Analysis of Gene Expression

Transcriptional signature of prion-induced neurotoxicity in a Drosophila model of transmissible mammalian prion disease.

Niek Scholten

2022-05-26

Contents

1 Setup

```
# Options for all chunks
knitr::opts_chunk$set(echo = TRUE)
knitr::opts_chunk$set(cache = TRUE)

# Load the packages & register the amount of workers
packages <- c("affy", "scales",
              "DESeq2", "BiocParallel",
              "pheatmap", "PoiClaClu",
              "ggplot2", "edgeR",
              "knitr", "pander",
              "EnhancedVolcano", "crayon")
invisible(lapply(packages, library, character.only = TRUE))
register(MulticoreParam(12))

# Load the data into a data frame
data <- read.table("Data/GSE144028.txt")

# Define groups for the replicants
group <- c("X51D_5_NBH",
           "X51D_5_S",
           "X51D_30_NBH",
           "X51D_30_S",
           "PrPCyt_5_NBH",
           "PrPCyt_5_S",
           "PrPCyt_30_NBH",
           "PrPCyt_30_S",
           "PrPGPI_5_NBH",
           "PrPGPI_5_S",
           "PrPGPI_40_NBH",
           "PrPGPI_40_S")
groups <- factor(rep(1:12, each=3),
                 labels = group)

# Set color distributions for the graphs
colors12 <- hue_pal()(12)
colors36 <- rep(colors12, each=3)
```

This is the setup of the project. It loads all the necessary packages and sets values that are important for later.

2 Initial analysis

The initial analysis includes a summary of the data and a quick look at the visualisation of this data in a boxplot.

2.1 Summary

```r
# Disable intertable text
panderOptions('table.continues', '')
# Pretty print the output of the data summary
pander(summary(data), split.tables = 64)
```

X51D_30_NBH_1	X51D_30_NBH_2	X51D_30_NBH_3
Min. : 0	Min. : 0	Min. : 0
1st Qu.: 0	1st Qu.: 0	1st Qu.: 0
Median : 7	Median : 18	Median : 19
Mean : 377	Mean : 972	Mean : 1030
3rd Qu.: 75	3rd Qu.: 200	3rd Qu.: 209
Max. :3445037	Max. :8342368	Max. :8875291

X51D_30_S_1	X51D_30_S_2	X51D_30_S_3
Min. : 0	Min. : 0	Min. : 0
1st Qu.: 0	1st Qu.: 0	1st Qu.: 0
Median : 10	Median : 10	Median : 10
Mean : 568	Mean : 480	Mean : 509
3rd Qu.: 113	3rd Qu.: 104	3rd Qu.: 107
Max. :5560520	Max. :4122340	Max. :4386825

X51D_5_NBH_1	X51D_5_NBH_2	X51D_5_NBH_3
Min. : 0	Min. : 0.0	Min. : 0.0
1st Qu.: 0	1st Qu.: 0.0	1st Qu.: 0.0
Median : 26	Median : 21.0	Median : 21.5
Mean : 869	Mean : 688.3	Mean : 718.8
3rd Qu.: 388	3rd Qu.: 325.0	3rd Qu.: 337.0
Max. :3832490	Max. :2415360.0	Max. :2533918.0

X51D_5_S_1	X51D_5_S_2	X51D_5_S_3
Min. : 0.0	Min. : 0	Min. : 0
1st Qu.: 0.0	1st Qu.: 1	1st Qu.: 1
Median : 31.0	Median : 89	Median : 92
Mean : 722.4	Mean : 1999	Mean : 2092
3rd Qu.: 320.0	3rd Qu.: 925	3rd Qu.: 961
Max. :3111359.0	Max. :7272134	Max. :7625567

PrPCyt_30_NBH_1	PrPCyt_30_NBH_2	PrPCyt_30_NBH_3
Min. : 0	Min. : 0.0	Min. : 0.0
1st Qu.: 0	1st Qu.: 0.0	1st Qu.: 0.0
Median : 22	Median : 4.0	Median : 4.0
Mean : 855	Mean : 176.5	Mean : 181.1
3rd Qu.: 254	3rd Qu.: 52.0	3rd Qu.: 53.0
Max. :5261726	Max. :1059586.0	Max. :1096115.0

PrPCyt_30_S_1	PrPCyt_30_S_2	PrPCyt_30_S_3
Min. : 0	Min. : 0	Min. : 0
1st Qu.: 0	1st Qu.: 0	1st Qu.: 0
Median : 23	Median : 27	Median : 28
Mean : 793	Mean : 857	Mean : 914
3rd Qu.: 299	3rd Qu.: 351	3rd Qu.: 374
Max. :4058764	Max. :3769299	Max. :4079216

PrPCyt_5_NBH_1	PrPCyt_5_NBH_2	PrPCyt_5_NBH_3
Min. : 0	Min. : 0.0	Min. : 0.0
1st Qu.: 0	1st Qu.: 0.0	1st Qu.: 0.0
Median : 43	Median : 29.0	Median : 30.0
Mean : 828	Mean : 591.5	Mean : 603.4
3rd Qu.: 421	3rd Qu.: 286.0	3rd Qu.: 294.0
Max. :3163765	Max. :2692026.0	Max. :2734069.0

PrPCyt_5_S_1	PrPCyt_5_S_2	PrPCyt_5_S_3
Min. : 0	Min. : 0.0	Min. : 0.0
1st Qu.: 0	1st Qu.: 0.0	1st Qu.: 0.0
Median : 60	Median : 31.0	Median : 32.0
Mean : 1537	Mean : 764.7	Mean : 821.5
3rd Qu.: 838	3rd Qu.: 403.8	3rd Qu.: 435.0
Max. :4603176	Max. :2386987.0	Max. :2556960.0

