Volcano Plot Visualisation: Differential Gene Expression BY4742 vs SY14

Marwa

December 2025

Purpose

This script generates a volcano plot to visualise differential gene expression between wild-type (BY4742) and SY14 chromosome fusion strains.

Input Data

- `sig_DESeq_with_protein_coding.csv`: DESeq2 results

Output

- `volcano_plot.png`: Unlabeled volcano plot
- `volcano_plot_labeled.png`: Volcano plot with top gene labels

Software Versions

```
# R version
R.version.string
```

```
## [1] "R version 4.5.1 (2025-06-13)"
```

```
# Package versions
packageVersion("ggplot2")
```

```
## [1] '4.0.0'
```

```
packageVersion("dplyr")
```

```
## [1] '1.1.4'
```

```r
packageVersion("ggrepel")
```

```
## [1] '0.9.6'
```

- R: 4.5.1 (2025-06-13)
- ggplot2: 4.0.0
- dplyr: 1.1.4
- ggrepel: 0.9.6

Load libraries

```r
library(ggplot2)
library(dplyr)
```

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```r
library(ggrepel)
```

```r
#Load and inspect data
results <- read.csv("sig_DESeq_with_protein_coding.csv", row.names = 1)

head(results)
```

```
##           baseMean log2FoldChange     lfcSE       stat       pvalue         padj
## YAL067C   84.65606       1.487849 0.2208065   6.738246 1.603097e-11 1.067591e-09
## YAR050W  107.21540      -3.509134 0.2571567 -13.645895 2.135756e-42 1.265862e-39
## YAR066W   15.40684      -7.436504 1.2480286  -5.958601 2.544071e-09 1.314795e-07
## YAR071W  253.73732     -11.476910 1.1840988  -9.692527 3.244012e-22 5.196556e-20
## YBL112C   18.33900      -7.684153 1.2388031  -6.202885 5.543747e-10 3.159403e-08
## YBL113C  343.42870     -10.951368 1.1821636  -9.263835 1.972187e-20 2.597589e-18
##          gene_name   gene_biotype
## YAL067C       SEO1 protein_coding
## YAR050W       FLO1 protein_coding
## YAR066W       <NA> protein_coding
## YAR071W      PHO11 protein_coding
## YBL112C       <NA> protein_coding
## YBL113C       <NA> protein_coding
```

```r
colnames(results)
```

```
## [1] "baseMean"       "log2FoldChange" "lfcSE"          "stat"
## [5] "pvalue"         "padj"           "gene_name"      "gene_biotype"
```

```r
str(results)
```

```
## 'data.frame':    103 obs. of  8 variables:
##  $ baseMean      : num  84.7 107.2 15.4 253.7 18.3 ...
##  $ log2FoldChange: num  1.49 -3.51 -7.44 -11.48 -7.68 ...
##  $ lfcSE         : num  0.221 0.257 1.248 1.184 1.239 ...
##  $ stat          : num  6.74 -13.65 -5.96 -9.69 -6.2 ...
##  $ pvalue        : num  1.60e-11 2.14e-42 2.54e-09 3.24e-22 5.54e-10 ...
##  $ padj          : num  1.07e-09 1.27e-39 1.31e-07 5.20e-20 3.16e-08 ...
##  $ gene_name     : chr  "SEO1" "FLO1" NA "PHO11" ...
##  $ gene_biotype  : chr  "protein_coding" "protein_coding" "protein_coding" "protein_coding" ...
```

