Find differentially expressed genes using DESeq2 library functions and reference genome assembly SL5 from June 2022

Molly Davis and Ann Loraine

2023-10-03

Introduction

This Markdown analyzes RNA-Seq gene expression data from data file `results/muday-144-SL5_counts-salmon.txt` documented in `Documentation/muday-144_sample_sheet.xlsx` using differential expression analysis library DESeq2.

The input data file was generated by nf-core/rna-seq pipeline that aligned RNA-Seq data and produced RNA-Seq fragment counts for gene annotations from the June 2022 release of the tomato genome assembly SL5 reported in open access article Graph pangenome captures missing heritability and empowers tomato breeding.

Also, see the Sol Genomics Web page Tomato graph pangenome project.

In this Markdown, we aim to answer:

- How many genes were differentially expressed in treatment versus control comparisons?

Input data file summary:

The git hash (version) of `results/muday-144-SL5_counts-salmon.txt` was: `62fdd9c2ca32374188634d2d8ba60f9c9413b27d`

The data file contains heat-treated and non-heat-treated samples from four tomato genotypes: `are` (anthocyanin reduced mutant), VF36 (wild-type tomoato cultivar), and a VF36 line designated `F3H` containing a transgene encoding the `are` wild-type gene. Samples experienced the heat stress over a time course which included four time points.

In addition to answering the above question, we also aim to create a data file that we will use to compare differential expression results obtained here with differential expression results obtained using a different, but similar, R library called "edgeR." If the results are similar, we can be more confident that we are using these libraries correctly.

Analysis and Results

Load custom functions for analyzing and visualizing differential expression:

```
source("Common.R")
```

Load required counts data:

```
counts=getCounts(counts_fname,keep_description = T)
```

The table of RNA-Seq counts per gene loaded in the previous code chunk from file results/muday-144-SL5_counts-salmon.txt contained 36,648 rows corresponding to measured genes.

Define false discovery rate threshold for deciding whether a gene is differentially expressed:

```
Q=0.05
```

We will use 0.05, defined in the previous code chunk, to "call" a gene as differentially expressed.

As we test whether the treatment changed gene expression for each genotype and treatment duration combination, we will save results to a single data frame. At the end of the Markdown, we'll write this very large table to a file named `results/CvT-DESeq2-SL5.txt`.

Genotype *are*, anthocyanin-reduced mutant

```
group1_name = "A.28.15"
group2_name = "A.34.15"
results_table = getDeGenes(counts,group1_name,group2_name)
volcano_A1 <- volcano_plot(results_table, paste(group1_name,"vs",group2_name),
                           method=method)
all = results_table
```

Comparing group A.28.15 to A.34.15 found 23 genes with changed expression levels with false discovery rate Q less than or equal to 0.05.

```
group1_name = "A.28.30"
group2_name = "A.34.30"
results_table = getDeGenes(counts,group1_name,group2_name)
volcano_A2 <- volcano_plot(results_table, paste(group1_name,"vs",group2_name),
                           method=method)
all = rbind(all,results_table)
```

Comparing group A.28.30 to A.34.30 found 49 genes with changed expression levels with false discovery rate Q less than or equal to 0.05.

```
group1_name = "A.28.45"
group2_name = "A.34.45"
results_table = getDeGenes(counts,group1_name,group2_name)
volcano_A3 <- volcano_plot(results_table, paste(group1_name,"vs",group2_name),
                           method=method)
all = rbind(all,results_table)
```

Comparing group A.28.45 to A.34.45 found 63 genes with changed expression levels with false discovery rate Q less than or equal to 0.05.

