RNA seq analysis for Clinical Neuro samples: V2

Ephifania Geza

24/11/2022

How different is version 2 from version 1?

In version 2,

(1) three analysis groups (1,2 and 3) were considered (with three as the reference) as compared to 8 groups in version 1 (see, AnaGroups_RNAseq_V1_2.xlsx for more information regarding these groups for the both V1 and V2).
(2) there are 24 samples yet version one had 43.

Best practices for differentail expression analyses: DESeq2

We report the results for differential expression analysis using the **DESeq2** tool. This analysis involves 24 clinical samples. We aligned the raw sequence reads (paired) to the reference genome **GRCh37** using the *STAR* alignment tool. The number of reads that mapped to each gene were counted using a pseudo-count method, **SALMON**. Quality check, read trimming, mapping and counting was done using the https://github.com/nf-core/rnaseq pipeline. While count normalization (creation of the DESeqDataSet from a matrix), exploratory data analysis (identifying outliers & sources of variation in the data), estimation of size factors (estimateSizeFactors), estimation of dispersion (estimateDispersions), Negative Binomial GLM fitting and Wald statistics, checking of the dispersion estimates (plotDispEsts), creating contrasts to perform Wald testing on the shrunken log2 fold changes between specific conditions (where necessary), visualization of results (volcano plots, heatmaps, normalized counts plots of top genes, etc), and determining significant results was done in R.

DESeq2 can take the **tximport** object as input, as such we first convert the **SALMON** counts to this object.

All count data directories contain the pattern "_L001"

```
# Assign to a variable list all directories containing data
samples <- list.files(path = wkdir, full.names = F, pattern="_L001$")
## Obtain a vector of all filenames including the path for quant files
files <- file.path(paste(wkdir,samples, sep = ""), "quant.sf")
```

Since all count data (quant) files have the same name the *sample* variable should be the names of each file.

```
names(files) <-  samples
```

Using the *tximport* package we import transcript-level estimates from *SALMON*

```
txi <- tximport(files, type="salmon", txIn = TRUE, txOut = FALSE,
      tx2gene=tx2gene, ignoreTxVersion=TRUE, ignoreAfterBar=TRUE)
```

```
## reading in files with read_tsv
```

```
## 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24
## transcripts missing from tx2gene: 1
```

Table 1: Metadata for the first five samples: UN stands for UNLIKELY.

	NFS_sym	group
COVC02	UN	1
COVC06	UN	3
COVC08	UN	3
COVC09	UN	3
COVC10	UN	1

```
## summarizing abundance
## summarizing counts
## summarizing length
```

We use the provided clinical details to extract relevant metadata. The possible sample classification (conditions of interest) are given simple names: {*Possible*: *PO*, *Unlikely*: *UN*}. The structure of our metadata is as follows

```
str(meta)
```

```
## 'data.frame':    24 obs. of  2 variables:
##  $ NFS_sym: Factor w/ 2 levels "PO","UN": 2 2 2 2 2 2 2 2 2 2 ...
##  $ group  : Factor w/ 3 levels "1","2","3": 1 3 3 3 1 1 3 3 1 3 ...
```

As highlighted earlier, when determining the genes that are differently expressed between conditions, the analysis can be split into two:

- Exploratory analysis or Quality checking (normalization and unsupervised clustering)
- Differential expression analysis (involves modelling the raw counts for each gene, shrinking log2 fold changes and testing for differential analysis between the conditions).

Differential expression analysis given the three (3) analysis groups

Each of the clinical samples belongs to a specific group: *1, 2, 3*. In this section we assume that the base/reference group is 3.

We provide the first 5 rows of our metadata in the given table, so that we know if we have the right comparisons.

```
knitr::kable(head(meta, n = 5 , tidy=TRUE),
  caption = " Metadata for the first five samples: UN stands for UNLIKELY.") %>%
  kable_styling(full_width = F,
  bootstrap_options = c("striped", "hover", "condensed"), font_size = 8) %>%
    row_spec(0, font_size=7)
```

The number of genes we have before filtering genes with low/no count is

```
nrow(dds)
```

```
## [1] 55773
```

Upon filtering genes with ten (10) or less counts across all samples,

```
# Number of rows after filtering
nrow(dds)
```

```
## [1] 24554
```

