Genomic Data Science Capstone: Week 9 (Gene Set Analysis)

*For this assignment I performed an analysis of the promoter associated histone modification **H3K4me3** in differential gene expression between and fetal adult brains.*

Analysis steps:

- *First, I obtained each H3K4me3 Chip-seq datasets (in BED format) through query annotations using AnnotationHub (Bioconductor package), for three different samples: (1) fetal brain; (2) adult brain; and (3) adult liver (control).*
- *Then, the peak distances were evaluated for each sample.*
- *Furthermore, the overlap of the peaks between the three datasets were examined.*
- *Lastly, the percentage of promoters with H3K4me3 peaks was calculated for each dataset and compared. Based on the findings from these analyses, the questions listed below were answered.*

Algorithm:

```
## Downloading the Data

library(AnnotationHub)
# assign annotations records to the ah
ah <- AnnotationHub()

## fetal samples
### perform query to ID files corresonding to fetal brain samples and
H3K4me3
fetal_brain <- ah[["AH44720"]]

## adult sample
adult_brain <- ah[["AH43565"]]

## liver sample
liver <- ah[["AH44167"]]
```

```
## Descriptive Statistics of Peak Distances

summary(width(fetal_brain))
summary(width(adult_brain))
summary(width(liver))
```

```
### Venn Diagram for Overlapping Peaks
# finding the overlap in peaks in the overal dataset
library(ChIPpeakAnno)
ol_f_a_l <- findOverlapsOfPeaks(fetal_brain, adult_brain, liver)

makeVennDiagram(ol_f_a_l)
```

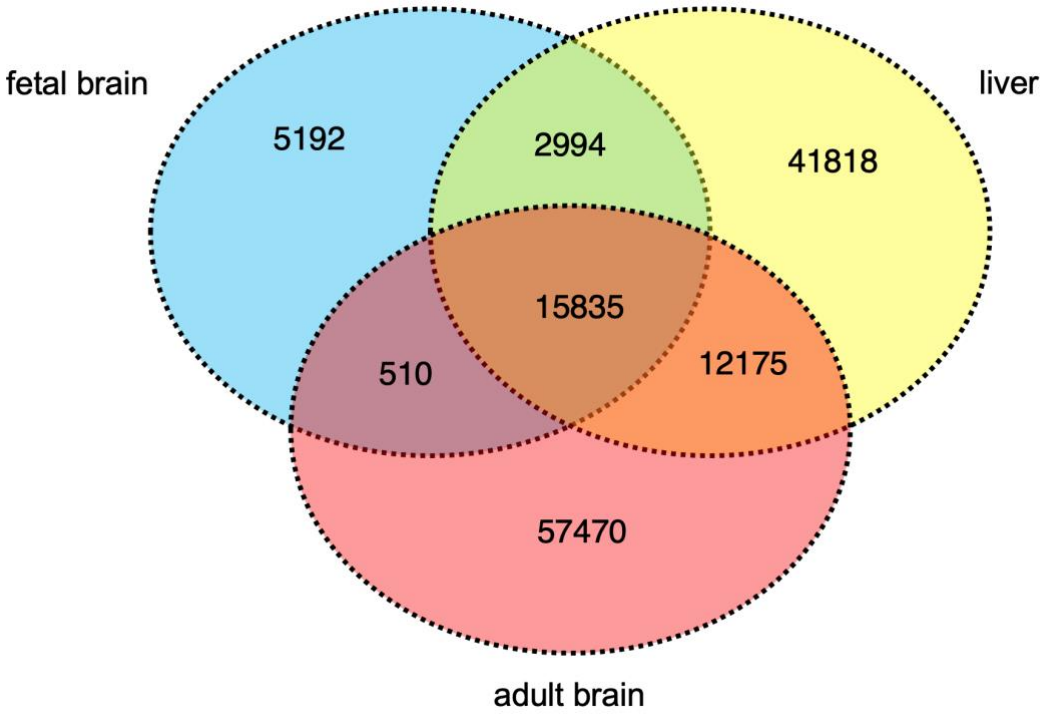
To evaluate the percentage of promoters having H3K4me3 peaks them, the code below was evaluated. First, a reference genome was used (hg19 as used throughout the genomic data science project), and the overlaps of promotes peaks with those corresponding to H3K4me3 determined.

