class12

Steven Gan

2022-11-04

Table of contents

1. Bioconductor and DESeq2 setup **1**

2. Import countData and colData **2**

3. Toy differential gene expression **2**

4. DESeq2 analysis **7**

5. Adding annotation data **9**

6. Data Visualization **11**

7. Pathway analysis **12**

OPTIONAL: Plotting counts for genes of interest **15**

1. Bioconductor and DESeq2 setup

```
# install.packages("BiocManager")
# BiocManager::install()

# BiocManager::install("DESeq2")

library(BiocManager)
library(DESeq2)
```

2. Import countData and colData

```
counts <- read.csv("airway_scaledcounts.csv", row.names=1)
metadata <-  read.csv("airway_metadata.csv")
```

Q1:

```
nrow(counts)
```

```
[1] 38694
```

Q2:

```
table(metadata$dex)
```

```
control treated
      4       4
```

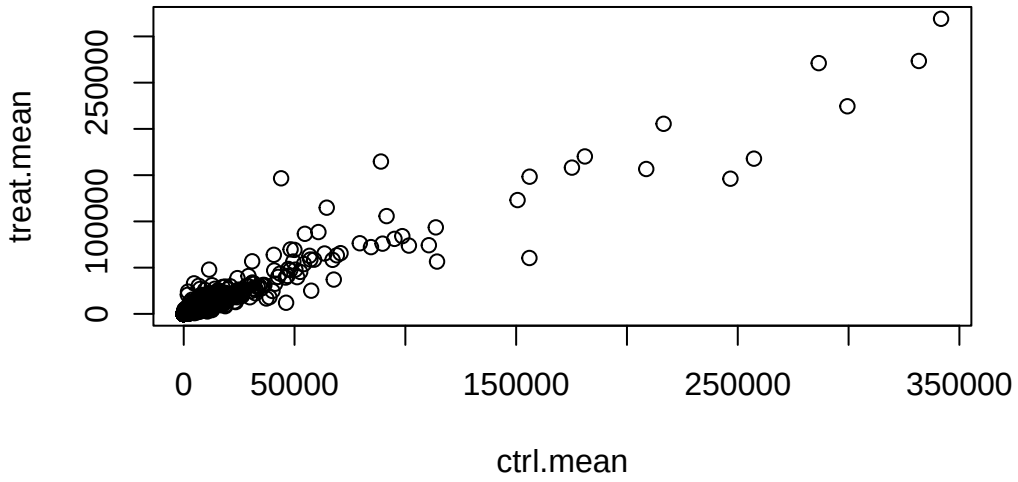
3. Toy differential gene expression

Q3 & Q4:

```
library(dplyr)

ctrl <- metadata %>% filter(dex == "control")
ctrl.mean <- counts %>%
  select(ctrl$id) %>%
  rowMeans()

treat <- metadata %>% filter(dex == "treated")
treat.mean <- counts %>%
  select(treat$id) %>%
  rowMeans()

meancounts <- data.frame(ctrl.mean, treat.mean)
```