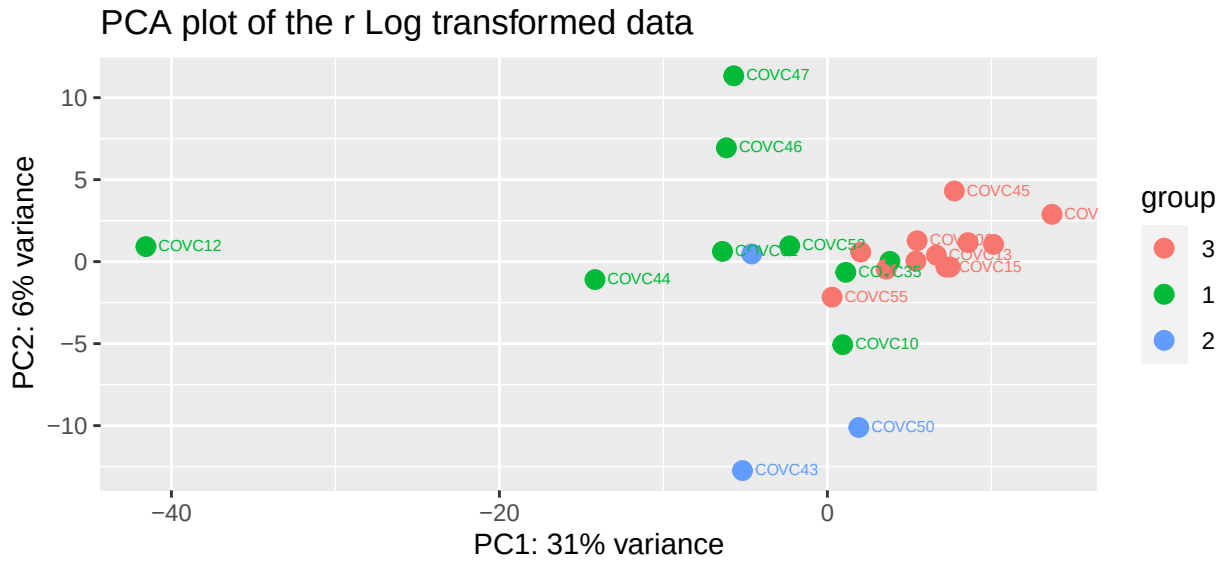
genes remained.

1. Exploratory Data Analysis

We first transform counts for data visualization. We use the *rlog* transform. We determine whether the differences between groups is greater than differences within groups using the PCA and the MDS plot.

```
plotPCA(rld, intgroup="condition", ntop = 500) +
  ggtitle("PCA plot of the r Log transformed data") +
  geom_text(aes(label=name), size=2, hjust=-0.2, vjust=0.4,
      check_overlap = T,fontface=0.5) +  geom_point(size=1)
```

a. The PCA plot



From the PCA plot, samples in same group do not cluster together, which possibly indicates that the samples in the same group vary greatly than the variations observed between samples in different groups, implying differential expression may not be greater than the variance and cannot be easily detected.

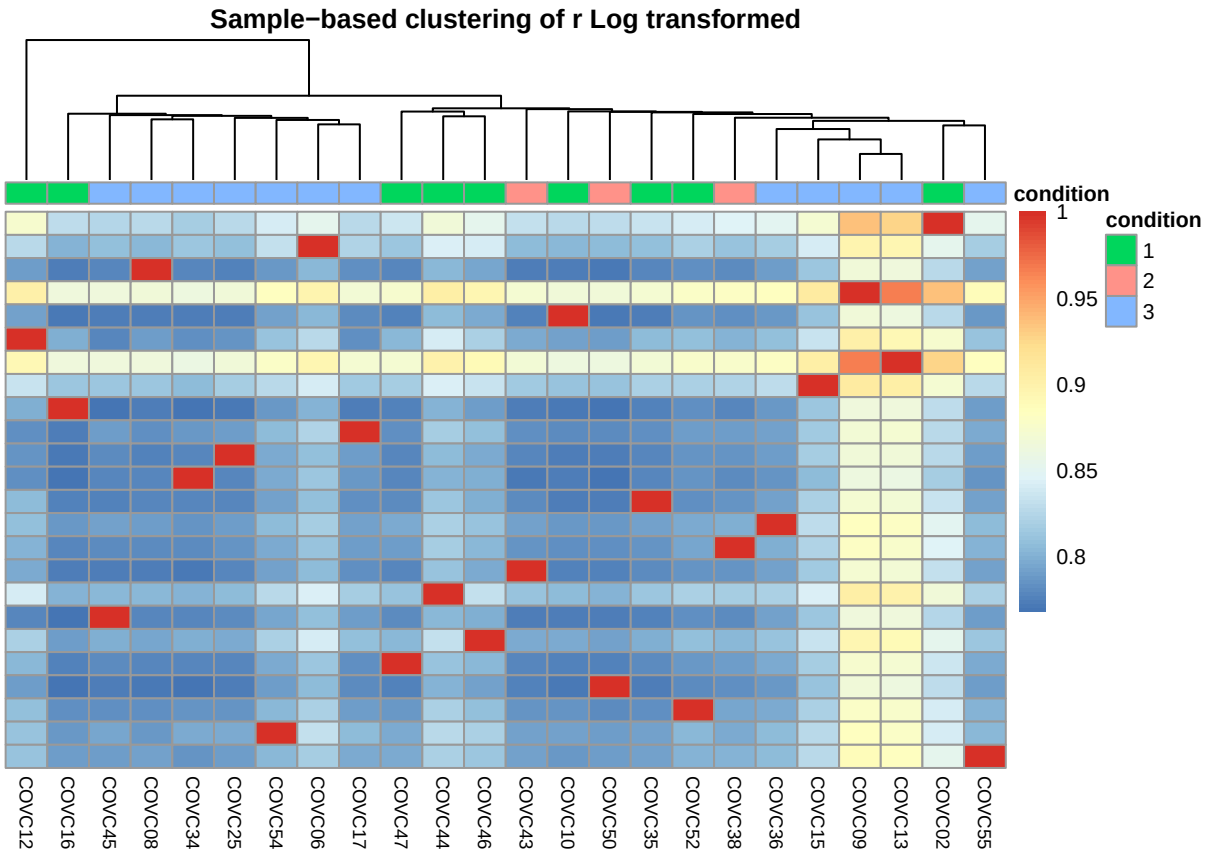
b. The MDS plot Attached to this report is interactive MDS plot in *.html* format, it was generated using the **GLIMMA** package.

```
# creates ma-plot.html in working directory
htmlwidgets::saveWidget(glimmaMDS(dds, groups = sampleTable,
    labels = rownames(sampleTable)), "mds-plot.html")
```

