Differences in Gene Expression: Fetal & Adult Human Brains

Genomic Data Science Capstone - Week 10 - Final Report

Cesar Benjumea

1/14/2022

Introduction

This project analyzes the transcriptome of fetal and adult brains, particularly the changes in H3K4me3 profiles between fetal and adult brains over promoters of differentially expressed genes. For this purpose, sequencing raw data was retrieved from the SRAs obtained in the scope of the published manuscript "Developmental regulation of human cortex transcription and its clinical relevance at base resolution", available at: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4281298/. Then, the alignment, mapping and counting of the reads was performed in https://usegalaxy.org/, and the data analysis performed in R.

The code of this R markdown, and the files used to build this report are publicly available at: https://github.com/cesarhbenjumea/CourseraCapstone_GenomicDataScience.git

Load Libraries

```
library(org.Hs.eg.db); library(TxDb.Hsapiens.UCSC.hg19.knownGene);
library(plyranges); library(GenomicRanges); library(cowplot)
library(rtracklayer); library(AnnotationHub)
library(knitr); library(scales); library(dplyr); library(ggplot2)
library(SummarizedExperiment); library(DESeq2)
library(reshape); library(PCAtools); library(edgeR)
```

Alignment and Gene Counts

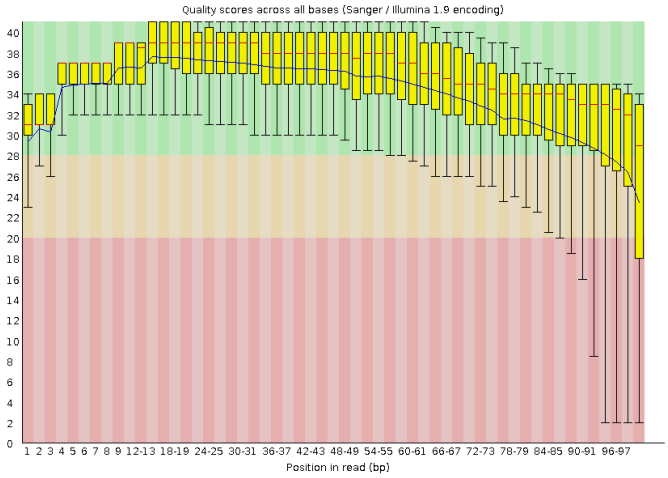
Three samples of fetal and adults brains were retrieved for processing. The median score of the sequences evaluated with FastQC were greater than 30, meaning that there were no critical problems with the data in terms of quality. However, at the end of each read the base-pair quality read significantly decreases (Figure 1, left). After performing the quality check, the HISAT2 algorithm was used for aligning the data. The detailed information about the specific versions of the methods and functions used for this analysis can be accessed through this link: https://usegalaxy.org/u/cesarhbenjumea/h/capstone-project

The phenotype information about the samples, quality measures and percentage of mapped reads are summarized in Table 1. 5 out of 6 the samples have very similar metrics in terms of quality and mapped reads, but the sample labeled Adult 1 (Table 1, and data 6 in Figure 1 -right) seems to have a lower percentage than the rest. Also, The Adult 2 sample has a lower RIN than the rest but this metric didn't seem to affect the quality scores and the percentage of mapped reads.

Name	ColName	SRX	SRA	age	sex	Mapped_Percentage	median_QCs	RIN
Fetal 1	data 126	SRX683826	SRS686996	-0.4986	male	0.926	33.94	8.0
Fetal 2	data 7	SRX683824	SRS686994	-0.4986	male	0.934	34.13	8.3
Fetal 3	data 26	SRX683825	SRS686995	-0.4027	male	0.913	35.03	8.6
Adult 1	data 6	SRX683793	SRS686963	41.5800	male	0.827	31.10	8.7
Adult 2	data 4	SRX683794	SRS686964	44.1700	female	0.918	34.76	5.3
Adult 3	data 5	SRX683797	SRS686967	36.5000	female	0.918	34.76	9.0

Table 1. Summary of phenotype information, quality measures and mapped reads of each sample.

A summary of the number of reads and mapped alignments is found in Figure 1.



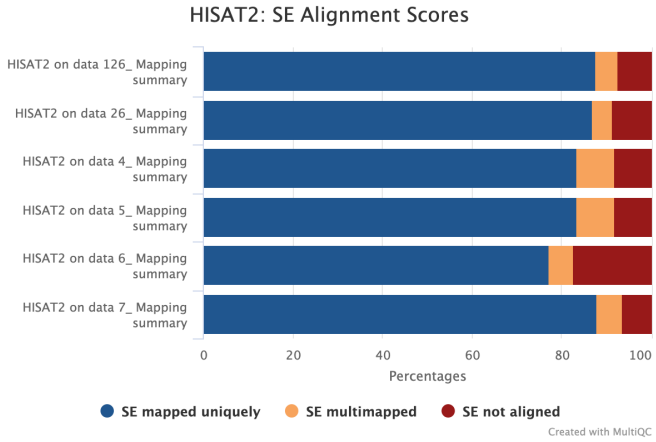


Figure 1: Examples of quality of the reads at base pair level (left) and Summary of mapped alignments (right).

