Volcano Plot Example

```
library(tidyverse)
library(ggrepel)
```

1. Introduction

- The Goal

This is an example of how to make and analyze a volcano plot.

```
counts_data <- read.csv('limma-voom_luminalpregnant-luminallactate.tsv',sep = '\t')

head(counts_data)
```

```
##   ENTREZID  SYMBOL
## 1    12992 Csn1s2b
## 2    13358 Slc25a1
## 3    11941  Atp2b2
## 4    20531 Slc34a2
## 5   100705   Acacb
## 6    13645     Egf
##                                                                    GENENAME
## 1                                                    casein alpha s2-like B
## 2 solute carrier family 25 (mitochondrial carrier, citrate transporter), member 1
## 3                                ATPase, Ca++ transporting, plasma membrane 2
## 4                          solute carrier family 34 (sodium phosphate), member 2
## 5                                       acetyl-Coenzyme A carboxylase beta
## 6                                                 epidermal growth factor
##       logFC   AveExpr         t      P.Value    adj.P.Val
## 1 -8.603611 3.5629500 -43.79650 3.830650e-15 6.053959e-11
## 2 -4.124175 5.7796989 -29.90785 1.758595e-13 1.389642e-09
## 3 -7.386986 1.2821431 -27.81950 4.836363e-13 2.432800e-09
## 4 -4.177812 4.2786290 -27.07272 6.157428e-13 2.432800e-09
## 5 -4.314320 4.4409137 -25.22357 1.499977e-12 4.741129e-09
## 6 -5.362664 0.7359047 -24.59930 2.116244e-12 5.574188e-09
```

- Understanding the dataset

This dataset is from a study,(Fu *et al* 2015) that explored the regulation of a pro-survival gene *Mcl-1* in mammopoiesis. The researchers analyzed expression levels of two different cell types, basal and luminal, under three different conditions, virgin, 18.5 day pregnancy, 2 day lactation in mice.

The RNA-seq data is available through the Gene Expression Omnibus database (GEO), accession number GSE 60450. The dataset I am using is the product of the counts data previously being processed with the limma package. The data is available through Zenodo.org.

In this example I am using the dataset of the genes under the condition of 18.5 day pregnant, and 2 day lactation in luminal cells.

2. Preparing the Dataset

- Adding the threshold values

I added the threshold values for significance and regulation based off the researchers using a threshold of 1% for a false discovery rate and a fold-change threshold of 1.5

```
counts_data$logFC <- -1 * counts_data$logFC

counts_data <- counts_data %>% mutate(Status  = case_when(
                    (counts_data$logFC > 0.58 & counts_data$adj.P.Val < 0.01) ~ 'Upregulated',
                    (counts_data$logFC < 0.58 & counts_data$adj.P.Val < 0.01) ~ 'Downregulated',
                    (counts_data$adj.P.Val > 0.01) ~ 'Non-significant'))




gene_data <- counts_data %>% arrange(adj.P.Val)

head(gene_data)
```

```
##   ENTREZID  SYMBOL
## 1    12992 Csn1s2b
## 2    13358 Slc25a1
## 3    11941  Atp2b2
## 4    20531 Slc34a2
## 5   100705   Acacb
## 6    13645     Egf
##                                                                                GENENAME
## 1                                                                  casein alpha s2-like B
## 2 solute carrier family 25 (mitochondrial carrier, citrate transporter), member 1
## 3                                    ATPase, Ca++ transporting, plasma membrane 2
## 4                            solute carrier family 34 (sodium phosphate), member 2
## 5                                                 acetyl-Coenzyme A carboxylase beta
## 6                                                              epidermal growth factor
##      logFC   AveExpr         t      P.Value    adj.P.Val      Status
## 1 8.603611 3.5629500 -43.79650 3.830650e-15 6.053959e-11 Upregulated
## 2 4.124175 5.7796989 -29.90785 1.758595e-13 1.389642e-09 Upregulated
## 3 7.386986 1.2821431 -27.81950 4.836363e-13 2.432800e-09 Upregulated
## 4 4.177812 4.2786290 -27.07272 6.157428e-13 2.432800e-09 Upregulated
## 5 4.314320 4.4409137 -25.22357 1.499977e-12 4.741129e-09 Upregulated
## 6 5.362664 0.7359047 -24.59930 2.116244e-12 5.574188e-09 Upregulated
```

