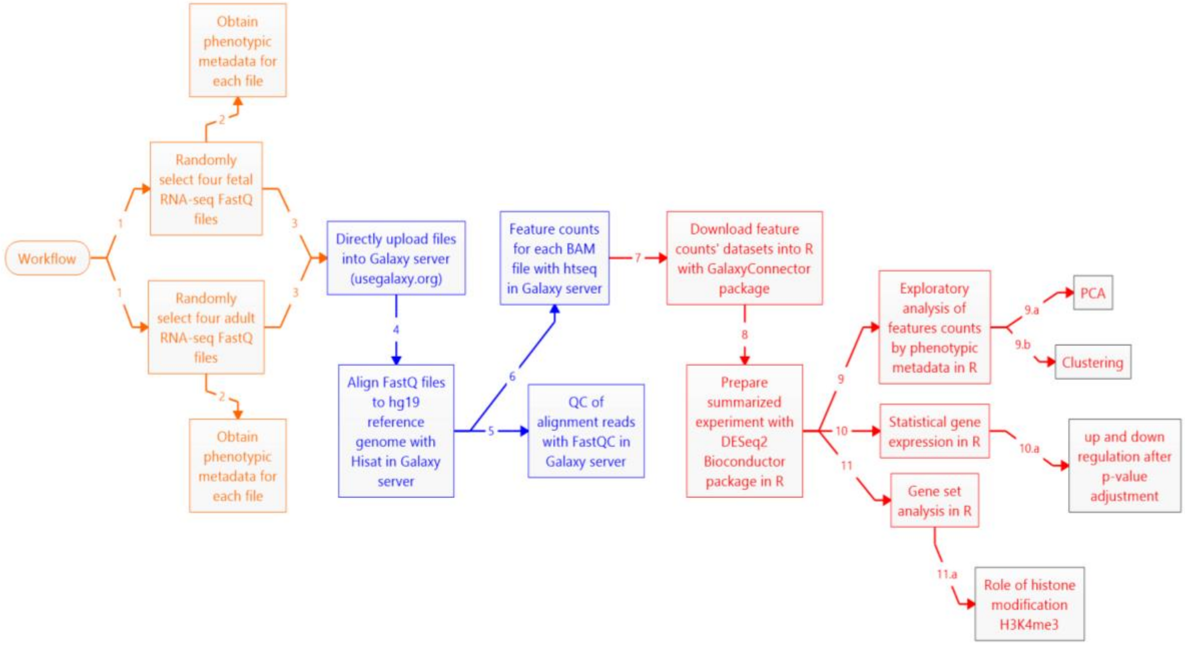
Genomic Data Science Capstone: Week 10 (Describe your Analysis)

1. INTRODUCTION

This project describes the RNA-seq data re-analysis workflow to evaluate differential gene expression between fetus and adult brains. The RNA-seq FastQ files have been previously analyzed and the result were published in the ***Nature Neuroscience*** *(Jaffe at al, 2015, vol 18(1), pages 154-161).*

The software, parameters, packages, and commands used in each step are briefly described in the following workflow:



Obtain phenotypic metadata for each file
2
Randomly select four fetal RNA-seq FastQ files
1
3
Workflow
1
3
Randomly select four adult RNA-seq FastQ files
2
Obtain phenotypic metadata for each file
Directly upload files into Galaxy server (usegalaxy.org)
4
Align FastQ files to hg19 reference genome with Hisat in Galaxy server
5
6
QC of alignment reads with FastQC in Galaxy server
Feature counts for each BAM file with htseq in Galaxy server
7
Download feature counts' datasets into R with GalaxyConnector package
8
Prepare summarized experiment with DESeq2 Bioconductor package in R
9
10
11
Exploratory analysis of features counts by phenotypic metadata in R
9.a
PCA
9.b
Clustering
Statistical gene expression in R
10.a
up and down regulation after p-value adjustment
Gene set analysis in R
11.a
Role of histone modification H3K4me3

Note: *All the files and codes associated with this genomic data science report can be found in this Github Repository.*

2. METHODS AND RESULTS

2.1 Obtaining RNA-seq and Phenotypic Metadata from the Samples

The files for all the samples analyzed in the article, including the phenotypic metadata for each sample, were uploaded directly into Galaxy server using the following mirror website: https://www.ebi.ac.uk/ena/browser/view/PRJNA245228. This website provides: ***(1)*** *links to download FastQ files into a local drive;* ***(2)*** *FTP links for each FastQ files; and* ***(3)*** *Links to directly upload FastQ files into a Galaxy instance. The latter option was used in this genomic data analysis report. The phenotypic metadata for the selected samples can be observed in* ***Table 1****.*