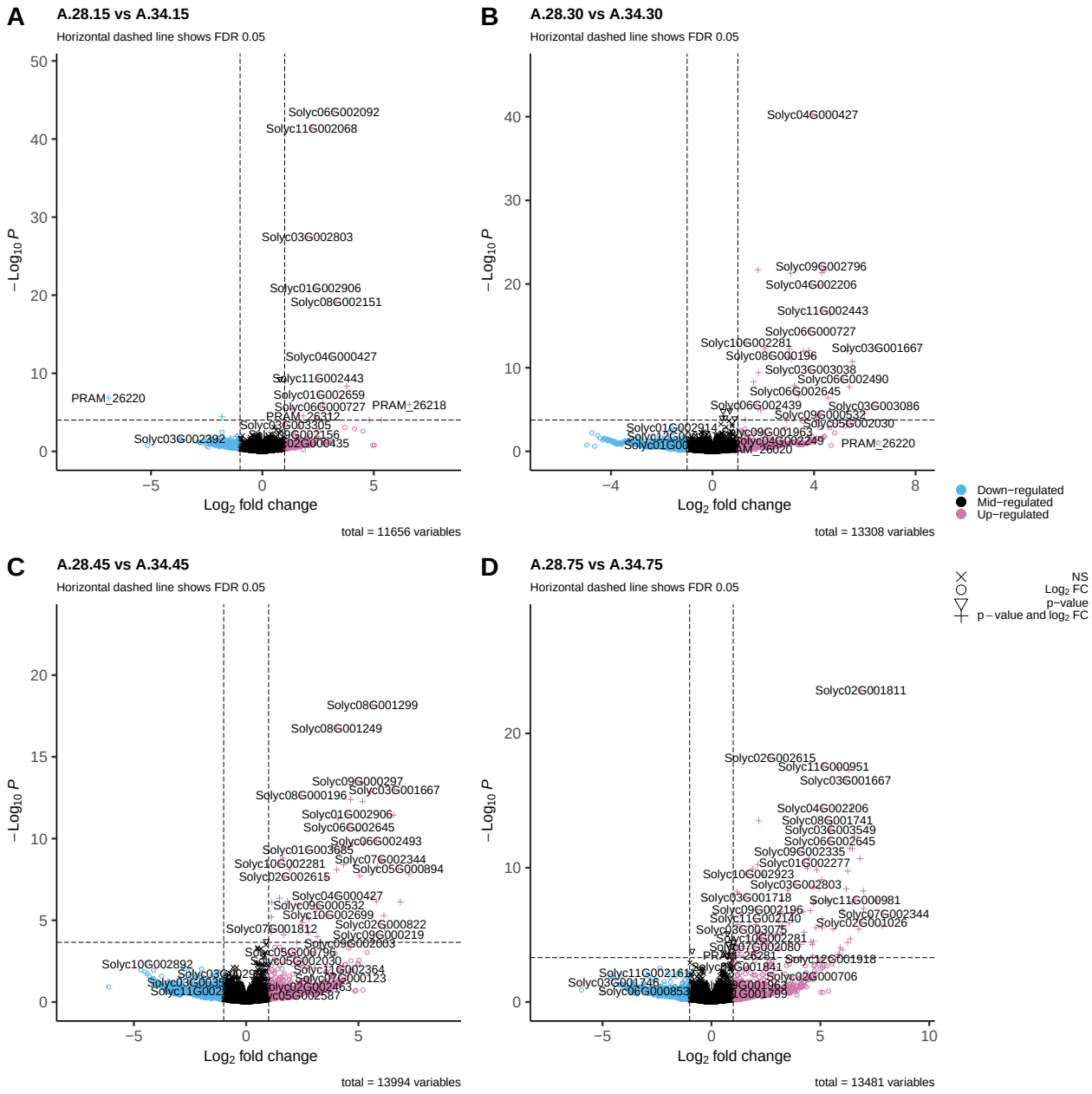
```
group1_name = "A.28.75"
group2_name = "A.34.75"
results_table = getDeGenes(counts,group1_name,group2_name)
volcano_A4 <- volcano_plot(results_table, paste(group1_name,"vs",group2_name),
                           method=method)
all = rbind(all,results_table)
```

Comparing group A.28.75 to A.34.75 found 134 genes with changed expression levels with false discovery rate Q less than or equal to 0.05.

