# Analysis Report (Part5): Differential Expression Analysis

(Last updated: 11/05/21)

Differential expression analysis between the sample age groups (adult vs. fetal) while adjusting for RIN was performed using DESeq2_1.24.0.
For a particular gene, a log2 fold change of -1 for **age_group** (adult vs. fetal) means that being an adult induces a multiplicative change in observed gene expression level of $2^{-1}$=0.5 compared to the being a fetal. If the variable of interest is continuous-valued, then the reported log2 fold change is per unit of change of that variable.

## 1. Loading the Necessary Libraries and Data

```
library(SummarizedExperiment)
library(DESeq2)
library(ggplot2)
library(ggthemes)
# read a merged_counts file
merged_counts = read.table("../merged_counts-v2.tsv", quote = "", sep = '\t')
# Read phenotype sample data
pheno_data = read.csv("../phenotype_data.tsv", quote = "", sep = '\t')
```

## 2. Create DESeq2 Object and Get the Results (Differentially_Expressed_Genes with FDR < 0.05)

```
deseq.dat = DESeqDataSetFromMatrix(countData = merged_counts, colData = pheno_data,
                                   design = ~ RIN + age_group)
# pre-filter
keep = rowSums(counts(deseq.dat)) >= 10
deseq.dat = deseq.dat[keep,]
dds = DESeq(deseq.dat)
# DESeq2 results
res_deseq2 = results(dds, contrast = c("age_group", "adult", "fetal"))
res_deseq2_shrunk = lfcShrink(dds=dds, contrast = c("age_group", "adult", "fetal"),
                              res = res_deseq2, type = "ashr")
# identify genes with FDR < 0.05
res_deseq2_shrunk_sig = subset(res_deseq2_shrunk, res_deseq2_shrunk$padj < 0.05)
# add gene symbol to res
gene_symbol = read.table("../Ensembl.symbols.txt", header = TRUE, na.strings = "n/a",
                         col.names = c("gene", "symbol"))
gene_symbol = gene_symbol[!duplicated(gene_symbol$symbol), ]
res_deseq2_shrunk_sig = as.data.frame(res_deseq2_shrunk_sig)
```

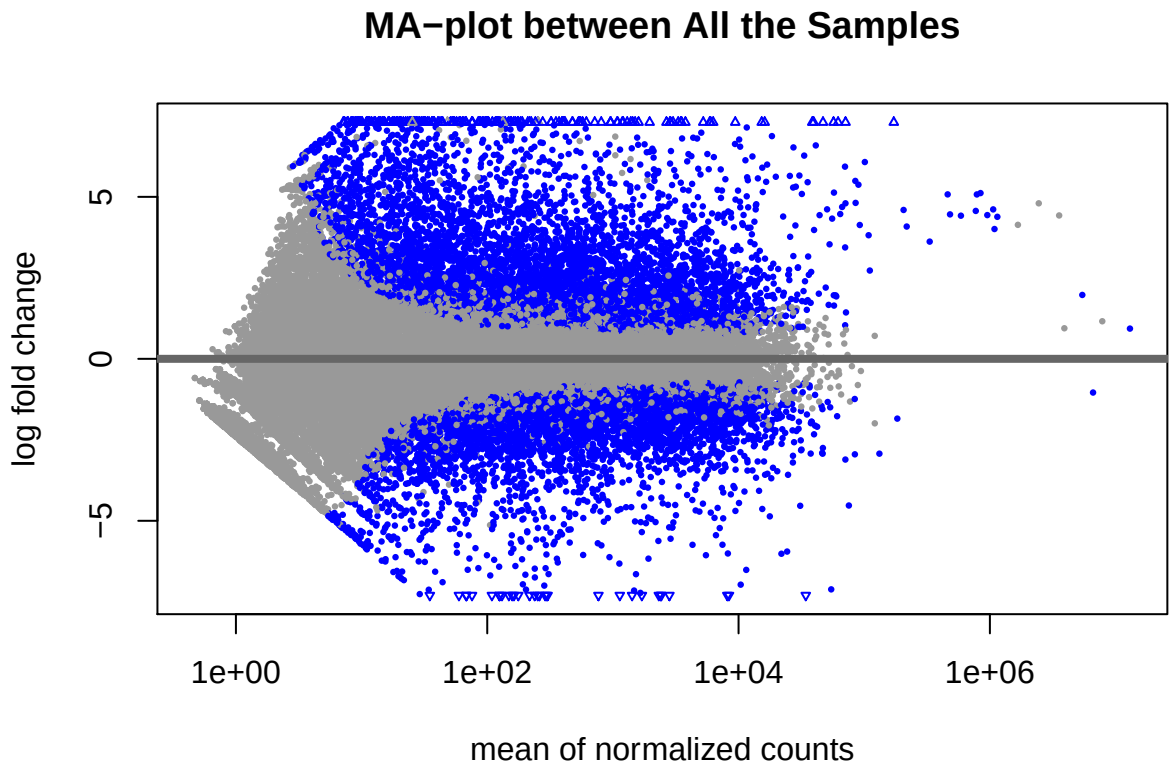
```
res_deseq2_shrunk_sig$row = rownames(res_deseq2_shrunk_sig)
res_deseq2_shrunk_annotated = merge(res_deseq2_shrunk_sig, gene_symbol, by.x = "row", by.y = "gene")
# Sort by increasing log2FoldChange
sorted_significant_differ_res_deseq2_shrunk = res_deseq2_shrunk_annotated[
  sort.list(res_deseq2_shrunk_annotated$log2FoldChange, decreasing = TRUE),]
# upload the results to a file

upload_df = data.frame(gene = sorted_significant_differ_res_deseq2_shrunk$row, name = sorted_significant
                       log2fc = sorted_significant_differ_res_deseq2_shrunk$log2FoldChange, pval = sorte
                       padj = sorted_significant_differ_res_deseq2_shrunk$padj)
write.table(upload_df, "../Differentially_Expressed_Genes-results.tsv", quote = FALSE, sep = '\t')
```

