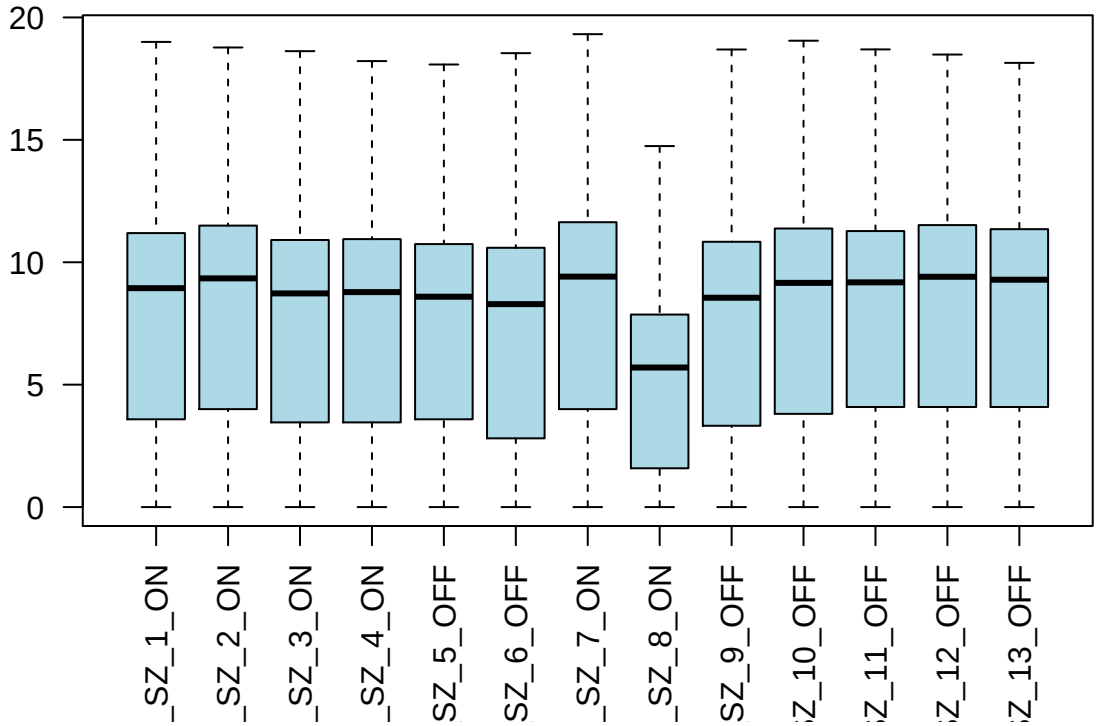

```
boxplot(log2(counts[case_ca3] + 1), outline = F, cex.names = 0.2)
```



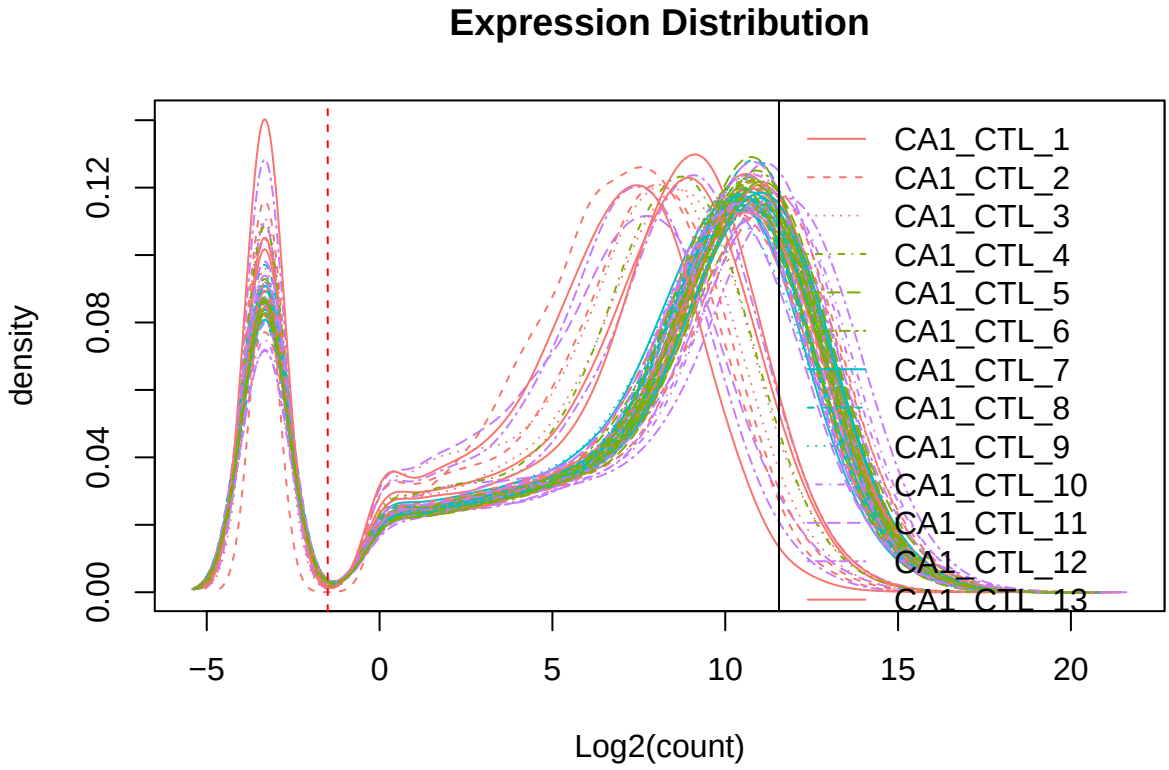

```
boxplot(log2(counts[control_dg] + 1), outline = F, cex.names = 0.2)
```



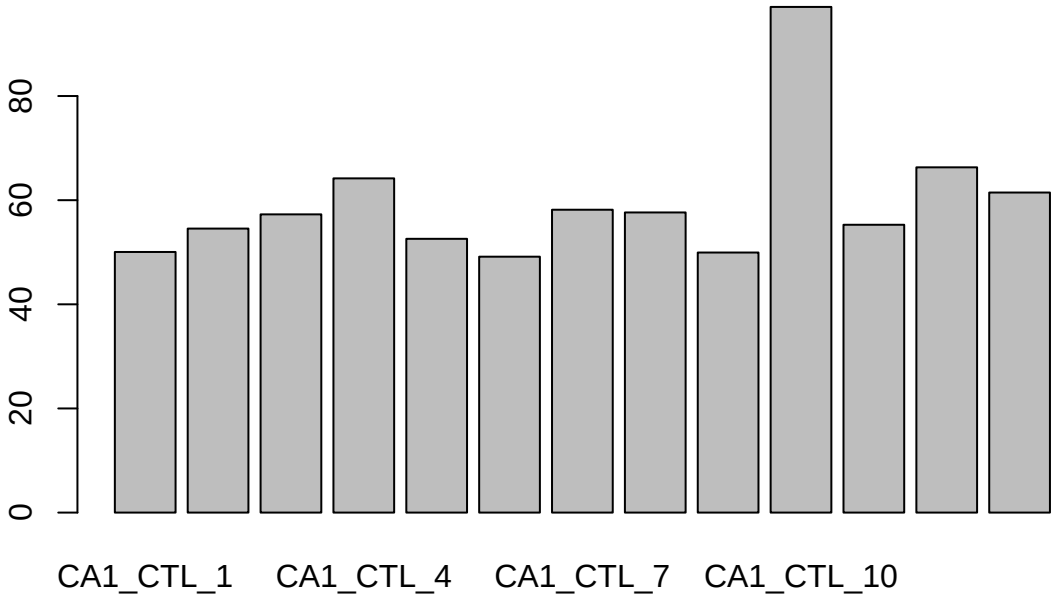

```
boxplot(log2(counts[case_dg] + 1), outline = F, cex.lab = 0.2, col = "lightblue", las = 2)
```



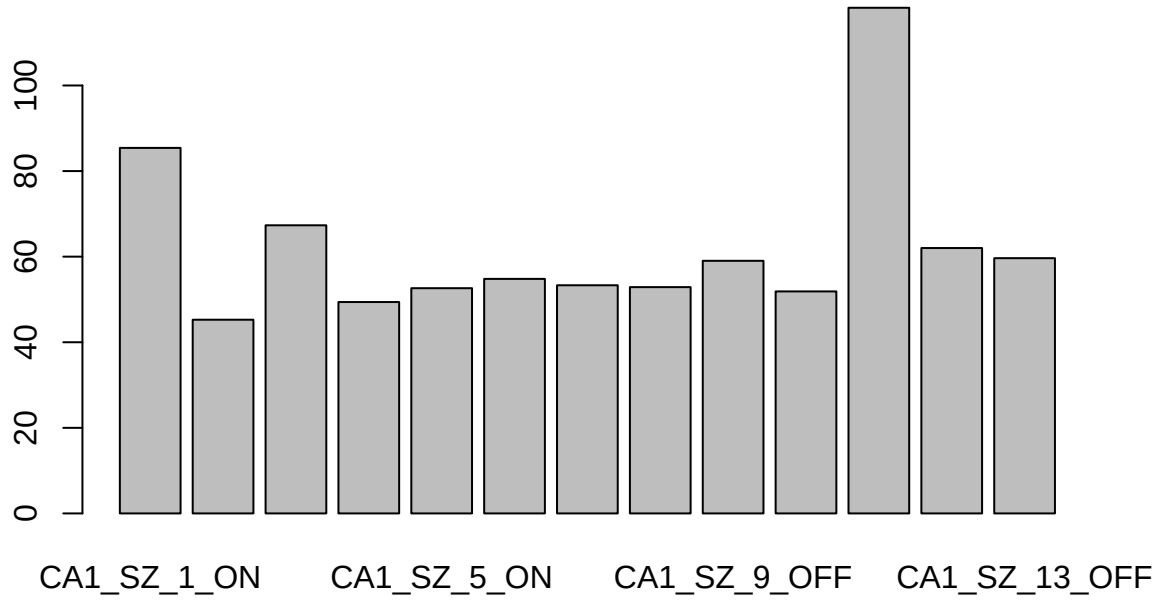

```
myColors <- hue_pal()(4)

plotDensity(log2(counts + 0.1), col=rep(myColors, each=3),
            lty=c(1:ncol(counts)), xlab='Log2(count)',