c. Sample- and gene-based clustering

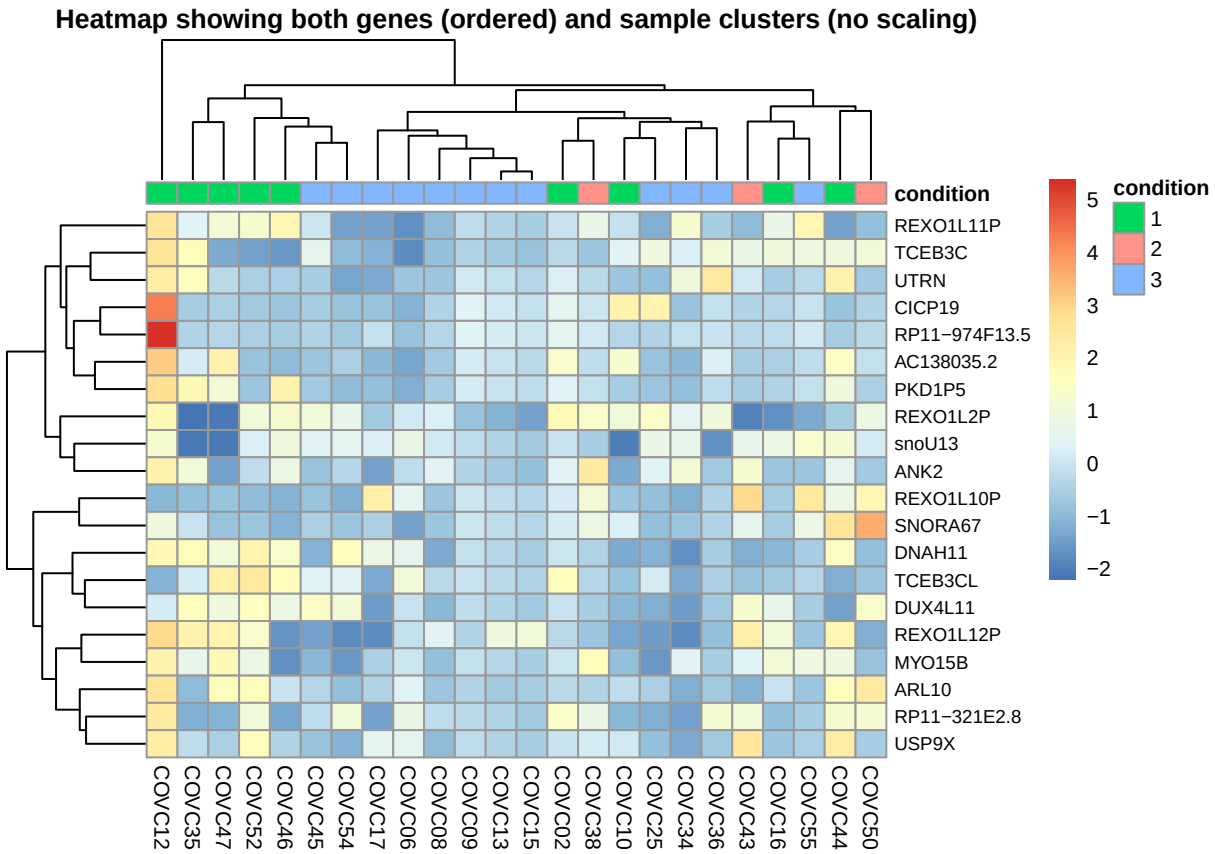
Sample-based clustering (Clustering of samples by gene expression) We cluster samples using the distance between samples based on the Pearson's correlation. Highest correlation value shows most correlated samples, those with more similar expression profiles for all transcripts.

```
rld_mat <- assay(rld)
## Compute pairwise correlation values (Pearson, the default)
rld_cor <- cor(rld_mat)
pheatmap(rld_cor, cluster_rows = FALSE, clustering_distance_cols = sampleDists,
    show_rownames = F, fontsize = 8, fontsize_row = 7, fontsize_col = 7,
    annotation_col = sampleTable, main="Sample-based clustering of r Log transformed")
```