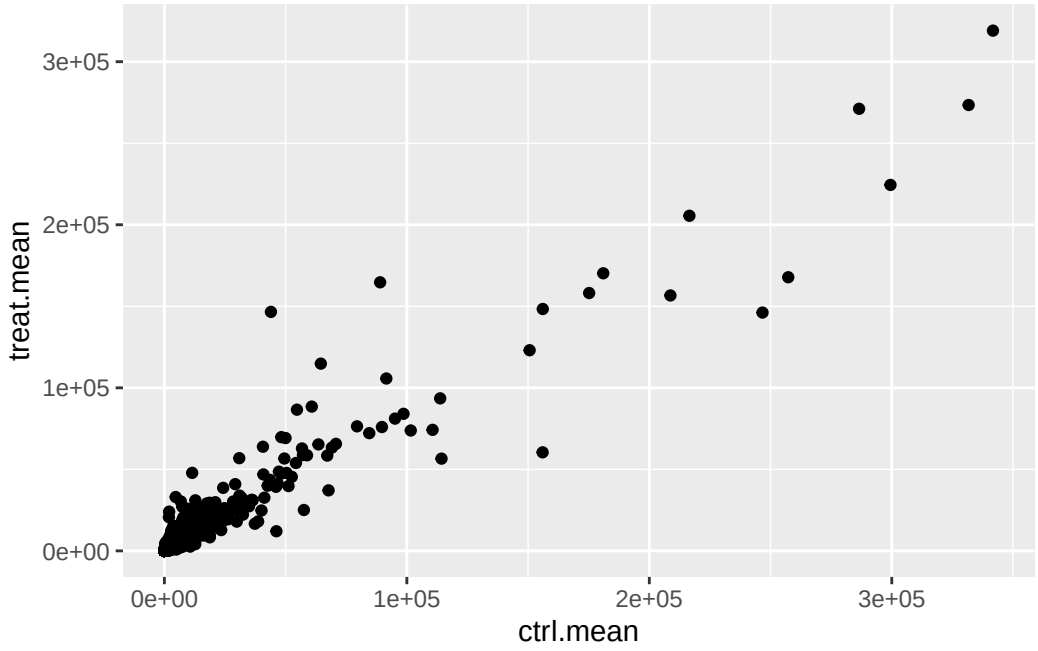
Q5 (a):

```
plot(meancounts)
```



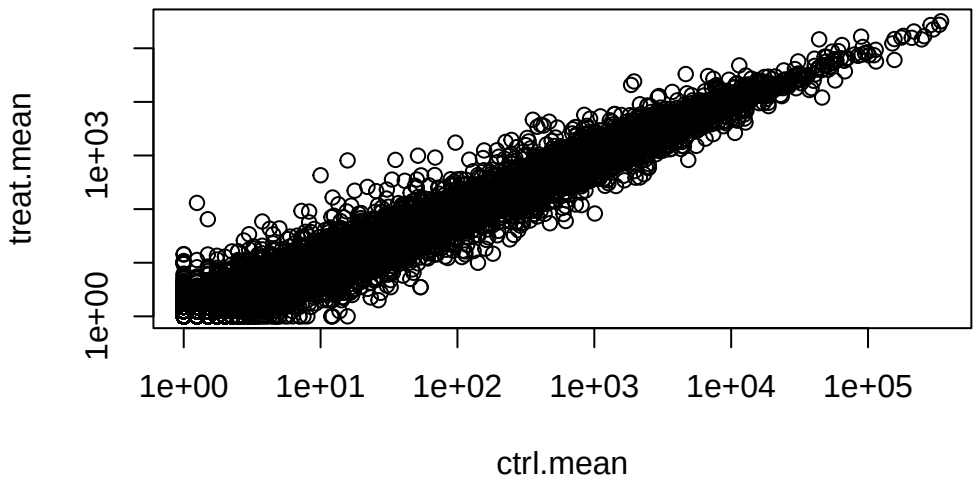
Q5 (b):

```
library(ggplot2)

ggplot(meancounts) +
  aes(x = ctrl.mean, y = treat.mean) +
  geom_point()
```

