Re-analysis of fetal and adult brain raw RNA-seq data from the study "Developmental regulation of human cortex transcription and its clinical relevance at base resolution" (Jaffe et al, 2015 Jan PMID:25501035)

Task 6: Statistical analysis of differential expression

For this step, I started from the results of previous Exploratory analysis (https://www.coursera.org/learn/genomic-data-science-project/peer/YQkBz/exploratory-analysis/review/tEmcJDY9EeumzA6b1BYH7w), that denoted reads quality, according to QC, and RIN level as the main confounders when studying differential gene expression between Fetal and Adult groups of samples.

It seems more reasonable to exclude poor quality samples from the analysis rather than to adjust for the quality level, so I re-performed PCA analysis (here not shown) considering just the good quality samples (5 fetal, 8 adult, see https://www.coursera.org/learn/genomic-data-science-project/peer/1UJTe/qc-the-alignment/review/j52jJjNdEeunYBLc-HtZXw) and obtained the following correlation coefficients:

```
##             First PC     Second PC
## Series    0.32774734  0.3960369357
## LStage   -0.93211181 -0.9890698718
## Age       0.90730878  0.9046919437
## Gender   -0.03895199  0.0002164819
## RINlevel  0.01605282 -0.0300253440
```

Now, the correlation with RIN as disappeared. Excluding the Age variable (which is already correlated with our regressor of interest), the only confounder appears to be the sample Series (the first or second technical replicate of the sequencing experiment, considered as a factor). Therefore, the null hypothesis is:

H0= "Expression of a feature is equal in Fetal and Adult samples, after adjusting for the set of technical replicates".

The alternative hypothesis is:

H1= "A feature is differentially expressed in Fetal and Adult samples, after adjusting for the set of technical replicates".

The significance level will be measured with p-value adjusted for multiple testing using the Benjamini-Hockberg correction for FDR.

I'm using the DESeq2 package in R 4.0.2 and raw counts data obtained with featureCounts on Galaxy. I'm not using normalized data as, according to the package vignette, DESeq2 runs normalisation behind the scenes, thus working on the same normalized data I used in my exploratory analysis.

Data won't even be tranformed. Indeed, although helpful for exploration and visualisation, an (e.g. log) transform is unuseful for statistical purposes and would lead to difficult interpretation of negative binomial modelling results (this modelling already involves computation of a log2 fold changes value). Filtering for feature means > 0 was conversely maintained, because it only removes genes that aren't expressed in any of the samples.

```
dim(counts)
```

```
## [1] 82960    24
```

```
countgood<-counts[,pdata$Quality=="Good"]
dim(countgood)
```

```
## [1] 82960    13
```

```
pdatagood<-pdata[pdata$Quality=="Good",]
dim(pdatagood)
```

```
## [1] 13 11
```

```
countgoodfilt<-countgood[rowMeans(countgood)>0,]
dim(countgoodfilt)
```

```
## [1] 36726    13
```

As it's more intuitive to observe changes in Adult samples with respect to Fetal samples, I first releveled the "Life Stage" factor:

```
pdatagood$LStage<-relevel(pdatagood$LStage,"Fetal")
pdatagood$LStage
```

```
##  [1] Fetal Fetal Fetal Fetal Fetal Adult Adult Adult Adult Adult Adult Adult
## [13] Adult
## Levels: Fetal Adult
```

Then I performed differential expression analysis and plotted the p-adjusted frequencies:

```
de<-DESeqDataSetFromMatrix(countgoodfilt, pdatagood, design= ~ LStage + Series)
de<-DESeq(de)
resultsNames(de)
```

```
## [1] "Intercept"             "LStage_Adult_vs_Fetal" "Series_2_vs_1"
```

```
res<-results(de,name="LStage_Adult_vs_Fetal")
res<-res[!is.na(res$padj),]
dim(res)[1]
```

```
## [1] 28039
```

```
dim(res[res$padj<0.05,])[1]
```

```
## [1] 11424
```