The number of DE genes called at Q less than 0.05increased with treatment duration time.

Display volcano plots that summarize the above results:

```
A_combined<- ggpubr::ggarrange(volcano_A1, volcano_A2, volcano_A3, volcano_A4, # list of plots
                  labels = "AUTO",
                  font.label = list(size = 30),
                  common.legend = T, # COMMON LEGEND
                  legend = "right", # legend position
                  align = "hv", # Align them both, horizontal and vertical
                  nrow = 2,
                  ncol = 2)
A_combined
```



The volcano plots confirm the previous observation that the number of genes found to be differentially expressed in response to the heat treatment increased with treatment duration.

They also show that most of the changes were in the positive direction. Most of the genes that met the significance threshold (the horizontal line) were in the upper right quadrant of the plots, indicating that their expression levels increased in response to the treatment.

Genotype *VF36*, wild-type

```
group1_name = "V.28.15"
group2_name = "V.34.15"
results_table = getDeGenes(counts,group1_name,group2_name)
volcano_A1 <- volcano_plot(results_table, paste(group1_name,"vs",group2_name),
                           method=method)
all = rbind(all,results_table)
```

Comparing group V.28.15 to V.34.15 found 0 genes with changed expression levels with false discovery rate Q less than or equal to 0.05.

```
group1_name = "V.28.30"
group2_name = "V.34.30"
results_table = getDeGenes(counts,group1_name,group2_name)
volcano_A2 <- volcano_plot(results_table, paste(group1_name,"vs",group2_name),
                           method=method)
all = rbind(all,results_table)
```

Comparing group V.28.30 to V.34.30 found 5 genes with changed expression levels with false discovery rate Q less than or equal to 0.05.

```
group1_name = "V.28.45"
group2_name = "V.34.45"
results_table = getDeGenes(counts,group1_name,group2_name)
volcano_A3 <- volcano_plot(results_table, paste(group1_name,"vs",group2_name),
                           method=method)
all = rbind(all,results_table)
```

Comparing group V.28.45 to V.34.45 found 26 genes with changed expression levels with false discovery rate Q less than or equal to 0.05.

```
group1_name = "V.28.75"
group2_name = "V.34.75"
results_table = getDeGenes(counts,group1_name,group2_name)
volcano_A4 <- volcano_plot(results_table, paste(group1_name,"vs",group2_name),
                           method=method)
all = rbind(all,results_table)
```

