# 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)

## Task7: Gene set analysis

For this step, first, informations about Chip-seq analysis for H3K4me3 histon modification and the available samples were retrieved at https://www.ncbi.nlm.nih.gov/geo/roadmap/epigenomics with the search matrix, then the narrow peak data for the chosen samples were accessed via AnnotationHub on R 4.0.2.

For the choice of the most appropriate fetal sample (as the exact fetal brain area investigated in the Jaffe et al. study, called PFC, was not analysed), mean fetal age expressed in years (from re-analysis phenotype table) was converted in pregnancy weeks:

```
40+mean(pdata$Age[pdata$LStage=="Fetal"])*52
```

```
## [1] 17.7284
```

Here are the AnnotationHub titles, with some information:

"BI.Brain_Mid_Frontal_Lobe.H3K4me3.112.narrowPeak.gz" Brodmann area 9/46, dorsolateral prefrontal cortex F Age 75y

"UCSF-UBC.Fetal_Brain.H3K4me3.HuFNSC02.narrowPeak.gz" brain fetal week17 F

"BI.Adult_Liver.H3K4me3.3.narrowPeak.gz" liver F Age NA

GRanges corresponding to UCSC known gene transcripts (that were used in the .gtf annotation file provided to featureCounts) were also retrieved via AnnotationHub. Then, the dataset was subsetted by gene IDs of differentially expressed transcripts obtained in previous Statistical analysis, and a GRanges object with the promoters was created (from 2000bp downstream each transcription start site to 200bp upstream).

```
ucsc<-query(hub,"TxDb.Hsapiens.UCSC.hg19.knownGene.sqlite")
ucsc<-ucsc[[1]]
tx<-transcripts(ucsc)
tx<-tx[tx$tx_name%in%row.names(sign),]
names(tx)<-tx$tx_name
tx<-tx[row.names(sign)]
prom<-promoters(tx)
```

Then, I calculated how many of those 11424 promoters overlap with peaks from ChipSeq in fetal, adult and liver samples:

```
##                                       Number        %
## Promoters with H3K4m3 in fetal brain   7436 65.09104
## Promoters with H3K4m3 in adult brain   7979 69.84419
## Promoters with H3K4m3 in liver         6765 59.21744
```

From this table, we can see that there are differences in the modification rate of the promoters of interest, but they are not large. For example, in liver the % of modified promoters is lower than in brain, but nonetheless remarkable.

Fisher's exact test was performed on an appropiate contingency table to determine if the difference in the number of promoters overlapping adult brain and liver peaks is statistically significant:

```
##             with H3K4m3 without H3K4m3
## Adult brain        7979           3445
## Liver              6765           4659
```

```
##
##  Fisher's Exact Test for Count Data
##
## data:  n
## p-value < 2.2e-16
## alternative hypothesis: true odds ratio is not equal to 1
## 95 percent confidence interval:
##  1.509656 1.685353
## sample estimates:
## odds ratio
##   1.595015
```

The result is definitely significant. That means that the observed distribution is not due to chance and so, to answer the second question in the assignment, even if we can't say that our promoters are not modified in liver, their modification level is lower than in brain cells.

I also used Fisher's test to analyse the difference in histon methylation of DE genes between adult and fetal brain:

```
##             with H3K4m3 without H3K4m3
## Adult brain        7979           3445
## Fetal brain        7436           3988
```

```
fisher.test(o)$p.value
```

```
## [1] 1.915566e-14
```

Once again, the result is significant, so there is a difference in histon methylation of our differentially expressed promoters between fetal and adult brain cell. Nonetheless, this analysis doesn't look very insightful: in fact, the list of promoters includes both genes overexpressed in fetuses (negative log2FC according to my analysis pipeline) and genes overexpressed in adults (log2FC>0), so it doesn't make much sense to analyze them together. Therefore, I divided differentially expressed genes in 2 groups and investigated their H3K4m3 modification as summarized here:

```
##                          Fetal brain peaks Adult brain peaks
## Prom overexpr in fetuses              3671              3644
## Prom overexpr in adults               3765              4335
```

```
fisher.test(p)$p.value
```

```
## [1] 4.560091e-06
```