Table 1: Percentge of Promoters with H3K4me3 peaks

```
ref_seq <- query(ah, "RefSeq")
ref_seq_hg19 <- ref_seq[ref_seq$genome == "hg19" & ref_seq$title==
"RefSeq Genes"]

## download the information
ref_seq_hg19 <- ref_seq_hg19[[1]]

## calculate percent of promoters with H3K4me3 peaks
promoters_ref_seq_hg19 <- promoters(ref_seq_hg19)

### find overlaps of fetal brain promoters
fetal_brain_ovlps <- findOverlaps(promoters_ref_seq_hg19, fetal_brain)

### find overlaps of adult brain promoters
adult_brain_ovlps <- findOverlaps(promoters_ref_seq_hg19, adult_brain)

### find overlaps of liver promoters
liver_ovlps <- findOverlaps(promoters_ref_seq_hg19, liver)

### calculate the percentages with promoters with H3K4me3 peaks
fetal_prcnt_pm_pk <- length(unique(subjectHits(fetal_brain_ovlps))) /
length(promoters_ref_seq_hg19)

adult_prcnt_pm_pk <- length(unique(subjectHits(adult_brain_ovlps))) /
length(promoters_ref_seq_hg19)

liver_prcnt_pm_pk <- length(unique(subjectHits(liver_ovlps))) /
length(promoters_ref_seq_hg19)

prcnt_pm_pk_df <- data.frame(fetal_prcnt_pm_pk, adult_prcnt_pm_pk,
liver_prcnt_pm_pk)

names(prcnt_pm_pk_df) <-c("fetal", "adult", "liver")

library(tidyr)
prcnt_pm_pk_df_tidy <- gather(prcnt_pm_pk_df, sample, H3K4me3_pcnt_pks,
fetal:liver)

pander(prcnt_pm_pk_df_tidy)

# obtain gene information from reference genome hg19
ref_seq <- query(ah, "RefSeq")
ref_seq_hg19 <- ref_seq[ref_seq$genome == "hg19" & ref_seq$title==
"RefSeq Genes"]

## download the information
ref_seq_hg19 <- ref_seq_hg19[[1]]

## calculate percent of promoters with H3K4me3 peaks
promoters_ref_seq_hg19 <- promoters(ref_seq_hg19)
```

```
fetal_brain_ovlps <- findOverlaps(promoters_ref_seq_hg19, fetal_brain)

adult_brain_ovlps <- findOverlaps(promoters_ref_seq_hg19, adult_brain)

liver_ovlps <- findOverlaps(promoters_ref_seq_hg19, liver)

fetal_prcnt_pm_pk <- length(unique(subjectHits(fetal_brain_ovlps))) /
length(promoters_ref_seq_hg19)
adult_prcnt_pm_pk <- length(unique(subjectHits(adult_brain_ovlps))) /
length(promoters_ref_seq_hg19)
liver_prcnt_pm_pk <- length(unique(subjectHits(liver_ovlps))) /
length(promoters_ref_seq_hg19)

prcnt_pm_pk_df <- data.frame(fetal_prcnt_pm_pk, adult_prcnt_pm_pk,
liver_prcnt_pm_pk)

names(prcnt_pm_pk_df) <-c("fetal", "adult", "liver")

library(tidyr)
prcnt_pm_pk_df_tidy <- gather(prcnt_pm_pk_df, sample, H3K4me3_pcnt_pks,
fetal:liver)

pander(prcnt_pm_pk_df_tidy)
```

Venn Diagram of the calculation with promoters with H3K4me3 peaks



Conclusions:

Finally, based on the findings from these analyses, the questions listed below were answered:

Questions 1: Are there changes in H3K4me3 between fetal and adult brain over promoters for 2 genes differentially expressed between fetal and adult brain?

Answer: *Yes, there are changes in H3K4me4 between fetal brain and adult brain samples. There is a considerable difference between the means of the peaks' width fetal brain and adult brain samples. This suggests that there are more peaks in the adult samples, and it is consistent with differences in overall sum of peaks, as shown in Ven Diagram, whereas adult brains are nearly 8 times more frequent than fetal brains.*

Furthermore, when considering the promoters with H3K4me4 peaks, adult brain samples had 14% higher proportion compared to fetal samples. Therefore, there are more peaks and a higher proportion of promoters with H3K4me3 in adult brain samples.

Questions 2: Are promoters of genes differentially expressed between adult and fetal brain marked by H3K4me3 in liver?

Answer: *No, changes in H3K4me3 between fetal brain and adult brains is not marked by H3K4me3 in the liver. This is due to the comparable mean widths between fetal brains and adult brain, suggesting that more peaks in the adult brain sample compared to the liver sample.*

Besides, as seen in the Venn Diagram, overall sum of peaks in the adult brain samples are nearly 2 times the sum of peaks in the liver sample. Furthermore, of the total peaks of adult brain, only 13% overlap with peaks in the liver sample, which suggests that overlap of peaks between adult brain and liver is minimal.

Finally, the percentage of peaks with H3K4me3 in the liver is 11% lower than the adult brain sample, while the percentage in the liver is comparable to that of the percentage of fetal sample. This further supports that 3K4me3 in promoters in the adult brain is not marked by 3K4me3 in the liver.