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

Ann Loraine and Molly Davis

2023-09-25

Prelude - Define variables

```
counts_fname = "results/muday-144-SL5_counts-salmon.txt"
sample_sheet_fname = "Documentation/muday-144_sample_sheet.xlsx"
out_fname = "results/CvT-edgeR-SL5.txt"
all = NULL
library(git2r)
r = revparse_single(counts_fname,"HEAD")
hash = sha(r)
method="edgeR"
```

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 edgeR.

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 known as "SL5," reported in open access article titled "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:

- `aba64266e33c8ecf5a6c33c8c6d4420a7b0f3b30`

The data file contains heat-treated and non-heat-treated samples from four tomato genotypes: `are` (anthocyanin reduced mutant), VF36 (wild-type tomato 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: 15 minutes, 30 minutes, 45 minutes, and 75 minutes. Two temperatures were tested: 28 degrees C, the control, and 34 degrees C, the heat treatment.

Samples and sample groups are named as follows:

- [genotype].[temperature].[treatment duration].[replicate number]

where genotype is A, V, or F; temperature is 28 or 34; and treatment duration is 15, 30, 25, or 75. There were three replicates per treatment group.

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 "DESeq2." If the results are similar, we can be more confident that we are using these libraries correctly.

Results

Load custom functions:

```
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.

Compute and tabulate differential expression

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-edgeR-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,method=method)
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 29 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,method=method)
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 57 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,method=method)
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 71 genes with changed expression levels with false discovery rate Q less than or equal to 0.05.

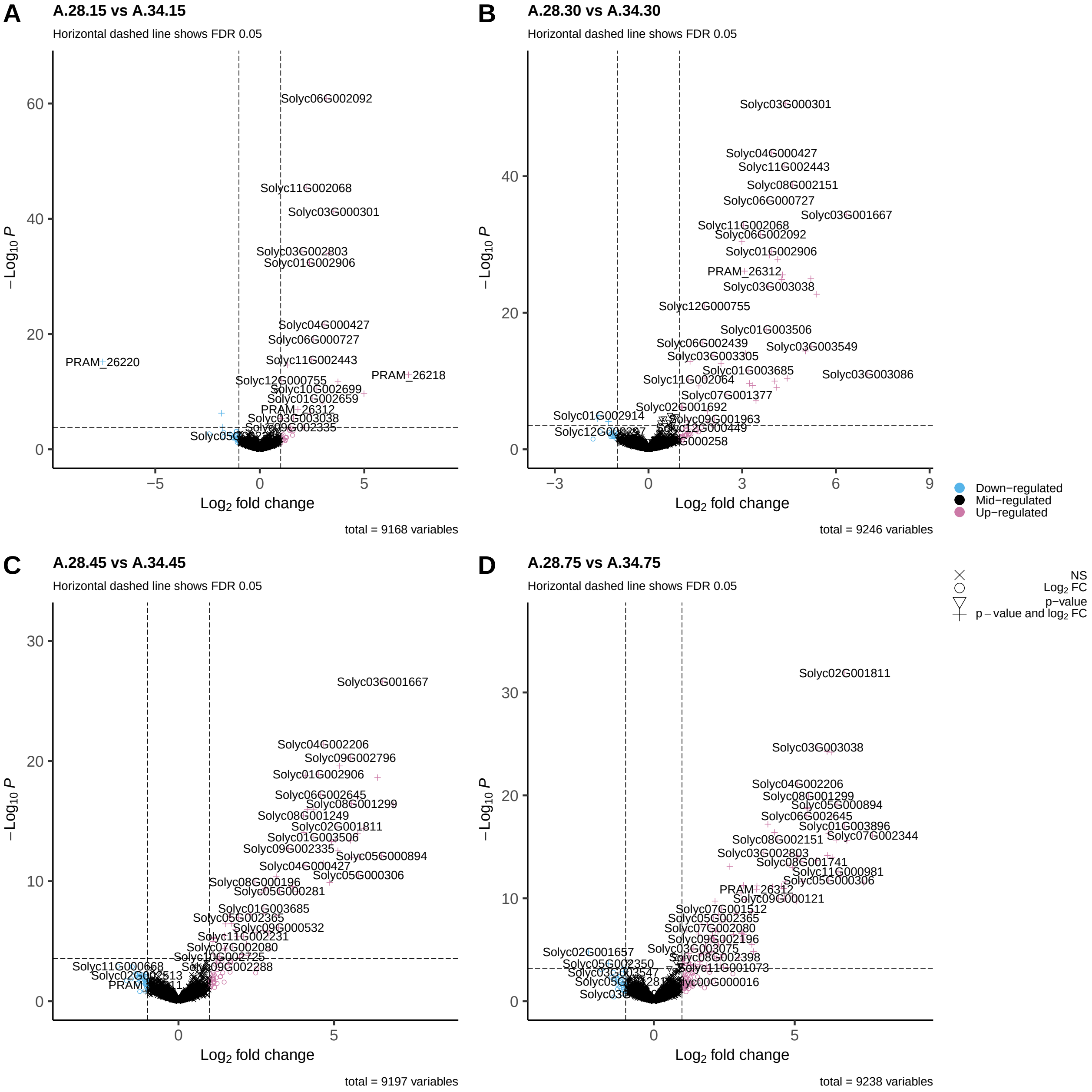
```
group1_name = "A.28.75"
group2_name = "A.34.75"
results_table = getDeGenes(counts,group1_name,group2_name,method=method)
volcano_A4 <- volcano_plot(results_table, paste(group1_name,"vs",group2_name),
                           method=method)
all = rbind(all,results_table)
```

Comparing group A.28.75 to A.34.75 found 129 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 above plot confirms that the number of genes found to differentially expressed within a time point increased with treatment duration.

Also, most of the genes that changed were higher in the treatment versus control.

Genotype *VF36*, wild-type