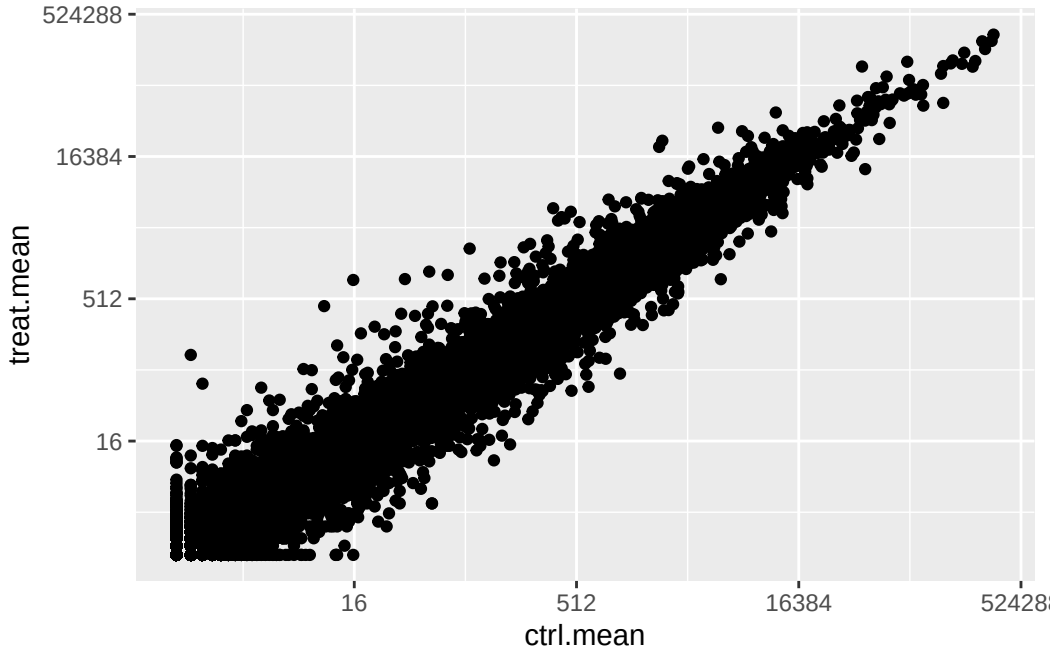


Q6:

```
plot(meancounts + 1, log = "xy")
```




```
ggplot(meancounts + 1) +
  aes(x = ctrl.mean, y = treat.mean) +
  geom_point() +
  scale_x_continuous(trans = "log2") +
  scale_y_continuous(trans = "log2")
```




```
meancounts$log2fc <- log2(meancounts[,"treat.mean"] / meancounts[,"ctrl.mean"])

zero.vals <- which(meancounts[,1:2] == 0, arr.ind = T)
to.rm <- unique(zero.vals[,1])
mycounts <- meancounts[-to.rm,]

mycounts2 <- meancounts[meancounts$ctrl.mean != 0 & meancounts$treat.mean != 0,]
identical(mycounts, mycounts2)
```

```
[1] TRUE
```

Q7: The first column is the row indices of rows with 0. Since it could be possible that both ctrl and treat mean contains 0, row indices might be count for two times. Unique can get rid of the duplicates.

Q8 & 9:

```
up.ind <- mycounts$log2fc > 2
down.ind <- mycounts$log2fc < -2

sum(up.ind)
```

```
[1] 250
```

```
sum(down.ind)
```

```
[1] 367
```

Q10:

Not completely trustworthy, as the significany are not evaluated.

4. DESeq2 analysis

```
dds <- DESeqDataSetFromMatrix(countData = counts,
                              colData = metadata,
                              design = ~dex)
```

```
converting counts to integer mode
```

```
Warning in DESeqDataSet(se, design = design, ignoreRank): some variables in
design formula are characters, converting to factors
```

```
dds
```

```
class: DESeqDataSet
dim: 38694 8
metadata(1): version
assays(1): counts
rownames(38694): ENSG00000000003 ENSG00000000005 ... ENSG00000283120
  ENSG00000283123
rowData names(0):
colnames(8): SRR1039508 SRR1039509 ... SRR1039520 SRR1039521
colData names(4): id dex celltype geo_id
```

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

```
estimating size factors
```

```
estimating dispersions
```

```
gene-wise dispersion estimates
```

```
mean-dispersion relationship
```

```
final dispersion estimates
```

```
fitting model and testing
```

```
dds
```

```
class: DESeqDataSet
dim: 38694 8
metadata(1): version
assays(4): counts mu H cooks
rownames(38694): ENSG00000000003 ENSG00000000005 ... ENSG00000283120
  ENSG00000283123
rowData names(22): baseMean baseVar ... deviance maxCooks
colnames(8): SRR1039508 SRR1039509 ... SRR1039520 SRR1039521
colData names(5): id dex celltype geo_id sizeFactor
```

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

```
out of 25258 with nonzero total read count
adjusted p-value < 0.1
LFC > 0 (up)       : 1563, 6.2%
LFC < 0 (down)     : 1188, 4.7%
outliers [1]       : 142, 0.56%
low counts [2]     : 9971, 39%
(mean count < 10)
[1] see 'cooksCutoff' argument of ?results
[2] see 'independentFiltering' argument of ?results
```

```
res05 <- results(dds, alpha = 0.05)
summary(res05)
```

```
out of 25258 with nonzero total read count
adjusted p-value < 0.05
LFC > 0 (up)       : 1236, 4.9%
LFC < 0 (down)     : 933, 3.7%
outliers [1]       : 142, 0.56%
low counts [2]     : 9033, 36%
(mean count < 6)
[1] see 'cooksCutoff' argument of ?results
[2] see 'independentFiltering' argument of ?results
```

