PulsePAT-RNAseq

Victoria Ruiz & Thomas W. Battaglia

Contents

1. Introduction

This is an Rmarkdown document which contains the R code for the RNA-sequencing analysis found within the manuscript "Ruiz et al. (2017)". Details about the tools and steps for processing the ileal transcriptome dataset can be found within the **'Methods'** section of the respective manuscript. This fastq files used to generate this data set is hosted on ArrayExpress under the ascension number E-MTAB-5101.

1a. Load the necessary libraries

```
# For RNAseq-related data
library(DESeq2)
library(ggplot2)
library(fdrtool)
library(pheatmap)
library(RColorBrewer)
library(tidyverse)
library(org.Mm.eg.db)
```

1b. Import the DESeq2 object containing the counts and metadata

```
load("data/deseq2_obj.rda")
```

Section 2: Compare Control vs. PAT within pups at day 52 of life

To find any significant genes altered due to early life antibiotic perturbation, we will subset to compare the two treatment groups.

```
# Subset to only analyze pups
ddPups <- ddsMat[ ,which(ddsMat$Breeder_Pup == "pup")]

# Run DESEq2
ddPups <- DESeq(ddPups)
```

```
## estimating size factors

## estimating dispersions

## gene-wise dispersion estimates

## mean-dispersion relationship

## final dispersion estimates

## fitting model and testing
```

```r
# Get results from testing with FDR adjust pvalues
ddPups_res <- results(ddPups, pAdjustMethod = "fdr", alpha = 0.05)

# Find out FC directionality
## Tylosin / Control
mcols(ddPups_res, use.names = T)
```

```
## DataFrame with 6 rows and 2 columns
##                         type
##                  <character>
## baseMean       intermediate
## log2FoldChange      results
## lfcSE               results
## stat                results
## pvalue              results
## padj                results
##                                                 description
##                                                 <character>
## baseMean              mean of normalized counts for all samples
## log2FoldChange log2 fold change (MAP): Treatment PAT vs Control
## lfcSE                   standard error: Treatment PAT vs Control
## stat                    Wald statistic: Treatment PAT vs Control
## pvalue              Wald test p-value: Treatment PAT vs Control
## padj                                      fdr adjusted p-values
```

```r
# Generate summary of testing.
summary(ddPups_res)
```

```
##
## out of 29106 with nonzero total read count
## adjusted p-value < 0.05
## LFC > 0 (up)     : 568, 2%
## LFC < 0 (down)   : 755, 2.6%
## outliers [1]     : 128, 0.44%
## low counts [2]   : 12486, 43%
## (mean count < 7)
## [1] see 'cooksCutoff' argument of ?results
## [2] see 'independentFiltering' argument of ?results
```

2a. Add gene annotations

```r
# Load mouse gene annotation
library(org.Mm.eg.db)

# Add gene annotation
ddPups_res$description <- mapIds(x = org.Mm.eg.db,
                                 keys = row.names(ddPups_res),
                                 column = "GENENAME",
                                 keytype = "SYMBOL",
                                 multiVals = "first")
```

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

```r
# Add gene symbol
ddPups_res$symbol <- row.names(ddPups_res)
```

```
# Add ENTREZ ID
ddPups_res$entrez <- mapIds(x = org.Mm.eg.db,
                            keys = row.names(ddPups_res),
                            column = "ENTREZID",
                            keytype = "SYMBOL",
                            multiVals = "first")
```

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

```
# Add GO terms
ddPups_res$GO <- mapIds(x = org.Mm.eg.db,
                        keys = row.names(ddPups_res),
                        column = "GO",
                        keytype = "SYMBOL",
                        multiVals = "first")
```

