Differential Gene Expression Analysis Report

Bioinformatics Analysis

PH. Fouad Mohsen

٢٠٢٦-٠١-١٨

Table of Contents

```
# 1. Setup options and load libraries
knitr::opts_chunk$set(echo = TRUE, warning = FALSE, message = FALSE, fig.widt
h = 8, fig.height = 6)
library(DESeq2)
```

```
## Loading required package: S4Vectors
```

```
## Loading required package: stats4
```

```
## Loading required package: BiocGenerics
```

```
## Loading required package: generics
```

```
##
## Attaching package: 'generics'
```

```
## The following objects are masked from 'package:base':
##
##     as.difftime, as.factor, as.ordered, intersect, is.element, setdiff,
##     setequal, union
```

```
##
## Attaching package: 'BiocGenerics'
```

```
## The following objects are masked from 'package:stats':
##
##     IQR, mad, sd, var, xtabs
```

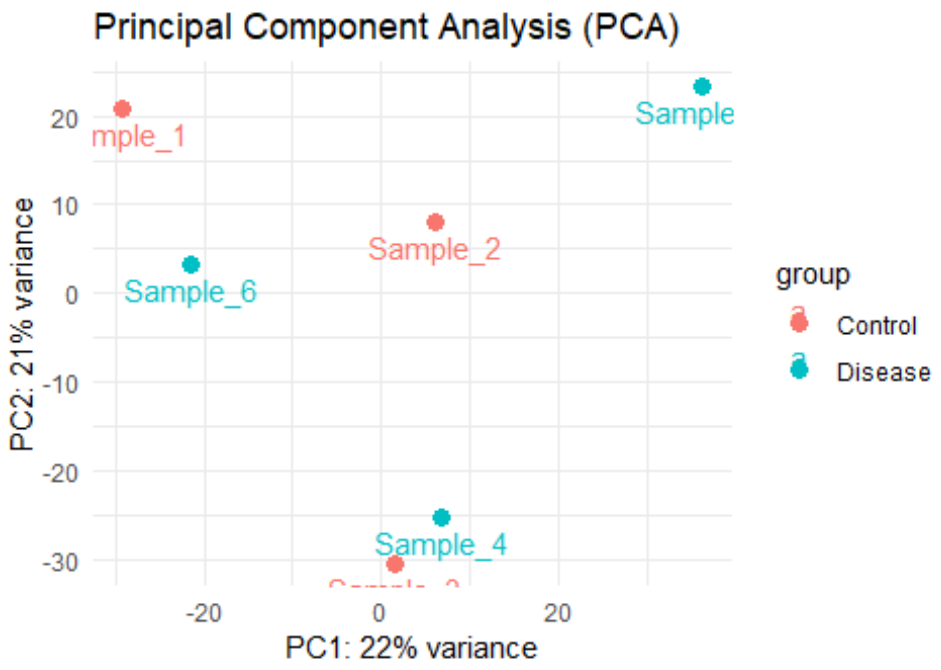
```
## The following objects are masked from 'package:base':
##
##     anyDuplicated, aperm, append, as.data.frame, basename, cbind,
##     colnames, dirname, do.call, duplicated, eval, evalq, Filter, Find,
##     get, grep, grepl, is.unsorted, lapply, Map, mapply, match, mget,
##     order, paste, pmax, pmax.int, pmin, pmin.int, Position, rank,
##     rbind, Reduce, rownames, sapply, saveRDS, table, tapply, unique,
##     unsplit, which.max, which.min
```

```
##
## Attaching package: 'S4Vectors'

## The following object is masked from 'package:utils':
##
##     findMatches

## The following objects are masked from 'package:base':
##
##     expand.grid, I, unname

## Loading required package: IRanges

##
## Attaching package: 'IRanges'

## The following object is masked from 'package:grDevices':
##
##     windows

## Loading required package: GenomicRanges

## Loading required package: Seqinfo

## Loading required package: SummarizedExperiment

## Loading required package: MatrixGenerics

## Loading required package: matrixStats

##
## Attaching package: 'MatrixGenerics'

## The following objects are masked from 'package:matrixStats':
##
##     colAlls, colAnyNAs, colAnys, colAvgsPerRowSet, colCollapse,
##     colCounts, colCummaxs, colCummins, colCumprods, colCumsums,
##     