5. Adding annotation data

Q11:

```
# BiocManager::install("AnnotationDbi")
# BiocManager::install("org.Hs.eg.db")

library("AnnotationDbi")
library("org.Hs.eg.db")

res$symbol <- mapIds(org.Hs.eg.db,
                     keys = row.names(res),
                     keytype = "ENSEMBL",
                     column = "SYMBOL",
                     multiVals = "first")

res$entrez <- mapIds(org.Hs.eg.db,
                     keys = row.names(res),
                     keytype = "ENSEMBL",
                     column = "ENTREZID",
                     multiVals = "first")

res$uniprot <- mapIds(org.Hs.eg.db,
                     keys = row.names(res),
                     keytype = "ENSEMBL",
                     column = "UNIPROT",
                     multiVals = "first")

res$genename <- mapIds(org.Hs.eg.db,
                     keys = row.names(res),
```

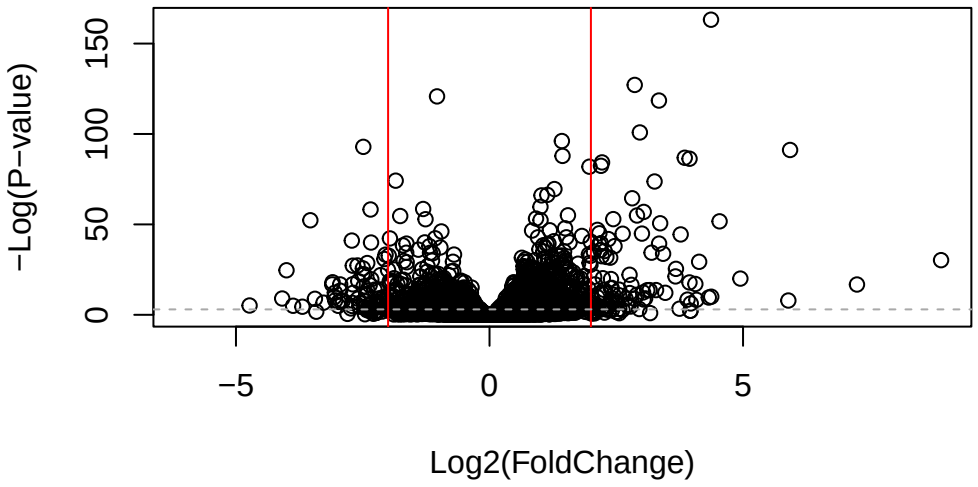
```
                    keytype = "ENSEMBL",
                    column = "GENENAME",
                    multiVals = "first")

head(res)
```

```
log2 fold change (MLE): dex treated vs control
Wald test p-value: dex treated vs control
DataFrame with 6 rows and 10 columns
                  baseMean log2FoldChange     lfcSE      stat    pvalue
                 <numeric>      <numeric> <numeric> <numeric> <numeric>
ENSG00000000003 747.194195     -0.3507030  0.168246 -2.084470 0.0371175
ENSG00000000005   0.000000             NA        NA        NA        NA
ENSG00000000419 520.134160      0.2061078  0.101059  2.039475 0.0414026
ENSG00000000457 322.664844      0.0245269  0.145145  0.168982 0.8658106
ENSG00000000460  87.682625     -0.1471420  0.257007 -0.572521 0.5669691
ENSG00000000938   0.319167     -1.7322890  3.493601 -0.495846 0.6200029
                     padj      symbol      entrez     uniprot
                <numeric> <character> <character> <character>
ENSG00000000003  0.163035      TSPAN6        7105  A0A024RCI0
ENSG00000000005        NA        TNMD       64102      Q9H2S6
ENSG00000000419  0.176032        DPM1        8813      O60762
ENSG00000000457  0.961694       SCYL3       57147      Q8IZE3
ENSG00000000460  0.815849    C1orf112       55732  A0A024R922
ENSG00000000938        NA         FGR        2268      P09769
                          genename
                       <character>
ENSG00000000003          tetraspanin 6
ENSG00000000005            tenomodulin
ENSG00000000419 dolichyl-phosphate m..
ENSG00000000457 SCY1 like pseudokina..
ENSG00000000460 chromosome 1 open re..
ENSG00000000938 FGR proto-oncogene, ..
```

```
ord <- order(res$padj)
write.csv(res[ord,], "deseq_results.csv")
```