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

```
# Subset for only significant genes (q < 0.05)
ddPups_res_sig <- subset(ddPups_res, padj < 0.05)
summary(ddPups_res_sig)
```

```
##
## out of 1323 with nonzero total read count
## adjusted p-value < 0.05
## LFC > 0 (up)     : 568, 43%
## LFC < 0 (down)   : 755, 57%
## outliers [1]     : 0, 0%
## low counts [2]   : 0, 0%
## (mean count < 7)
## [1] see 'cooksCutoff' argument of ?results
## [2] see 'independentFiltering' argument of ?results
```

```
# Remove any gene without an ENTREZ-id annotation (cannot be used in any pathway)
ddPups_res_sig_filter <- subset(ddPups_res_sig, is.na(entrez) == F & symbol != "Gm5739" & symbol != "29
summary(ddPups_res_sig_filter)
```

```
##
## out of 1298 with nonzero total read count
## adjusted p-value < 0.05
## LFC > 0 (up)     : 552, 43%
## LFC < 0 (down)   : 746, 57%
## outliers [1]     : 0, 0%
## low counts [2]   : 0, 0%
## (mean count < 7)
## [1] see 'cooksCutoff' argument of ?results
## [2] see 'independentFiltering' argument of ?results
```

2c. Plot Volcano Plot

```
# Check directionality of the log2 fold change values
# PAT / Control
mcols(ddPups_res)
```

```
## DataFrame with 10 rows and 2 columns
##            type                                          description
```

```
##     <character>                                       <character>
## 1  intermediate        mean of normalized counts for all samples
## 2       results log2 fold change (MAP): Treatment PAT vs Control
## 3       results         standard error: Treatment PAT vs Control
## 4       results         Wald statistic: Treatment PAT vs Control
## 5       results      Wald test p-value: Treatment PAT vs Control
## 6       results                          fdr adjusted p-values
## 7            NA                                             NA
## 8            NA                                             NA
## 9            NA                                             NA
## 10           NA                                             NA
```

