Genomic Capstone Final Report

Dongyuan Wu

2020-06-24

1 Introduction

This report displays the workflow for RNA-seq analysis to evaluate differential gene expression between fetus and adult brains. The detailed background can be found from this paper: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4281298. There are totally 12 samples (6 fetal vs. 6 adult) available. However, due to the resource limitation, I only chose half of them (3 fetal vs. 3 adult) to analyze. I selected 3 samples for fetal group: R3452, R3462, and R3485; and 3 samples for adult group: R2869, R3098, R3467. In addition, there are 2 runs (SRR15545xx and SRR20713xx) for each sample. Only SRR15545xx runs were used. Below is the workflow used.

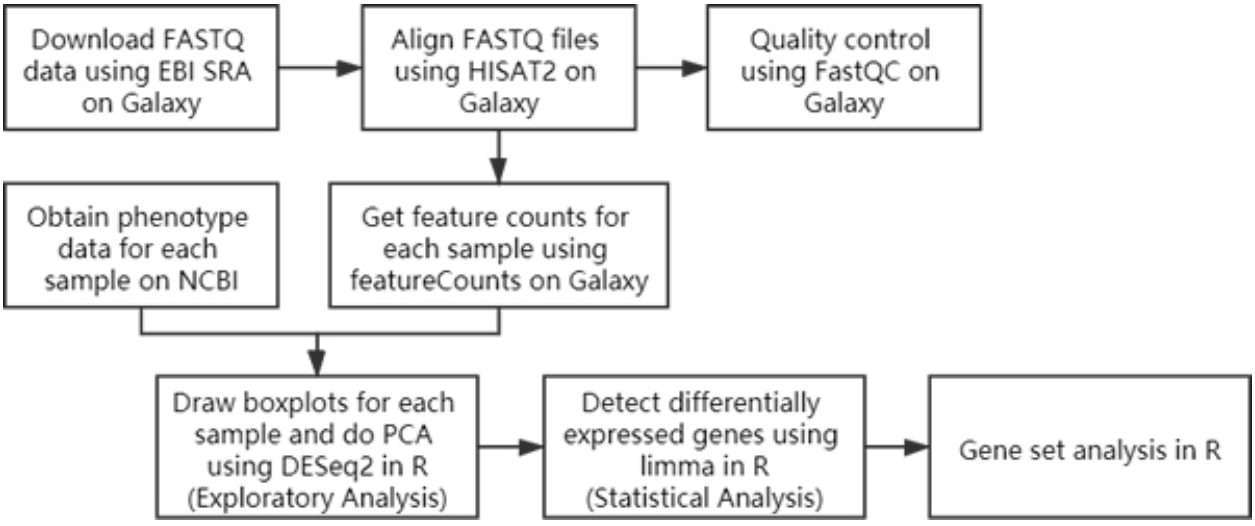


Figure 1: Workflow

Although I really want to display all the code for each step, I have to satisfy the five page limitation. So the codes to organize downloaded data in alignment, quality control, and feature counts will be hidden in R Markdown.

2 Alignment

First of all, I downloaded FASTQ data using EBI SRA on Galaxy. Files are in fastq.gz format and the library layout is paired. Each sample has two files (e.g.: SRR1554537 contains SRR1554537_1 and SRR1554537_2). Then I used HISAT2 on Galaxy to align reads to BAM format. The reference genome was set as `hg19` and the library layout was chosen as `Paired-end`. Other parameters were default. Below is the alignment result.

run	sample	age	sex	RIN	group	total_reads	aligned_reads
SRR1554537	R3452	-0.3836	female	9.6	fetal	55133946	54659776 (99.14%)
SRR1554538	R3462	-0.4027	female	6.4	fetal	68026190	67476885 (99.19%)
SRR1554541	R3485	-0.3836	male	5.7	fetal	69278357	68716869 (99.19%)
SRR1554535	R2869	41.5800	male	8.7	adult	38063721	37530244 (98.6%)
SRR1554536	R3098	44.1700	female	5.3	adult	21450348	21308422 (99.34%)
SRR1554539	R3467	36.5000	female	9.0	adult	33742728	33239440 (98.51%)

3 Quality Control

I used FastQC on Galaxy to do quality control. All the parameters were default. In this way, I got two kinds of results for each sample: a webpage report with summary graphs and a text file with statistical results. Due to page limitation, I directly summarize information from the text file using R (without showing any graphs).