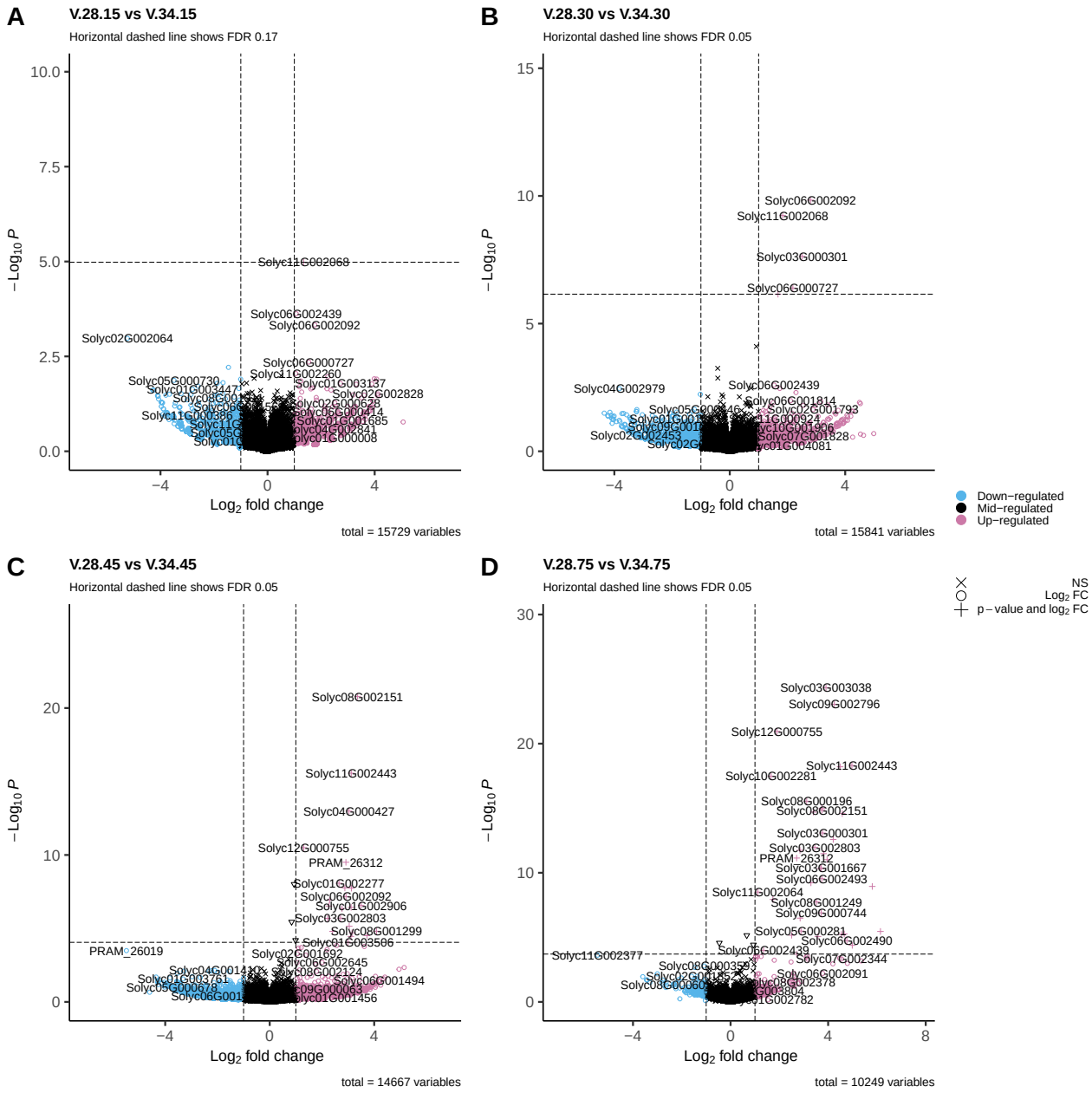
Comparing group V.28.75 to V.34.75 found 41 genes with changed expression levels with false discovery rate Q less than or equal to 0.05.

The number of DE genes called at Q less than or equal to 0.05 increased with treatment duration time.

Display volcano plots that summarize the above results:

```
A_combined<- ggpubr::ggarrange(volcano_A1, volcano_A2, volcano_A3, volcano_A4, # list of plots
                  labels = "AUTO",
                  font.label = list(size = 30),
                  common.legend = T, # COMMON LEGEND
                  legend = "right", # legend position
                  align = "hv", # Align them both, horizontal and vertical
                  nrow = 2,
```

```
                  ncol = 2)
A_combined
```



The volcano plots recapitulate the previous observation that the number of genes found to be differentially expressed in response to the heat treatment increased with treatment duration.

Second, nearly all changes that did meet achieve the target false discovery rate of 0.05 were changes in the positive direction with the average estimated treatment expression higher than control. This can be seen by observing how most of the points above the horizontal line, when it was drawn, were also the in the right half of the plot.

Note that plot A draws at line at a different FDR, a value larger than 0.05. If no gene achieved FDR of 0.05 or higher, then our code picks a new Q indicating the p value for the gene with the smallest adjusted p value observed.