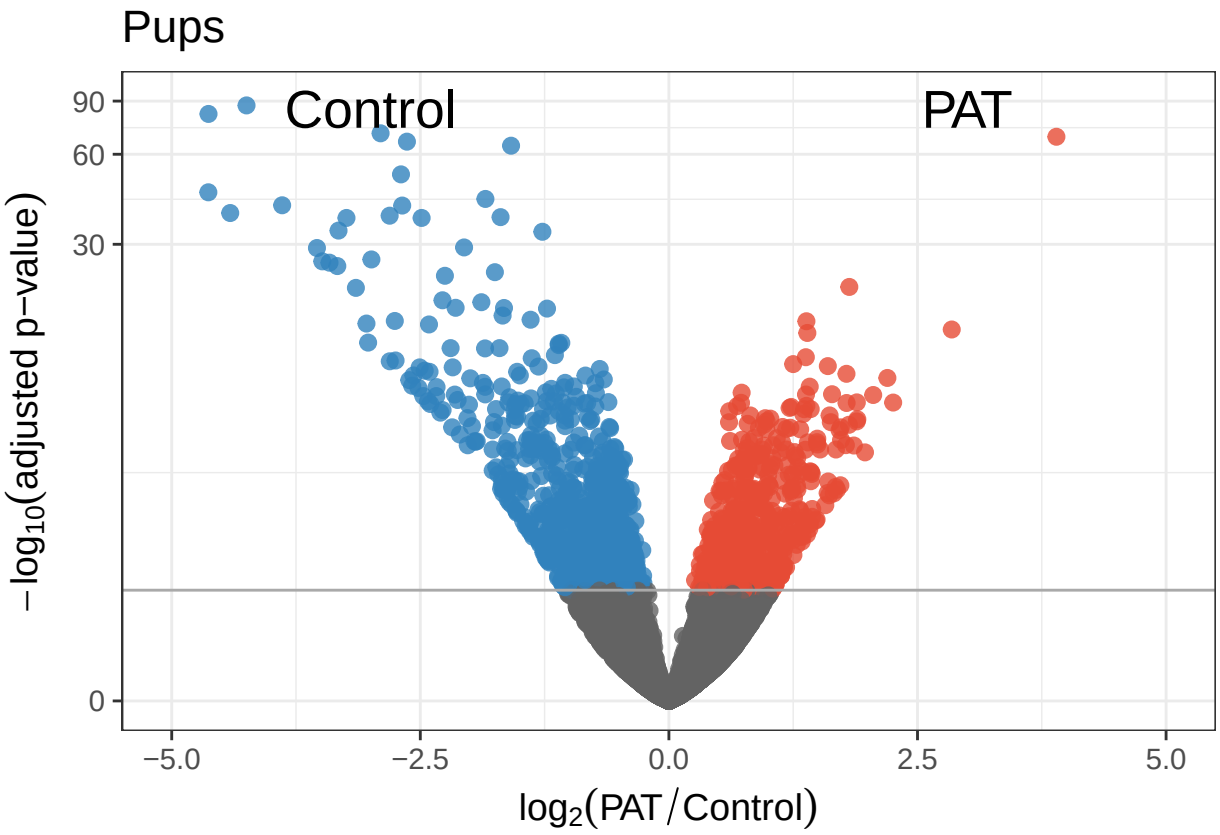
```r
# Gather Log-fold change and FDR-corrected pvalues from DESeq2 results
pups_data <- data.frame(gene = row.names(ddPups_res),
                        pval = -log10(ddPups_res$padj),
                        lfc = ddPups_res$log2FoldChange)

# Remove any rows that have NA as an entry
pups_data <- na.omit(pups_data)

# Color the points which are up or down
## If fold-change > 0 and pvalue > 1.3 (Increased significant)
## If fold-change < 0 and pvalue > 1.3 (Decreased significant)
pups_data <- mutate(pups_data, color = case_when(pups_data$lfc > 0 & pups_data$pval > 1.3 ~ "PAT",
                                                 pups_data$lfc < 0 & pups_data$pval > 1.3 ~ "Control",
                                                 pups_data$pval < 1.3 ~ "nonsignificant"))

# Make a basic ggplot2 object with x-y values
pups_vol <- ggplot(pups_data, aes(x = lfc, y = pval, color = color))

# Add ggplot2 layers
pups_vol = pups_vol +
  ggtitle(label = "Pups") +
  geom_point(size = 2.5, alpha = 0.8, na.rm = T) +
  annotate("text", label = "Control", x = -3, y = 85, size = 7, colour = "black") +
  annotate("text", label = "PAT", x = 3, y = 85, size = 7, colour = "black") +
  scale_color_manual(values = c("PAT" = "#E64B35", "Control" = "#3182bd", "nonsignificant" = "#636363")) +
  scale_x_continuous(limits = c(-5, 5)) +
  scale_y_continuous(limits = c(0, 90), trans = "log1p") +
  theme_bw(base_size = 14) +
  theme(legend.position = "none") +
  xlab(expression(log[2]("PAT" / "Control"))) +
  ylab(expression(-log[10]("adjusted p-value"))) +
  geom_hline(yintercept = 1.3, colour = "darkgrey")
pups_vol
```



