Genomic Data Science Capstone: Week 8 (Statistical Analysis)

For this assignment I performed a statistical analysis of the gene expression data across two different groups based on age: the fetal brain vs adult brain, where each group contains 3 different samples.

The first step was to develop my hypotheses:

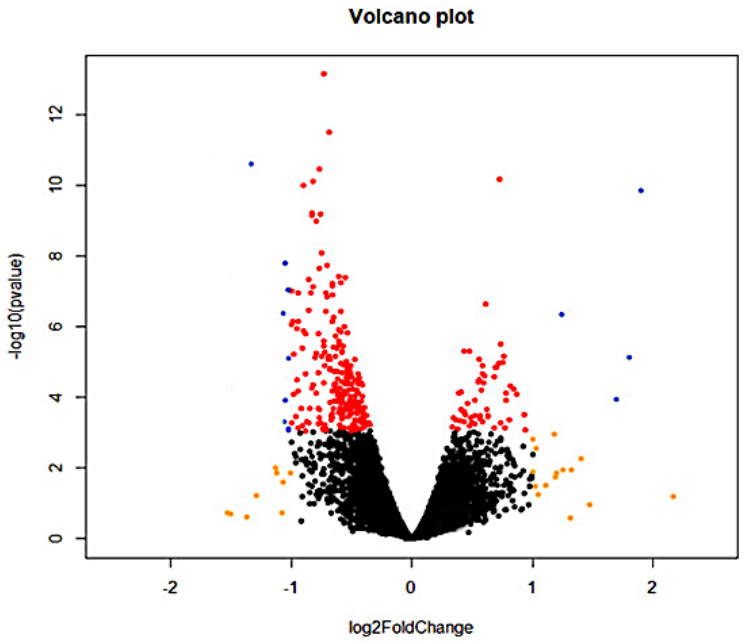
1- *Null hypothesis: there is no differential gene expression across fetal and adult brain tissues (log2 fold change == 0)*

2- *Alternative hypothesis: there is a differential gene expression across fetal and adult brain tissues (log2 fold change > 0 or < 0).*

*In order to test the hypothesis, I used a statistical test of differential expression analysis to determine which hypothesis is true based on the selected data. To test the differentially expressed genes between age groups, a Generalized Linear Model (GLM) was constructed using the package **DESeq**. Using DESeq with the expression data is possible to get probability value (p-value), log2 foldchange estimates, the adjusted p-value (padj) for each gene uploaded in the .txt file obtained within the last assignment. The results are shown in the GLM_results.txt previously annexed.*

*Then, the resulting p-values were adjusted using the Benjamini-Hochberg (BH) method. As a result, I found **472** differentially expressed genes between the groups, with error rate of ± 12 (.*

Finally, the p-values and fold change of the GLM analysis were used to create a volcano plot (Figure 1). P-values are shown in y-axis (-log10(pvalue)), while fold change is represented in x- axis (log2FoldChange).



Together, these results indicate a clearly differential gene expression between the fetal and adult brains, which corroborates with my alternative hypothesis.

Algorithm

```
      pdata = colData(capstone) edata = assay(capstone) summary(edata)
sum(is.na(edata))

      edata = as.matrix(edata[rowMeans(edata)>10,])
```

```
# Fitting the generalized linear models.
      de2= DESeqDataSetFromMatrix(edata, pdata, ~age_group+sex)
glm_all_nb2 = DESeq(de2)
```

```
# Adjusting the p-values
      fp_bonf = p.adjust(result_nb2$pvalue,method="bonferroni")
hist(fp_bonf,col=3)
```

```
# Writing the table
      res = data.frame(Gene = row.names(result_nb2), log2FoldChange =
result_nb2$log2FoldChange,

      pvalue = result_nb2$pvalue, padj = fp_bh) head(res)

      write.table(res, GLM_results.txt', sep = '\t', row.names = F)
```

```
# Make the volcano plot
      with(res, plot(log2FoldChange, -log10(pvalue), pch=20,
main="Volcano plot", xlim = c(-25,25))) with(subset(res, padj<.05 ),
points(log2FoldChange, -log10(pvalue), pch=20, col="red"))

      with(subset(res, abs(log2FoldChange)>1), points(log2FoldChange,
-log10(pvalue), pch=20, col="orange"))

      with(subset(res, padj<.05 & abs(log2FoldChange)>1),
points(log2FoldChange, -log10(pvalue), pch=20, col="blue"))
```