Genotype *F3H-OX3*, F3H overexpression genotype

```
group1_name = "F.28.15"
group2_name = "F.34.15"
results_table = getDeGenes(counts,group1_name,group2_name)
volcano_A1 <- volcano_plot(results_table, paste(group1_name,"vs",group2_name),
                           method=method)
all = rbind(all,results_table)
```

Comparing group F.28.15 to F.34.15 found 1 genes with changed expression levels with false discovery rate Q less than or equal to 0.05.

```
group1_name = "F.28.30"
group2_name = "F.34.30"
results_table = getDeGenes(counts,group1_name,group2_name)
volcano_A2 <- volcano_plot(results_table, paste(group1_name,"vs",group2_name),
                           method=method)
all = rbind(all,results_table)
```

Comparing group F.28.30 to F.34.30 found 14 genes with changed expression levels with false discovery rate Q less than or equal to 0.05.

```
group1_name = "F.28.45"
group2_name = "F.34.45"
results_table = getDeGenes(counts,group1_name,group2_name)
volcano_A3 <- volcano_plot(results_table, paste(group1_name,"vs",group2_name),
                           method=method)
all = rbind(all,results_table)
```

Comparing group F.28.45 to F.34.45 found 22 genes with changed expression levels with false discovery rate Q less than or equal to 0.05.

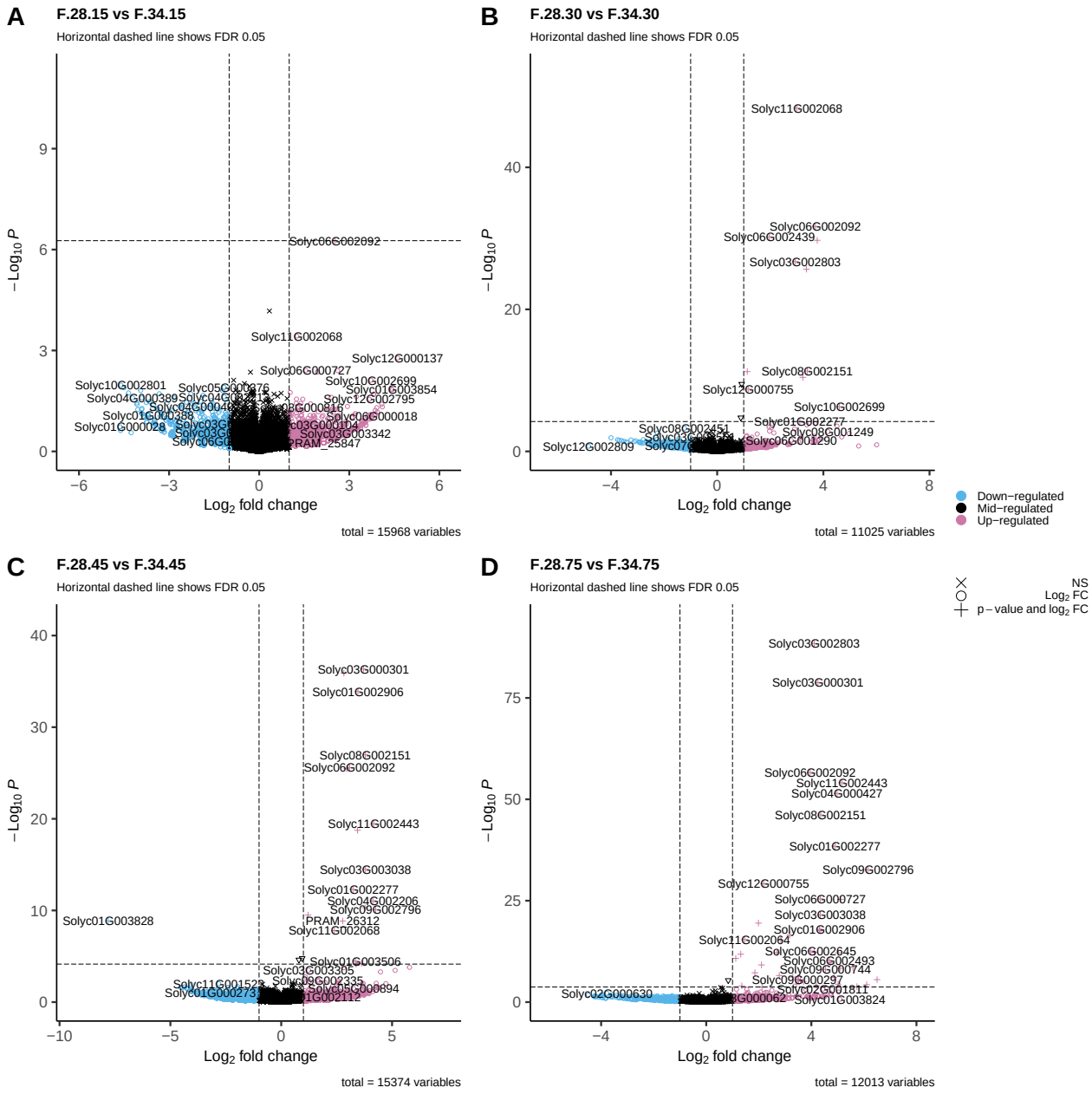
```
group1_name = "F.28.75"
group2_name = "F.34.75"
results_table = getDeGenes(counts,group1_name,group2_name)
volcano_A4 <- volcano_plot(results_table, paste(group1_name,"vs",group2_name),
                           method=method)
all = rbind(all,results_table)
```