25702 genes were obtained after using the annotateMyIDs function in useGalaxy and matching the GeneIDs with the reads resulting from the featureCounts function.

Exploratory Analysis

The gene count and phenotype data was downloaded from the galaxy server and imported into R.

```
# Import gene count data
data<- read.csv("Feature_Counts_final.csv", header=TRUE)
# Import phenotype data
ColInfo <- read.table("ColInfo.txt", header=TRUE, sep="\t")
# Remove rows without counts
keep <- rowSums(data[,5:10]) > 1
data <- data[keep,]; rm(keep)
```

After removing the rows without counts from the data, there was a total of 7614 genes.

In order to perform a Gene Differential Expression analysis, the Null hypothesis was defined as the case where the gene expression in adult brains is the same as in fetal brains. As alternative hypothesis, the gene expression in fetal brains is different from adult brains. To prepare the data for analysis, the dge package was used to normalize the data (calcNormFactors function).

```
# Counts data are transgormed to a DGE object
dge <- DGEList(counts=data[,5:10])
# Apply TMM normalization
dge <- calcNormFactors(dge)
```

To stabilize the variance across the mean the Voom method was applied. This transforms the count data to log2-counts per million, estimates the mean-variance relationship and use it to compute appropriate observational-level weights.

At this point of the processing, a design matrix was built to test the hypothesis. The Covariates used to adjust the model are Percentage of Mapped Reads (MAP); and, RNA Integrity Number (RIN). These variables are added in an effort to adjust the potential variance of the data created by notorious differences in the phenotype data (Table 1).

```
# Covariates are created in vectors
RIN <- ColInfo[,"RIN"]
FETAL <- c(1,1,1,0,0,0)
MAP <- ColInfo[,"Mapped_Percentage"]
# Design matrix to test the hypothesis
design <- model.matrix(~1+FETAL+MAP+RIN)
```

After building the matrix, the Voom method was applied to the model. The variance-trend graph is shown in Annex 1.

```
v <- voom(dge, design, plot=FALSE)
```

After the transformation, the Boxplot revealed that the general expression level of each sample is similar (Annex 2). This indicated that the data is appropriate for the differential analysis. The Principal Component Analysis (PCA) deteremined that

the first and second PCA explain almost 90% of the variation of the data. Together with the third PCA explain about 99% of the variance (Annex 3).

A biplot for PCA1 and PCA2, including Age as metadata showed a clear cluster of fetal brain samples at the negative variation side of PC1, while adult samples are at the positive side (Figure 3). These clusters that differentiate Fetal and Adult brain samples is expected. Interestingly, two adult samples are very similar and are clearly correlated (The adult samples with 44.17 and 36.5 years), however, the adult sample of 41.58 years is more distant. In fact, in this case it seems that the PC2 mostly explains the difference in variation of the 41.58 years old adult brain sample from the rest of the samples, including the fetal brain samples.

```
p <- pca(v$E, metadata = ColInfo, removeVar = 0.01)
biplot(p, showLoadings = FALSE,
       lab = paste0(p$metadata$age, ' yrs'),
       hline = 0, vline = 0,
       legendPosition = 'right',
       sizeLoadingsNames = 5)
```

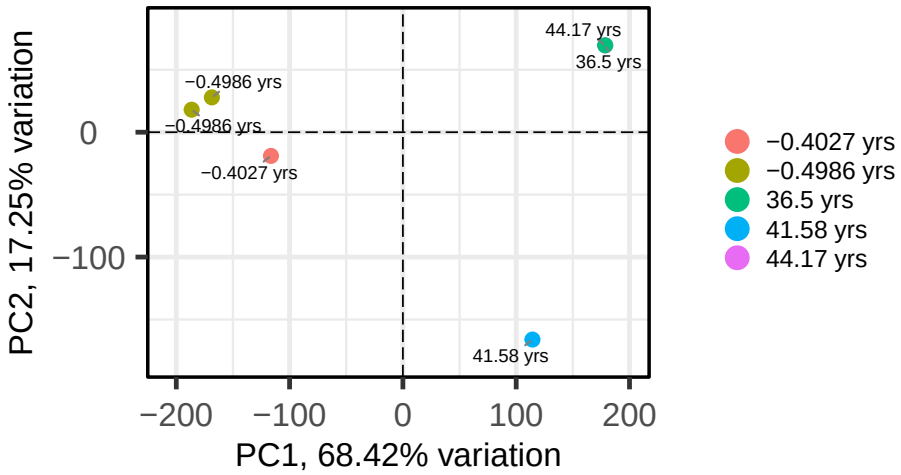


Figure 3. Biplot of PCA2 vs. PCA1.