PrPGPI_40_NBH_1	PrPGPI_40_NBH_2	PrPGPI_40_NBH_3
Min. : 0	Min. : 0	Min. : 0
1st Qu.: 0	1st Qu.: 0	1st Qu.: 0
Median : 14	Median : 13	Median : 10
Mean : 1556	Mean : 1521	Mean : 1116
3rd Qu.: 160	3rd Qu.: 150	3rd Qu.: 115
Max. :18885278	Max. :18935887	Max. :13407360

PrPGPI_40_S_1	PrPGPI_40_S_2	PrPGPI_40_S_3
Min. : 0	Min. : 0	Min. : 0
1st Qu.: 0	1st Qu.: 0	1st Qu.: 0
Median : 14	Median : 17	Median : 13

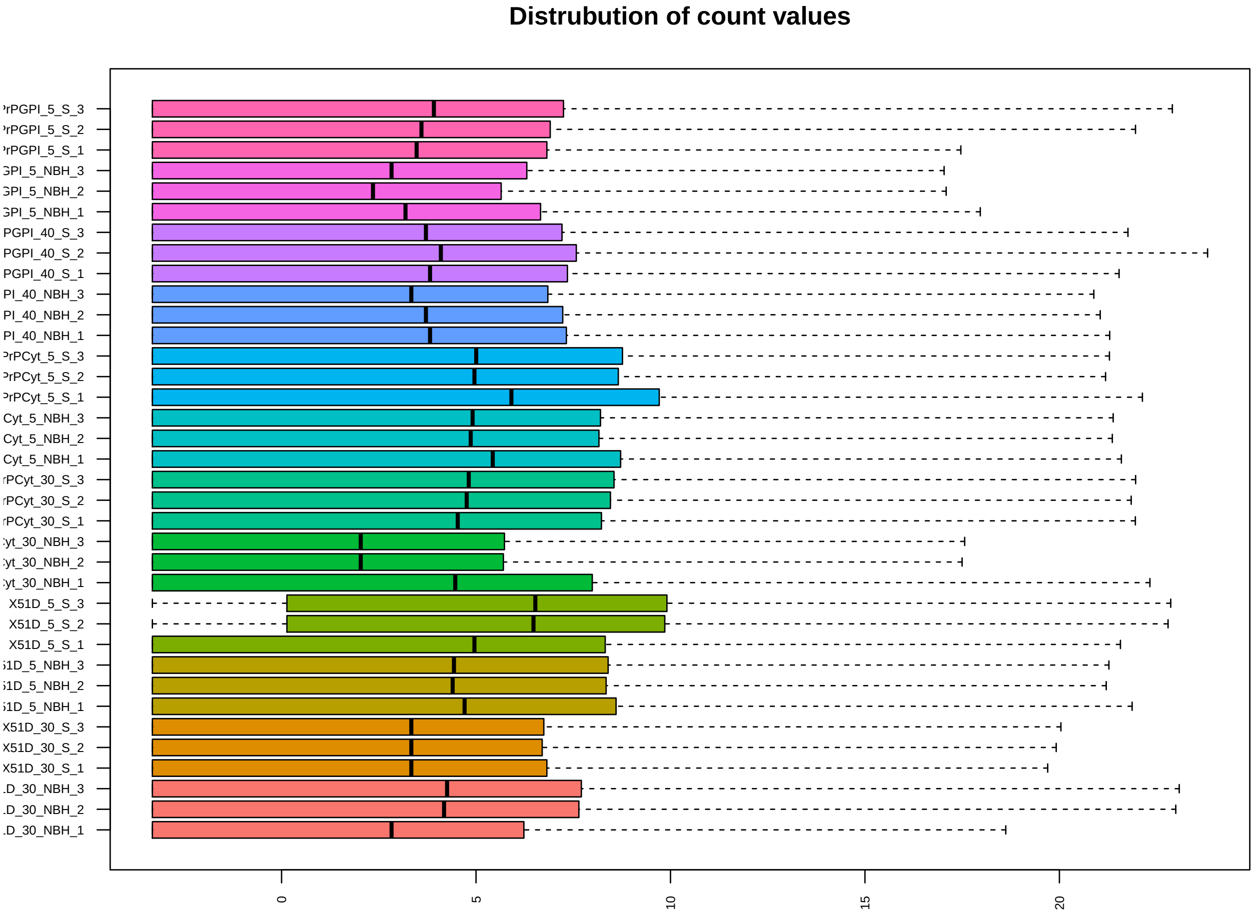
PrPGPI_40_S_1	PrPGPI_40_S_2	PrPGPI_40_S_3
Mean : 1235	Mean : 1318	Mean : 979
3rd Qu.: 163	3rd Qu.: 191	3rd Qu.: 148
Max. :14289546	Max. :14709751	Max. :9635362

PrPGPI_5_NBH_1	PrPGPI_5_NBH_2	PrPGPI_5_NBH_3
Min. : 0	Min. : 0	Min. : 0
1st Qu.: 0	1st Qu.: 0	1st Qu.: 0
Median : 9	Median : 5	Median : 7
Mean : 1077	Mean : 629	Mean : 968
3rd Qu.: 101	3rd Qu.: 50	3rd Qu.: 79
Max. :11252267	Max. :6579166	Max. :10120453

PrPGPI_5_S_1	PrPGPI_5_S_2	PrPGPI_5_S_3
Min. : 0	Min. : 0	Min. : 0
1st Qu.: 0	1st Qu.: 0	1st Qu.: 0
Median : 11	Median : 12	Median : 15
Mean : 793	Mean : 782	Mean : 1018
3rd Qu.: 113	3rd Qu.: 120	3rd Qu.: 152
Max. :6111197	Max. :6215874	Max. :7851434