Comparing group F.28.75 to F.34.75 found 47 genes with changed expression levels with false discovery rate Q less than or equal to 0.05.

The number of DE genes called at Q less than or equal to 0.05 increased with treatment duration time.

Display volcano plots that summarize the above results:

```
A_combined<- ggpubr::ggarrange(volcano_A1, volcano_A2, volcano_A3, volcano_A4, # list of plots
                  labels = "AUTO",
                  font.label = list(size = 30),
                  common.legend = T, # COMMON LEGEND
                  legend = "right", # legend position
                  align = "hv", # Align them both, horizontal and vertical
                  nrow = 2,
                  ncol = 2)
A_combined
```



The volcano plots recapitulate the previous observations that the number of genes found to be differentially expressed in response to the heat treatment increased with treatment duration.

Second, nearly all changes that did meet achieve the target false discovery rate of 0.05 were changes in the positive direction with the average estimated treatment expression higher than control. This can be seen by observing how most of the points above the horizontal line were also the in the right half of the plot.

Write results to a file for analysts' convenience

All DE results were saved to a large data frame, saved to data frame `all`, with 163305 rows and 10 columns.

Organize results and round numeric results to 3 significant digits:

```
all = all[,c("gene","group1","group2","baseMean","padj","pvalue","log2FoldChange","lfcSE","stat","descri
for (i in 4:9) {
```

```
  all[,i]=signif(all[,i],3)
}
```

Add SL4 gene name as a new column:

```
SL4_gene_names = getSL4GeneNames(all$description)
all$SL4 = SL4_gene_names
```

Write the data file:

```
write.table(all,file=out_fname,quote=F,row.names = F,sep="\t")
```

A file was created named results/CvT-DESeq2-SL5.txt that contains all the results.

All numeric values are rounded to three significant digits.

Explanation of columns:

- gene - SL5 gene measured
- group 1 - control group
- group 2 - treatment group
- baseMean - mean across samples
- padj - false discovery rate; adjusted p-value computed using method of Benjamini and Hochberg
- log2FoldChange - log2(group 2 average/group 1 average)
- lfcsE - log2FoldChange standard error
- stat - test statistic used to assess significance
- description - gene description
- SL4 - putative SL4 (June 2019 assembly release) gene counterpart

---

Discussion

Within each genotype, the number of genes exhibiting expression changes increased with treatment duration.

The different genotypes exhibited different numbers of differentially expressed genes, with the `are` genotype exhibiting the greatest number. This is consistent with the known `are` lower fertility mutant phenotype.

