Genomic Data Science Capstone – Project summary

This report 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 results published in the "Nature Neuroscience (2015, vol 18(1), pages 154-161)", and the published article can be found (http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4281298/). The softwares, parameters, packages, and commands used in each step are described. Briefly, phenotypic metadata and FastQ files from three fetal and three adult brains RNA samples were selected, and their corresponding FastQ files uploaded into the galaxy local host server. Each FastQ file was aligned against a human reference genome (hg38), and quality scores of the reads evaluated.

- **Alignment and Quality control:**
- The alignment was done using HISAT2 (v 2.1.0) with the hg38 human reference genome for which the index was downloaded and built. The samples contained paired and unpaired reads. FASTQC (v 0.11.8) was used to do quality control on the BAM files.
    - Quality scores:

Sample	Runs	Per Base sequence quality	GC content (%)
SRX683794 (Adult)	SRR1554536	32.812	46.5
SRX683797 (Adult)	SRR1554539	34.65495324	48.5
SRX683792 (Adult)	SRR1554534	34.69436006	52
SRX683795 (Fetal)	SRR1554537	31.47136605	49.5
SRX683824 (Fetal)	SRR1554566	32.27080346	48.5
SRX683825 (Fetal)	SRR1554567	32.70195493	47.5

- **Adult:**

Values	Mean	Maximum value	Minimum Value	1st Quartile	Median	3rd Quartile
Per base sequence quality	33.31176603	34.69436006	32.22108614	32.46574408	32.87645	34.66480495
GC content	49	52	46.5	46.875	48.75	51.25

- **Fetal:**

Values	Mean	Maximum value	Minimum Value	1st Quartile	Median	3rd Quartile
Per base sequence quality	32.08882371	32.70195493	31.01888734	31.35824637	32.35163915	32.65358046
GC content	48.41666667	49.5	47	47.375	48.75	49.125

Sample	Run	Total Reads	Alignment Exactly once	Alignment more than once	Alignment zero times	Total Alignment Rate
SRX683794 (Adult)	SRR1554536	4070571 (Unpaired)	3626723 (89.10%)	421962 (10.37%)	21886 (0.54%)	99.46%
		21450348 (Paired)	20218602 (94.26%)	1014202 (4.73%)	217544 (1.01%)	98.99%
SRX683797 (Adult)	SRR1554539	9648705 (Unpaired)	9022038 (93.51%)	560019 (5.80%)	66648 (0.69%)	99.31%
		33742728 (Paired)	31805200 (94.26%)	1445879 (4.29%)	491649 (1.46%)	98.54%
SRX683792 (Adult)	SRR1554534	7208155 (Unpaired)	6639723 (92.11%)	511524 (7.10%)	56908 (0.79%)	99.21%
		28181772 (Paired)	26212392 (93.01%)	1564206 (5.55%)	405174 (1.44%)	98.56%
SRX683795 (Fetal)	SRR1554537	11724434 (Unpaired)	10677707 (91.07%)	963788 (8.22%)	82939 (0.71%)	99.29%
		55133946 (Paired)	51619039 (93.62%)	3133968 (5.68%)	380939 (0.69%)	99.31%
SRX683824 (Fetal)	SRR1554566	9962800 (Unpaired)	9043045 (90.77%)	842515 (8.46%)	77240 (0.78%)	99.22%
		53161501 (Paired)	49591395 (93.28%)	3267149 (6.15%)	302957 (0.57%)	99.43%
SRX683825 (Fetal)	SRR1554567	12466551 (Unpaired)	11423388 (91.63%)	957580 (7.68%)	85583 (0.69%)	99.31%
		61922935 (Paired)	58193574 (93.98%)	3325963 (5.37%)	403398 (0.65%)	99.35%

- **Get Feature counts:**
  The reference gtf file was obtained from UCSC and built into the featureCounts software (v 1.6.4). A tab delimited text file was generated that is formatted with one gene per row and one sample per column. The count for that gene in that sample is in the corresponding cell.

