Task 7 – Coursera Genomic Data Science Capstone

After performing expression mapping the RNA samples to human genome version hg19 using *HISAT 2.1* and *featureCounts* 2.0.1 (features were combined using *Multi-Join* 1.1.1) at *Galaxy Server 2.10.8*, data were downloaded and loaded at *R* 4.0.3 for analyses using *Bioconductor* 2.5. The counts data were first filtered to exclude genes with counts smaller than 10 when considering all six samples. To normalize the data, the raw counts values were transformed by *log2* conversion. The counts were then analyzed using a linear model with *limma* 3.46, using the *lmFit* function with group age as predictor. This model was then analyzed using the *eBayes* function to compute several statistics of the differential expression by empirical Bayes moderation of standard errors. Results of the *eBayes* function were then summarized using the *topTable* function to extract the top-ranked genes and perform *p-value* adjustments for multiple testing by the Benjamini & Hochberg (1995) method.

Of the 3119 genes with expression counts after normalization, 2261 showed adjusted *p-values* lower than 0.05, consistent with differential expression. Of these, 1733 were downregulated, and 205 were upregulated. The fetal samples presented more differentially expressed genes, especially downregulated genes, while the adult samples presented more upregulated genes. The *AnnotationHub* package was used to obtain methylation patterns for the H3K4me3 marker between brain of fetal and adult samples, as well as liver methylation (control). Queries were performed using combinations of the key-words “Brain”, “Fetal”, “H3K4me3”, and “Liver”. As there are several samples with markers for these key words, there were selected ones for narrow peaks and with the tag consolidated, which resulted in the selection of data named as AH30471 (fetal), AH30413 (adult), and AH30367 (liver), which were then retrieved from the *AnnotationHub* objects for each search. Overall patters of overlap in methylation for each of the downloaded H3K4me3 datasets were obtained using the *ChIPpeakAnno* package and the function *findOverlapsOfPeaks*, with the resulting object being used as input to the function *makeVennDiagram* to generate the Fig. 1 (left).

The differentially expressed gene names/symbols used in the *limma* analysis were then converted to Entrez gene IDs using the *mygene* library and the *queryMany* function. Using the *TxDb.Hsapiens.UCSC.h19.knowGene* database and its *gene* function the human genes were extracted for annotation of the promoters of the differentially expressed genes using the *promoters* function with default settings. This resulted in a total of 24531 markers for fetal samples, 72822 for adult samples, and 85990 for liver samples. Then the differentially expressed genes/promoters were compared to the H3K4me3 markers obtained for fetal and adult brain tissues, and liver samples using the *subsetByOverlaps* function of the *GenomicRanges* package. Of these, 37 were expressed only in the fetal brain samples, 442 only in the adult brain samples, and 238 only in the liver samples (Fig. 1, right), and a total of 1618 were expressed in all sample types. A hypergeometric test (equivalent to a one-tailed Fischer’s exact test) implemented along with the function *makeVennDiagram*, with 3500 replicates, returned significant p-values for all overlaps found between samples (Table 1). Additionally, a comparison of the counts generated by the *makeVennDiagram* were used to perform a Pearson chi-squared test implemented with the *chisq.test* function, with rescaling of the *p-values*, which were simulated based on 2000 replicates, returning a X^2 of 478.79 and a *p-value* of 0.0009995, further reinforcing the significant differences between the occurrence of H3K4me3 markers in the different groups of samples.

These results suggests that there are relevant changes between the promoters of the differentially expressed genes in the fetal and adult samples, which is show by the expression of some genes uniquely in some sample groups or subgroups. Adult brain tissues showed more overlapping genes/promoters with adult liver tissues than with fetal brain tissues, what demonstrates the effects of age in gene expression. Even so fetal and adult brain tissues shared a number of genes/promoters larger than the ones expressed only in the fetal samples, but more than four times smaller than the ones expressed only in the adult brain tissues. Some promoters of the brain tissues overlapped with the ones from the liver tissues, what is expected, as all tissues share some basic functions that needed to be expressed over all the body.

Table 1. Results of the Hypergeometric test comparing the significance of the overlap between the H3K4me3 markers in the fetal and adult brain samples, and the liver tissue.

Overlap between	**Hypergeometric test p-value**
Fetal x Adult	2.802317e-152
Fetal x Liver	5.288929e-150
Adult x Liver	2.528220e-96

Table 2. Binary table (0 = absent, 1 = present) indicating the overlaps found between the H3K4me3 markers in the fetal and adult brain samples, and the liver tissue, and its total count.

Fetal	**Adult**	**Liver**	**Total counts**
0	0	0	348
0	0	1	238
0	1	0	442
0	1	1	718
1	0	0	37
1	0	1	2
1	1	0	96
1	1	1	1619

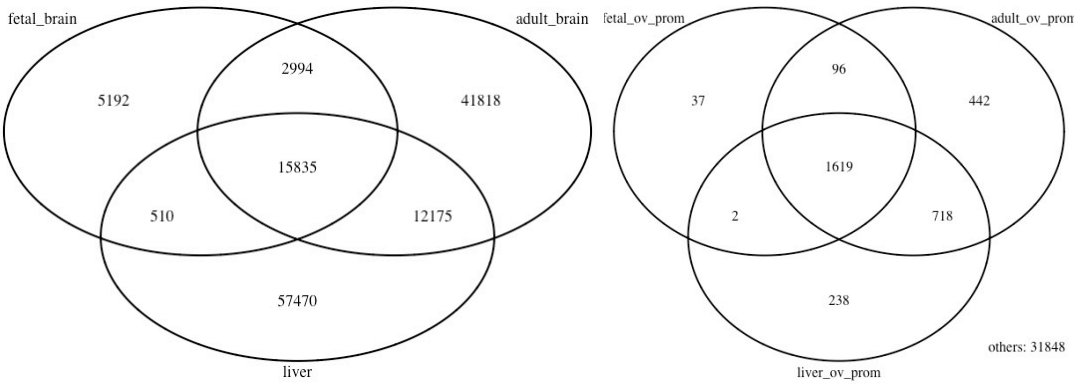

Fig. 1. At right: Venn diagram of the relation between the counts of the H3K4me3 markers in fetal and adult brain samples and liver tissue. At left: Venn diagram of the relation between

the counts of the H3K4me3 markers in the genes and promoters found to be differentially expressed genes in the current analysis.