3. Making the Volcano Plot

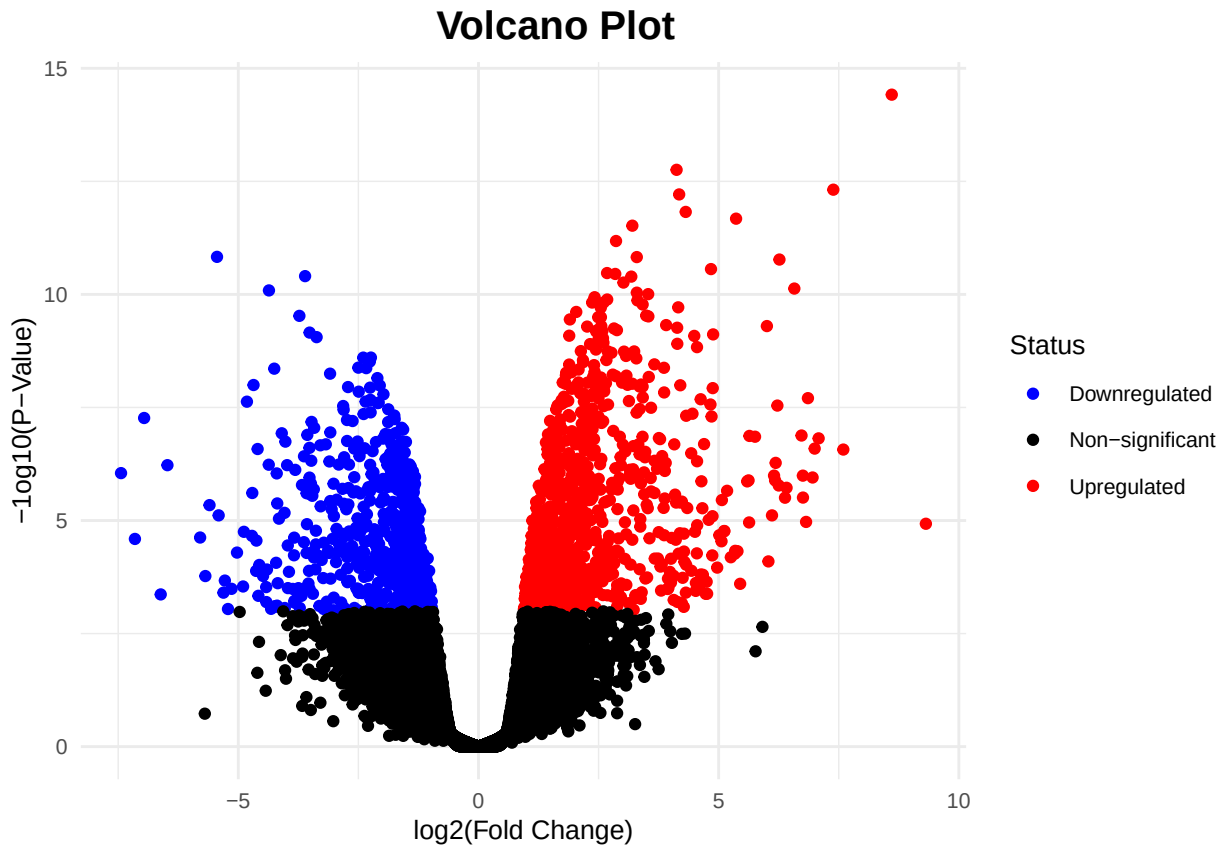
- Making an easy to read Plot

Now the first volcano plot can be created, using colors to identify the genes that are upregulated, downregulated, and non-significant based on the thresholds.

```
plot1 <- gene_data %>% ggplot(aes(x = logFC, y = -log10(P.Value), col = Status )) +
              geom_point() +
              theme_minimal()+
```

```
                scale_color_manual(values = c('blue','black','red')) +
                theme(text = element_text(size = 20)) +
                labs(x = 'log2(Fold Change)',
                     y = '-1og10(P-Value)',
                     title =  'Volcano Plot')+
                theme(plot.title = element_text(hjust = 0.55, face = 'bold', size = 15),
                      text =  element_text(size = 10))
```

```
plot1
```



4. Making the Visualization more Informative

- Threshold lines

It is important to visualize the thresholds I used for significance and regulation.

```
min_upregulated <- gene_data %>%
                    subset(Status == 'Upregulated') %>%
                    arrange(logFC) %>%
                    slice(1L) %>%
                    select(logFC) %>%
                    pull(1)
```

```
min_downregulated <- gene_data %>%
                     subset(Status == 'Downregulated') %>%
                     arrange(desc(logFC)) %>%
                     slice(1L) %>%
                     select(logFC) %>%
                     pull(1)



vline_threshold <- min(c(round(abs(min_downregulated) - 0.1,1) ,
                         round(abs(min_upregulated) - 0.1,1)))



hline_threshold <-  gene_data %>%
                     subset(Status == 'Non-significant') %>%
                     arrange(P.Value) %>%
                     slice(1L) %>%
                     select(P.Value) %>%
                     pull(1)
```

- Add Gene names

The dataset has a column of the gene symbols allowing me to label some genes with their symbols. I chose to label the top 10 most expressed genes.