```
group1_name = "V.28.15"
group2_name = "V.34.15"
results_table = getDeGenes(counts,group1_name,group2_name,method=method)
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 2 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,method=method)
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 6 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,method=method)
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 34 genes with changed expression levels with false discovery rate Q less than or equal to 0.05.

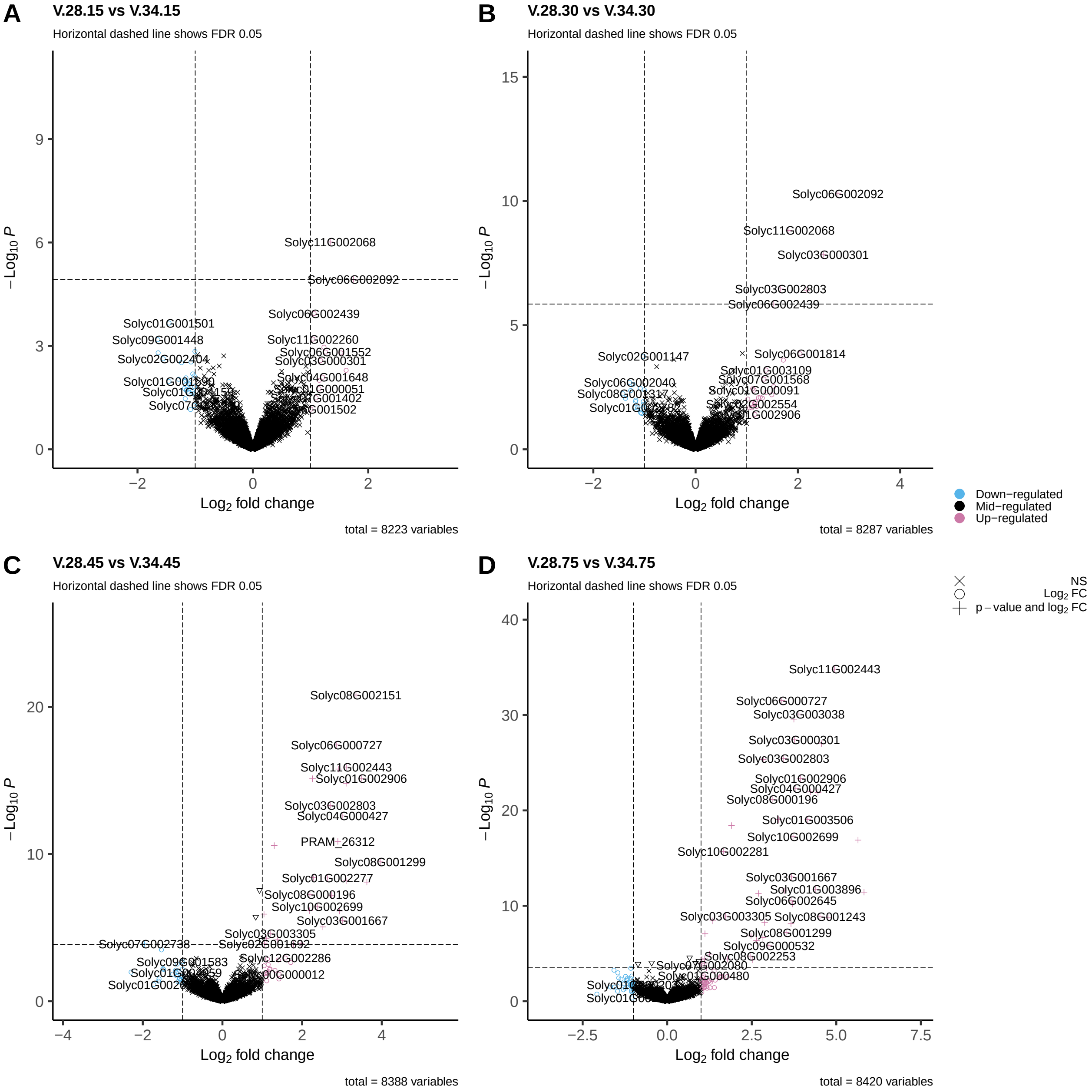
```
group1_name = "V.28.75"
group2_name = "V.34.75"
results_table = getDeGenes(counts,group1_name,group2_name,method=method)
volcano_A4 <- volcano_plot(results_table, paste(group1_name,"vs",group2_name),
                           method=method)
all = rbind(all,results_table)
```