            main='Expression Distribution')

legend('topright', names(counts), lty=c(1:ncol(counts)),
       col=rep(myColors, each=3))
abline(v=-1.5, lwd=1, col='red', lty=2)
```



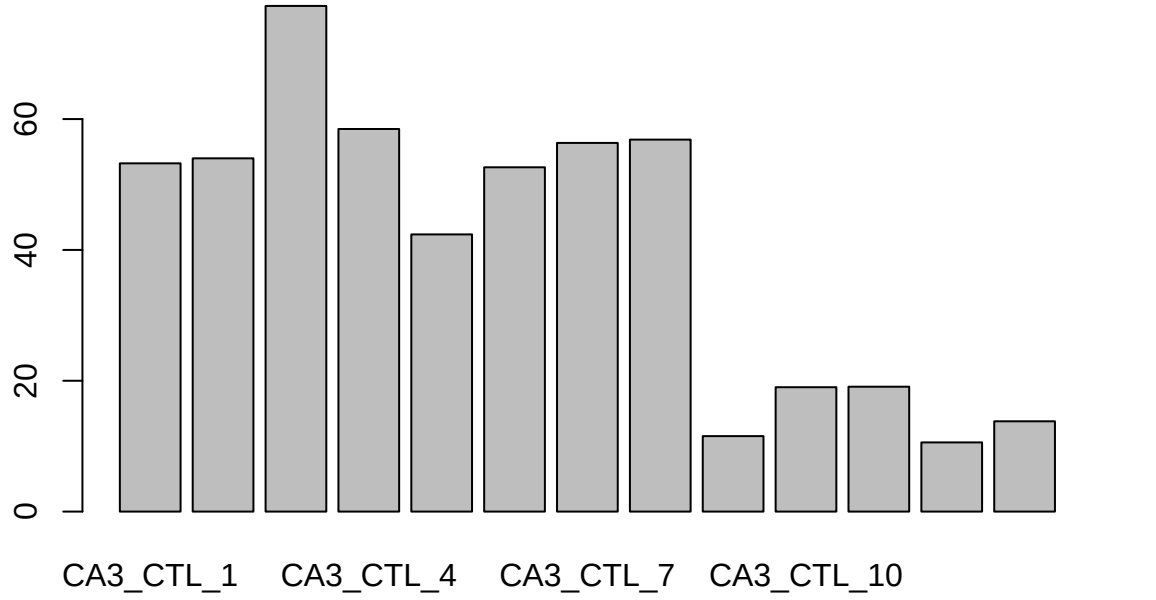

```
barplot(colSums(counts[control_ca1]) / 1e6, col = )
```



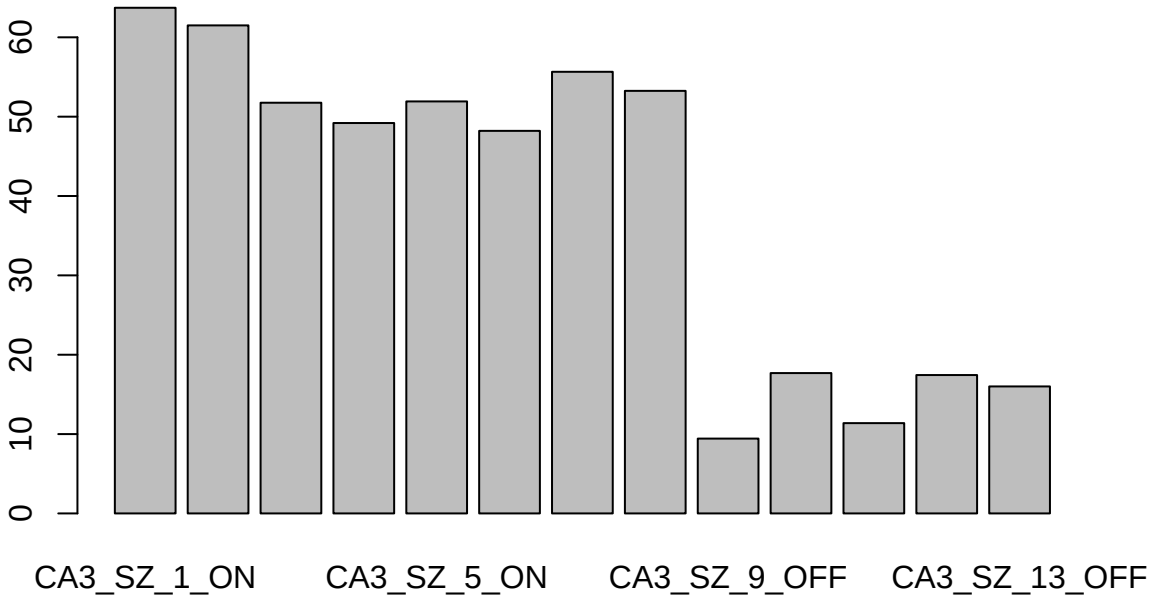

```
barplot(colSums(counts[case_ca1]) / 1e6, col = )
```



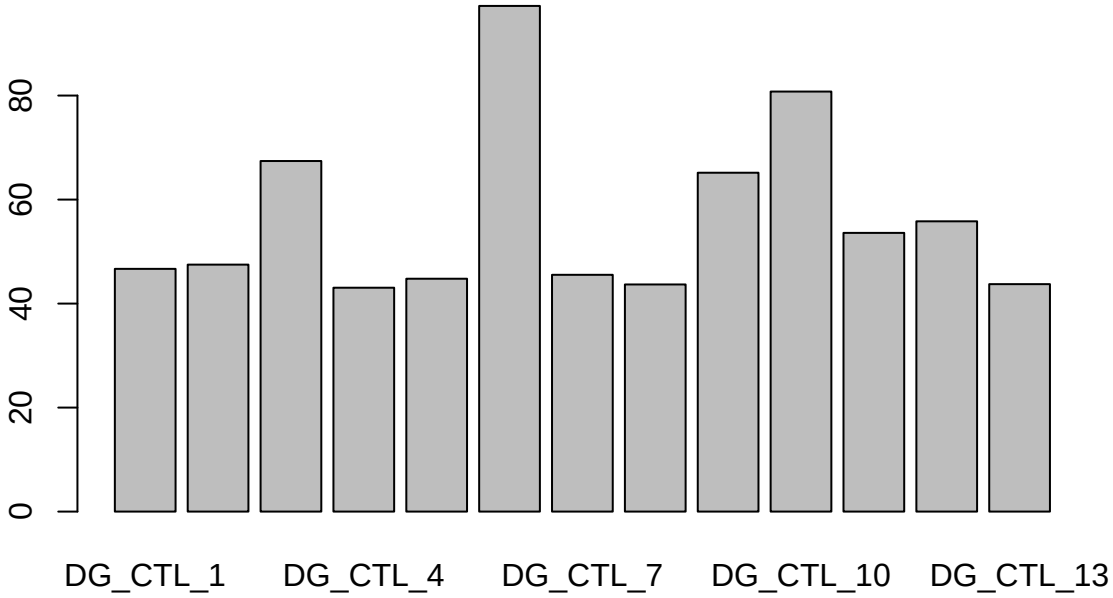

```
barplot(colSums(counts[control_ca3]) / 1e6, col = )
```



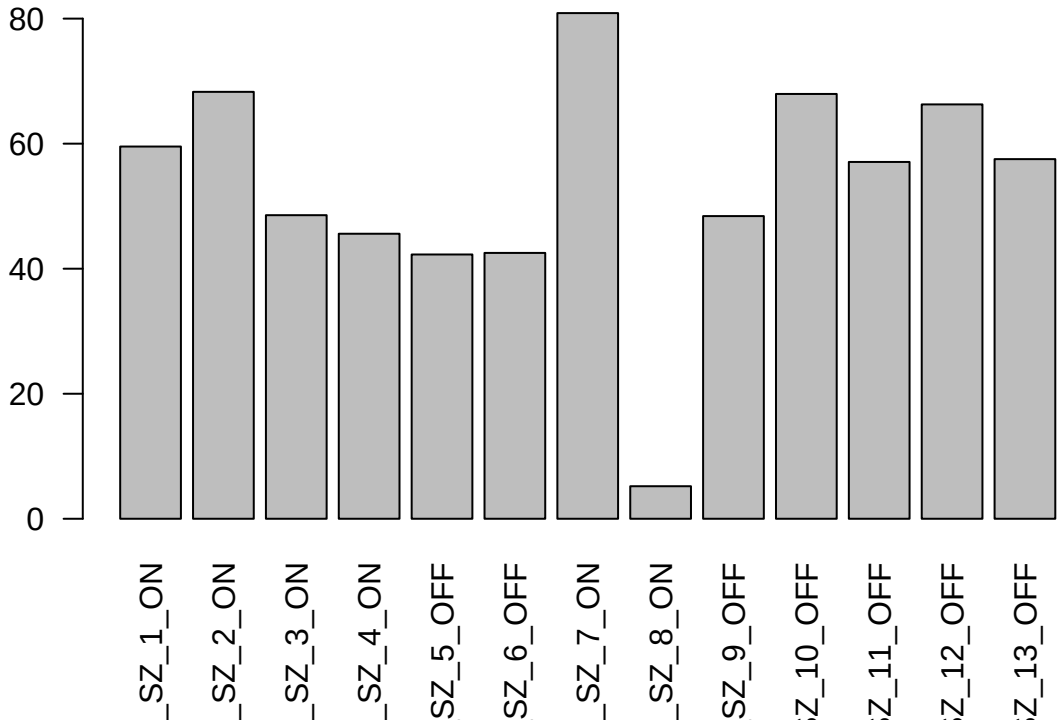

```