I am not sure what kind of number from report represents *"average quality score of mapped reads"*, so I calculated two numbers:

1. **mean_perbase**: I extracted the `Mean` column from `Per base sequence quality` table, and calculated the avarage of this column.
2. **mean_perseq**: This one is computed from `Per sequence quality scores` table. The rule is: Calculated the sum of `Quality` × `Count`, and then divided by the sum of `Count`.

run	sample	age	sex	RIN	group	perc_align	mean_perbase	mean_perseq
SRR1554537	R3452	-0.3836	female	9.6	fetal	99.14	34.09632	33.72967
SRR1554538	R3462	-0.4027	female	6.4	fetal	99.19	34.79535	34.44037
SRR1554541	R3485	-0.3836	male	5.7	fetal	99.19	33.61498	33.26189
SRR1554535	R2869	41.5800	male	8.7	adult	98.60	33.43514	33.02989
SRR1554536	R3098	44.1700	female	5.3	adult	99.34	34.22886	33.87449
SRR1554539	R3467	36.5000	female	9.0	adult	98.51	35.84102	35.49709

I used two-sided t-test to compare the mapping rates and the average quality score of mapped reads between fetal group and adult group.

Mapping Rates:

```
t.test(dat$perc_align[dat$group == "fetal"], dat$perc_align[dat$group == "adult"])$p.value
```

```
## [1] 0.3076247
```

According to the p-value calculated, we can know that alignment rates for fetal group and adult group are similar under the 0.05 significance level (the p-value is larger than 0.05).

Average Quality Score of Mapped Reads:

```
t.test(dat$mean_perbase[dat$group == "fetal"], dat$mean_perbase[dat$group == "adult"])$p.value
```

```
## [1] 0.7016731
```

```r
t.test(dat$mean_perseq[dat$group == "fetal"], dat$mean_perseq[dat$group == "adult"])$p.value
```

```
## [1] 0.7148898
```

The results show that neither **mean_perbase** nor **mean_perseq** has difference between two groups under the 0.05 significance level (their p-values are all larger than 0.05).

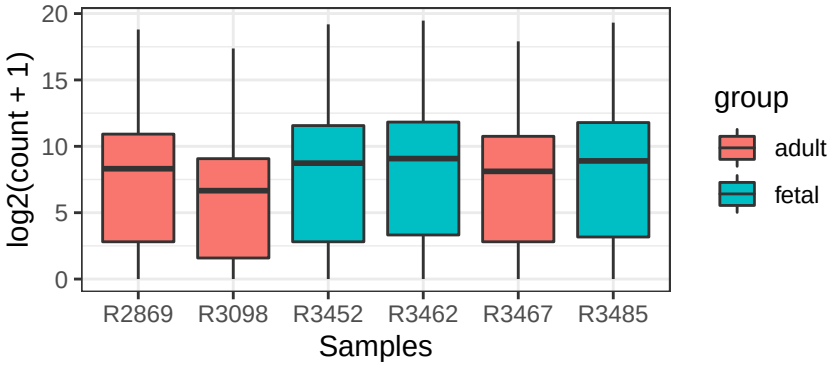
4 Get Feature Count

I used featureCounts on Galaxy to get gene expression from BAM file for each sample. The gene annotation genome was hg19. I found the hg19 GTF file from Galaxy -> Shared Data -> Data Libraries -> iGenomes, and then imported it into History. In this way, we can get a `.tabular` file for one sample (so there are totally 6 files). Then I used R to combine them together.

5 Exploratory Analysis

I used DESeq2 to do the exploratory analysis. When drawing boxplots for each sample, log2 + 1 transformation was used because there are some zeros in raw expression matrix.

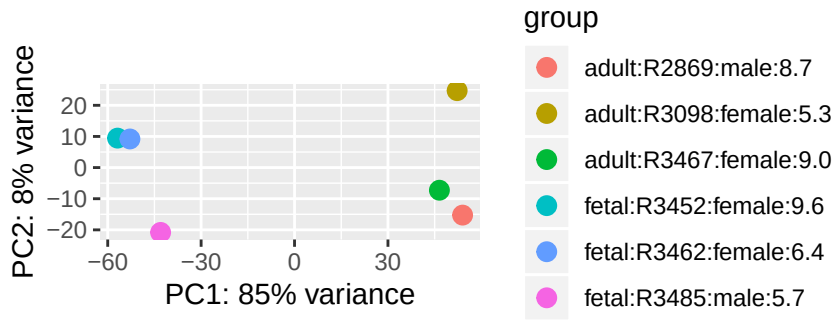
```r
library(DESeq2); library(ggplot2); library(tidyverse)

rawcount <- read.table("featurecounts.txt", header=TRUE, sep="\t")
phenotype <- read.table("phenotype.txt", header=TRUE, sep="\t")
dds <- DESeqDataSetFromMatrix(countData=rawcount, colData=phenotype, design=~group)
log2count <- as.data.frame(log2(counts(dds) + 1))

datforplot <- log2count %>% gather(sample, value, colnames(log2count)[1]:colnames(log2count)[6])
datforplot <- left_join(datforplot, phenotype[, c("sample", "group")], by="sample")
ggplot(data=datforplot) + aes(x=sample, y=value, fill=group) +
  geom_boxplot() + theme_bw() + labs(x="Samples", y="log2(count + 1)")
```



As we can see, the expression levels of fedal group may be a little larger than adult group.

