Genomics Data Science Capstone Week-8

Statistical analysis was performed(on the Gene expression Counts obtained for Featurecounts in week6, using the DESEQ2 method. (Version1.30.1). A DESEQ object is created from the SummarizedExperiment object created in week 7.

Normalization is in-built in DESEQ2 method and it is recommended to use raw counts while performing DESEQ2. So, Normalization was not performed. As a clear distinction between adult and fetal sample was seen during PCA analysis in week 7. Thus, no covariates were adjusted during analysis.

The results from DESEQ2 were sorted for p-values and the genes were annotated using the org.Hs.eg.db package version 3.12.2. A dataframe object containing the Gene name, logFoldChange, p=value and p-adj was saved into a textfile.

The Null hypothesis states that there is no significant difference in gene expression between the Fetal and Adult samples.

Hypothesis testing - Wald test is used for hypothesis testing in DESEQ2, when comparing two groups. Here the LogFoldChange is used to derive the p-value. If the p value of gene expression data is small, we can reject the null hypothesis.

Correction for multiple testing - As a very large number of genes(25,702) are being tested here creating a Multiple testing problem. For significance with a pvalue cutoff of 0.05, 5% of genes found to be significant would actually be false positives. (To correct for this, DESEQ2 uses the Benjamini-Hochberg principle, which uses FDR to limit the number of false positives. An FDR of 0.05 indicates that out of 1000 genes determined to be significant, 50 (5%) of them are expected to be false positives.

FDR cutoff of 0.05 and LFC= 0.58 was used to extract significantly differential expressed genes. The following table represents the combinations tried out for obtaining the Up and Down regulated genes.

	Up	**Down**

FDR = 0.01
LFC =0.58
	2212
	9.5%
		2950
		13%

FDR = 0.05
LFC =0.58
	2756
	12%
		3551
		15%

FDR = 0.01
LFC =1
	1221
	5.3%
		1952
		8.4%

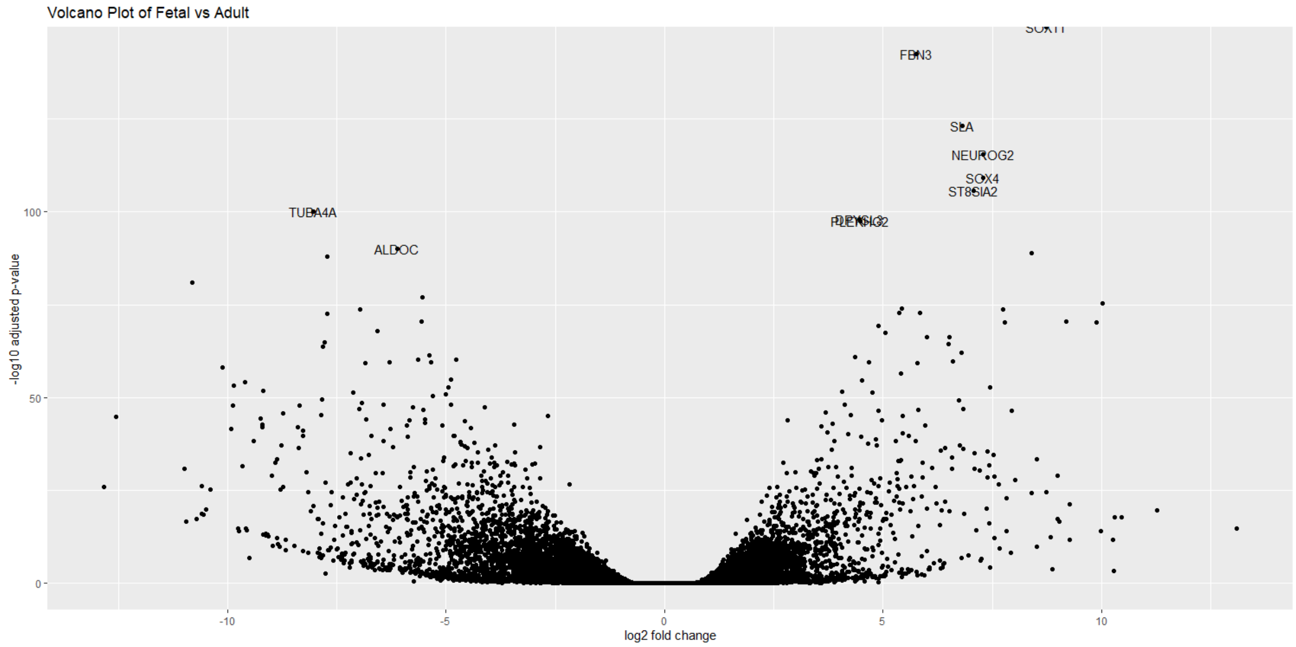
FDR = 0.05
LFC =0.58
	1567
	6.8%
		2401
		10%

Table 1 – Number of significant genes obtained, based on FDR and LFC

Results – A very large number of genes was seen to be differentially expressed and the null hypothesis can be rejected in this case. With an FDR of 0.05, 27 % genes were found to be up or down regulated by a fold change of 1.5 times and ~17% showed a fold change of 2. With an FDR of 0.01, 23% genes showed Log change of 1.5 and ~14% showed a log change of 2.

Visualization of the Differential Expression

A volcano plot of the differentially expressed genes was made using ggplot2 ver3.3.3.



The top 10 differentially expressed genes were labelled on the plot.

SOX11

FBN3

SLA

NEUROG2

SOX4

ST8SIA2

TUBA4A

DPYSL3

PLEKHG2

ALDOC