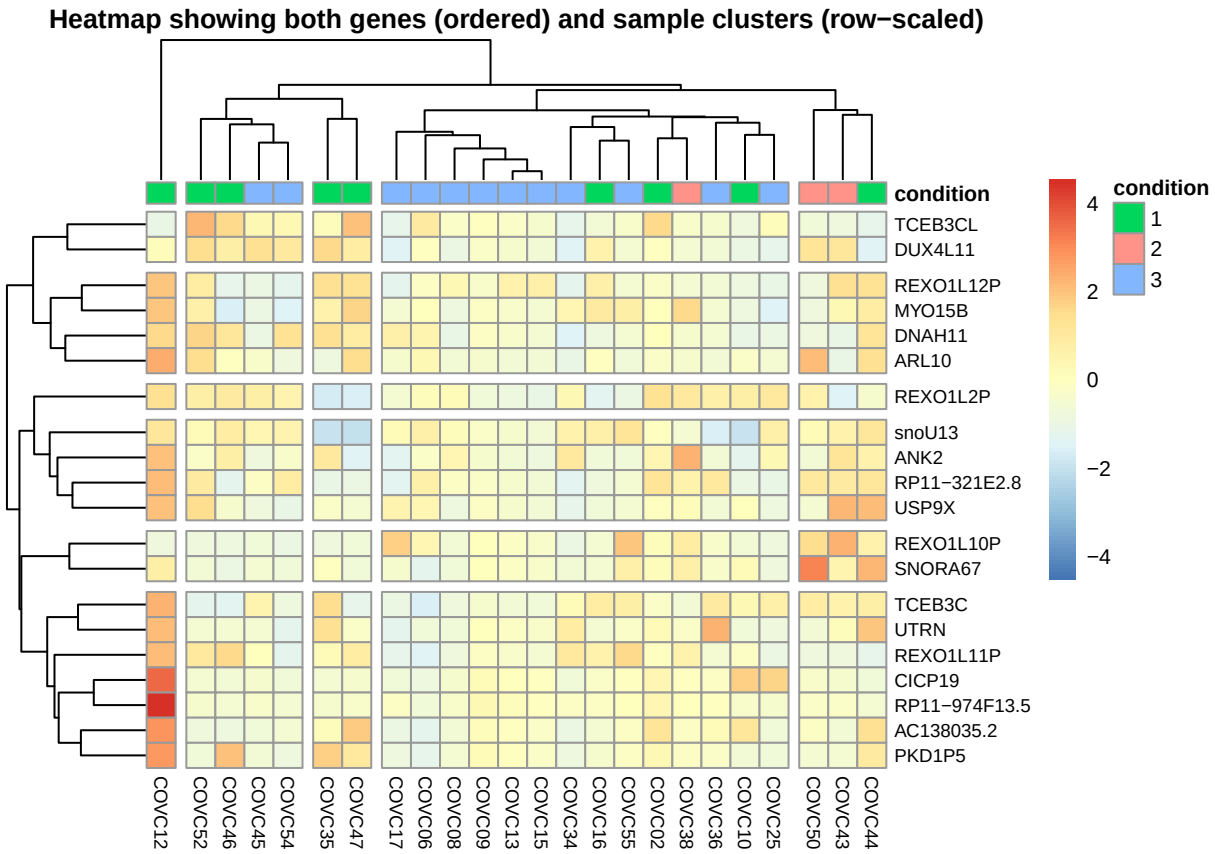
Gene- and sample-based clustering Gene-and sample-based clustering combines the heatmap (clustering of genes with similar expression patterns) & dendrogram of samples (how samples with similar gene expression cluster).

```
topVarGenes <- head(order(rowVars(assay(rld)), decreasing = T), 20)
mat  <- assay(rld)[ topVarGenes, ]
mat  <- mat - rowMeans(mat)
# Dendrogram and Heapmap
pheatmap(mat , annotation_col = sampleTable, fontsize = 8, fontsize_row = 7,
    main = "Heatmap showing both genes (ordered) and sample clusters (no scaling)")
```



Scaling makes it easier to observe differences in values for each of the variables. In RNA seq, we scale by row since individuals are listed across col. Here we also customize annotation colors.

```
pheatmap(mat, fontsize = 8, fontsize_row = 7, fontsize_col = 7, scale = "row",
    cutree_rows = 6, annotation_col = sampleTable, cutree_cols=5,
    main = "Heatmap showing both genes (ordered) and sample clusters (row-scaled)")
```



3. Library normalisation, dispersion estimation and the Wald test

In this section we

(a) normalize the count data by library size by estimating the *size factor*,
(b) estimate dispersion for the negative binomial model, and
(c) fit models and get statistics for each gene for the design specified the data is imported.

When considering how genes are expressed between different analysis groups, we have the following contrasts

```
resultsNames(dds)
```

```
## [1] "Intercept"        "condition_1_vs_3" "condition_2_vs_3"
```

Testing different hypotheses, creating contrasts

DESeq2 adjusts the p value by different methods including the "holm", "hochberg", "hommel", "bonferroni", "BY" and Benjamini and Hochberg ("BH") which is default. We adjusted the p-values using the "BH" method. All the compared groups: **(1 vs 3** and **2 vs 3** have genes that are expressed differently (up-regulated).

We test the following hypothesis:

- **H0:** Each gene in group 1 and 3 have equal expression distribution (each gene is not differentially expressed) vs
- **H1:** Genes in group 1 and 3 show significant expression distribution (they are differentially expressed).

```
res13DF <- as.data.frame(res_1vs3[order(res_1vs3$padj),])
table(res13DF$padj< 0.05)
```

```
##
## FALSE  TRUE
```

```
##  9703  1041
```

1041 genes have padj < 0.05.