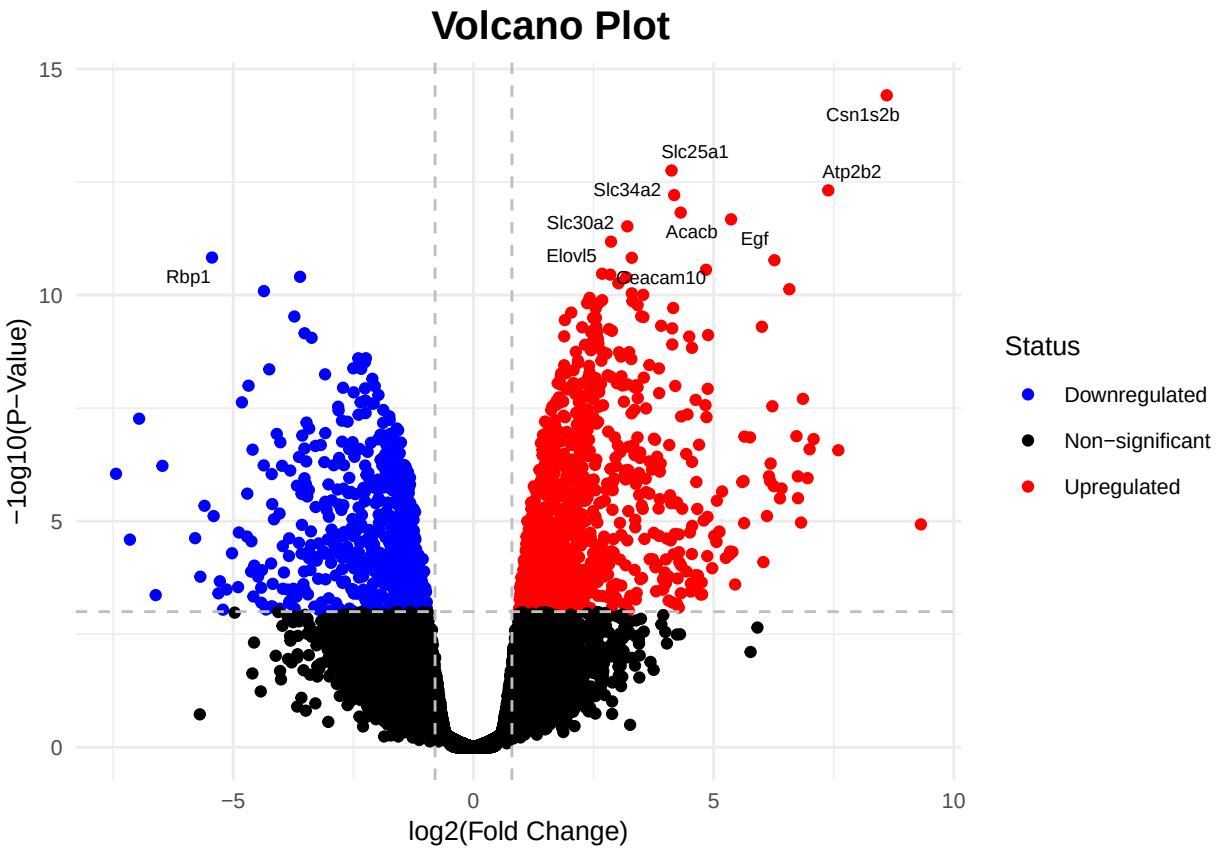
The gene_data dataset has already been arranged by adjusted p-value, so I could use a splice to label the top 10 genes

```
plot_volcano <- plot1 +
                geom_hline(yintercept = -log10(round(hline_threshold,3)),
                   linetype = 2,
                   color = 'grey')+
                geom_vline(xintercept = c(-vline_threshold,vline_threshold),
                   linetype = 2,
                   color = 'grey') +

                geom_text_repel(data = subset(gene_data[1:10,]),
                  aes(label = SYMBOL), color = 'black', size = 2.5)




plot_volcano
```



5. Analysis

- General Analysis

The Volcano Plot shows the statistical significance on the y-axis and the magnitude of change of the gene expression on the x-axis. The right side will show the up-relegated genes, while the left side will show the the down-relegated genes.

- Comparing My Graph with the Researchers' Analysis

The graph shows that more genes are significantly upregulated in the lumnial cells during lactation than when the mice were pregnant, matching what the researchers found. They found that upregulation of *ECF* had an impact on *Mcl1* upregulation during lactation. Plots like this are useful in highlighting specific genes that could be investigated further.