The variables that are indicated by arrows indicate their 'weight' in different directions. The numbers within the rectangles are the GeneIDs (Annex 4). From this analysis, an interesting finding is GeneID 8404, which represents the SPARC like 1 (SPARCL1) gene. SPARC-like 1 is a member of the SPARC family of matricellular proteins that has been implicated in the regulation of processes such as cell migration, proliferation, and differentiation. There have been studies that show SPARCL1 exhibits remarkably diverse and dynamic expression in the developing and adult nervous system. This is a gene that would be expected to be found while exploring the differences betweeen the fetal and adult cell brain samples.

Another interesting finding is GeneID 83259, which represents the PCDH11Y, a gene unique to Homo males that encodes Protocadherin 11Y, a protein that guides the development of nerve cells. GeneID: 9283 represents the G protein-coupled receptor 37 like 1 (GPR37L1), that has been shown to modulate astrocyte glutamate transporters and neuronal NMDA receptors and is neuroprotective in ischemia.

However, not all the genes found in the biplot are as significant and intuitive as SPARCL1. For example, there is still no gene linked to the sequence represented by GeneID: 84663; GeneID 183 represents angiotensinogen (AGT). The AGT gene provides instructions for making a protein called angiotensinogen. This protein is part of the renin-angiotensin system, which regulates blood pressure and the balance of fluids and salts in the body. These finding may not be intuitively correlated to the differences in development and adult brain samples and require a deeper inspection.

Differential Gene Expression Analysis

The Limma statistical test was applied to the Voom-preprocessed data.

```
# Limma linear fit
fit <- lmFit(v,design)
# eBayes to compute the relevant statistics
tmp <- eBayes(fit)
```

The correction of the pvalues after accounting for multiple testing was the False Discovery Rate (FDR). This correction is embedded in the top.table method (Annex 5).

```
tmp2 <- contrasts.fit(fit, coef = 2) # Second coefficient represents the FETAL variable
tmp2 <- eBayes(tmp2)
top.table <- topTable(tmp2, sort.by = "P", n = Inf)
```

The differentially expressed genes using adjusted pvalues ($p < 0.05$) is 3285. Regarding the other covariates, RIN does not have a significant effect on the linear model, however, the 'Mapped percentage'variable seems to explain some of the variance of certain genes (Annex 6).

After the Limma+Voom processing, the volcano plot of the data (Annex 7) depicts the classic v-type shape, and regions color coded by significance (padj<0,.05, logFC>5) which is expected and a good indication that the data processing had adequately been carried-out. The log Fold Changes (logFC) variables is directly taken from the top.table results.

At this point, a simplified table is build for further processing:

```
table.genes <- data.frame("Geneid" = row.names(top.table),
                  "logFC" = top.table$logFC,
                  "pvalue" = top.table$P.Value,
                  "padj" = top.table$adj.P.Val)
table.genes <- merge(table.genes, data[,c("Geneid", "SYMBOL")], by.y = 'Geneid')
```

Then we take the Differentially Expressed Genes (DEG) and then we group the up-regulated and down-regulated genes:

```
# Filter the genes that are deferentially expressed.
table.genes.f <- subset(table.genes, padj<.05)
# Filter the genes that don't have a gene name in the SYMBOL format
table.genes.f <- table.genes.f %>% filter(!is.na(SYMBOL))
# Filter the up-regulated genes in fetal samples
table.genes.up <- table.genes.f %>% filter(logFC > 0)
# Filter the down-regulated genes in fetal samples
table.genes.down <- table.genes.f %>% filter(logFC < 0)
```

After applying the filters, there are 3054 differentially expressed genes, 1508 are up-regulated, and 1546 are down-regulates, in fetal samples.

Analysis of H3K4me3 profiles in DEG promoters

To analyze the promoter sequences of the differentially expressed genes with the information contained in the Roadmap Epigenetics, it's necessary to use transcript annotation packages to retrieve the genomic coordinates of such genes.

```
txdb <- TxDb.Hsapiens.UCSC.hg19.knownGene
cds <- cdsBy(txdb, by="tx", use.names=TRUE)
```

And the genomic coordinates of each gene were retrieved and this information stored in a vector.

```
# for up-regulated genes
genes.up<- c()
for (i in 1:length(table.genes.up[,5])) {
  a <- table.genes.up[i,5]
  if (a %in% keys(org.Hs.eg.db, "SYMBOL")) {
    eid <- AnnotationDbi::select(org.Hs.eg.db, a, "ENTREZID", "SYMBOL")[["ENTREZID"]]
    if (eid %in% keys(txdb, "GENEID")) {
      txid <- AnnotationDbi::select(txdb, eid, "TXNAME", "GENEID")[["TXNAME"]]
      tryCatch({
        gene.cds <- cds[names(cds) %in% txid]
        genes.up<-c(genes.up,gene.cds[[1]])
      }, error=function(e){cat("ERROR :",conditionMessage(e), "\n")})
    }
  }
}
```

From 1508 genes, it was possible to retrieve the genomic coordinates of 1363.

```
# For down-regulated genes
genes.down<- c()
```

```
for (i in 1:length(table.genes.down[,5])) {
  a <- table.genes.down[i,5]
  if (a %in% keys(org.Hs.eg.db, "SYMBOL")) {
    eid <- AnnotationDbi::select(org.Hs.eg.db, a, "ENTREZID", "SYMBOL")[["ENTREZID"]]
    if (eid %in% keys(txdb, "GENEID")) {
      txid <- AnnotationDbi::select(txdb, eid, "TXNAME", "GENEID")[["TXNAME"]]
      tryCatch({
        gene.cds <- cds[names(cds) %in% txid]
        genes.down<-c(genes.down,gene.cds[[1]])
      }, error=function(e){cat("ERROR :",conditionMessage(e), "\n")})
    }
  }
}
```

And from 1546 genes, it was possible to retrieve the genomic coordinates of 1336.

