Task 6 – Coursera Genomic Data Science Capstone

After performing expression mapping the RNA samples to human genome version hg19 using *HISAT 2.1* and *featureCounts* 2.0.1 (features were combined using *Multi-Join* 1.1.1) at *Galaxy Server 2.10.8*, data were downloaded and loaded at *R* 4.0.3 for analyses using *Bioconductor* 2.5. The counts data were first filtered to exclude genes with counts smaller than 10 when considering all six samples. To normalize the data, the raw counts values were transformed by *log2* conversion. The counts were then analyzed using a linear model with *limma* 3.46, using the *lmFit* function with group age as predictor. This model was then analyzed using the *eBayes* function to compute several statistics of the differential expression by empirical Bayes moderation of standard errors. Results of the *eBayes* function were then summarized using the *topTable* function to extract the top-ranked genes and perform *p-value* adjustments for multiple testing by the Benjamini & Hochberg (1995) method.

Results of the *topTable* function were utilized to generate a Volcano plot (Fig. 1) of the gene expression values. Of the 3119 genes with expression counts after normalization, 2261 showed adjusted *p-values* lower than 0.05, consistent with differential expression. Of these, 1733 were downregulated, and 205 were upregulated. The fetal samples presented more differentially expressed genes (Fig. 2), especially downregulated genes, while the adult samples presented more upregulated genes, with the samples of the two age groups being easily separated in a scatter plot of upregulated versus downregulated genes.

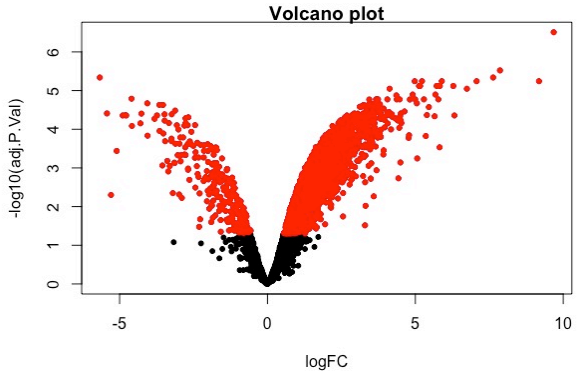

Fig. 1. Volcano plot generated of the *limma* analysis of differential expression, with the genes with an adjusted *p* value lower than 0.05 show in red.

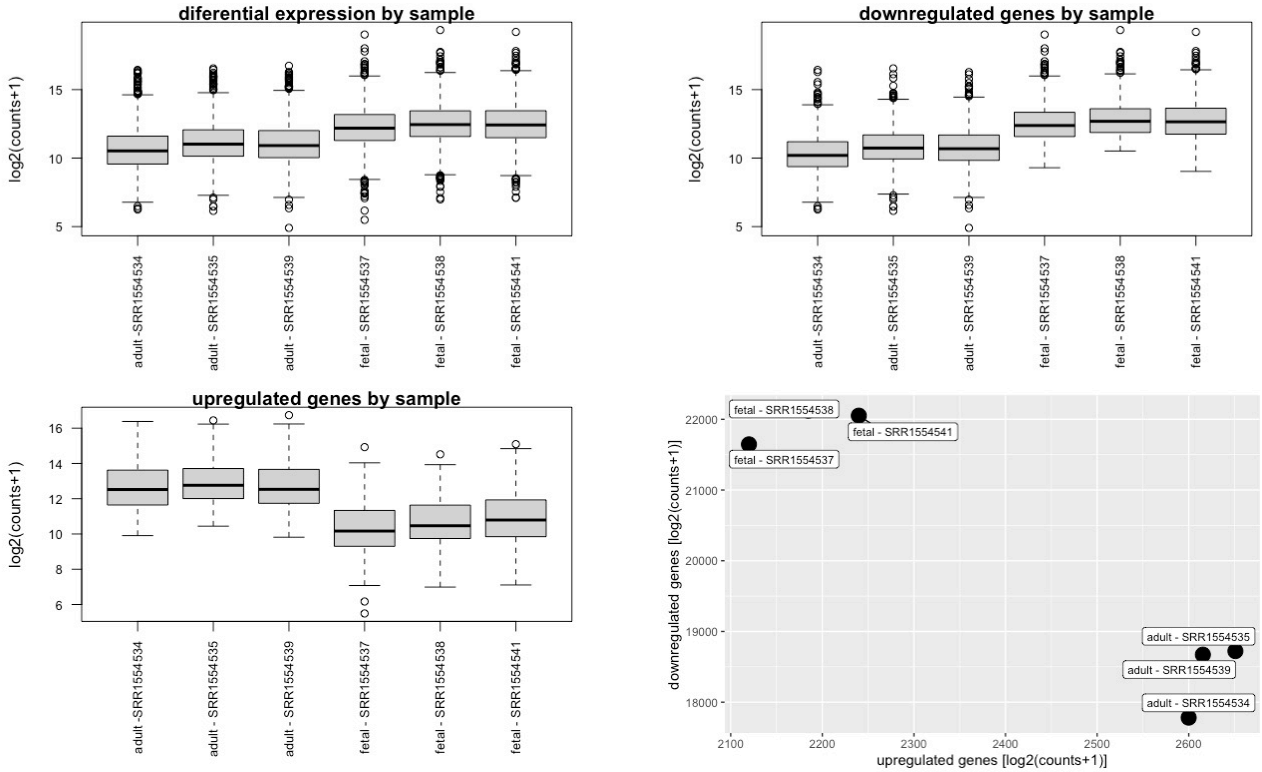

Fig. 2. Boxplots and scatter plot of the results of the *limma* differential expression analysis. At top left are show the counts of all differentially expressed genes by sample. At top right are show the counts of downregulated genes by sample. At bottom left are show the counts of the upregulated genes by sample. At bottom right are show a scatter plot of the number of counts of up and downregulated genes by sample.