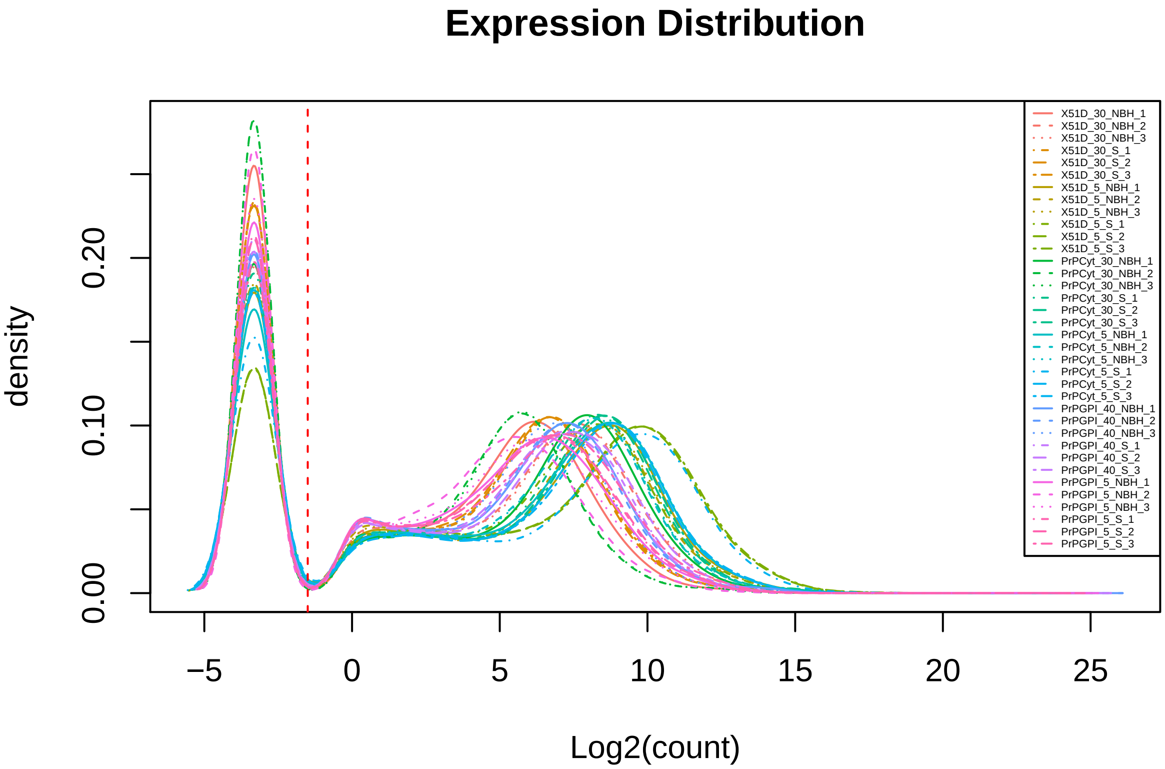
2.2 Boxplot

```r
# Create a boxplot for initial analysis
boxplot(log2(data+0.1),
        outline = FALSE,
        col = colors36,
        horizontal = TRUE,
        las = 2,
        main = "Distrubution of count values",
        cex.axis= 0.6)
```



2.3 Density plot

```
myColors <- hue_pal()(12)

plotDensity(log2(data + 0.1), col=colors36,
            lty= seq_len(ncol(data)), xlab="Log2(count)",
            main="Expression Distribution")

legend('topright', names(data), lty= seq_len(ncol(data)),
       col=colors36,
       cex=0.32)  # Fix scale for knitted output
abline(v=-1.5, lwd=1, col='red', lty=2)
```