barplot(colSums(counts[case_ca3]) / 1e6, col = )
```




```
barplot(colSums(counts[control_dg]) / 1e6, col = )
```




```
barplot(colSums(counts[case_dg]) / 1e6, col = , las = 2)
```




```
(ddsMat <- DESeqDataSetFromMatrix(countData = counts,
                                  colData = data.frame(samples = names(counts)),
                                  design = ~ 1))
```

```
## class: DESeqDataSet
## dim: 20268 78
## metadata(1): version
## assays(1): counts
## rownames(20268): A1BG A1CF ... ZZEF1 ZZZ3
## rowData names(0):
## colnames(78): CA1_CTL_1 CA1_CTL_2 ... DG_SZ_12_OFF DG_SZ_13_OFF
## colData names(1): samples
```

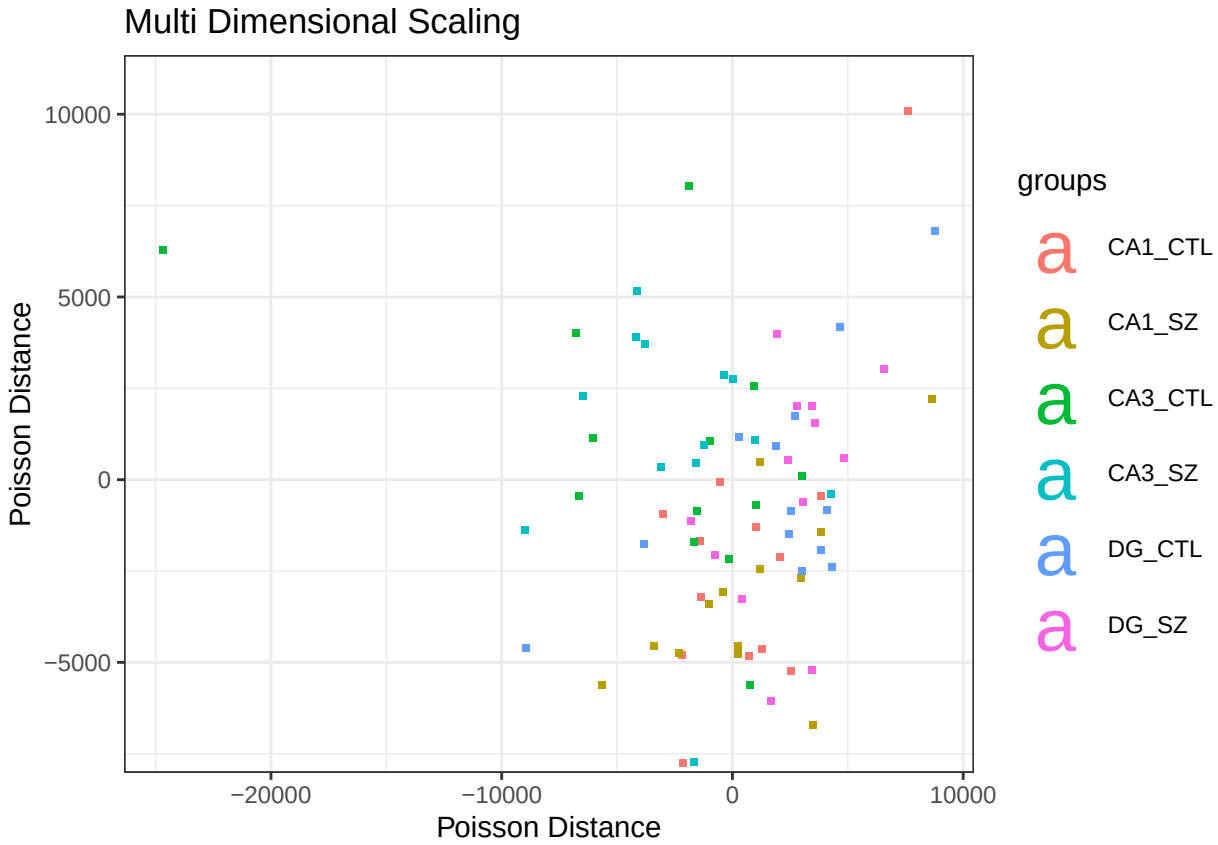
```
# Perform normalization
rld.dds <- vst(ddsMat)
# 'Extract' normalized values
rld <- assay(rld.dds)
sampledists <- dist( t( rld ))
```

```
annotation <- data.frame(subfield = factor(rep(1:3, each = 26),
                                           labels = c("CA1", "CA3", "DG")),
                         type = factor(rep(rep(1:2, each = 13), 3),
                                       labels = c("control", "schizophrenia")))