## 3. MA-plot between All the Samples

MA-plot between the sample age groups (adult vs. fetal) while adjusting for RIN: plot log2 fold-changes (on the y-axis) versus the mean of normalized counts (on the x-axis).

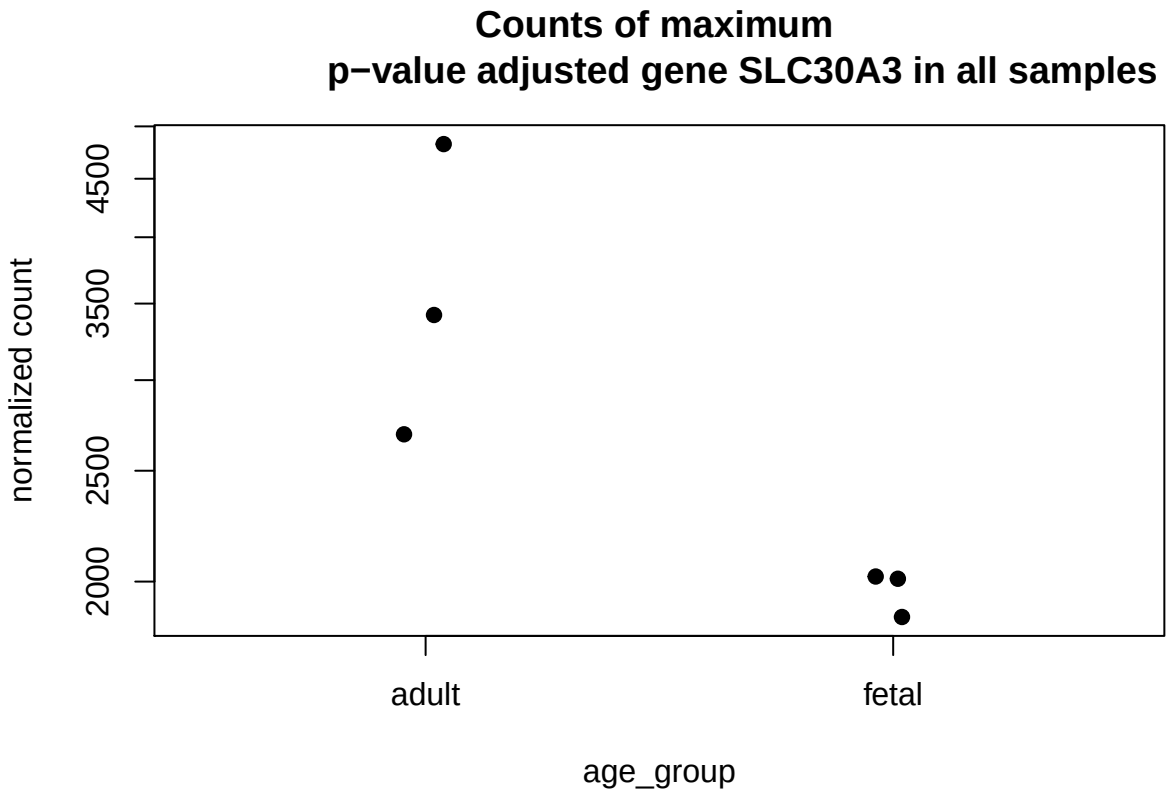
```
pval_threshold = 10e-3
plotMA(res_deseq2, alpha = pval_threshold, main = "MA-plot between All the Samples")
```