Table1. Phenotypic metadata for the fetal and adult brain samples data

Sample	SRA	AGE	SEX	RACE	FRACTION
SRR1554539	SRS686967	36,5	FEM	AA	Total
SRR1554534	SRS686962	40,4	MAL	AA	Total

SRR1554536	SRS686964	44,1	FEM	AA	Total
SRR1554541	SRS686969	-0,38	MAL	AA	Total
SRR1554537	SRS686965	-0,38	FEM	AA	Total
SRR1554567	SRS686995	-0,4	MAL	AA	Total

2.2 Alignment

*The FastQ files for the selected samples in **Table 1** were uploaded directly into Galaxy. To access raw data and aligning specific set of samples, the files were imported, and the alignments performed with HISAT2 (v.2.1.0+galaxy5). I have aligned to the build human genome hg19 (b37) with the paired-end from single interleaved dataset and threated the data as unstranded.*

2.3 Quality Control

*On Galaxy, I have used the **Samtools flagstat tabulate descriptive stats for BAM dataset** (Galaxy Version 2.0.3) tool to find the percentage of mapped reads per sample (Table 3). Next, I have used the FastQC Read Quality reports (Galaxy Version 0.72+galaxy1) tool to perform a Quality Control of the alignments **(Table 2)**.*

Table 2. Data analysis of percent mapped, average of per sequence quality, quality scores and GC percentage of fetus and adult brain samples.

Sample	Pct_mapped	psqallmean	psqscallmean	qc_content
SRR1554539	0.9861171	34.80674	3965141	0.47
SRR1554534	0.9818536	33.63508	3960700	0.46
SRR1554536	0.9829014	34.00784	3128929	0.49
SRR1554541	0.97899783	33.46695	2027034	0.47
SRR1554537	0.977613	35.81316	1571383	0.51
SRR1554567	0.09757721	35.01644	2182210	0.52

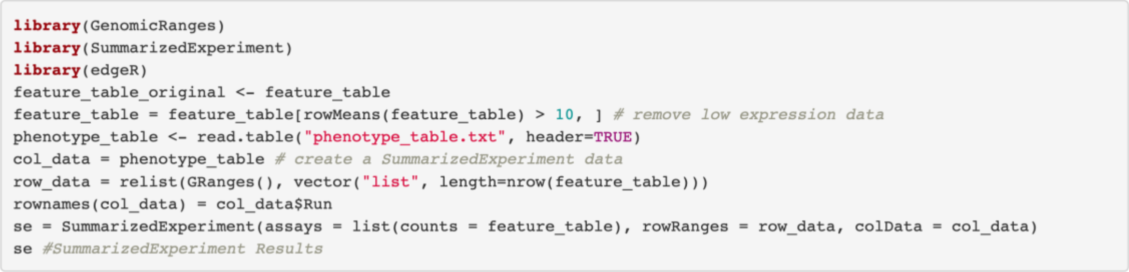
Pct_mapped = percent mapped; **psqallmean** = average of per sequence quality; **psqscallmean** = quality scores; **and qc_content** = GC percentage

2.4 Gene Expression Analysis

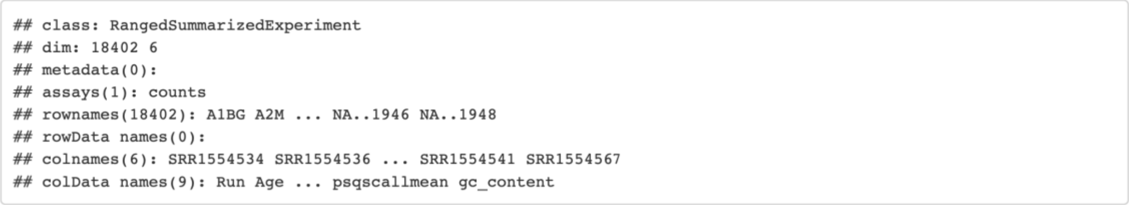
*Using Galaxy, I performed gene expression analysis **with featureCounts measure gene expression in RNA-Seq experiments from SAM or BAM files** (Galaxy Version 1.6.4+galaxy2) tool. I have calculated the abundance of every gene in every sample. This count approximates the expression level for each gene and can be accessed here.*