Regarding the Roadmap Epigenomics Data, the AnnotationHub package was used to retrieve the information of H3K4me3 from fetal brain, adult brain and liver samples. Liver tissue is taken into consideration for control purposes.

```
ah <- AnnotationHub()
ah <- subset(ah, species == "Homo sapiens")
qhs <- query(ah, "H3K4me3")
# Fetal Brain H3K4me3 data
qhs1 <- query(qhs, "Fetal_Brain"); qhs1$title; qhs1$dataprovider
gr1 <- subset(qhs1, title == "BI.Fetal_Brain.H3K4me3.UW_H-22510.narrowPeak.gz")[[1]]
# Adult Brain H3K4me3 data
# There are several types of cells, here we randomly picked Mid Frontal Lobe
qhs2 <- query(qhs, "Brain"); qhs2$title; qhs2$dataprovider
gr2 <- subset(qhs2, title == "BI.Brain_Mid_Frontal_Lobe.H3K4me3.149.narrowPeak.gz")[[1]]
# Liver H3K4me3 data
qhs3 <- query(qhs, "Liver"); qhs3$title; qhs3$dataprovider
gr3 <- subset(qhs3, title == "BI.Adult_Liver.H3K4me3.4.narrowPeak.gz")[[1]]
```

The definition used for promoter sequences was 2kb upstream and 200bp downstream of the start site. Then we searched if there is overlap between the H3K4me3 in promoter regions of differentially expressed genes using the findOverlaps() function. The code is detailed in Annex 8 and the results summarized in Table 2.

	fetal.brain	adult.brain	liver
upregulated	52%	66%	52%
downregulated	38%	71%	55%

Table 2. Summary of percentage of overlap H3K4me3 in promoter genes for different tissues

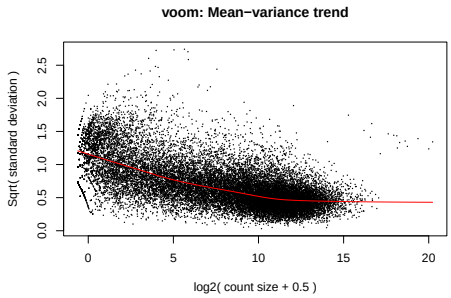
Conclusions

There is evidence that there changes in H3K4me3 between fetal and adult brain over promoters for genes differentially expressed between fetal and adult brain. The H3K4me3 marks are more present in the differentially expressed genes of adult brains, considering both upregulated and downregulated genes. It seems that the H3K4me3 marks are gained during the development, reason why in fetal samples the H3K4me3 marks in key gene promoters are less common.

Some of the promoteers of these genes are also marked by H3K4me3 in the liver, but when analyzing the differences between upregulated and downregulated genes, it's possible to observe that there is no difference in the overlap, as occurs with the brain samples. Thus, the H3K4me3 marks in the liver don't appear to be related with differences in gene expression between fetal and adult brain samples, as expected.

ANNEXES

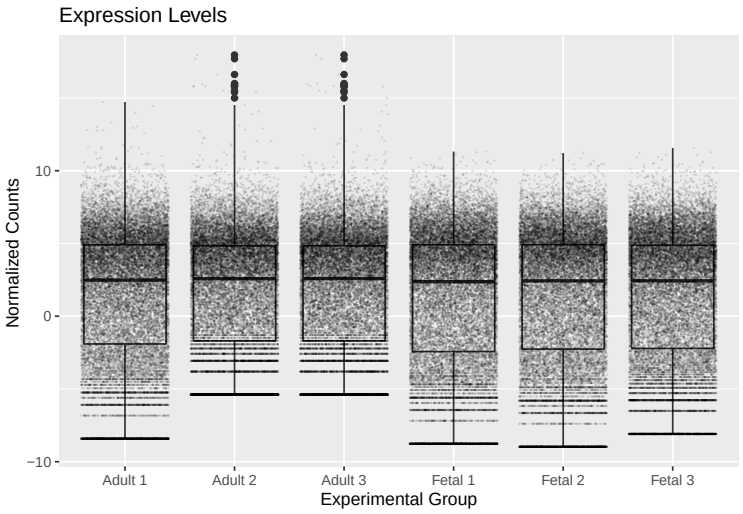
```
v <- voom(dge, design, plot=TRUE)
```