- **Exploratory analysis:**
  I used DESeq2 package to do the exploratory analysis. On the PCA plot, the fetal samples are on the same side so are the adult samples. The adult samples are very distant whereas the fetal samples are grouped together. The count data and phenotype table containing the metadata was uploaded in R as a data-frame. The lowly expressed genes were filtered using the log transform.

```
library("DESeq2")
```

Create a summarized experiment object:

```
ddsMat <- DESeqDataSetFromMatrix(countData = countdata, colData =
coldata,  design = ~ Age + RIN)
rs <- rowSums(counts(dds))
```

Boxplot:

```
boxplot(log2(counts(dds)[rs > 0,] + 1))
```

PCA plot:

```
rld <- rlog(dds, blind=FALSE)
plotPCA(rld, intgroup = c("Age", "RIN")
```

- **Statistical Analysis:**

I used the DESeq2 package to perform the gene expression analysis. For the gene expression statistical analysis, adjusted p-value rather than p-value was used to subset up-regulated and down-regulated genes between fetal brain and adult brains samples. The reason for this is to adjust for multiple testing between the three fetal brain and three adult brains samples. padj < 0.05 and fold change > 1 or < -1 were selected for up-regulation and down-regulation, respectively. 346 genes were up-regulated and 92 were down-regulated.

Get gene expression results:

```
dds <- DESeq(dds)
res <- results(dds)
gene_exp_results_df <- as.data.frame(res)
```

Calculate Differentially expressed genes:

```
sum(gene_exp_results_df$padj < 0.05
&gene_exp_results_df$log2FoldChange > 1, na.rm=TRUE)
```

```
sum(gene_exp_results_df$padj < 0.05
&gene_exp_results_df$log2FoldChange < -1, na.rm=TRUE)
```

Generate Volcano plot:

```
with(gene_exp_results_df, plot(log2FoldChange, -log10(pvalue), pch=20,
main="Volcano plot"))
```

Create dataframe of differentially expressed genes:

```
library (dplyr)
DE_genes < -  gene_exp_results_df %>% filter (padj <0.05 &(log2FoldChange >
1) | log2FoldChange < -1) %>% arrange(padj)
```

- **Gene set analysis:**

  The Bioconductor package AnnotationHub() is used for this gene set analysis. The human hg38 annotation database is downloaded and saved as a TxDB object. The differentially expressed genes from the statistical analysis was loaded and the gene ids were converted to entrez ids using the biomart package. I found that about 15% of differentially expressed genes overlapped with the fetal brain and about 20% of differentially expressed genes are found in adult brain. Changes in H3K4me3 between fetal brain and adult brains is not marked by H3K4me3 in the liver

  Initialize annotation hub and get granges object:

  ```
  library(AnnotationHub)
  library(TxDb.Hsapiens.UCSC.hg38.knownGene)
  ah <- AnnotationHub()
  ah_fetal <- query(ah, c("EpigenomeRoadMap", "H3K4me3", "E081"))
  ah_adult <- query(ah, c("EpigenomeRoadMap", "H3K4me3", "E073"))
  ah_liver <- query(ah, c("EpigenomeRoadMap", "H3K4me3", "Liver"))
  fetal_gr <- ah_fetal[[2]]
  adult_gr <- ah_adult[[2]]
  liver_gr <- ah_liver[[2]]
  ```

  Conversion of gene_ids:

  ```
  library(bioMart)
  mart<- useDataset("hsapiens_gene_ensembl", useMart("ensembl"))
  ```

```
BM <- getBM(filters="refseq_mrna", attributes=c("refseq_mrna",
"entrezgene"), values=res1$gene_id, mart=mart)
```

Initialize txdb object and find overlap:

```
txdb <- TxDb.Hsapiens.UCSC.hg19.knownGene
gx <- genes(txdb)
Deg.gr <- promoters(gx[BM$entrezgene %in% gx$gene_id])
fetal_brain_overlap <- length(subsetByOverlaps(Deg.gr, fetal_gr))/
length(de_genes) *100
adult_brain_overlap <- length(subsetByOverlaps(Deg.gr, adult_gr))/
length(DE_genes) *100
liver_overlap <- length(subsetByOverlaps(Deg.gr,
liver_gr))/length(DE_genes) *100
```

Sites to access the 1) original article, 2) RNA-seq data, 3) phenotype meta-data

- Find in this link the related publication: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4281298/
- Find in this link the RNA-seq data: https://www.dropbox.com/s/m0qgo3mo1jiiyyp/list-of-possible-samples-to-use.pdf?dl=0
- Find in this link the phenotype meta-data for the samples: http://www.ncbi.nlm.nih.gov/bioproject/?term=PRJNA245228