2.5 Exploratory Analysis

Then, I performed an exploratory analysis technique called Principal Component Analysis (PCA) on the Gene Expression results. The exploratory analysis was performed with the following R (v. 4.0.2) code:



```
library(GenomicRanges)
library(SummarizedExperiment)
library(edgeR)
feature_table_original <- feature_table
feature_table = feature_table[rowMeans(feature_table) > 10, ] # remove low expression data
phenotype_table <- read.table("phenotype_table.txt", header=TRUE)
col_data = phenotype_table # create a SummarizedExperiment data
row_data = relist(GRanges(), vector("list", length=nrow(feature_table)))
rownames(col_data) = col_data$Run
se = SummarizedExperiment(assays = list(counts = feature_table), rowRanges = row_data, colData = col_data)
se #SummarizedExperiment Results
```



```
## class: RangedSummarizedExperiment
## dim: 18402 6
## metadata(0):
## assays(1): counts
## rownames(18402): A1BG A2M ... NA..1946 NA..1948
## rowData names(0):
## colnames(6): SRR1554534 SRR1554536 ... SRR1554541 SRR1554567
## colData names(9): Run Age ... psqscallmean gc_content
```

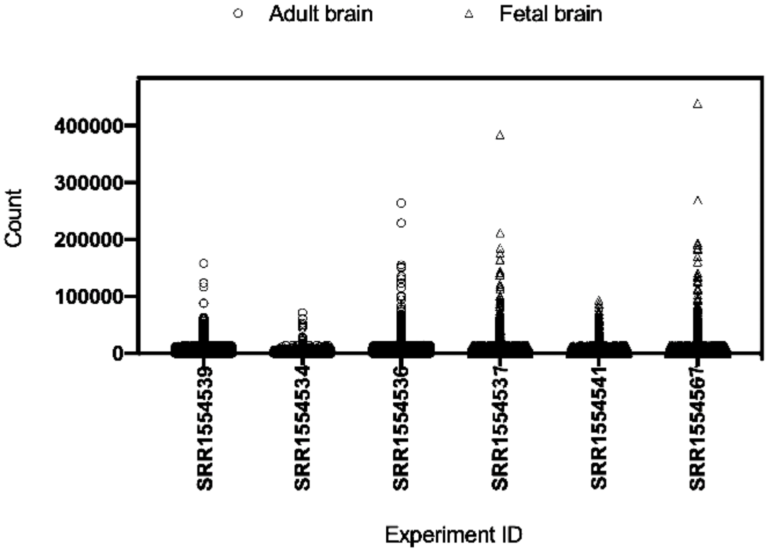
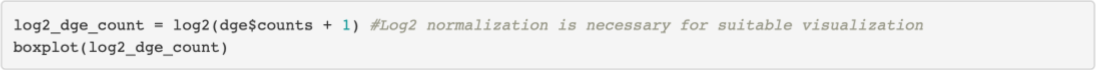


Figure 1. Raw gene counts for each sample.



```
log2_dge_count = log2(dge$counts + 1) #Log2 normalization is necessary for suitable visualization
boxplot(log2_dge_count)
```

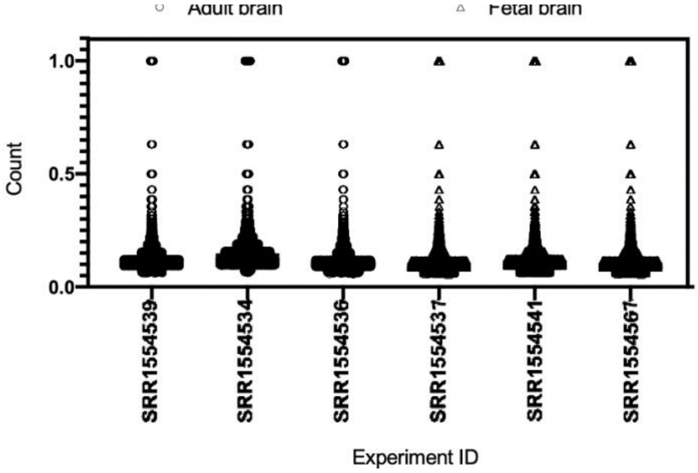
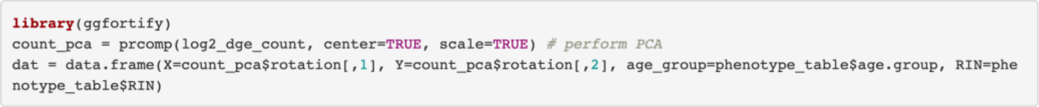
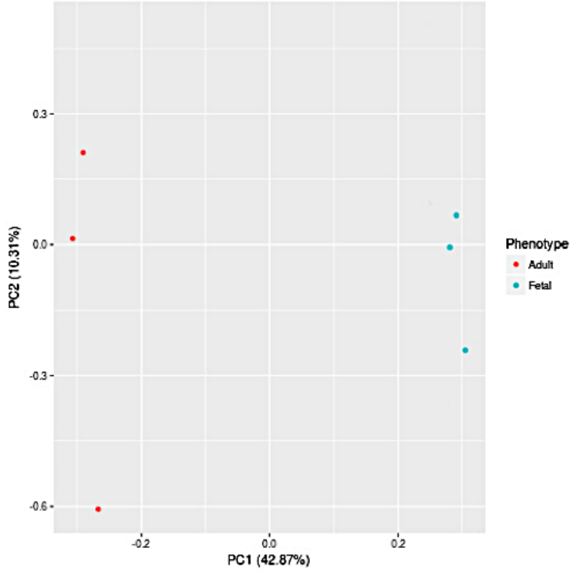