The number of the genes that are not significant, up- or down-regulated given $padj < 0.05$ is given below.

```
summary(res_1vs3, alpha=0.05)
```

```
##
## out of 24554 with nonzero total read count
## adjusted p-value < 0.05
## LFC > 0 (up)       : 1041, 4.2%
## LFC < 0 (down)     : 0, 0%
## outliers [1]       : 7, 0.029%
## low counts [2]     : 13803, 56%
## (mean count < 1)
## [1] see 'cooksCutoff' argument of ?results
## [2] see 'independentFiltering' argument of ?results
```

Now we tabulate the genes whose p-value < 0.05 considering group 2 and 3.

```
res23DF <- as.data.frame(res_2vs3[order(res_2vs3$padj),])
print(table(res23DF$padj < 0.05))
```

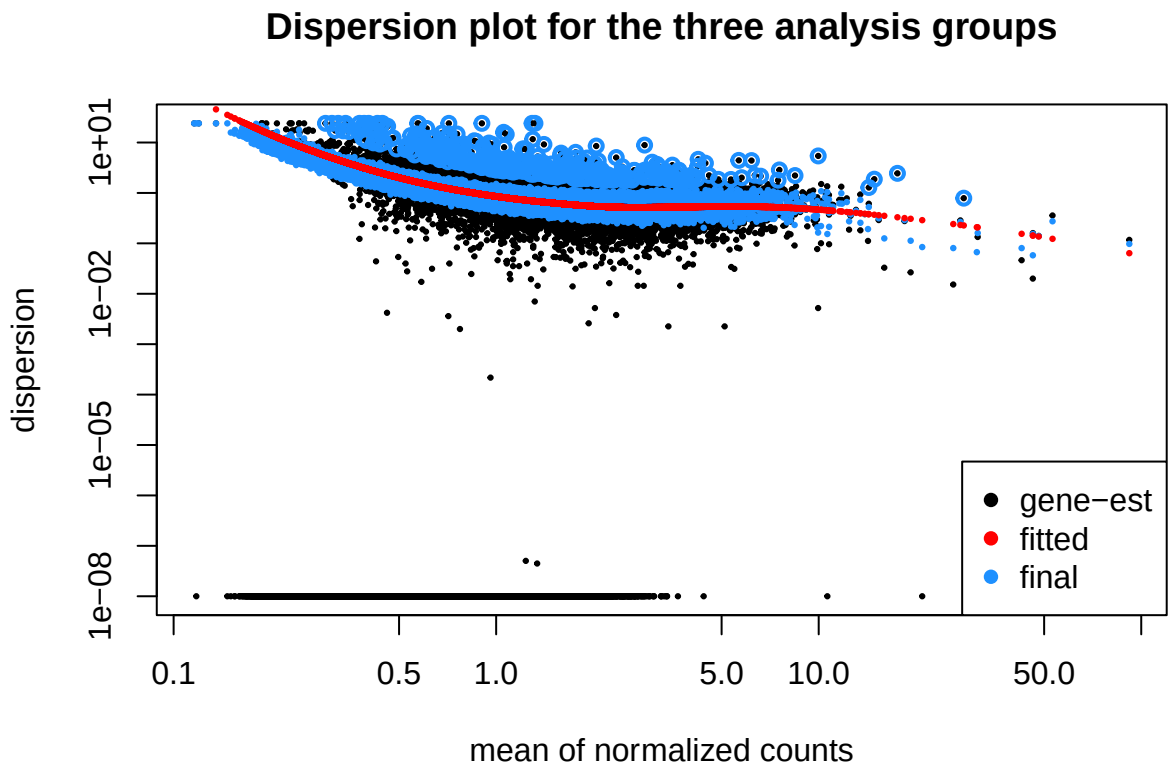
```
##
## FALSE  TRUE
##  6452     8
```

The summary of non-significant, up- and down-regulated genes between groups 2 and 3 is given by

```
summary(res_2vs3,alpha=0.05)
```

```
##
## out of 24554 with nonzero total read count
## adjusted p-value < 0.05
## LFC > 0 (up)       : 8, 0.033%
## LFC < 0 (down)     : 0, 0%
## outliers [1]       : 7, 0.029%
## low counts [2]     : 18087, 74%
## (mean count < 2)
## [1] see 'cooksCutoff' argument of ?results
## [2] see 'independentFiltering' argument of ?results
```

```
plotDispEsts(dds, main="Dispersion plot for the three analysis groups")
```