colDiffs, colIQRDiffs, colIQRs, colLogSumExps, colMadDiffs,
##     colMads, colMaxs, colMeans2, colMedians, colMins, colOrderStats,
##     colProds, colQuantiles, colRanges, colRanks, colSdDiffs, colSds,
##     colSums2, colTabulates, colVarDiffs, colVars, colWeightedMads,
##     colWeightedMeans, colWeightedMedians, colWeightedSds,
##     colWeightedVars, rowAlls, rowAnyNAs, rowAnys, rowAvgsPerColSet,
##     rowCollapse, rowCounts, rowCummaxs, rowCummins, rowCumprods,
##     rowCumsums, rowDiffs, rowIQRDiffs, rowIQRs, rowLogSumExps,
##     rowMadDiffs, rowMads, rowMaxs, rowMeans2, rowMedians, rowMins,
##     rowOrderStats, rowProds, rowQuantiles, rowRanges, rowRanks,
##     rowSdDiffs, rowSds, rowSums2, rowTabulates, rowVarDiffs, rowVars,
##     rowWeightedMads, rowWeightedMeans, rowWeightedMedians,
##     rowWeightedSds, rowWeightedVars

## Loading required package: Biobase
```

```
## Welcome to Bioconductor
##
##     Vignettes contain introductory material; view with
##     'browseVignettes()'. To cite Bioconductor, see
##     'citation("Biobase")', and for packages 'citation("pkgname")'.

##
## Attaching package: 'Biobase'

## The following object is masked from 'package:MatrixGenerics':
##
##     rowMedians

## The following objects are masked from 'package:matrixStats':
##
##     anyMissing, rowMedians

library(ggplot2)
library(pheatmap)
library(dplyr)

##
## Attaching package: 'dplyr'

## The following object is masked from 'package:Biobase':
##
##     combine

## The following object is masked from 'package:matrixStats':
##
##     count

## The following objects are masked from 'package:GenomicRanges':
##
##     intersect, setdiff, union

## The following object is masked from 'package:Seqinfo':
##
##     intersect

## The following objects are masked from 'package:IRanges':
##
##     collapse, desc, intersect, setdiff, slice, union

## The following objects are masked from 'package:S4Vectors':
##
##     first, intersect, rename, setdiff, setequal, union

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

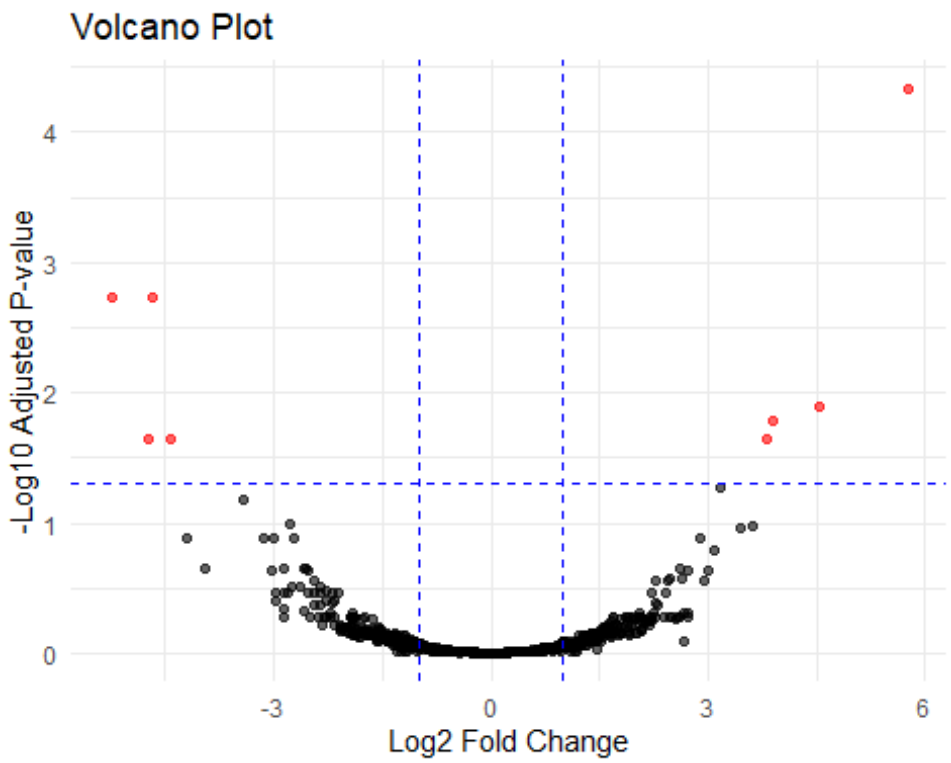
```
## The following object is masked from 'package:generics':
##
##     explain