```r
sum(is.na(results$padj))
```

```
## [1] 0
```

```r
sum(is.na(results$log2FoldChange))
```

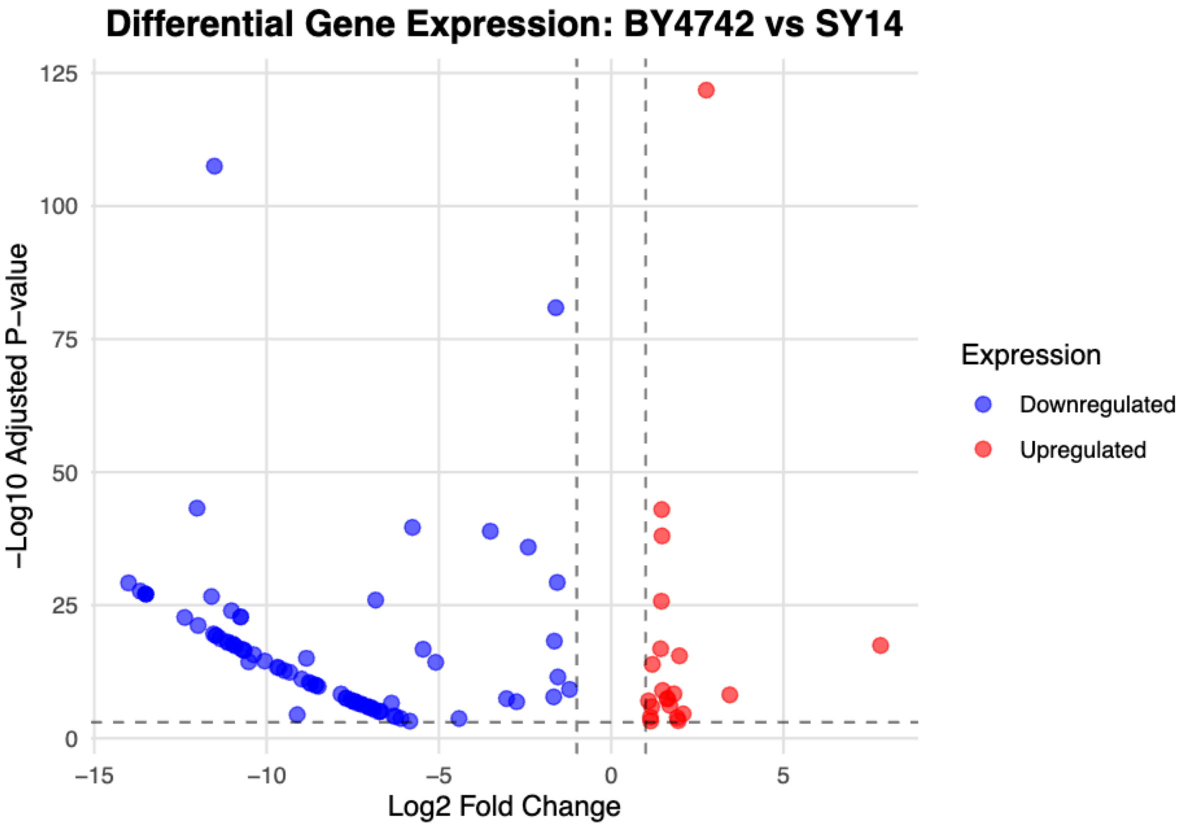
```
## [1] 0
```

```r
# Set thresholds
padj_threshold <- 0.001
log2fc_threshold <- 1

# Creating significance categories
results <- results %>%
  mutate(
    gene_type = case_when(
      padj < padj_threshold & log2FoldChange > log2fc_threshold ~ "Upregulated",
      padj < padj_threshold & log2FoldChange < -log2fc_threshold ~ "Downregulated",
      TRUE ~ "Not Significant"
    )
  )
```

```r
#Creating volcano plot
volcano_plot <- ggplot(results, aes(x = log2FoldChange, y = -log10(padj), color = gene_type)) +
  geom_point(alpha = 0.6, size = 2.5) +
  scale_color_manual(
    values = c("Upregulated" = "red",
               "Downregulated" = "blue",
               "Not Significant" = "grey"),
    name = "Expression"
  ) +
  geom_vline(xintercept = c(-log2fc_threshold, log2fc_threshold),
             linetype = "dashed", color = "black", alpha = 0.5) +
  geom_hline(yintercept = -log10(padj_threshold),
             linetype = "dashed", color = "black", alpha = 0.5) +
  labs(
    title = "Differential Gene Expression: BY4742 vs SY14",
    x = "Log2 Fold Change",
    y = "-Log10 Adjusted P-value"
  ) +
  theme_minimal() +
  theme(
    plot.title = element_text(hjust = 0.5, size = 14, face = "bold"),
    legend.position = "right",
    panel.grid.major = element_line(color = "grey90"),
    panel.grid.minor = element_blank()
  )