## 4. Plot Counts

Plot the counts of the gene which had the smallest adjusted p value from the results table created above.

```
par(pch = 19)
minGene = which.min(res_deseq2_shrunk_annotated$padj)
plotCounts(dds, gene=minGene, intgroup="age_group", main = paste0("Counts of maximum
                p-value adjusted gene ", res_deseq2_shrunk_annotated[minGene, 7],
                " in all samples"), xlab = "age_group")
```



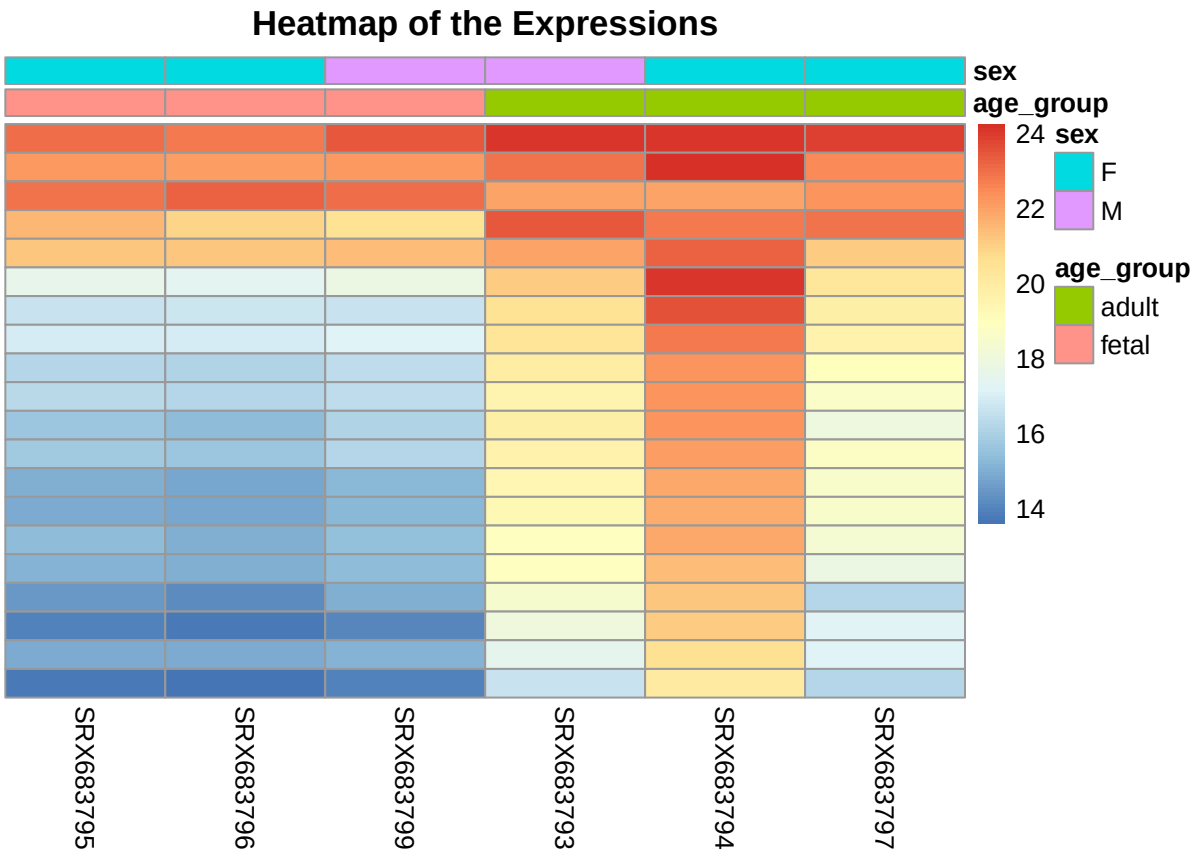

## 5. Heatmaps Using pheatmap Package (Pretty Heatmaps)

Heatmaps are a useful method to explore large multivariate data sets. Response variables (e.g., abundances) are visualised using colour gradients or colour schemes. With the right transformation, and row and column clustering, interesting patterns within the data can be seen. They can also be used to show the results after statistical analysis, for example, to show those variables that differ between treatment groups.