Comparing group V.28.75 to V.34.75 found 54 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
```



As with the `are` genotype, the above plot confirms that the number of genes found to differentially expressed within a time point increased with treatment duration.

Also, most of the genes that changed were higher in the treatment versus control.

Genotype *F3H-OX3*, F3H overexpression genotype

```
group1_name = "F.28.15"
group2_name = "F.34.15"
results_table = getDeGenes(counts,group1_name,group2_name,method=method)
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,method=method)
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 24 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,method=method)
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 28 genes with changed expression levels with false discovery rate Q less than or equal to 0.05.

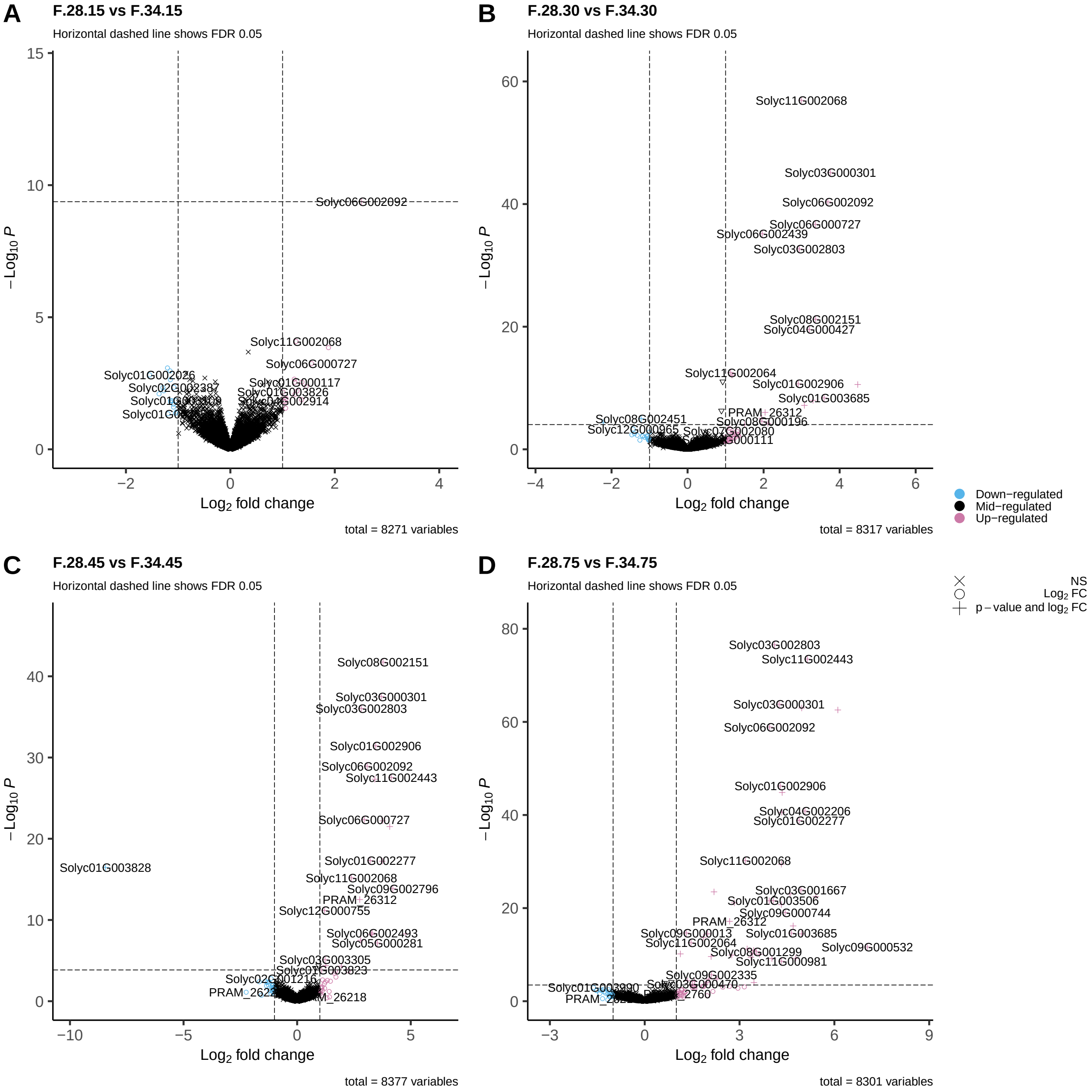
```
group1_name = "F.28.75"
group2_name = "F.34.75"
results_table = getDeGenes(counts,group1_name,group2_name,method=method)
volcano_A4 <- volcano_plot(results_table, paste(group1_name,"vs",group2_name),
                           method=method)
all = rbind(all,results_table)
```