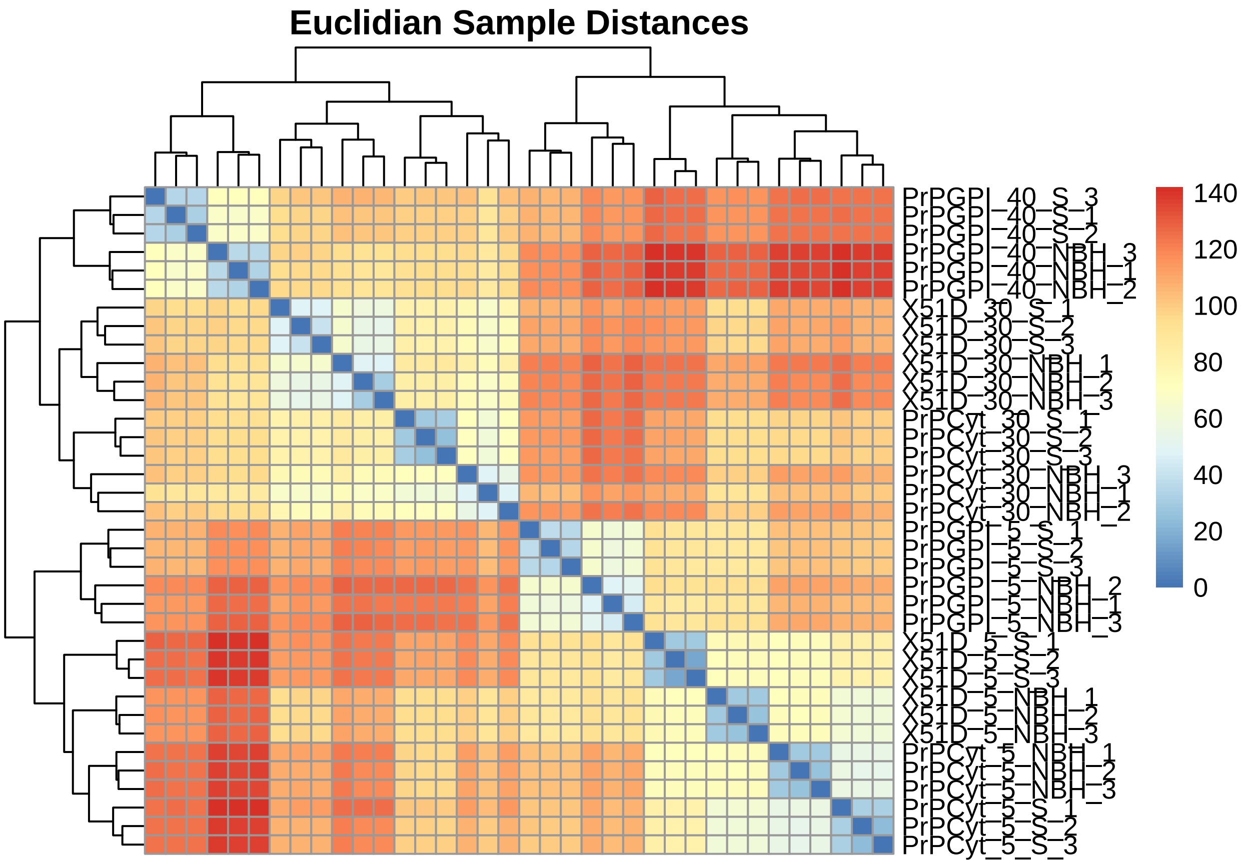
2.4 Heatmap

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

```
## class: DESeqDataSet
## dim: 17742 36
## metadata(1): version
## assays(1): counts
## rownames(17742): FBgn0000003 FBgn0000008 ... __not_aligned
##   __too_low_aQual
## rowData names(0):
## colnames(36): X51D_30_NBH_1 X51D_30_NBH_2 ... PrPGPI_5_S_2 PrPGPI_5_S_3
## colData names(1): samples
```

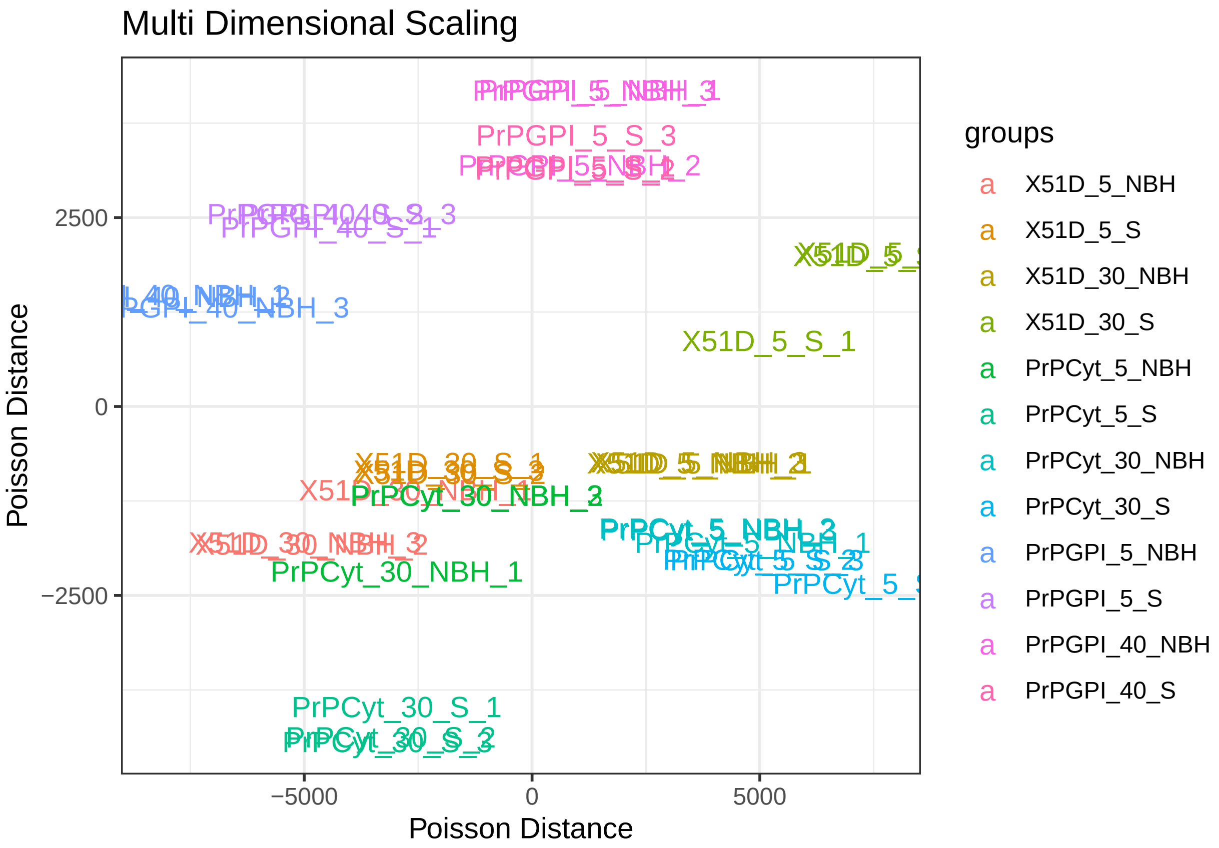
```
rld.dds <- vst(ddsMat)
rld <- assay(rld.dds)

sampledists <- dist( t( rld ))
```

```
sampleDistMatrix <- as.matrix(sampledists)

annotation <- data.frame(Type = factor(rep(rep(1:2, each = 3), each = 6),
                                          labels = c("Normal Brain Homogenate",
                                                     "Scrapie")))

rownames(annotation) <- names(counts)

pheatmap(sampleDistMatrix, show_colnames = FALSE,
         # annotation_col = annotation,  # Gives an error
         clustering_distance_rows = sampledists,
         clustering_distance_cols = sampledists,
         main = "Euclidian Sample Distances")
```


Euclidian Sample Distances
PrPGPI_40_S_3
PrPGPI_40_S_1
PrPGPI_40_S_2
PrPGPI_40_NBH_3
PrPGPI_40_NBH_1
PrPGPI_40_NBH_2
X51D_30_S_1
X51D_30_S_2
X51D_30_S_3
X51D_30_NBH_1
X51D_30_NBH_2
X51D_30_NBH_3
PrPCyt_30_S_1
PrPCyt_30_S_2
PrPCyt_30_S_3
PrPCyt_30_NBH_3
PrPCyt_30_NBH_1
PrPCyt_30_NBH_2
PrPGPI_5_S_1
PrPGPI_5_S_2
PrPGPI_5_S_3
PrPGPI_5_NBH_2
PrPGPI_5_NBH_1
PrPGPI_5_NBH_3
X51D_5_S_1
X51D_5_S_2
X51D_5_S_3
X51D_5_NBH_1
X51D_5_NBH_2
X51D_5_NBH_3
PrPCyt_5_NBH_1
PrPCyt_5_NBH_2
PrPCyt_5_NBH_3
PrPCyt_5_S_1
PrPCyt_5_S_2
PrPCyt_5_S_3
140
120
100
80
60
40
20
0