6. Data Visualization

```
plot(res$log2FoldChange, -log(res$padj),
     xlab = "Log2(FoldChange)",
     ylab = "-Log(P-value)")

abline(v = c(-2, 2), col = "red")
abline(h = -log(0.05), col = "darkgray", lty = 2)
```

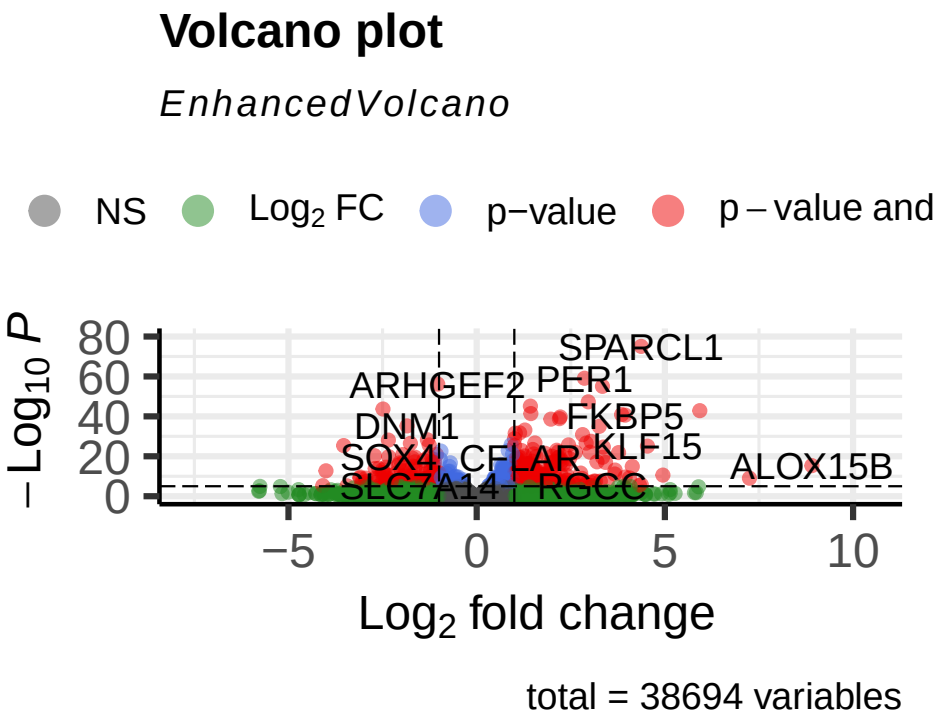



```
# BiocManager::install("EnhancedVolcano")
library(EnhancedVolcano)
```

```
Loading required package: ggrepel
```

```
x <- as.data.frame(res)

EnhancedVolcano(x,
    lab = x$symbol,
    x = 'log2FoldChange',
```

```
    y = 'pvalue')
```