### 5.1. Heatmap of the Expressions

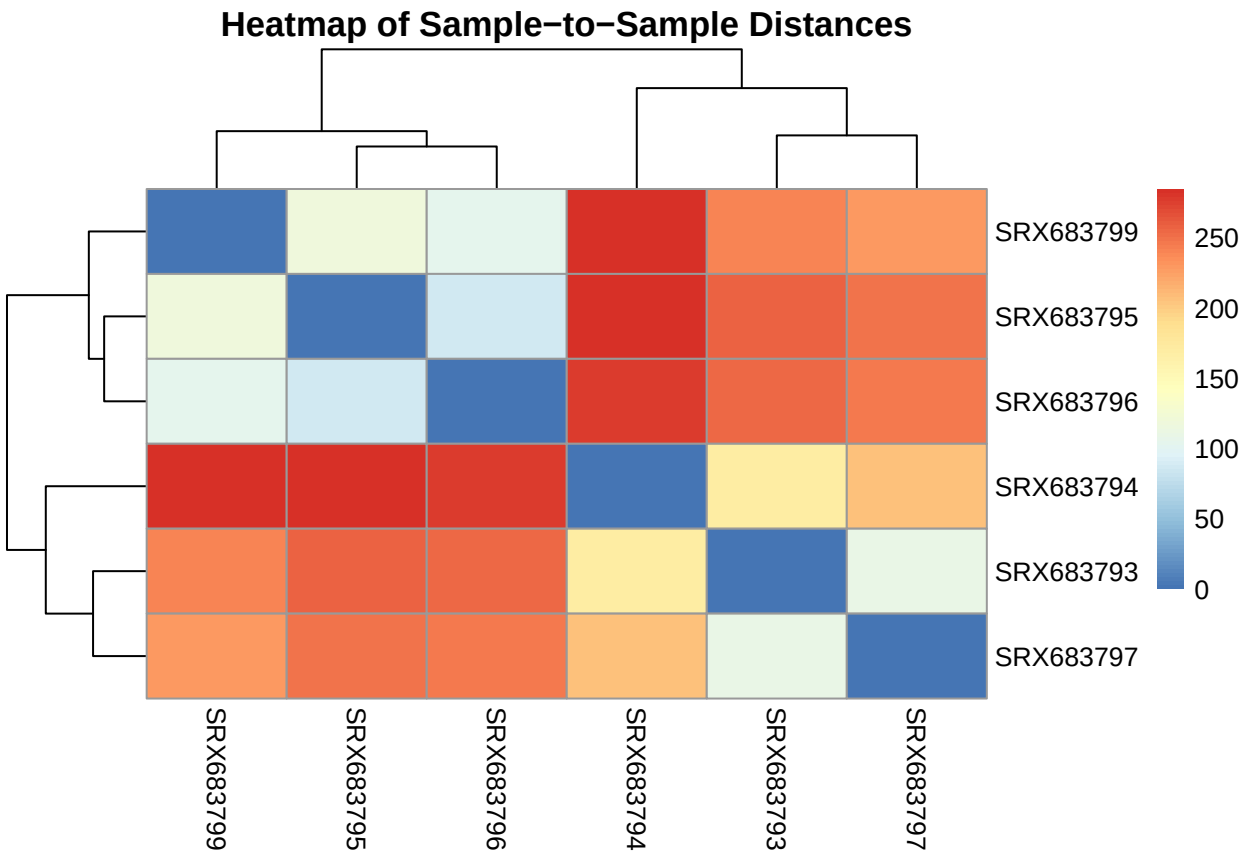
This heatmap gives us an overview over the similarities and dissimilarities between samples' expressions. For example, from the heatmap of the 20 highest gene expressions below we notice that the **age_groupe** variable has an effect on the gene expression, where the *adult* group has higher expressions than the *fetal* group. Whereas, the **sex** variable has no noticeable effect.

```
library("pheatmap")
select = order(rowMeans(counts(dds,normalized=TRUE)), decreasing=TRUE)[1:20]
df = as.data.frame(pheno_data[, c("age_group", "sex")])
rownames(df) = colnames(dds)
ntd = normTransform(dds)
pheatmap(assay(ntd)[select, ], cluster_rows=FALSE, show_rownames=FALSE, cluster_cols=FALSE,
         annotation_col=df, main = "Heatmap of the Expressions")
```