The scatter plot of PCA is displayed below. In this plot, samples of fetal group gather on the left while adult group on the right, which means the value of PC1 in fetal group is lower than in adult group. In addition, male may have lower value of PC2 than female.

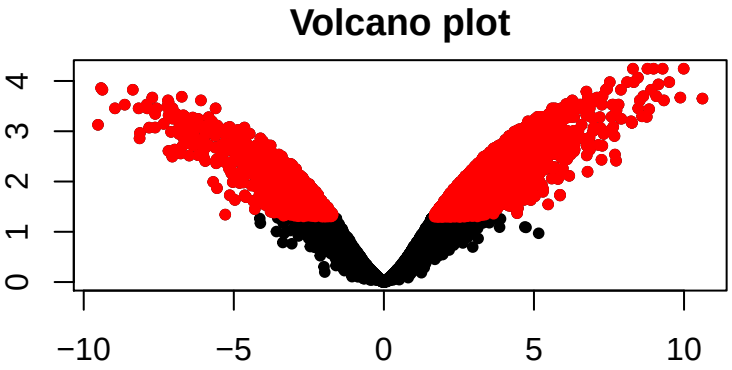
```
plotPCA(rlog(dds, blind=FALSE), intgroup=c("group", "sample", "sex", "RIN"))
```



6 Statistical Analysis

I used limma to do the statistical analysis. Genes with adjusted p-value smaller than 0.05 would be viewed as the differential expressed genes.

```
library(limma)
rawcount <- subset(rawcount, rowSums(rawcount) > 0) # remove genes with zeros in all samples
normcount <- log2(as.matrix(rawcount) + 1) # make log2 + 1 transformation

fit <- lmFit(normcount, model.matrix(~ phenotype$group))
fit <- eBayes(fit)
toptable <- topTable(fit, number=nrow(normcount))
deresult <- toptable[, c(1, 4, 5)]
deresult$gene <- rownames(deresult)
deresult <- deresult[, c(4, 1:3)]

par(mar = c(2,2,2,1))
with(deresult, plot(logFC, -log10(adj.P.Val), pch=20, main="Volcano plot"))
with(subset(deresult, adj.P.Val < 0.05), points(logFC, -log10(adj.P.Val), pch=20, col="red"))
```



The red points in volcano plot represent those differential expressed genes.

```r
sum(deresult$adj.P.Val < 0.05)
```

```
## [1] 7094
```

As we can see, there are totally 7094 differential expressed genes detected.

7 Gene Set Analysis

The fetal brain, adult brain and adult liver datasets from roadmap epigenomics project with narrow peaks are downloaded using AnnotationHub. TxDb.Hsapiens.UCSC.hg19.knownGene is the hg19 annotation database will be used.

```r
library(TxDb.Hsapiens.UCSC.hg19.knownGene); library(AnnotationHub); library(mygene)
degene <- subset(deresult, adj.P.Val < 0.05)
ah <- AnnotationHub()
ah <- subset(ah, species == "Homo sapiens")
fetal <- AnnotationHub::query(ah, c("EpigenomeRoadMap", "H3K4me3", "E081"))[[2]]
adult <- AnnotationHub::query(ah, c("EpigenomeRoadMap", "H3K4me3", "E073"))[[2]]
liver <- AnnotationHub::query(ah, c("EpigenomeRoadMap", "H3K4me3", "E066"))[[2]]
txdb <- TxDb.Hsapiens.UCSC.hg19.knownGene
txdb_gene <- genes(txdb)
```

Gene symbols were used to identify genes in previous tasks, while hg19 database uses gene entrez id. So I used `mygene::queryMany()` to transfer gene symbol into entrez id.

```r
degene_entrez <- queryMany(degene$gene, scopes="symbol", fields="entrezgene", species="human")$entrezge
```

```
## Finished
## Pass returnall=TRUE to return lists of duplicate or missing query terms.
```

```r
promoter <- promoters(txdb_gene[degene_entrez %in% txdb_gene$gene_id])
(fetal_perc <- length(subsetByOverlaps(fetal, promoter)) / length(fetal) * 100)
```

```
## [1] 30.7914
```

```r
(adult_perc <- length(subsetByOverlaps(adult, promoter)) / length(adult) * 100)
```

```
## [1] 20.54879
```

```r
(liver_perc <- length(subsetByOverlaps(liver, promoter)) / length(liver) * 100)
```

```
## [1] 17.07825
```

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

Yes. We can find there are 30.79% of DE genes in fetal brain while 20.55% of DE genes in adult brain, which means there are some genes expressed in fetal brain but not in adult brain.

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

Yes. There are 17.08% of DE genes in liver, which is different from adult and fetal brain.