7. Pathway analysis

```
# BiocManager::install( c("pathview", "gage", "gageData") )
library(pathview)
library(gage)
library(gageData)
```

```
data(kegg.sets.hs)

head(kegg.sets.hs, 2)
```

```
$`hsa00232 Caffeine metabolism`
[1] "10"   "1544" "1548" "1549" "1553" "7498" "9"
```

```
$`hsa00983 Drug metabolism - other enzymes`
 [1] "10"     "1066"   "10720"  "10941"  "151531" "1548"   "1549"   "1551"
 [9] "1553"   "1576"   "1577"   "1806"   "1807"   "1890"   "221223" "2990"
[17] "3251"   "3614"   "3615"   "3704"   "51733"  "54490"  "54575"  "54576"
[25] "54577"  "54578"  "54579"  "54600"  "54657"  "54658"  "54659"  "54963"
[33] "574537" "64816"  "7083"   "7084"   "7172"   "7363"   "7364"   "7365"
[41] "7366"   "7367"   "7371"   "7372"   "7378"   "7498"   "79799"  "83549"
[49] "8824"   "8833"   "9"      "978"
```

```
foldchanges = res$log2FoldChange
names(foldchanges) = res$entrez
head(foldchanges)
```

```
      7105       64102        8813       57147       55732        2268
-0.35070302          NA  0.20610777  0.02452695 -0.14714205 -1.73228897
```

```
keggres = gage(foldchanges, gsets = kegg.sets.hs)
```

```
attributes(keggres)
```

```
$names
[1] "greater" "less"    "stats"
```

```
head(keggres$less, 3)
```

```
                                      p.geomean stat.mean        p.val
hsa05332 Graft-versus-host disease 0.0004250461 -3.473346 0.0004250461
hsa04940 Type I diabetes mellitus  0.0017820293 -3.002352 0.0017820293
hsa05310 Asthma                    0.0020045888 -3.009050 0.0020045888
                                        q.val set.size         exp1
hsa05332 Graft-versus-host disease 0.09053483       40 0.0004250461
hsa04940 Type I diabetes mellitus  0.14232581       42 0.0017820293
hsa05310 Asthma                    0.14232581       29 0.0020045888
```

```
pathview(gene.data = foldchanges, pathway.id = "hsa05310")
```

```
'select()' returned 1:1 mapping between keys and columns
```

```
Info: Working in directory /Users/stevengan/UCSD/Biology/4 Foundations of Bioinformatics/12 T
Info: Writing image file hsa05310.pathview.png
```

```
pathview(gene.data = foldchanges, pathway.id = "hsa05310", kegg.native = FALSE)
```

```
'select()' returned 1:1 mapping between keys and columns
Info: Working in directory /Users/stevengan/UCSD/Biology/4 Foundations of Bioinformatics/12 T
Info: Writing image file hsa05310.pathview.pdf
```

Q12:

```
pathview(gene.data = foldchanges, pathway.id = "hsa05332")
```

```
'select()' returned 1:1 mapping between keys and columns
Info: Working in directory /Users/stevengan/UCSD/Biology/4 Foundations of Bioinformatics/12 T
Info: Writing image file hsa05332.pathview.png
```

```
pathview(gene.data = foldchanges, pathway.id = "hsa05332", kegg.native = FALSE)
```

```
'select()' returned 1:1 mapping between keys and columns
Warning in .subtypeDisplay(object): Given subtype 'missing interaction' is not found!
Info: Working in directory /Users/stevengan/UCSD/Biology/4 Foundations of Bioinformatics/12 T
Info: Writing image file hsa05332.pathview.pdf
```

```
pathview(gene.data = foldchanges, pathway.id = "hsa04940")
```

```
'select()' returned 1:1 mapping between keys and columns
Info: Working in directory /Users/stevengan/UCSD/Biology/4 Foundations of Bioinformatics/12 T
Info: Writing image file hsa04940.pathview.png
```

```
pathview(gene.data = foldchanges, pathway.id = "hsa04940", kegg.native = FALSE)
```

```
'select()' returned 1:1 mapping between keys and columns
```

```
Info: Working in directory /Users/stevengan/UCSD/Biology/4 Foundations of Bioinformatics/12 T
```

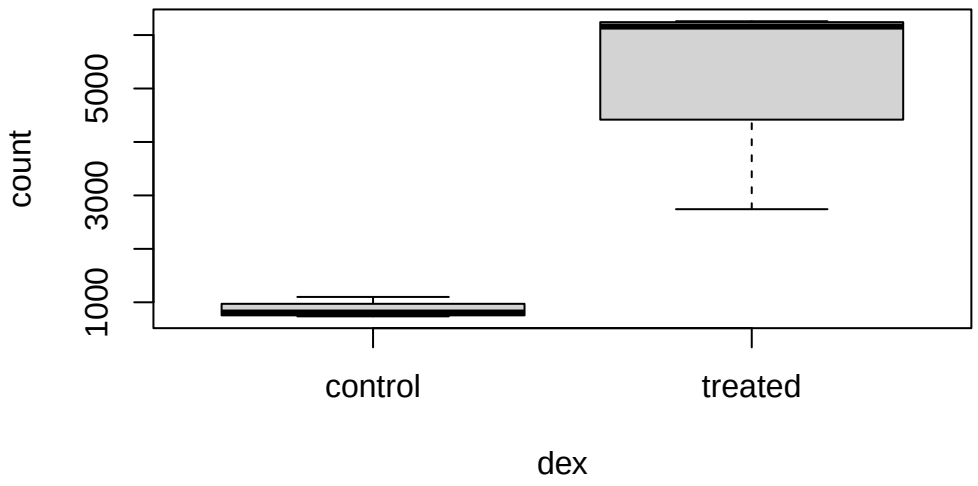
```
Info: Writing image file hsa04940.pathview.pdf
```