Annex 1. Voom- mean variance trend plot

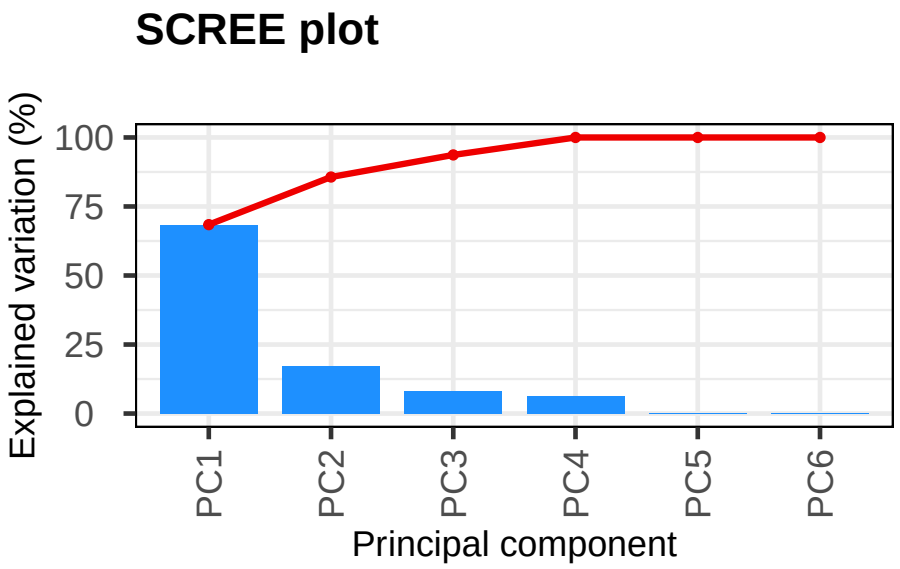
```
d <- as.data.frame(v$E)
colnames(d) <- c("Fetal.1", "Fetal.2", "Fetal.3", "Adult.1","Adult.2","Adult.3")
boxplot <- ggplot(as.data.frame(d)) +
  geom_boxplot(aes(x="Fetal 1", y=Fetal.1)) + geom_jitter(size=0.01,aes(x="Fetal 1", y=Fetal.1), alpha=0.1) +
  geom_boxplot(aes(x="Fetal 2", y=Fetal.2)) + geom_jitter(size=0.01,aes(x="Fetal 2", y=Fetal.2), alpha=0.1) +
  geom_boxplot(aes(x="Fetal 3", y=Fetal.3)) + geom_jitter(size=0.01,aes(x="Fetal 3", y=Fetal.3), alpha=0.1) +
  geom_boxplot(aes(x="Adult 1", y=Adult.1)) + geom_jitter(size=0.01,aes(x="Adult 1", y=Adult.1), alpha=0.1) +
  geom_boxplot(aes(x="Adult 2", y=Adult.2)) + geom_jitter(size=0.01,aes(x="Adult 2", y=Adult.2), alpha=0.1) +
  geom_boxplot(aes(x="Adult 3", y=Adult.3)) + geom_jitter(size=0.01,aes(x="Adult 3", y=Adult.3), alpha=0.1) +
  labs(x = "Experimental Group", y = "Normalized Counts ", title = "Expression Levels")
```

```
boxplot
```



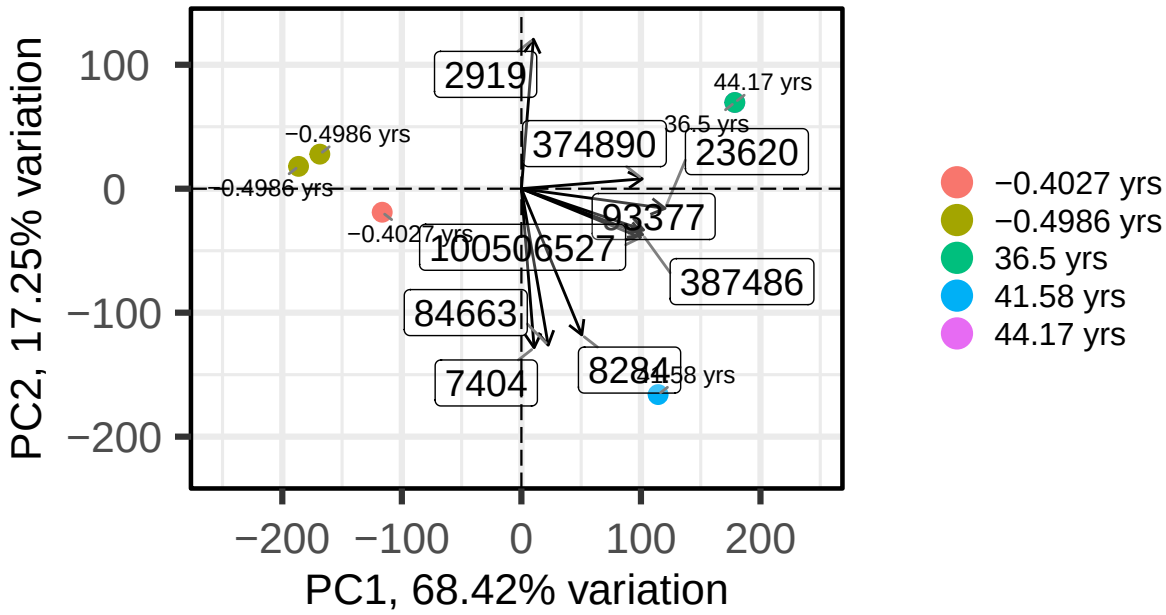
Annex 2. Boxplot showing normalized and log-transformed data for each sample

```
screeplot(p, axisLabSize = 18, titleLabSize = 22)
```