Section 3: Compare Control vs. PAT within dams at day 110 of life

```
# Subset to only analyze pups
ddDams <- ddsMat[ ,which(ddsMat$Breeder_Pup == "Breeder")]

# Run DESEq2
ddDams <- DESeq(ddDams)
```

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

```
# Get results from testing with FDR adjust pvalues
ddDams_res <- results(ddDams, pAdjustMethod = "fdr", alpha = 0.05)

# Find out FC directionality
## Tylosin (PAT) / Control
mcols(ddDams_res, use.names = T)
```

```
## DataFrame with 6 rows and 2 columns
```

```
##                       type
##                <character>
## baseMean       intermediate
## log2FoldChange      results
## lfcSE               results
## stat                results
## pvalue              results
## padj                results
##                                                  description
##                                                  <character>
## baseMean              mean of normalized counts for all samples
## log2FoldChange log2 fold change (MAP): Treatment PAT vs Control
## lfcSE                   standard error: Treatment PAT vs Control
## stat                    Wald statistic: Treatment PAT vs Control
## pvalue               Wald test p-value: Treatment PAT vs Control
## padj                                      fdr adjusted p-values
```

```
# Generate summary of testing.
summary(ddDams_res)
```

```
##
## out of 29831 with nonzero total read count
## adjusted p-value < 0.05
## LFC > 0 (up)     : 63, 0.21%
## LFC < 0 (down)   : 41, 0.14%
## outliers [1]     : 124, 0.42%
## low counts [2]   : 9471, 32%
## (mean count < 2)
## [1] see 'cooksCutoff' argument of ?results
## [2] see 'independentFiltering' argument of ?results
```