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

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

set.seed(123)
# 2. Generate Simulated Data
genes <- 1000
samples <- 6

counts <- matrix(
  rnbinom(genes * samples, mu = 100, size = 1),
  nrow = genes
)
rownames(counts) <- paste0("Gene_", 1:genes)
colnames(counts) <- paste0("Sample_", 1:samples)

metadata <- data.frame(
  condition = factor(c("Control", "Control", "Control", "Disease", "Disease",
"Disease")),
  row.names = colnames(counts)
)

# 3. Run DESeq2 Analysis
dds <- DESeqDataSetFromMatrix(countData = counts, colData = metadata, design
= ~ condition)

## converting counts to integer mode

dds <- DESeq(dds)

## estimating size factors

## estimating dispersions

## gene-wise dispersion estimates

## mean-dispersion relationship

## -- note: fitType='parametric', but the dispersion trend was not well captu
red by the
##    function: y = a/x + b, and a local regression fit was automatically sub
stituted.
##    specify fitType='local' or 'mean' to avoid this message next time.
```

```
## final dispersion estimates

## fitting model and testing

res <- results(dds)

# Data Cleaning (Remove NA values)
res_df <- as.data.frame(res) %>% filter(!is.na(padj))
vsd <- vst(dds, blind = FALSE)

## -- note: fitType='parametric', but the dispersion trend was not well captu
red by the
##    function: y = a/x + b, and a local regression fit was automatically sub
stituted.
##    specify fitType='local' or 'mean' to avoid this message next time.

# Principal Component Analysis to check sample clustering
plotPCA(vsd, intgroup = "condition") +
  theme_minimal() +
  geom_text(aes(label = name), vjust = 1.5) +
  labs(title = "Principal Component Analysis (PCA)")