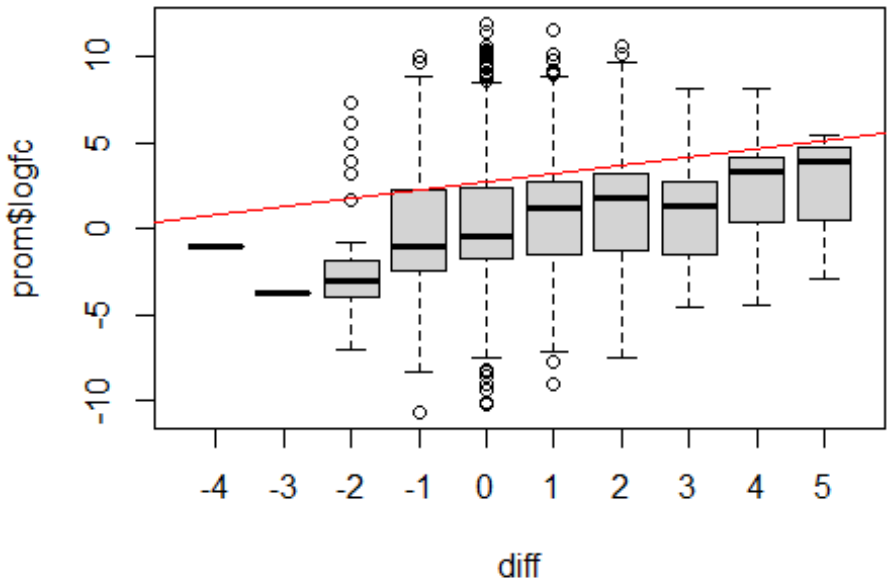
We can see that, as hoped, histons bound to promoters of genes detected as overexpressed in fetuses (according to my Capstone project) tend to be more methylated in fetal brain than in adult brain, and analougously for adult overexpressed genes. The differences are statistically significant, though not big.

So, to answer the first assignment question: yes, differentially expressed genes show changes in H3K4m3 modification between fetal and adult brain, but those changes are quite small and definitely can't explain all the differences in expression levels. This is reasonable, because transcription is regulated by so many elements: other histon modifications, CpG methylation, levels of a variety of transcription factors and so on.

To confirm and complete my conclusions, I decided to test the relationship between the log2FoldChange in expression of DE genes in my analysis and the difference in the number of overlaps, for each of them, with fetal and adult brain peaks. Here is the R code used:

```
t<-table(queryHits(promad)) #counting overlaps with adult peaks for each
gene, from the result of findOverlaps
t<-as.data.frame(t)
ad←numeric(dim(sign)[1]) #sign is a matrix with adj p< 0.05 genes resulting
from DESeq2
for(i in 1:length(ad)) {
if(any(t$Var1==i)) ad[i]<-t[t$Var1==i,2]
}
for(i in 1:length(ad)) {
if(any(t$Var1==i)) ad[i]<-t[t$Var1==i,2]
}
t<-table(queryHits(promfet)) #same procedure for fetal peaks
t<-as.data.frame(t)
fet<-numeric(dim(sign)[1])
for(i in 1:length(fet)) {
if(any(t$Var1==i)) fet[i]<-t[t$Var1==i,2]
}
```

```
diff<-ad-fet
boxplot(prom$logfc~diff)
mod<-lm(prom$logfc~diff)
abline(mod$coefficients,col="red")
```




The boxplot of the log2FC against this difference shows in fact that when the difference is >0 (more overlaps with adult peaks than with fetal ones), the log2FC tends to be >0 too (gene overexpression in adults) and vice-versa. But it also reveals that variability is huge, so the linear model (shown in red) is not reliable.

Indeed, the correlation coefficient between the two variables is quite low:

```
cor(diff,prom$logfc)
```

```
## [1] 0.1192327
```