2a. Add gene annotations

```
# Add gene annotation
ddDams_res$description <- mapIds(x = org.Mm.eg.db,
                                 keys = row.names(ddDams_res),
                                 column = "GENENAME",
                                 keytype = "SYMBOL",
                                 multiVals = "first")
```

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

```
# Add gene symbol
ddDams_res$symbol <- row.names(ddDams_res)

# Add ENTREZ ID
ddDams_res$entrez <- mapIds(x = org.Mm.eg.db,
                            keys = row.names(ddDams_res),
                            column = "ENTREZID",
                            keytype = "SYMBOL",
                            multiVals = "first")
```

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

```
# Add GO terms
ddDams_res$GO <- mapIds(x = org.Mm.eg.db,
```

```
                        keys = row.names(ddDams_res),
                        column = "GO",
                        keytype = "SYMBOL",
                        multiVals = "first")
```

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

```
# Subset for only significant genes (q < 0.05)
ddDams_res_sig <- subset(ddDams_res, padj < 0.05)
summary(ddDams_res_sig)
```

```
##
## out of 104 with nonzero total read count
## adjusted p-value < 0.05
## LFC > 0 (up)     : 63, 61%
## LFC < 0 (down)   : 41, 39%
## outliers [1]     : 0, 0%
## low counts [2]   : 0, 0%
## (mean count < 2)
## [1] see 'cooksCutoff' argument of ?results
## [2] see 'independentFiltering' argument of ?results
```

```
# Remove any gene without an ENTREZ-id annotation (cannot be used in any pathway)
ddDams_res_sig_filter <- subset(ddDams_res_sig, is.na(entrez) == F)
summary(ddDams_res_sig_filter)
```

```
##
## out of 103 with nonzero total read count
## adjusted p-value < 0.05
## LFC > 0 (up)     : 62, 60%
## LFC < 0 (down)   : 41, 40%
## outliers [1]     : 0, 0%
## low counts [2]   : 0, 0%
## (mean count < 2)
## [1] see 'cooksCutoff' argument of ?results
## [2] see 'independentFiltering' argument of ?results
```

3b. Plot Volcano Plot

```
# Check directionality of the log2 fold change values
# PAT / Control
mcols(ddDams_res)
```

```
## DataFrame with 10 rows and 2 columns
##            type                                       description
##     <character>                                       <character>
## 1  intermediate        mean of normalized counts for all samples
## 2       results log2 fold change (MAP): Treatment PAT vs Control
## 3       results          standard error: Treatment PAT vs Control
## 4       results          Wald statistic: Treatment PAT vs Control
## 5       results      Wald test p-value: Treatment PAT vs Control
## 6       results                          fdr adjusted p-values
## 7            NA                                             NA
## 8            NA                                             NA
## 9            NA                                             NA
```

```
## 10           NA                                                    NA
```

```
# Gather Log-fold change and FDR-corrected pvalues from DESeq2 results
dams_data <- data.frame(gene = row.names(ddDams_res),
                        pval = -log10(ddDams_res$padj),
                        lfc = ddDams_res$log2FoldChange)

# Remove any rows that have NA as an entry
dams_data <- na.omit(dams_data)

# Color the points which are up or down
## If fold-change > 0 and pvalue > 1.3 (Increased significant)
## If fold-change < 0 and pvalue > 1.3 (Decreased significant)
dams_data <- mutate(dams_data, color = case_when(dams_data$lfc > 0 & dams_data$pval > 1.3 ~ "PAT",
                                                 dams_data$lfc < 0 & dams_data$pval > 1.3 ~ "Control",
                                                 dams_data$pval < 1.3 ~ "nonsignificant"))

# Make a basic ggplot2 object with x-y values
dams_vol <- ggplot(dams_data, aes(x = lfc, y = pval, color = color))

# Add ggplot2 layers
dams_vol = dams_vol +
  ggtitle(label = "Dams") +
  geom_point(size = 2.5, alpha = 0.8, na.rm = T) +
  annotate("text", label = "Control", x = -3, y = 85, size = 7, colour = "black") +
  annotate("text", label = "PAT", x = 3, y = 85, size = 7, colour = "black") +
  scale_color_manual(values = c("PAT" = "#E64B35", "Control" = "#3182bd", "nonsignificant" = "#636363")) +
  scale_x_continuous(limits = c(-5, 5)) +
  scale_y_continuous(limits = c(0, 90), trans = "log1p") +
  theme_bw(base_size = 14) +
  theme(legend.position = "none") +
  xlab(expression(log[2]("PAT" / "Control"))) +
  ylab(expression(-log[10]("adjusted p-value"))) +
  geom_hline(yintercept = 1.3, colour = "darkgrey")
dams_vol
```