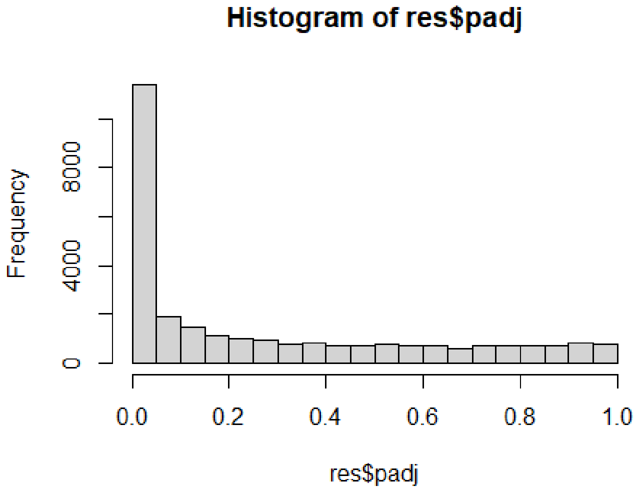
```
sign<-(res[res$padj<0.05,])
dim(sign[sign$log2FoldChange>0,])[1]
```

```
## [1] 6010
```

```
dim(sign[sign$log2FoldChange<0,])[1]
```

```
## [1] 5414
```

```
hist(res$padj)
```



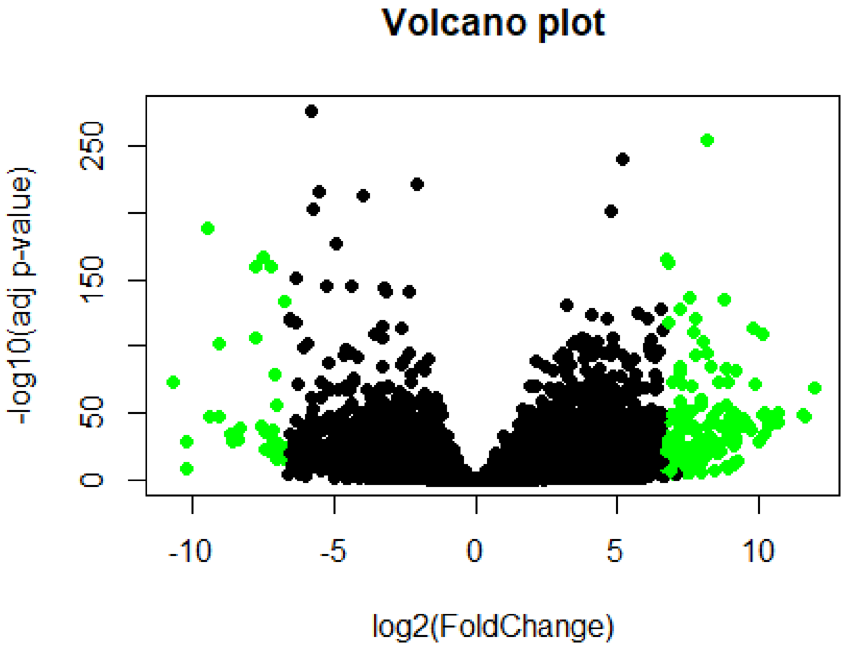
At a max false discovery rate of 0.05, there are 11424 differentially expressed features (out of 82960 total features in the UCSC known gene hg19 assembly database). Of those, 6010 are overexpressed in adults and 5414 are overexpressed in fetuses.

Furthermore, at the same significance level, the majority (10123) of those features shows an adult expression level at least halved or doubled with respect to fetal expression. So, we observe a really considerable change in gene expression.

The histogram shows a good distribution of adjusted p-values.

In the following volcano plot, genes with both high FC (>100 times) and low FDR (<0.01) are labeled in green.

```
par(pch=19)
plot(res$log2FoldChange,-log(res$padj), main="Volcano plot",
xlab="log2(FoldChange)", ylab="-log10(adj p-value)",
col=ifelse(abs(res$log2FoldChange)>log2(100) & res$padj<.01,"green","black"))
```



As a check, I picked up one of the genes with the lowest p-value and looked at its counts (first 5 samples are fetal, last 8 samples are adult):

```
res["uc002mjf.3",]
```

```
## log2 fold change (MLE): LStage Adult vs Fetal
## Wald test p-value: LStage Adult vs Fetal
## DataFrame with 1 row and 6 columns
##             baseMean log2FoldChange     lfcSE      stat       pvalue
##            <numeric>      <numeric> <numeric> <numeric>    <numeric>
## uc002mjf.3   2059.96       -5.82328    0.2449  -23.7782 5.60834e-125
##                    padj
##               <numeric>
## uc002mjf.3 7.86261e-121
```

```
countgoodfilt["uc002mjf.3",]
```

```
##            SRR1554537 SRR1554538 SRR1554566 SRR1554567 SRR1554568 SRR1554535
## uc002mjf.3       9823      12144       9769       9818       8786         58
##            SRR1554536 SRR2071347 SRR1554539 SRR1554556 SRR1554561 SRR1554534
## uc002mjf.3         21         57         93        110        132         49
##            SRR2071345
## uc002mjf.3         58
```

As shown, DESeq2 computing is consistent with count data (negative log2FC and gene underexpressed in adults).