2.5 Multi dimensional scaling

```r
dds <- assay(ddsMat)
poisd <- PoissonDistance( t(dds) )

samplePoisDistMatrix <- as.matrix(poisd$dd)

mdsPoisData <- data.frame( cmdscale(samplePoisDistMatrix) )

names(mdsPoisData) <- c('x_coord', 'y_coord')

coldata <- names(data)

ggplot(mdsPoisData, aes(x_coord, y_coord, color = groups, label = coldata)) +
  geom_text(size = 4) +
  ggtitle('Multi Dimensional Scaling') +
  labs(x = "Poisson Distance", y = "Poisson Distance") +
  theme_bw()
```



Some samples clearly deviate from the other 2 in the group. This is especially clear with X51D_5_S, PrPCyt_30_NBH, X51D_30_NBH & PrPCyt_5_S. Strangely, these samples are all the first one in their respective group. This could indicate that the first tests were less accurate. Since 3 samples must remain in each group, no data will be removed from the set.

3 Further processing

```
counts.fpm <- log2( fpm(ddsMat, robust = TRUE) + 1 )
dds <- DESeq(ddsMat, parallel = TRUE)
```

```
## Warning in DESeq(ddsMat, parallel = TRUE): the design is ~ 1 (just an
## intercept). is this intended?

## estimating size factors

## estimating dispersions

## gene-wise dispersion estimates: 12 workers

## mean-dispersion relationship

## final dispersion estimates, fitting model and testing: 12 workers

## -- replacing outliers and refitting for 147 genes
## -- DESeq argument 'minReplicatesForReplace' = 7
## -- original counts are preserved in counts(dds)

## estimating dispersions

## fitting model and testing
```

```
res <- results(dds)
```

3.1 Preprocessing

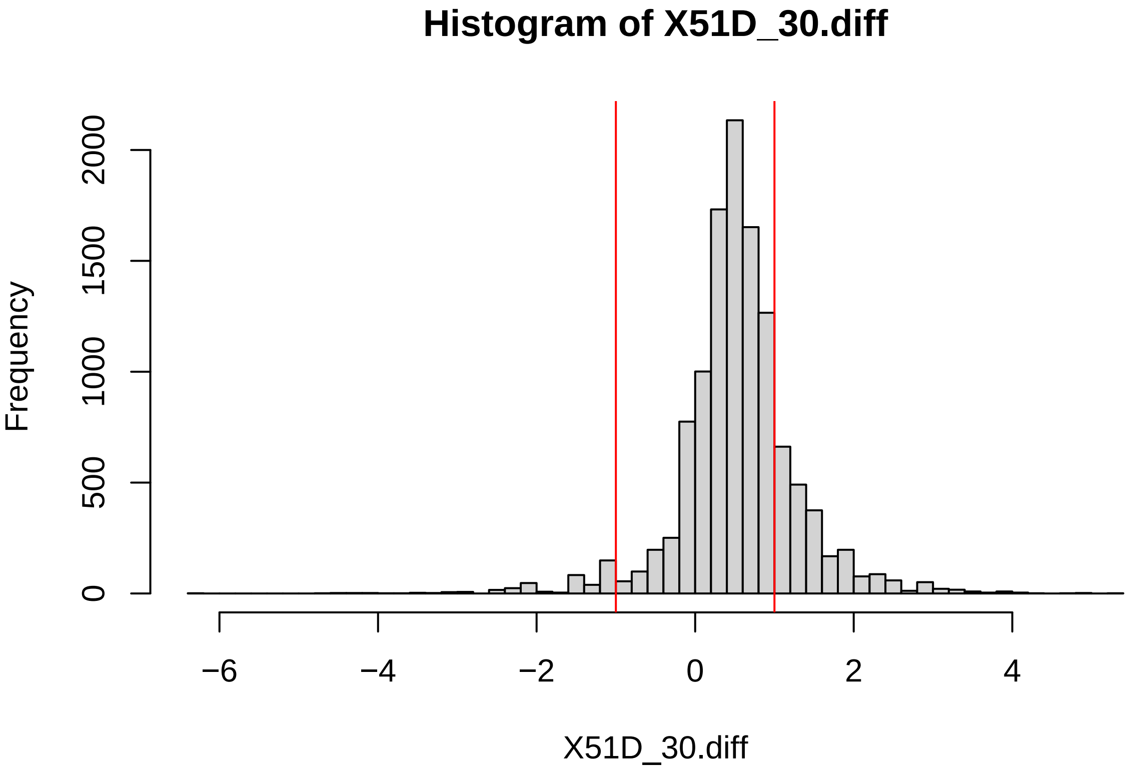
```
beforeCounts <- counts(dds)
keep <- rowSums(beforeCounts) >= 10
dds <- dds[keep,]
afterCounts <- counts(dds)

countCompare <- data.frame(nrow(beforeCounts),
                           nrow(afterCounts),
                           nrow(beforeCounts) - nrow(afterCounts))
colnames(countCompare) <- c("Counts before filtering",
                            "Counts after filtering",
                            "Difference in counts")
kable(countCompare)
```

Counts before filtering	Counts after filtering	Difference in counts
17742	13618	4124

The dataset has been trimmed to filter out genes with count values lower than 10. This results in a smaller dataset because more than 4000 genes have been removed.