## using ntop=500 top features by variance
```

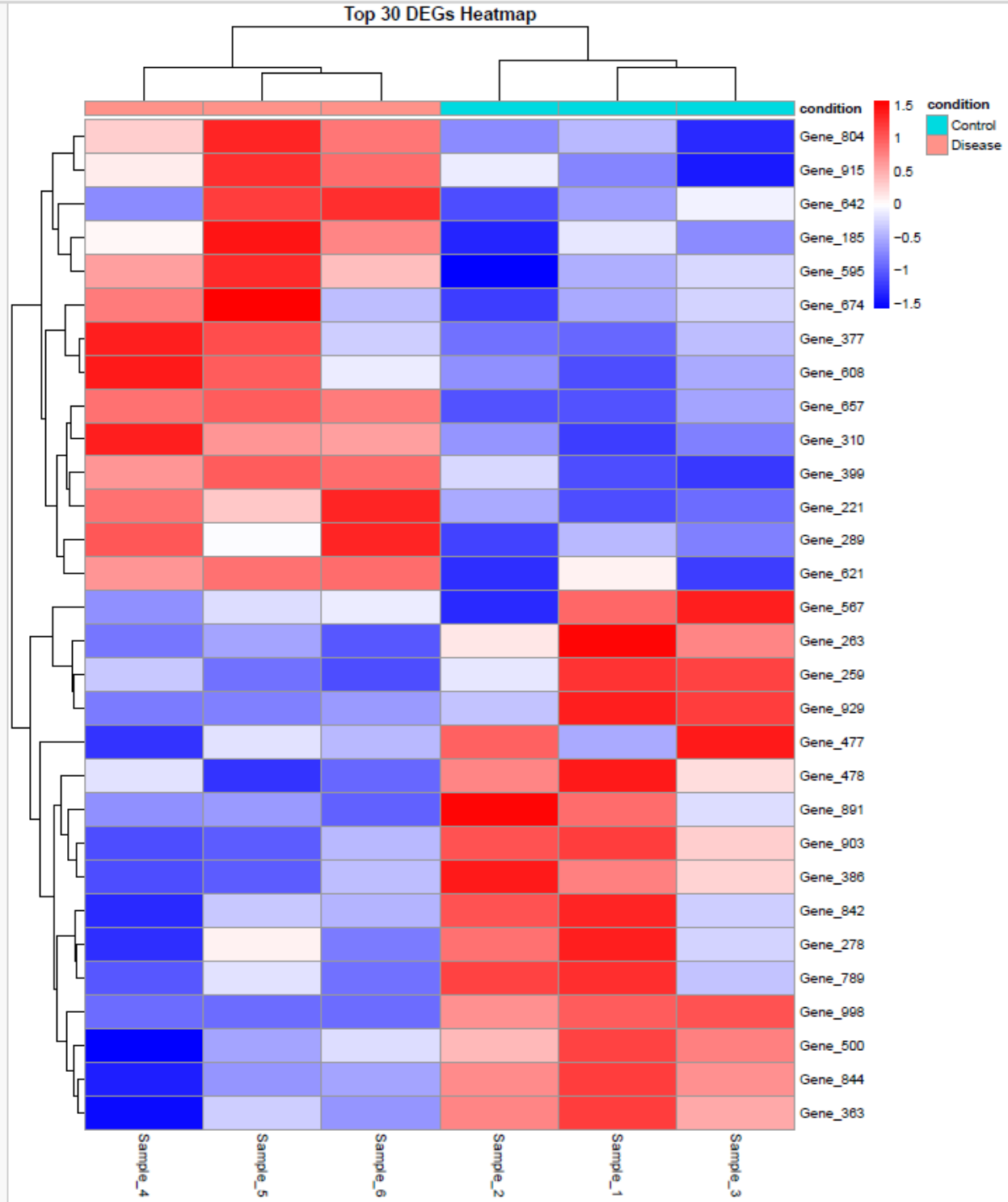



```
# Visualize significant genes based on Fold Change and P-value
ggplot(res_df, aes(x = log2FoldChange, y = -log10(padj))) +
  geom_point(aes(color = padj < 0.05 & abs(log2FoldChange) > 1), alpha = 0.6)
+
```

```
  scale_color_manual(values = c("black", "red")) +
  geom_hline(yintercept = -log10(0.05), linetype = "dashed", color = "blue")
+
  geom_vline(xintercept = c(-1, 1), linetype = "dashed", color = "blue") +
  theme_minimal() +
  labs(title = "Volcano Plot", x = "Log2 Fold Change", y = "-Log10 Adjusted P
-value") +
  theme(legend.position = "none")
```




```
# Heatmap of the top 30 most significantly differentially expressed genes
top_genes <- head(order(res$padj), 30)
heatmap_data <- assay(vsd)[top_genes, ]

pheatmap(
  heatmap_data,
  scale = "row",
  annotation_col = metadata,
  main = "Top 30 DEGs Heatmap",
  color = colorRampPalette(c("blue", "white", "red"))(150)
)
```




```r
# Create directory and save results to CSV
if (!dir.exists("DESeq2_Results")) dir.create("DESeq2_Results")
write.csv(res_df, "DESeq2_Results/Full_Results.csv")
```

```r
# Filter and display the top 10 significant genes
sig_results <- res_df %>% filter(padj < 0.05 & abs(log2FoldChange) > 1)
knitr::kable(head(sig_results, 10), caption = "Top 10 Significant Genes")
```

Top 10 Significant Genes

	baseMean	log2FoldChange	lfcSE	stat	pvalue	padj
Gene_310	٣٢,٨٥٠٤٥	٣,٨٩٥٠٨٦	٠,٩٩٠٤٢٤	٣,٩٣٢٧٤٦	٠,٠٠٠٠٨٤٠	٠,٠١٦٤٤٣٥
Gene_377	٦٥,١٠٨٠٩	٤,٥٣٠٤٧٦	١,١١٩٨٢٢	٤,٠٤٥٧١١	٠,٠٠٠٠٥٢٢	٠,٠١٢٧٦٧٣
Gene_399	٥٢,٥٣٥٠٤	٣,٨١٥٥٤٩	١,٠٢٠٧٩٦	٣,٧٣٧٨١٦	٠,٠٠٠١٨٥٦	٠,٠٢٢٧١٥٩
Gene_478	٧٨,٠٥٧٥٣	-٤,٤١٣٢٧٦	١,١٧٧٠٩٨	-٣,٧٤٩٢٨٤	٠,٠٠٠١٧٧٣	٠,٠٢٢٧١٥٩
Gene_657	٧٤,٣٢٤٥٠	٥,٧٥٠٦١٦	١,٠٥٣١٦٤	٥,٤٦٠٣٢٥	٠,٠٠٠٠٠٠	٠,٠٠٠٤٦٥
Gene_842	٥٩,٨٣١٦٢	-٤,٧٣٦٢١٣	١,٢٥٠٧٢٢	-٣,٧٨٦٧٨٥	٠,٠٠٠١٥٢٦	٠,٠٢٢٧١٥٩
Gene_844	٨٩,٤٦١٤٥	-٤,٦٨٦٣٩٠	١,٠٣٣٠٩٩	-٤,٥٣٦٢٤٥	٠,٠٠٠٠٠٥٧	٠,٠٠١٨٦٨٧
Gene_891	٧٢,٧٧٤٧٠	-٥,٢٤٢٧٠٩	١,١٥٥٣٨١	-٤,٥٣٧٦٤٦	٠,٠٠٠٠٠٥٧	٠,٠٠١٨٦٨٧