# Set the rownames of the annotation dataframe to the sample names (required)
rownames(annotation) <- names(counts)
```

```
library('PoiClaClu')
# Note: uses the raw-count data, PoissonDistance performs normalization
# set by the 'type' parameter (uses DESeq)
dds <- assay(ddsMat)
poisd <- PoissonDistance( t(dds), type = "deseq")
```

```
# Extract the matrix with distances
samplePoisDistMatrix <- as.matrix(poisd$dd)
# Calculate the MDS and get the X- and Y-coordinates
mdsPoisData <- data.frame( cmdscale(samplePoisDistMatrix) )

# And set some better readable names for the columns
names(mdsPoisData) <- c('x_coord', 'y_coord')
```

```
groups <- factor(rep(1:6, each=13),
                 labels = c("CA1_CTL", "CA1_SZ", "CA3_CTL", "CA3_SZ", "DG_CTL", "DG_SZ"))
coldata <- names(counts)

# Create the plot using ggplot
ggplot(mdsPoisData, aes(x_coord, y_coord, color = groups, label = ".")) +
  geom_text(size = 10) +
  ggtitle('Multi Dimensional Scaling') +
  labs(x = "Poisson Distance", y = "Poisson Distance") +
  theme_bw()
```



TODOs voor kwaliteitscontrole: - normaliseren met DESeq2 vst() - annotation dataframe maken (zie 3.4.3) - MDS (eerst in 1 plot)