3.2 Fold change value

```
X51D_30_NBH.means <- data.frame(X51D_30_NBH.means=rowMeans(afterCounts[,1:3]))
X51D_30_S.means <- data.frame(X51D_30_S.means=rowMeans(afterCounts[,4:6]))
X51D_5_NBH.means <- data.frame(X51D_5_NBH.means=rowMeans(afterCounts[,7:9]))
X51D_5_S.means <- data.frame(X51D_5_S.means=rowMeans(afterCounts[,10:12]))

PrPCyt_30_NBH.means <- data.frame(PrPCyt_30_NBH.means=rowMeans(afterCounts[,13:15]))
PrPCyt_30_S.means <- data.frame(PrPCyt_30_S.means=rowMeans(afterCounts[,16:18]))
PrPCyt_5_NBH.means <- data.frame(PrPCyt_5_NBH.means=rowMeans(afterCounts[,19:21]))
PrPCyt_5_S.means <- data.frame(PrPCyt_5_S.means=rowMeans(afterCounts[,22:24]))

PrPGPI_40_NBH.means <- data.frame(PrPGPI_40_NBH.means=rowMeans(afterCounts[,25:27]))
PrPGPI_40_S.means <- data.frame(PrPGPI_40_S.means=rowMeans(afterCounts[,28:30]))
PrPGPI_5_NBH.means <- data.frame(PrPGPI_5_NBH.means=rowMeans(afterCounts[,31:33]))
PrPGPI_5_S.means <- data.frame(PrPGPI_5_S.means=rowMeans(afterCounts[,34:36]))

X51D_30.diff <- na.omit(log2(X51D_30_NBH.means) - log2(X51D_30_S.means))
X51D_30.diff <- X51D_30.diff[is.finite(rowSums(X51D_30.diff)),]
X51D_30.diff <- as.numeric(X51D_30.diff)

hist(X51D_30.diff, breaks=60)
abline(v = 1, col = "red")
abline(v = -1, col = "red")
```



This histogram shows that there are some significant changes to the fold values, especially up-regulated. The data compared is that of the X51D fly after 30 days with a Scrapie pathogen and without.

3.3 Discovering DEG's

```
species <- factor(rep(seq(1:3), each = 12), labels = c("X51D", "PrP_Cyt", "PrP_GPI"))
replicates <- rep(seq(1:3), 12)
time <- factor(c(1,1,1,1,1,1,2,2,2,2,2,2,1,1,1,1,1,1,2,2,2,2,2,2,1,1,1,1,1,1,2,2,2,2,2,2), labels = c("
type <- factor(c(1,1,1,2,2,2,1,1,1,2,2,2,1,1,1,2,2,2,1,1,1,2,2,2,1,1,1,2,2,2,1,1,1,2,2,2), labels = c("
design <- data.frame(species, row.names = colnames(data))
design <- cbind(design, replicates, time, type)

dds <- DESeqDataSetFromMatrix(countData = data, colData = design, design = ~ species)
dds <- DESeq(dds, parallel = TRUE)
```

```
## estimating size factors
```

```
## estimating dispersions
```

```
## gene-wise dispersion estimates: 12 workers
```

```
## mean-dispersion relationship
```

```
## final dispersion estimates, fitting model and testing: 12 workers
```

```
## -- replacing outliers and refitting for 82 genes
## -- DESeq argument 'minReplicatesForReplace' = 7
## -- original counts are preserved in counts(dds)
```

```
## estimating dispersions
```

```
## fitting model and testing
```

```
res <- results(dds, alpha = 0.05)

group <- c(1,1,1,2,2,2,3,3,3,4,4,4,5,5,5,6,6,6,7,7,7,8,8,8,9,9,9,10,10,10,11,11,11,12,12,12)
time <- factor(c(1,1,1,1,1,1,2,2,2,2,2,2,1,1,1,1,1,1,2,2,2,2,2,2,1,1,1,1,1,1,2,2,2,2,2,2))
type <- factor(c(1,1,1,2,2,2,1,1,1,2,2,2,1,1,1,2,2,2,1,1,1,2,2,2,1,1,1,2,2,2,1,1,1,2,2,2))
model <- model.matrix(~ group + replicates + time + type)

d <- DGEList(counts=afterCounts, group = species)
d <- calcNormFactors(d)

output <- estimateDisp(d, design = model)
fit <- glmQLFit(output, design = model)

test <- glmQLFTest(fit, coef=5)

LRT <- glmLRT(fit)

kable(topTags(LRT))
```

This table shows the genes with the most significant differences.

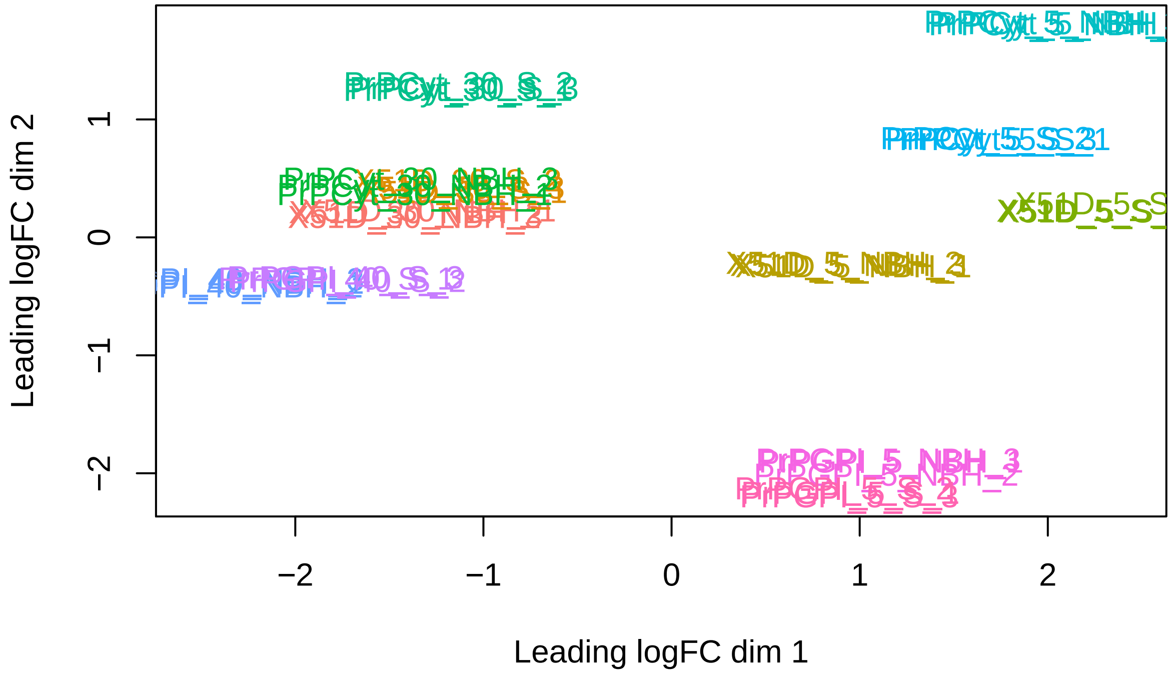
	logFC	logCPM	LR	PValue	FDR
FBgn0004240	-2.591225	5.888021	218.80958	0	0
FBgn0034407	-2.297217	5.040921	196.72457	0	0
FBgn0010388	-2.168257	5.894200	173.42085	0	0
FBgn0041579	-1.947014	7.018217	141.99141	0	0
FBgn0036600	1.860137	5.242363	113.00308	0	0
FBgn0019661	1.846219	4.174026	107.82879	0	0
FBgn0266405	-2.267430	3.318964	105.49745	0	0
FBgn0013279	-1.485205	6.641321	91.35769	0	0
FBgn0041581	-1.397773	4.213615	83.73515	0	0
FBgn0014865	-1.395028	7.194669	79.72194	0	0