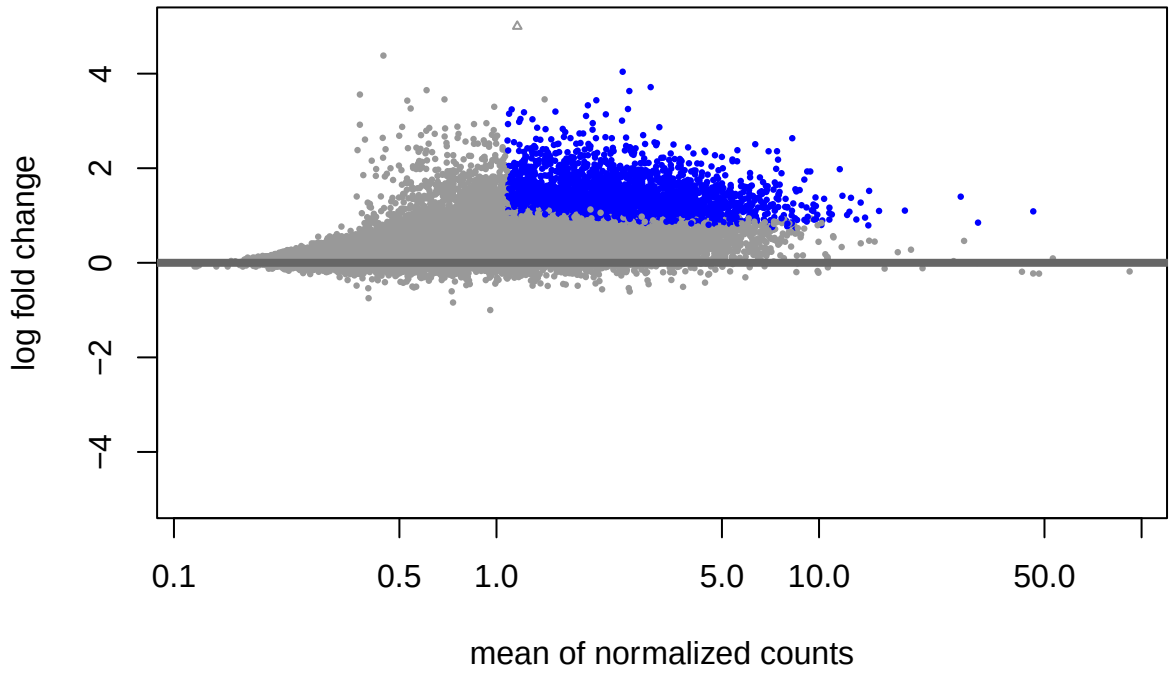


Shrinkage of effect size (LFC) is used for visualization and to rank the genes. We use the *apeglm* method for effect size shrinkage as it improves the estimator when specified.

```
MAres_1vs3 <- lfcShrink(dds, coef="condition_1_vs_3", type="apeglm")
```

```
## using 'apeglm' for LFC shrinkage. If used in published research, please cite:
##     Zhu, A., Ibrahim, J.G., Love, M.I. (2018) Heavy-tailed prior distributions for
##     sequence count data: removing the noise and preserving large differences.
##     Bioinformatics. https://doi.org/10.1093/bioinformatics/bty895
```

```
DESeq2::plotMA(MAres_1vs3, ylim=c(-5, 5), main="MA plot of Analysis Group 1 vs 3")
```



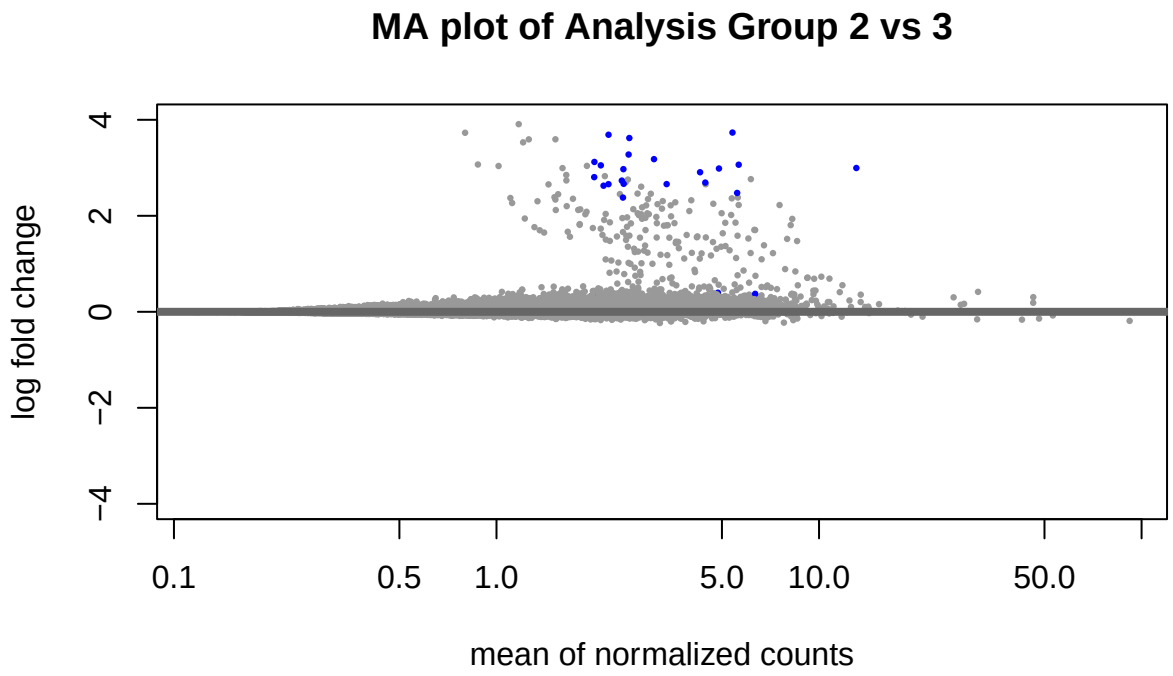
From the MA plot of group 1 and 3, the points in grey represent the genes that are not significant despite having $M \neq 0$. The points in blue with $M \geq 0$ are up-regulated. We do not have down-regulated genes since there are no blue points with $M \leq 0$, confirming the summary we gave earlier on number of genes that are up- and down-regulated and those that are not significant.