Annex 3. Scree plot from PCA results

```
biplot(p, showLoadings = TRUE,
       lab = paste0(p$metadata$age, ' yrs'),
       hline = 0, vline = 0,
       legendPosition = 'right',
       sizeLoadingsNames = 5)
```



Annex 4. Biplot of PCA2 vs. PCA1, gene weights shown.

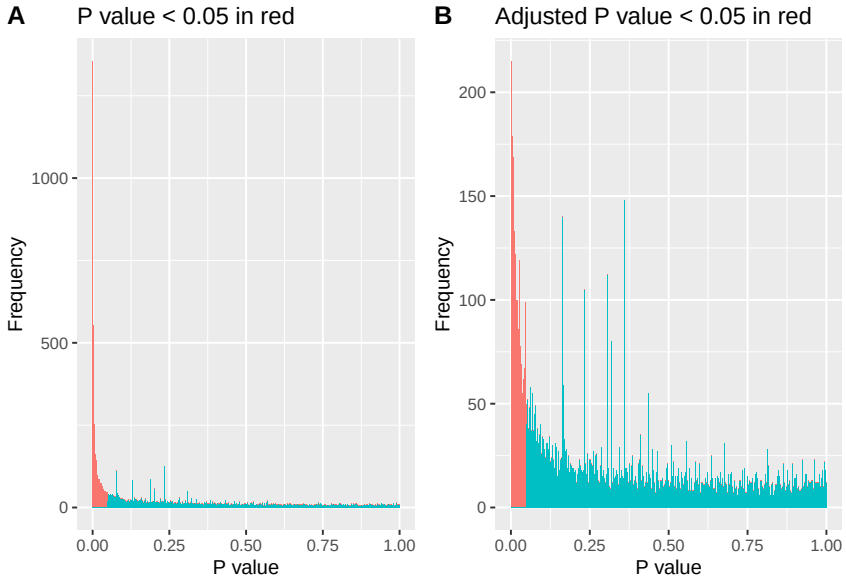
```
x <- data.frame("Pvalues" = top.table$P.Value,
                "adjPvalues" = top.table$adj.P.Val)
x$tag1 <- "B"; x$tag1[x[,1] < 0.05] <- "A"
x$tag2 <- "B"; x$tag2[x[,2] < 0.05] <- "A"

hist.c1 <- ggplot(x, aes(Pvalues, fill = tag1)) +
  geom_histogram(bins=1500) +
  labs(x = "P value", y = "Frequency", title = "P value < 0.05 in red") +
  theme(legend.position = "none")

hist.c2 <- ggplot(x, aes(adjPvalues, fill = tag2)) +
  geom_histogram(bins=1500) +
  labs(x = "P value", y = "Frequency", title = "Adjusted P value < 0.05 in red") +
  theme(legend.position = "none")
```

Counts of differentially expressed genes using unadjusted pvalues ($p < 0.05$):

```
length(x$tag2[x[,1] < 0.05])
```

Counts of differentially expressed genes using unadjusted pvalues ($p < 0.05$):