x
BH

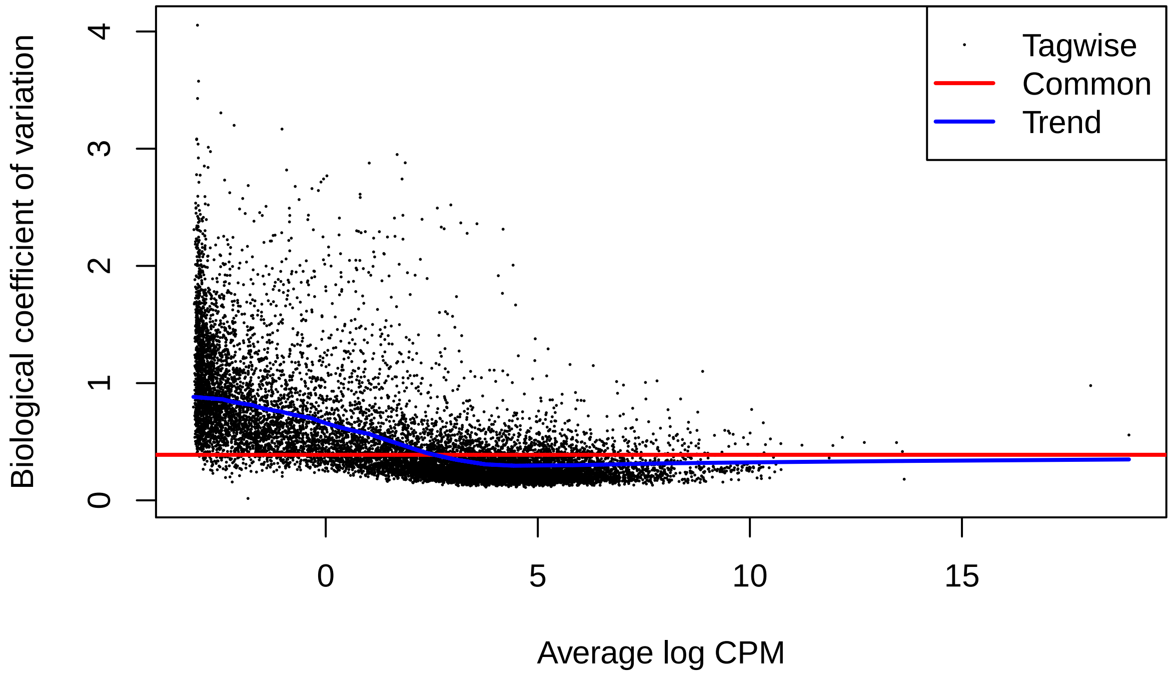
x
type2

x
glm

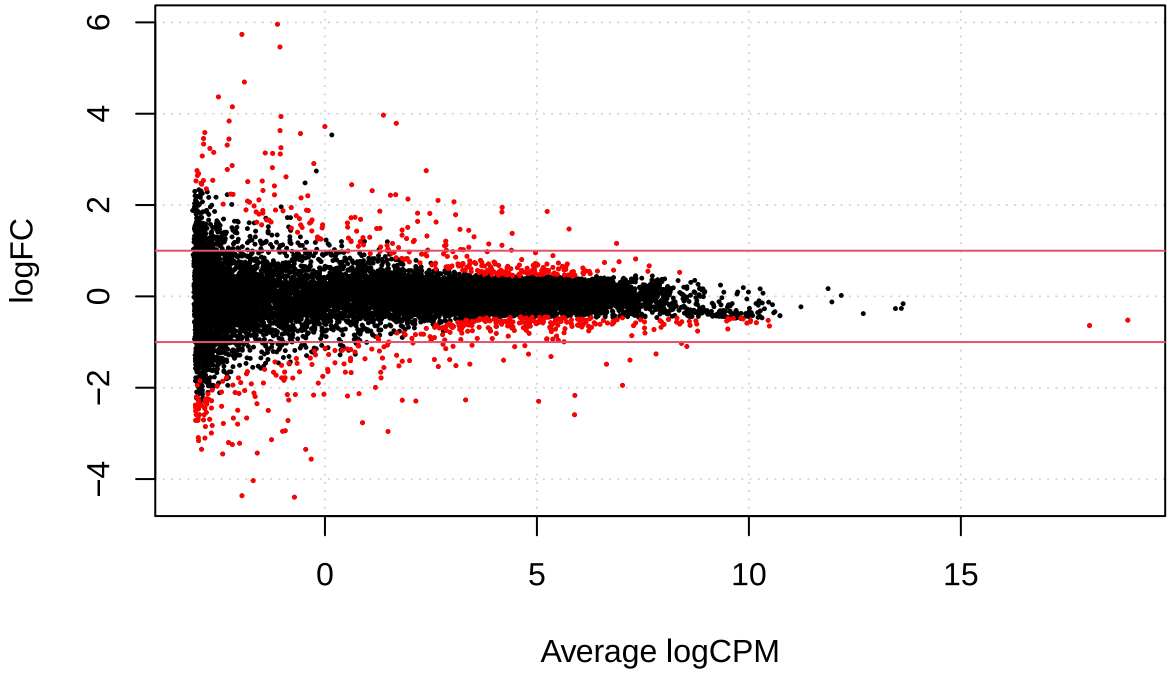
```r
plotMDS(calcNormFactors(output), col = colors36)
```



```r
plotBCV(calcNormFactors(output))
```



```
deGenes <- decideTestsDGE(LRT, p=0.05)
deGenes <- rownames(LRT)[as.logical(deGenes)]
plotSmear(LRT, de.tags=deGenes)
abline(h=c(-1, 1), col=2)
```