Between 2 and 3, most of the genes are not significant (just a few that are in blue).

```
MAres_2vs3 <- lfcShrink(dds, coef="condition_2_vs_3", type="apeglm")
```

```
## using 'apeglm' for LFC shrinkage. If used in published research, please cite:
##     Zhu, A., Ibrahim, J.G., Love, M.I. (2018) Heavy-tailed prior distributions for
##     sequence count data: removing the noise and preserving large differences.
##     Bioinformatics. https://doi.org/10.1093/bioinformatics/bty895
```

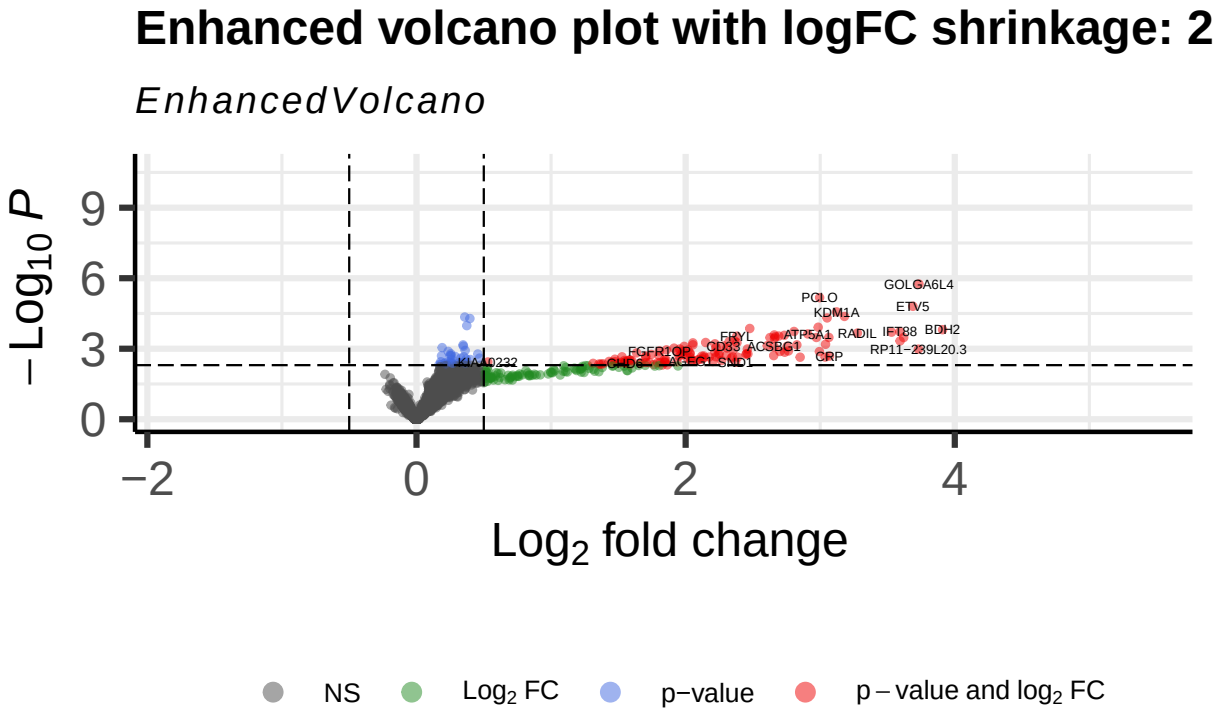
```
DESeq2::plotMA(MAres_2vs3, ylim=c(-4, 4), main="MA plot of Analysis Group 2 vs 3")
```



Volcano plot As shown on the legend, the green points are genes that are significant when we only consider the log fold change thresholds, while the grey are the genes that are not significant and the red (where a few are shown with annotated names) are the significant and up-regulated considering a cut-off line of $-\text{Log}_{10}(0.005)$, with the padj as the cutoff column and Log_2fold change $= \frac{1}{2}$.