## 5.2. Heatmap of Sample-to-Sample Distances

```
vsd  = vst(dds, blind = FALSE)
sampleDists = dist(t(assay(vsd)))
library("RColorBrewer")
sampleDistMatrix <- as.matrix(sampleDists)
colors <- colorRampPalette( rev(brewer.pal(9, "Blues")) )(255)
pheatmap(sampleDistMatrix,
         clustering_distance_rows=sampleDists,
         clustering_distance_cols=sampleDists,
         main = "Heatmap of Sample-to-Sample Distances")
```




## 6. Volcano Plot

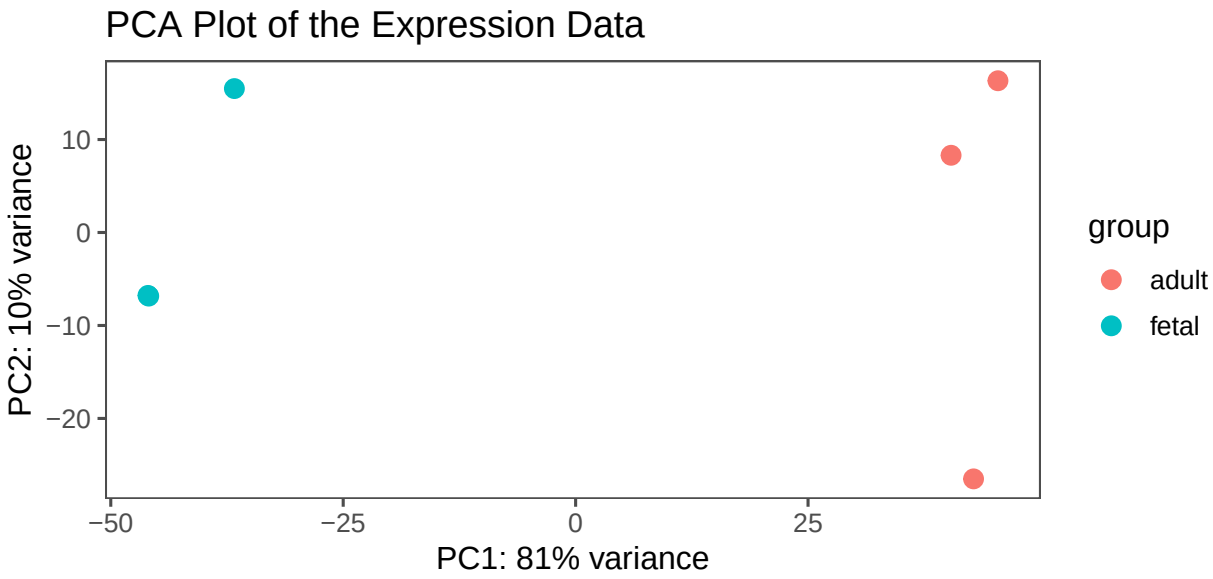
Volcano Plots are used to visualize the significance and the magnitude of changes in genes. For example, from the plot below, we can notice that the **MT1H** (Metallothionein-1H gene, which is associated with autism) is upregulated (23 log2 fold-change) in the adult group vs. the fetal group.

```
library(EnhancedVolcano)
EnhancedVolcano(sorted_significant_differ_res_deseq2_shrunk, FCcutoff = 1, pCutoff = pval_threshold,
                sorted_significant_differ_res_deseq2_shrunk$symbol,
                x = 'log2FoldChange', y = 'pvalue',
                legend = c("Non-significant", "Passed log2 fold-change threshold",
                           "Passed the p-value threshold", "Passed both thresholds"),
                legendPosition = 'right',
                legendLabSize = 9,
                legendIconSize = 2,
                widthConnectors = 0.2,
                colConnectors = 'grey30',
                colAlpha = 1)
```


## 7. Principal Components Analysis (PCA) plot

A PCA plot shows clusters of samples based on their similarity. It reduces the overwhelming number of dimensions by constructing principal components (PCs).

```
rld = rlogTransformation(dds)
plotPCA(rld, intgroup = "age_group") + ggtitle("PCA Plot of the Expression Data") + theme_few()
```


PCA Plot of the Expression Data
PC2: 10% variance
10
0
−10
−20
−50
−25
0
25
PC1: 81% variance
group
adult
fetal