Figure 2. Normalized logged base 2 gene counts for each sample.

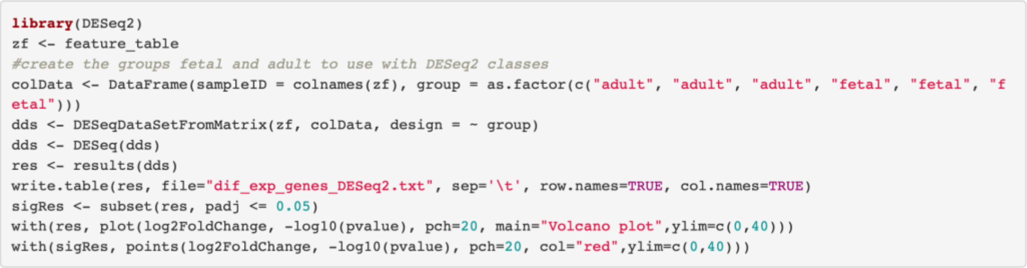


```
library(ggfortify)
count_pca = prcomp(log2_dge_count, center=TRUE, scale=TRUE) # perform PCA
dat = data.frame(X=count_pca$rotation[,1], Y=count_pca$rotation[,2], age_group=phenotype_table$age.group, RIN=phe
notype_table$RIN)
```