# Display the plot
print(volcano_plot)
```



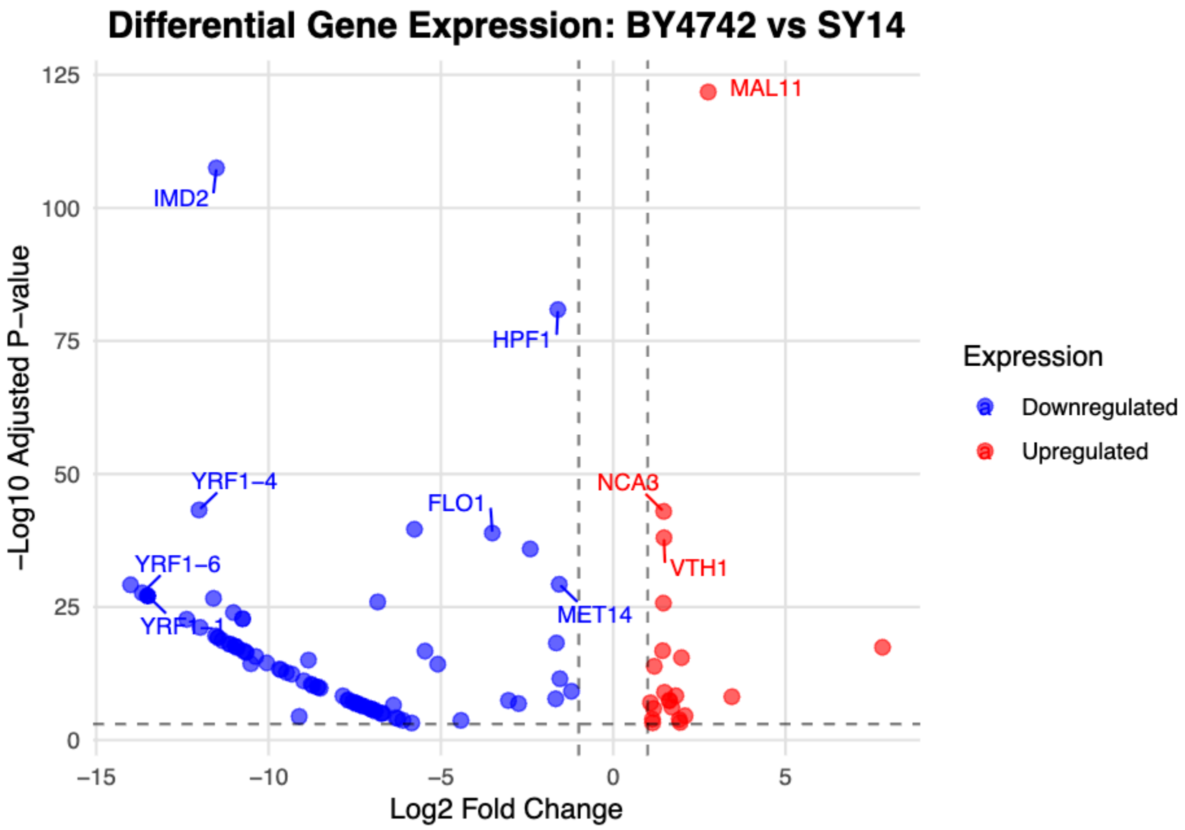
```
# Save the plot
ggsave("volcano_plot.png", plot = volcano_plot, width = 10, height = 8, dpi = 300)
```

```
#Adding gene labels for top significant genes
library(ggrepel)

#Top 10 most significant genes
top_genes <- results %>%
  filter(!is.na(gene_name)) %>%  # Remove genes without names
  arrange(padj) %>%
  head(10)

volcano_plot_labeled <- volcano_plot +
  geom_text_repel(
    data = top_genes,
    aes(label = gene_name),
    size = 3,
    max.overlaps = 15,
    box.padding = 0.5
  )

print(volcano_plot_labeled)
```



```
# Save labeled version
ggsave("volcano_plot_labeledFinal.png", plot = volcano_plot_labeled,
       width = 10, height = 8, dpi = 300)
```

```
# Summary
cat("Total protein-coding genes:", nrow(results), "\n")
```

```
## Total protein-coding genes: 103
```

```
cat("Upregulated in SY14:", sum(results$expression == "Upregulated"), "\n")
```

```
## Upregulated in SY14: 0
```

```
cat("Downregulated in SY14:", sum(results$expression == "Downregulated"), "\n")
```

```
## Downregulated in SY14: 0
```

```
cat("Not significant:", sum(results$expression == "Not significant"), "\n")
```

```
## Not significant: 0
```