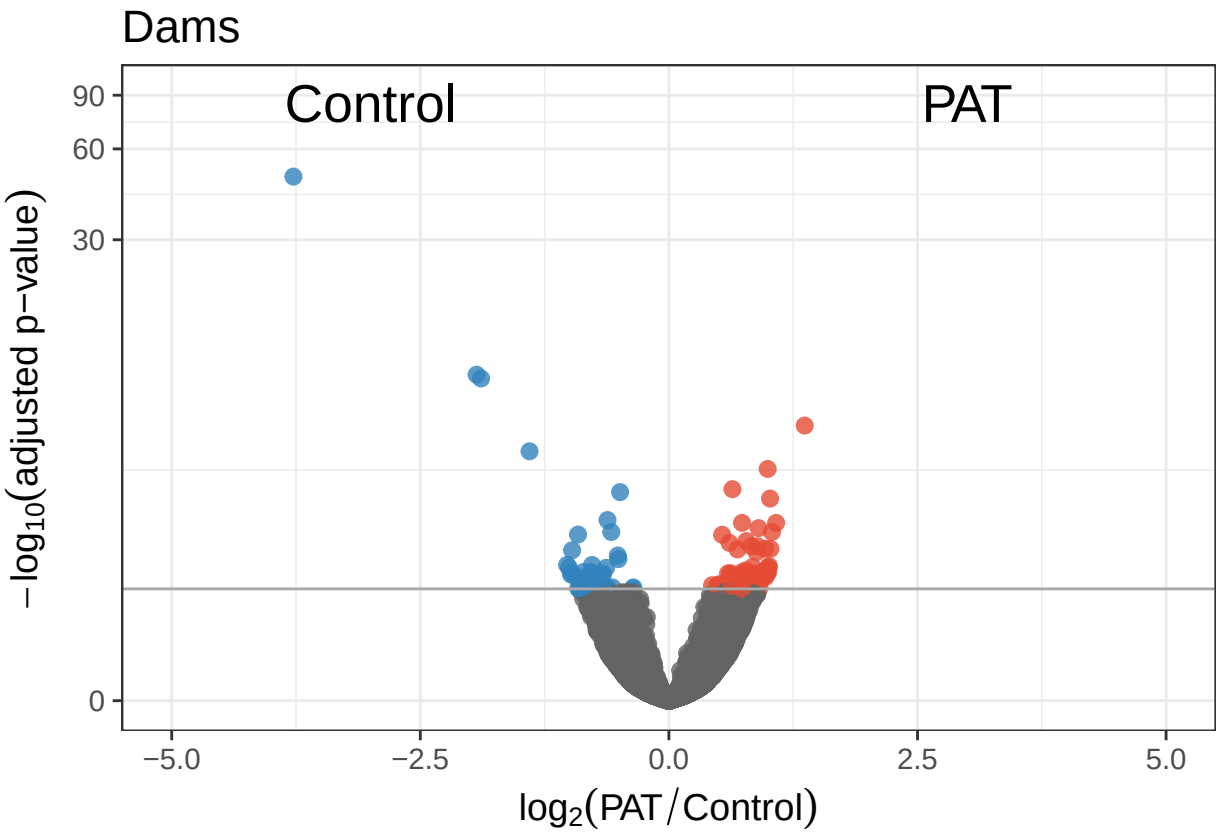


Figure 2f. Plot Volcano Plot

```
# Use both volcano plots to create a new single side-by-side figure
## using the package Rmisc
Rmisc::multiplot(pups_vol, dams_vol, cols = 2)
```

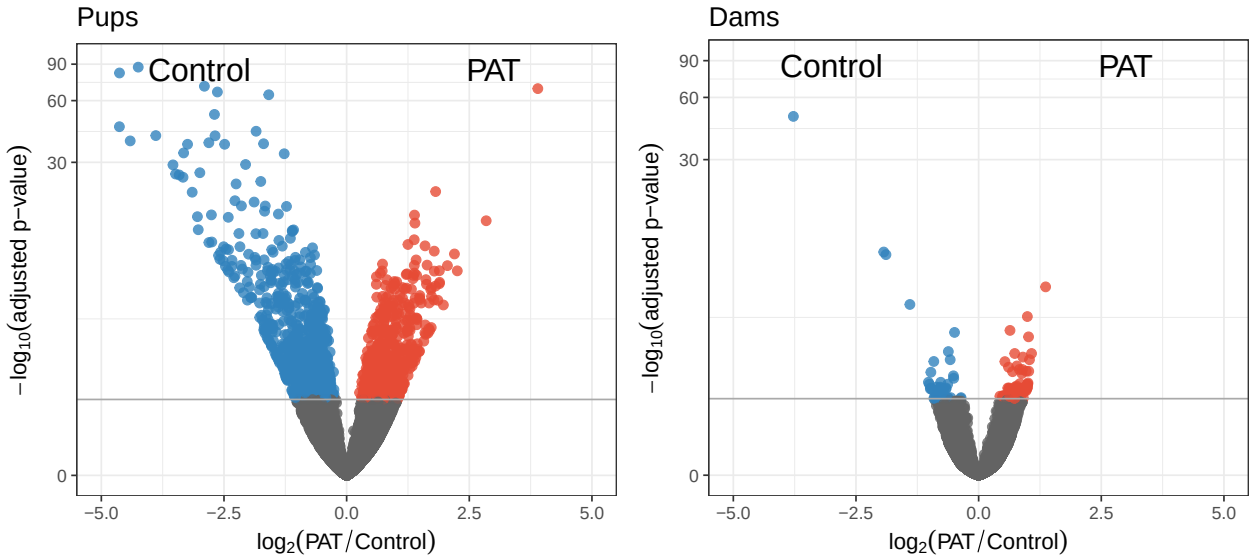


Figure 2g. Significant gene counts

This figure summarizes the number of genes found between the pups and the dams. It shows the dramatic difference between the two groups when antibiotics are given early in life.

```
# Number of siginicant genes (Pups)
## LFC > 0 (up)     : 552, 43%
## LFC < 0 (down)   : 746, 57%
summary(ddPups_res_sig_filter)
```