Conclusion

- The number of genes detected as changed increased with treatment duration.
- Most fold-changes for genes that changed in respond to the treatment were positive, indicating that expression levels increased.
- We observed more temperature-dependent gene expression differences in the `are` genotype than for the other two genotypes tested.

Session info

```
sessionInfo()
```

```
## R version 4.3.1 (2023-06-16)
## Platform: x86_64-apple-darwin20 (64-bit)
```

```
## Running under: macOS Big Sur 11.7.9
##
## Matrix products: default
## BLAS:   /Library/Frameworks/R.framework/Versions/4.3-x86_64/Resources/lib/libRblas.0.dylib
## LAPACK: /Library/Frameworks/R.framework/Versions/4.3-x86_64/Resources/lib/libRlapack.dylib;  LAPACK v
##
## locale:
## [1] en_US.UTF-8/en_US.UTF-8/en_US.UTF-8/C/en_US.UTF-8/en_US.UTF-8
##
## time zone: America/New_York
## tzcode source: internal
##
## attached base packages:
## [1] stats4    stats     graphics  grDevices utils     datasets  methods
## [8] base
##
## other attached packages:
##  [1] EnhancedVolcano_1.18.0      ggrepel_0.9.3
##  [3] ggplot2_3.4.3               DESeq2_1.40.2
##  [5] SummarizedExperiment_1.30.2 Biobase_2.60.0
##  [7] MatrixGenerics_1.12.3       matrixStats_1.0.0
##  [9] GenomicRanges_1.52.0        GenomeInfoDb_1.36.3
## [11] IRanges_2.34.1              S4Vectors_0.38.1
## [13] BiocGenerics_0.46.0         edgeR_3.42.4
## [15] limma_3.56.2                readxl_1.4.3
## [17] readr_2.1.4                 stringr_1.5.0
## [19] git2r_0.32.0
##
## loaded via a namespace (and not attached):
##  [1] gtable_0.3.4            xfun_0.40               rstatix_0.7.2
##  [4] lattice_0.21-8          tzdb_0.4.0              vctrs_0.6.3
##  [7] tools_4.3.1             bitops_1.0-7            generics_0.1.3
## [10] parallel_4.3.1          tibble_3.2.1            fansi_1.0.4
## [13] pkgconfig_2.0.3         Matrix_1.6-1            lifecycle_1.0.3
## [16] GenomeInfoDbData_1.2.10 farver_2.1.1            compiler_4.3.1
## [19] munsell_0.5.0           codetools_0.2-19        carData_3.0-5
## [22] htmltools_0.5.6         RCurl_1.98-1.12         yaml_2.3.7
## [25] car_3.1-2               tidyr_1.3.0             ggpubr_0.6.0
## [28] pillar_1.9.0            crayon_1.5.2            BiocParallel_1.34.2
## [31] DelayedArray_0.26.7     abind_1.4-5             tidyselect_1.2.0
## [34] locfit_1.5-9.8          digest_0.6.33           stringi_1.7.12
## [37] purrr_1.0.2             dplyr_1.1.3             labeling_0.4.3
## [40] cowplot_1.1.1           fastmap_1.1.1           grid_4.3.1
## [43] colorspace_2.1-0        cli_3.6.1               magrittr_2.0.3
## [46] S4Arrays_1.0.6          utf8_1.2.3              broom_1.0.5
## [49] withr_2.5.0             backports_1.4.1         scales_1.2.1
## [52] rmarkdown_2.24          XVector_0.40.0          gridExtra_2.3
## [55] ggsignif_0.6.4          cellranger_1.1.0        hms_1.1.3
## [58] evaluate_0.21           knitr_1.44              rlang_1.1.1
## [61] Rcpp_1.0.11             glue_1.6.2              rstudioapi_0.15.0
## [64] R6_2.5.1                zlibbioc_1.46.0
```