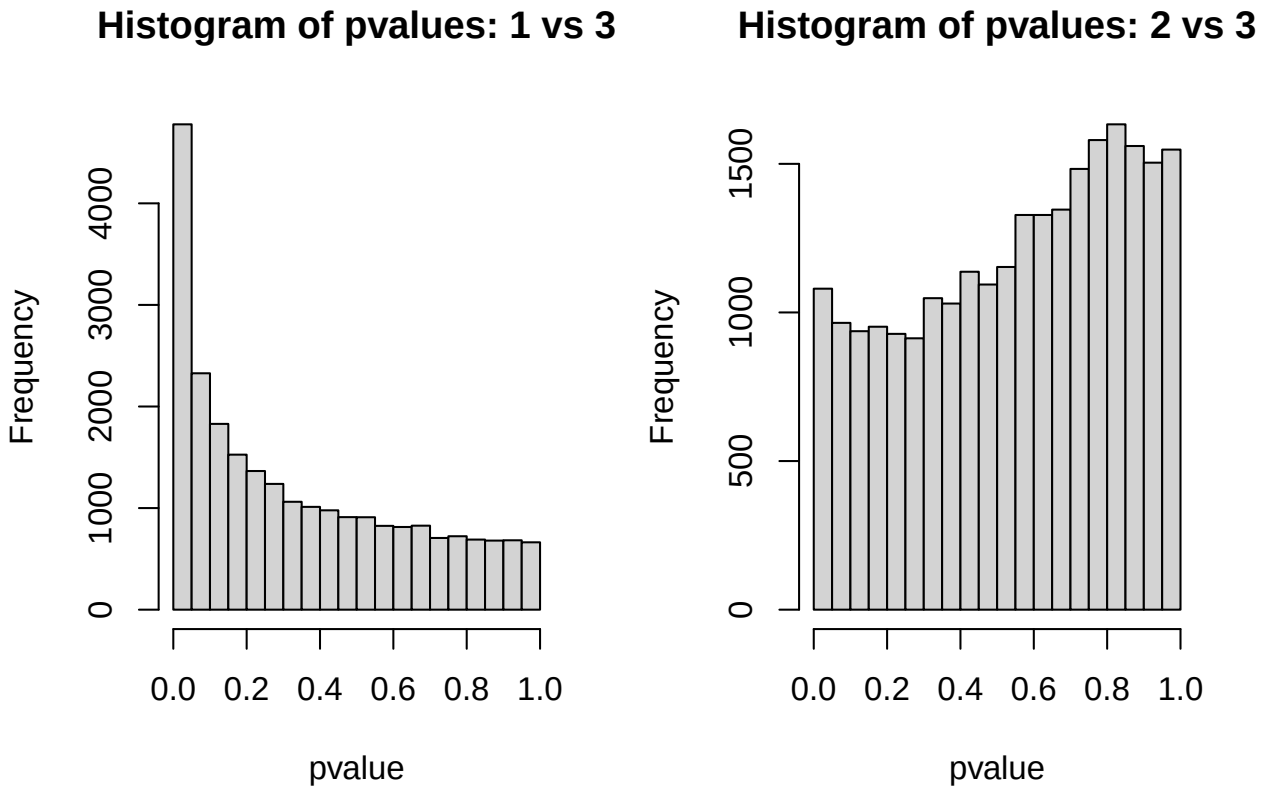
```
##WE CAN USE res in place of DGE.results
dev.new(width=15, height=15)
vp1 <- EnhancedVolcano(res_1vs3, lab = rownames(res_1vs3),
  x = 'log2FoldChange', y = 'pvalue', pointSize = 1.0, labSize = 2.0,
  pCutoff = 0.005, xlab = bquote(~Log[2] ~ "fold change"),
  ylab = bquote(~-Log[10] ~adjusted~italic(P)), pCutoffCol = 'padj', FCcutoff = 0.5,
  cutoffLineType = "longdash", cutoffLineCol = "black", cutoffLineWidth = 0.4,
  col = c("grey30", "yellow", "royalblue", "red2"), colAlpha=1, legendLabels =
    c("NS", expression(Log[2] ~ FC), "pvalue",   expression(pvalue ~ and ~ log[2] ~ FC)),
  legendPosition = 'bottom', legendLabSize = 10, legendIconSize = 3.0,
  title = 'Enhanced volcano plot without shrinkage')

vp2 <- EnhancedVolcano(MAres_1vs3, lab = rownames(MAres_1vs3), x = 'log2FoldChange',
  y = 'pvalue', pCutoff = 0.005, legendPosition = 'bottom', pointSize = 1.0, labSize = 1.75,
  legendLabSize = 10, legendIconSize = 3.0, FCcutoff = 0.5,
  title = "Enhanced volcano plot with logFC shrinkage: 1 vs 3")

vp3 <- EnhancedVolcano(MAres_2vs3, lab = rownames(MAres_2vs3), x = 'log2FoldChange',
  y = 'pvalue', pCutoff = 0.005, legendPosition = 'bottom', pointSize = 1.0, labSize = 1.75,
  legendLabSize = 10, legendIconSize = 3.0, FCcutoff = 0.5,
  title = "Enhanced volcano plot with logFC shrinkage: 2 vs 3")
```

```
print(vp3)
```



total = 24554 variables

```
par(mfrow=c(1,2))
hist(res13DF$pvalue,xlab = "pvalue", main = "Histogram of pvalues: 1 vs 3")
hist(res23DF$pvalue,xlab = "pvalue", main = "Histogram of pvalues: 2 vs 3")
```



The number of genes that are significantly different between group 1 and 3 considering a padj < 0.05 and abs(log2FoldChange) > 0.58 cutoff are:

```
print(table(res13DF$padj < 0.05 & abs(res13DF$log2FoldChange) > 0.58))
```

```
##
## FALSE  TRUE
## 14385  1041
```

We considered significant genes as those that have passed the following thresholds: padj < 0.05 and abs(log2FoldChange) > 0.58 and save these genes in a file under the name - 1. **res__sig__genes__1vs3.csv** and - 2. **res__sig__genes__2vs3.csv**.

Now, the summary of genes that are significantly different between group 2 and 3 given the same cutoffs is:

```
print(table(res23DF$padj < 0.05 & abs(res23DF$log2FoldChange) > 0.58))
```

```
##
## FALSE  TRUE
## 13460     8
```