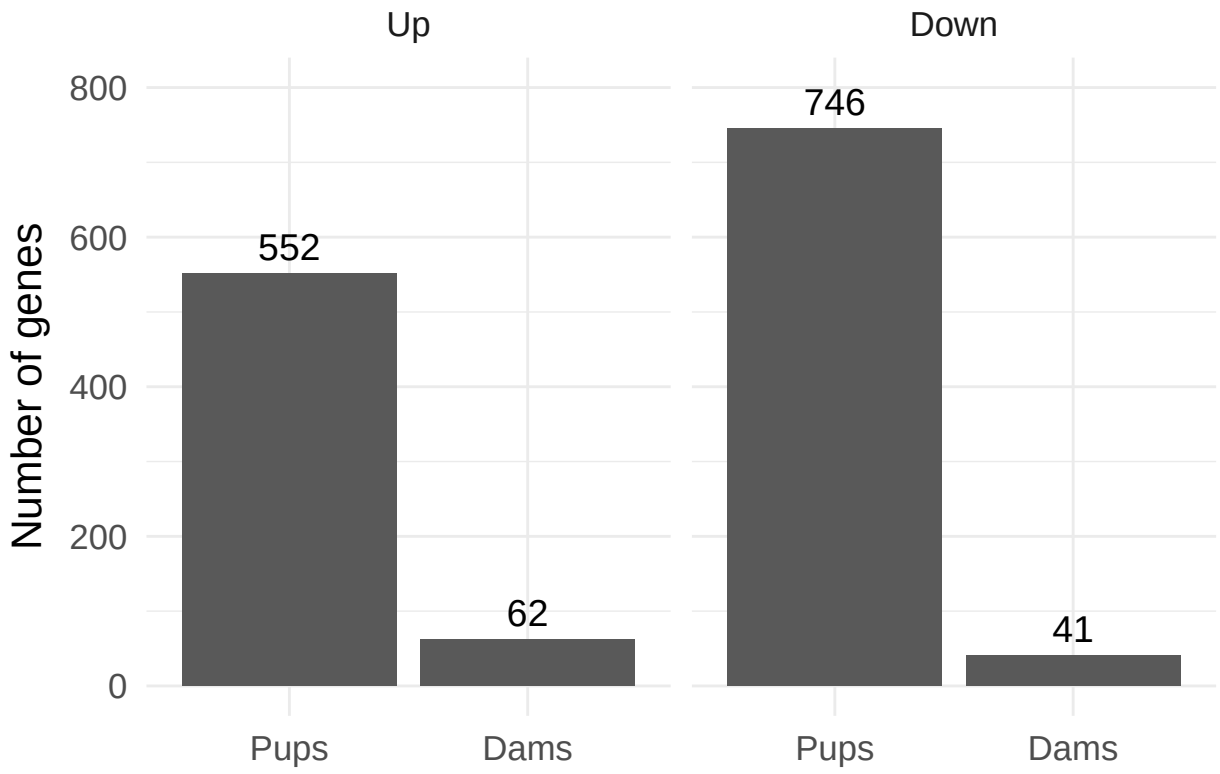
```
##
## out of 1298 with nonzero total read count
## adjusted p-value < 0.05
## LFC > 0 (up)     : 552, 43%
## LFC < 0 (down)   : 746, 57%
## outliers [1]     : 0, 0%
## low counts [2]   : 0, 0%
## (mean count < 7)
## [1] see 'cooksCutoff' argument of ?results
## [2] see 'independentFiltering' argument of ?results
```

```
# Number of siginicant genes (Dams)
## LFC > 0 (up)     : 62, 60%
## LFC < 0 (down)   : 41, 40%
summary(ddDams_res_sig_filter)
```

```
##
## out of 103 with nonzero total read count
## adjusted p-value < 0.05
## LFC > 0 (up)     : 62, 60%
## LFC < 0 (down)   : 41, 40%
## outliers [1]     : 0, 0%
## low counts [2]   : 0, 0%
## (mean count < 2)
## [1] see 'cooksCutoff' argument of ?results
## [2] see 'independentFiltering' argument of ?results
```

```
# Create dataframe to store numbers
gene_barplot = data.frame(Group = c("Pups", "Pups", "Dams", "Dams"),
                          Direction = c("Up", "Down", "Up", "Down"),
                          Number = c(552, 746, 62, 41))

# Plot using ggplot2
gene_barplot %>%
  mutate(Group = factor(Group, levels = c("Pups", "Dams"))) %>%
  mutate(Direction = factor(Direction, levels = c("Up", "Down"))) %>%
  ggplot(aes(x = Group, y = Number, group = "black")) +
  geom_col() +
  facet_grid(.~ Direction, switch = "y") +
  theme_minimal(base_size = 16) +
  xlab("") + ylab("Number of genes") +
  geom_text(aes(x = Group, y = Number, label = Number), size = 5, vjust = -0.5) +
  scale_y_continuous(limits = c(0, 800))
```




```
sessionInfo()
```

```
## R version 3.3.1 (2016-06-21)
## Platform: x86_64-apple-darwin13.4.0 (64-bit)
## Running under: OS X 10.10.5 (Yosemite)
##
## locale:
## [1] en_US.UTF-8/en_US.UTF-8/en_US.UTF-8/C/en_US.UTF-8/en_US.UTF-8
##
## attached base packages:
## [1] parallel  stats4    stats     graphics  grDevices utils     datasets
## [8] methods   base
##
## other attached packages:
##  [1] org.Mm.eg.db_3.4.0         AnnotationDbi_1.36.2
##  [3] dplyr_0.5.0                