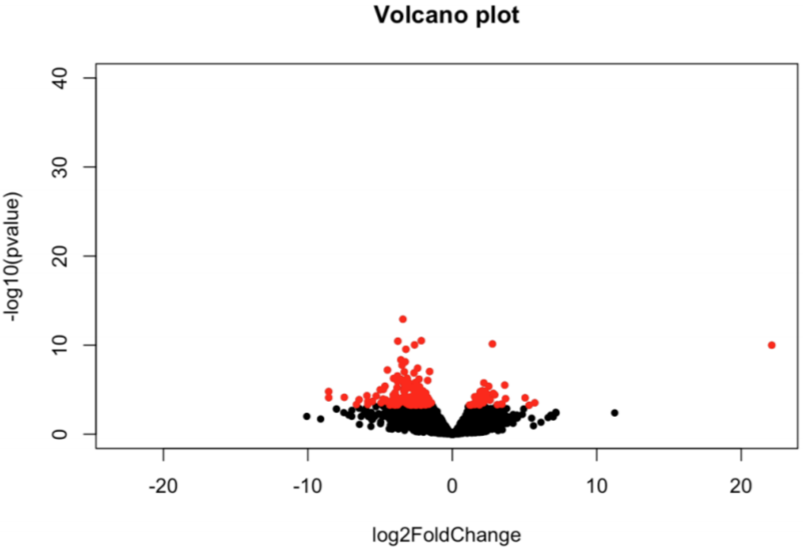


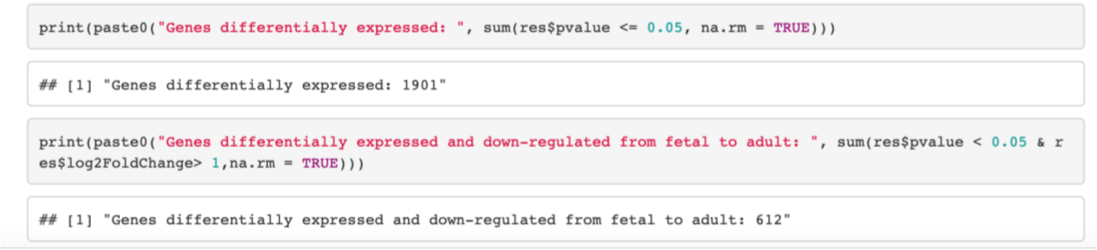
2.6 Statistical Analysis

Then, I performed a Statistical Analysis of the gene expression data across fetal and adult samples by using the following R (v. 4.0.2) code and the DESeq2 library.



```
library(DESeq2)
zf <- feature_table
#create the groups fetal and adult to use with DESeq2 classes
colData <- DataFrame(sampleID = colnames(zf), group = as.factor(c("adult", "adult", "adult", "fetal", "fetal", "f
etal")))
dds <- DESeqDataSetFromMatrix(zf, colData, design = ~ group)
dds <- DESeq(dds)
res <- results(dds)
write.table(res, file="dif_exp_genes_DESeq2.txt", sep='\t', row.names=TRUE, col.names=TRUE)
sigRes <- subset(res, padj <= 0.05)
with(res, plot(log2FoldChange, -log10(pvalue), pch=20, main="Volcano plot",ylim=c(0,40)))
with(sigRes, points(log2FoldChange, -log10(pvalue), pch=20, col="red",ylim=c(0,40)))
```





```
print(paste0("Genes differentially expressed: ", sum(res$pvalue <= 0.05, na.rm = TRUE)))
```

```
## [1] "Genes differentially expressed: 1901"
```

```
print(paste0("Genes differentially expressed and down-regulated from fetal to adult: ", sum(res$pvalue < 0.05 & r
es$log2FoldChange> 1,na.rm = TRUE)))
```

```
## [1] "Genes differentially expressed and down-regulated from fetal to adult: 612"
```

2.7 Gene Set Analysis

*Finally, I have proceeded to perform an analysis of the promoter associated histone modification **H3K4me3** in differential gene expression between and fetal adult brains.*

*The full algorithm for this analysis can be found at this Github Repository. The summary results are shown in **Figure 3**.*

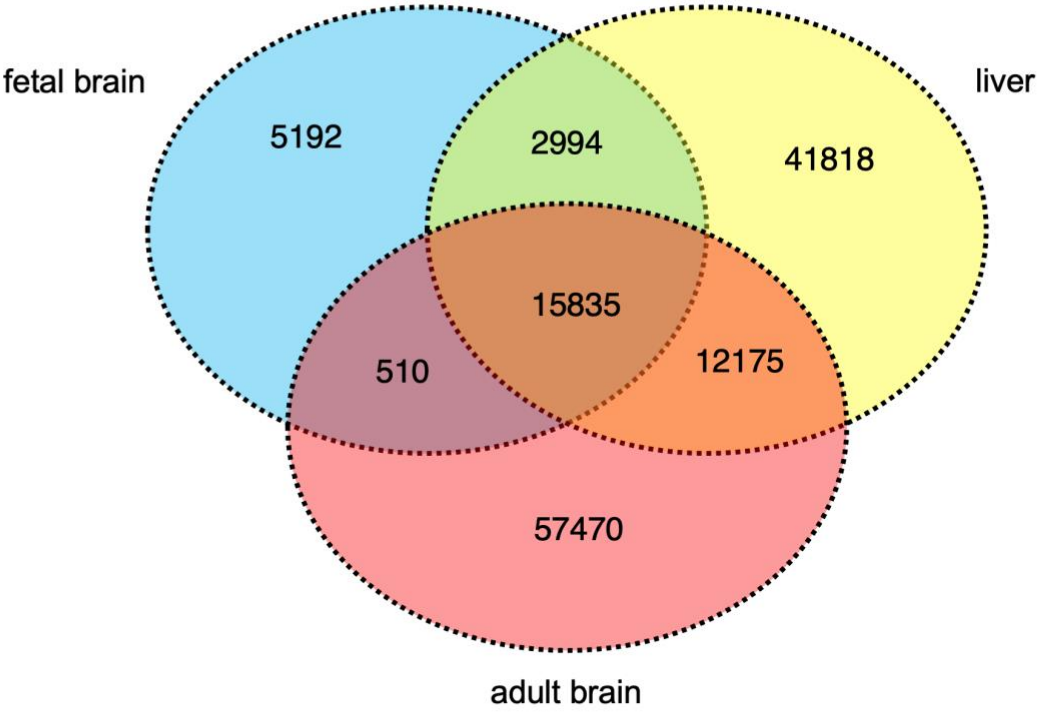


Figure 3. Venn Diagram of the calculation with promoters with H3K4me3 peaks

3. CONCLUSIONS

Together, my results suggest a differential expression of 1601 genes between the fetal and adult dorsolateral prefrontal cortex. The promoters of these genes have greater percentage of H3K4me3 histone modification in the adult tissues than in the fetal tissue, indicating that those genes are more expressed in adult tissue.