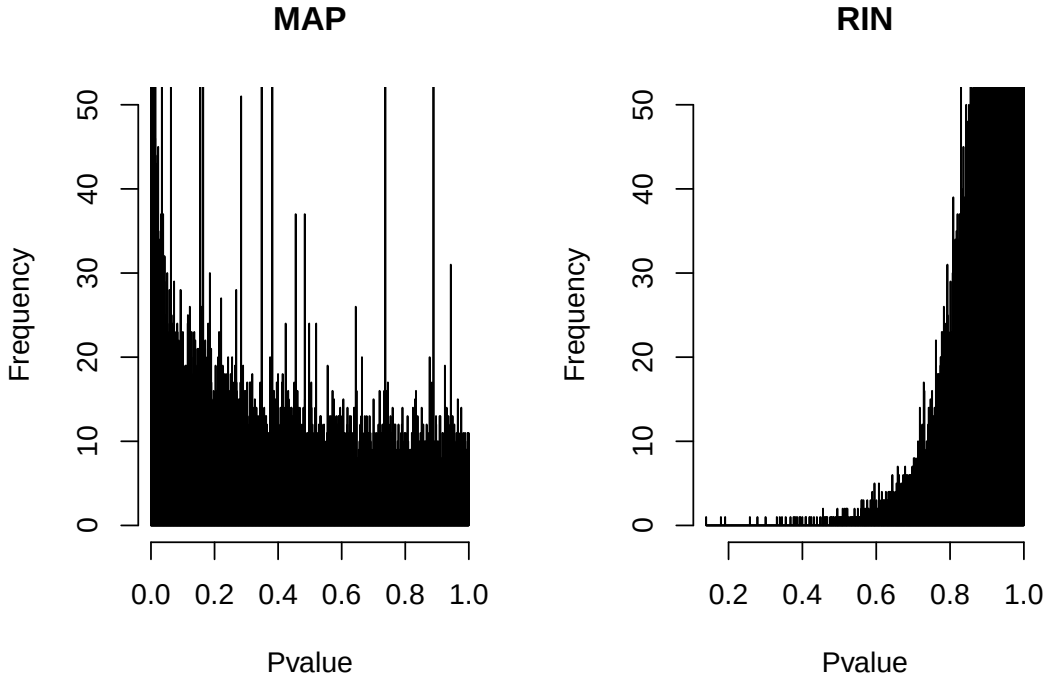
```
length(x$tag2[x[,2] < 0.05])
```



Annex 5. Code for histogram of pvalues before and after the correction for multiple testing.

```
tmp3 <- contrasts.fit(fit, coef = 3) # Third coefficient represents the MAP variable
tmp3 <- eBayes(tmp3)
top.table3 <- topTable(tmp3, sort.by = "P", n = Inf)
tmp4 <- contrasts.fit(fit, coef = 4) # fourth coefficient represents the RIN variable
tmp4 <- eBayes(tmp4)
```

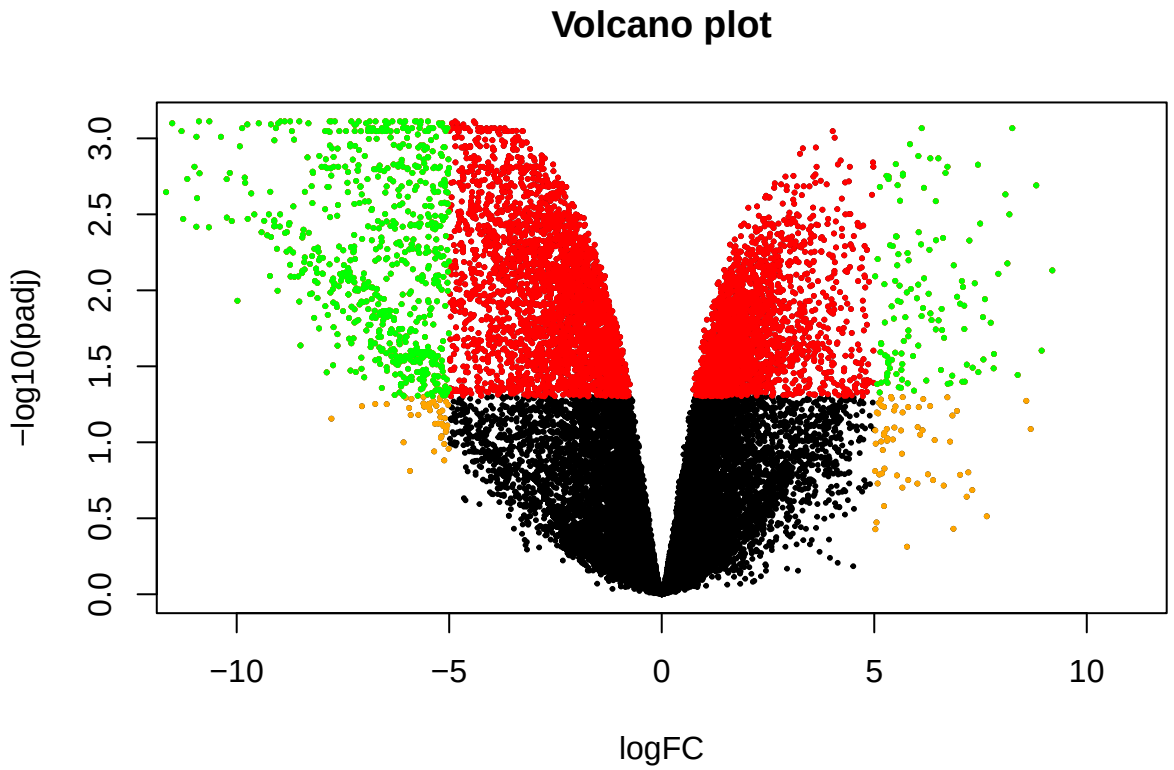
```
par(mfrow=c(1,2))
hist(top.table3$P.Value, breaks=3000, ylim=c(0,50), main = 'MAP', xlab = 'Pvalue')
top.table4 <- topTable(tmp4, sort.by = "P", n = Inf)
hist(top.table4$P.Value, breaks=3000, ylim=c(0,50), main = 'RIN', xlab = 'Pvalue')
```



Annex 6. - histogram of pvalue distribution for MAP and RIN variables.