Comparing group F.28.75 to F.34.75 found 55 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_table:

```
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
```



As with the `are` genotype, the above plot confirms that the number of genes found to differentially expressed within a time point increased with treatment duration.

Also, most of the genes that changed were higher in the treatment versus control. # Write results to a file for analysts' convenience

All DE results were saved to a large data frame, saved to data frame `all`, with 103433 rows and 9 columns.

Format and organize results:

```r
all$logFC=round(all$logFC,3)
all$logCPM=round(all$logCPM,3)
all$Q=signif(all$Q,3)
all$PValue=signif(all$PValue,3)
all = all[,c("gene","group1","group2","Q","PValue","logFC",
             "logCPM","description")]
```

Add SL4 gene name as a new column:

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

Write all the results to a data file:

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

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

Explanation of columns:

- gene - SL5 gene measured
- group 1 - control group
- group 2 - treatment group
- Q - false discovery rate; adjusted p-value computed using method of Benjamini and Hochberg
- PValue - nominal (unadjusted) p value
- logFC - log2(group 2 average/group 1 average)
- logCPM - ?
- description - gene description migrated from SL4 gene counterpart, if available
- SL4 - putative SL4 (June 2019 assembly release) gene counterpart

---

Discussion

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

Note that 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.

However, comparing numbers of differentially expressed across comparisons is not the best way to assess the influence of genotype on changes in gene expression detected between treatment and control samples. A better, more rigorous approach is to *simultaneously* test the interaction between genotype and treatment, by testing an interaction term in a linear model. We will do this in a different Markdown document.

Conclusion

- The number of genes detected as changed increased with treatment duration.

- The data suggest, but do not prove, that heat stress triggers greater changes in gene expression for 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
##
## 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] splines_4.3.1           cowplot_1.1.1           fastmap_1.1.1
## [43] grid_4.3.1              colorspace_2.1-0        cli_3.6.1
## [46] magrittr_2.0.3          S4Arrays_1.0.6          utf8_1.2.3
## [49] broom_1.0.5             withr_2.5.0             backports_1.4.1
## [52] scales_1.2.1            rmarkdown_2.24          XVector_0.40.0
## [55] gridExtra_2.3           ggsignif_0.6.4          cellranger_1.1.0
## [58] hms_1.1.3               evaluate_0.21           knitr_1.44
## [61] rlang_1.1.1             Rcpp_1.0.11             glue_1.6.2
## [64] rstudioapi_0.15.0       R6_2.5.1                zlibbioc_1.46.0
```