OPTIONAL: Plotting counts for genes of interest

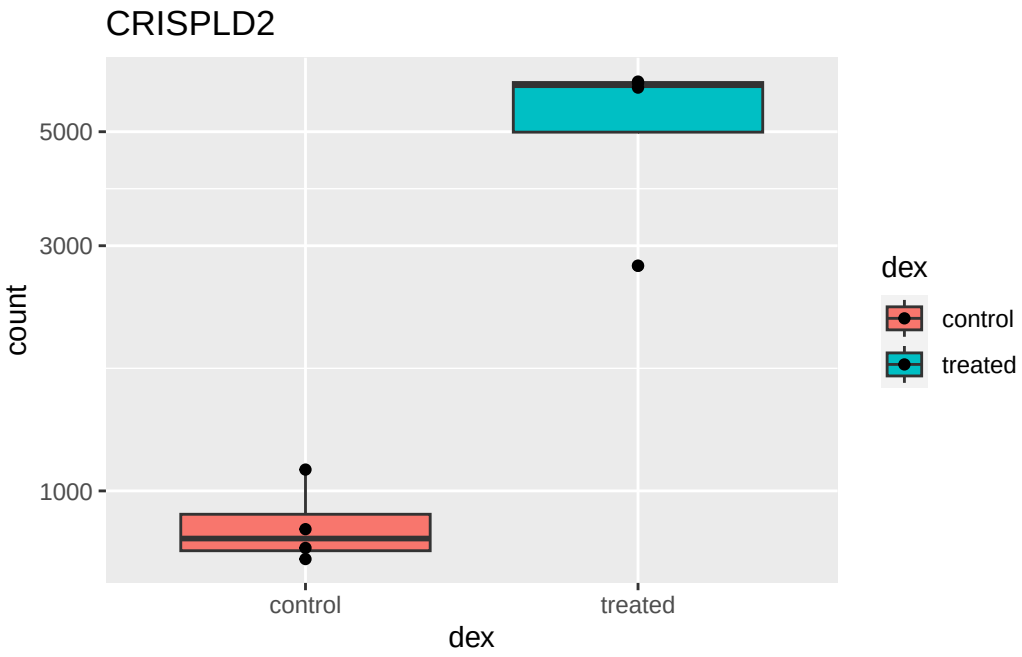
```
i <- grep("CRISPLD2", res$symbol)
res[i,]
```

```
log2 fold change (MLE): dex treated vs control
Wald test p-value: dex treated vs control
DataFrame with 1 row and 10 columns
                 baseMean log2FoldChange     lfcSE      stat      pvalue
                <numeric>      <numeric> <numeric> <numeric>   <numeric>
ENSG00000103196   3096.16        2.62603  0.267444   9.81899 9.32747e-23
                      padj      symbol      entrez     uniprot
                 <numeric> <character> <character> <character>
ENSG00000103196 3.36344e-20    CRISPLD2       83716  A0A140VK80
                           genename
                        <character>
ENSG00000103196 cysteine rich secret..
```

```
d <- plotCounts(dds, gene = row.names(res[i,]), intgroup = "dex", returnData = T)
boxplot(count ~ dex, data = d)
```




```
library(ggplot2)

ggplot(d, aes(dex, count, fill = dex)) +
  geom_boxplot() +
  geom_point() +
  scale_y_log10() +
  ggtitle("CRISPLD2")
```


CRISPLD2
5000
3000
1000
count
control
treated
dex
dex
control
treated