A volcano plot is made in order to better visualize major patterns among the observed differential expression:

```
res <- data.frame("Geneid" = row.names(top.table),
                  "logFC" = top.table$logFC,
                  "pvalue" = top.table$P.Value,
                  "padj" = top.table$adj.P.Val)

res <- merge(res, data, by.y = 'Geneid')

row.names(res) <- data$Geneid

# Make a basic volcano plot
with(res, plot(logFC, -log10(padj), pch=20, main="Volcano plot", xlim=c(-11,11), cex = 0.4))

# Add colored points: red if padj<0.05, orange of log2FC>1, green if both)
with(subset(res, padj<.05 ), points(logFC, -log10(padj), pch=20, col="red", cex = 0.4))
with(subset(res, abs(logFC)>5), points(logFC, -log10(padj), pch=20, col="orange", cex = 0.4))
with(subset(res, padj<.05 & abs(logFC)>5), points(logFC, -log10(padj), pch=20, col="green", cex = 0.4))
```



Annex 7 - Volcano plot and code

```
# For upregulated genes and H3K4me3 data in fetal samples:
up.fetal <- c()
for (i in 1:length(genes.up)) {
  promoters <- promoters(genes.up[[i]])
  ov <- findOverlaps(promoters, gr1)
  up.fetal <- c(up.fetal,length(unique(queryHits(ov))))
}
# For upregulated genes and H3K4me3 data in adult samples:
up.adult <- c()
for (i in 1:length(genes.up)) {
  promoters <- promoters(genes.up[[i]])
  ov <- findOverlaps(promoters, gr2)
  up.adult <- c(up.adult,length(unique(queryHits(ov))))
}
```

For upregulated genes, 66% peak overlap in adult vs 50% peak overlap in fetal samples.

```
# For downregulated genes and H3K4me3 data in fetal samples:
down.fetal <- c()
for (i in 1:length(genes.down)) {
  promoters <- promoters(genes.down[[i]])
  ov <- findOverlaps(promoters, gr1)
  down.fetal <- c(down.fetal,length(unique(queryHits(ov))))
}
# For downregulated genes and H3K4me3 data in adult samples:
down.adult <- c()
for (i in 1:length(genes.down)) {
  promoters <- promoters(genes.down[[i]])
  ov <- findOverlaps(promoters, gr2)
  down.adult <- c(down.adult,length(unique(queryHits(ov))))
}
```

For downregulated genes, 68% peak overlap in adult vs 35% peak overlap in fetal samples.

```
# For upregulated genes and H3K4me3 data in liver samples:
up.liver <- c()
for (i in 1:length(genes.up)) {
  promoters <- promoters(genes.up[[i]])
  ov <- findOverlaps(promoters, gr3)
  up.liver <- c(up.liver,length(unique(queryHits(ov))))
}
# For downregulated genes and H3K4me3 data in liver samples:
down.liver <- c()
for (i in 1:length(genes.down)) {
  promoters <- promoters(genes.down[[i]])
  ov <- findOverlaps(promoters, gr3)
  down.liver <- c(down.liver,length(unique(queryHits(ov))))
}
```

In liver samples, the upregulated genes overlap a 53% vs 50% of downregulated genes.

```
# Percentage of peaks in a promoter region

# For upregulated genes and H3K4me3 data in fetal samples:
up.fetal.pc <- (1-length(up.fetal[up.fetal==0])/length(up.fetal))
# For upregulated genes and H3K4me3 data in adult samples:
up.adult.pc <- (1-length(up.adult[up.adult==0])/length(up.adult))
```

For upregulated genes, 66% peak overlap in adult vs 51% peak overlap in fetal samples.

```
# Percentage of peaks in a promoter region

# For downregulated genes and H3K4me3 data in fetal samples:
down.fetal.pc <- (1-length(down.fetal[down.fetal==0])/length(down.fetal))
# For downregulated genes and H3K4me3 data in adult samples:
down.adult.pc <- (1-length(down.adult[down.adult==0])/length(down.adult))
```

For downregulated genes, 70% peak overlap in adult vs 35% peak overlap in fetal samples.

```
# Percentage of peaks in a promoter region

# For downregulated genes and H3K4me3 data in fetal samples:
up.liver.pc <- (1-length(up.liver[up.liver==0])/length(up.liver))
# For downregulated genes and H3K4me3 data in adult samples:
down.liver.pc <- (1-length(down.liver[down.liver==0])/length(down.liver))
```

In liver samples, the upregulated genes overlap a 52.6% vs 52.7% of downregulated genes.

```
fetal <- c(percent(up.fetal.pc), percent(down.fetal.pc))
adult <- c(percent(up.adult.pc), percent(down.adult.pc))
```

```
liver <- c(percent(up.liver.pc), percent(down.liver.pc))
dat <- data.frame("fetal brain" = fetal,
                  "adult brain" = adult,
                  "liver" = liver)
row.names(dat)<- c("upregulated", "downregulated")
kable(dat, align = 'c')
```

	fetal.brain	adult.brain	liver
upregulated	52%	66%	52%
downregulated	38%	71%	55%

Annex 8 - Code to calculate percentage of overlap H3K4me3 in promoter genes