purrr_0.2.2
##  [5] readr_1.1.0                tidyr_0.6.1
##  [7] tibble_1.3.0               tidyverse_1.1.1
##  [9] RColorBrewer_1.1-2         pheatmap_1.0.8
## [11] fdrtool_1.2.15             ggthemes_3.4.0
## [13] ggplot2_2.2.1              DESeq2_1.14.1
## [15] SummarizedExperiment_1.4.0 Biobase_2.34.0
## [17] GenomicRanges_1.26.3       GenomeInfoDb_1.10.3
## [19] IRanges_2.8.1              S4Vectors_0.12.1
## [21] BiocGenerics_0.20.0
##
## loaded via a namespace (and not attached):
##  [1] httr_1.2.1          jsonlite_1.4        splines_3.3.1
##  [4] modelr_0.1.0        Formula_1.2-1       assertthat_0.2.0
##  [7] latticeExtra_0.6-28 cellranger_1.1.0    yaml_2.1.14
```

```
## [10] RSQLite_1.1-2       backports_1.0.5     lattice_0.20-35
## [13] digest_0.6.12       XVector_0.14.0      checkmate_1.8.2
## [16] rvest_0.3.2         colorspace_1.3-2    htmltools_0.3.5
## [19] Matrix_1.2-8        plyr_1.8.4          psych_1.7.3.21
## [22] XML_3.98-1.6        broom_0.4.2         haven_1.0.0
## [25] genefilter_1.56.0   zlibbioc_1.20.0     xtable_1.8-2
## [28] scales_0.4.1        BiocParallel_1.8.1  htmlTable_1.9
## [31] annotate_1.52.1     nnet_7.3-12         lazyeval_0.2.0
## [34] mnormt_1.5-5        readxl_1.0.0        survival_2.41-3
## [37] magrittr_1.5        memoise_1.1.0       evaluate_0.10
## [40] nlme_3.1-131        xml2_1.1.1          forcats_0.2.0
## [43] foreign_0.8-68      tools_3.3.1         data.table_1.10.4
## [46] hms_0.3             stringr_1.2.0       munsell_0.4.3
## [49] locfit_1.5-9.1      cluster_2.0.6       grid_3.3.1
## [52] RCurl_1.95-4.8      Rmisc_1.5           htmlwidgets_0.8
## [55] labeling_0.3        bitops_1.0-6        base64enc_0.1-3
## [58] rmarkdown_1.5       codetools_0.2-15    gtable_0.2.0
## [61] DBI_0.6-1           reshape2_1.4.2      R6_2.2.0
## [64] lubridate_1.6.0     gridExtra_2.2.1     knitr_1.15.1
## [67] Hmisc_4.0-2         rprojroot_1.2       stringi_1.1.5
## [70] Rcpp_0.12.10        geneplotter_1.52.0  rpart_4.1-11
## [73] acepack_1.4.1
```