These plots contain information regarding the DEG's in the dataset.

3.4 Volcano plot

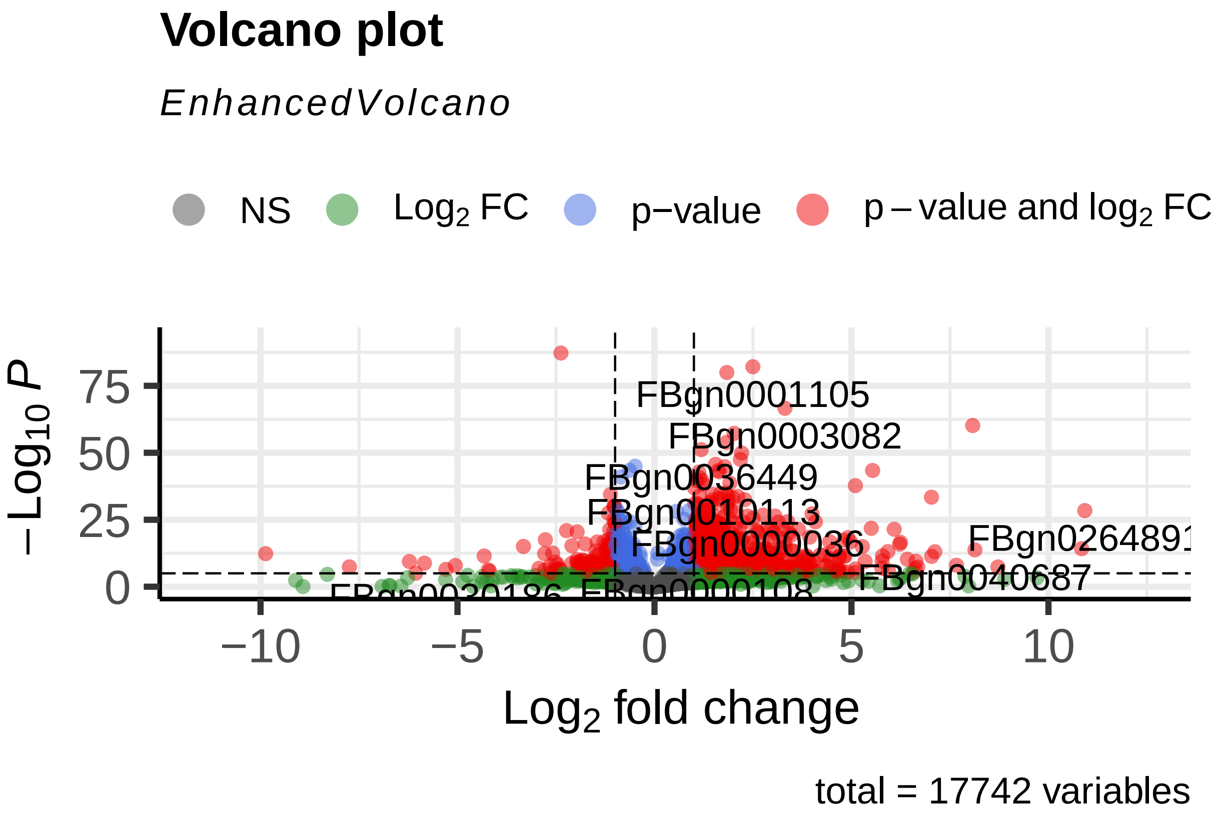
```
filtered <- res[!res$baseMean < 10,]
resultsNames(dds)
```

```
## [1] "Intercept"                "species_PrP_Cyt_vs_X51D"
## [3] "species_PrP_GPI_vs_X51D"
```

```
shrunk <- lfcShrink(dds, coef = "species_PrP_GPI_vs_X51D", res = res,
                    type = "apeglm")
```

```
## using 'apeglm' for LFC shrinkage. If used in published research, please cite:
##     Zhu, A., Ibrahim, J.G., Love, M.I. (2018) Heavy-tailed prior distributions for
##     sequence count data: removing the noise and preserving large differences.
##     Bioinformatics. https://doi.org/10.1093/bioinformatics/bty895
```

```
 EnhancedVolcano(shrunk,
    lab = rownames(shrunk),
    x = 'log2FoldChange',
    y = 'pvalue')
```


Volcano plot
EnhancedVolcano
● NS ● Log2 FC ● p−value ● p – value and log2 FC
−Log10 P
FBgn0001105
FBgn0003082
FBgn0036449
FBgn0010113
FBgn0000036
FBgn0264891
FBgn0040687
FBgn0020186
FBgn0000108
Log2 fold change
total = 17742 variables