Genomic Data Science Capstone Report

Ly Le

September 2021

1 Introduction

1.1 Background

Transcriptome analysis of human brain provides fundamental insight about development and disease, but largely relies on existing annotation. We sequenced transcriptomes of 72 prefrontal cortex samples across six life stages, and identified 50,650 differentially expression regions (DERs) associated with developmental and aging, agnostic of annotation. While many DERs annotated to non-exonic sequence (41.1%), most were similarly regulated in cytosolic mRNA extracted from independent samples. The DERs were developmentally conserved across 16 brain regions and within the developing mouse cortex, and were expressed in diverse cell and tissue types. The DERs were further enriched for active chromatin marks and clinical risk for neurodevelopmental disorders like schizophrenia. Lastly, we demonstrate quantitatively that these DERs associate with a changing neuronal phenotype related to differentiation and maturation. These data highlight conserved molecular signatures of transcriptional dynamics across brain development, some potential clinical relevance and the incomplete annotation of the human brain transcriptome.

1.2 Problem

In this project, I determine whether or not the gene expression level is different between fetals and adults. To address this problem, I will re-conduct the analysis which described in this paper.

2 Data

Although there are 48 different samples belonging to 6 different age groups from fetal to old in the original research, I only randomly select 3 samples for each fetal (¡0 years) and adult (20-50 years) group. This is because the data is significantly large and some following steps are extremely time-consuming.
I downloaded these 6 samples from European Nucleotide Archive. Three of them were fetal: SRR1554538, SRR1554566, SRR1554568. The rest three datasets

were adult: SRR1554536, SRR1554556, SRR1554561. Each file was paired-end library, and there were two fastq files for each sample. For example, SRR1554538 contains SRR1554538_1 and SRR1554538_2, each file was in fastq.gz format.

3 Experiment workflow

3.1 Alignment

I made use of galaxy to align all samples to the suitable reference genome. Firstly, all files were uploaded to the server with the following settings: file type was fastqs.gz; reference genome was hg19 (b37). Then, HISAT2 (Version 2.1.0+galaxy5) is chosen to run alignment: the reference genome was built-in genome Human (Homo Sapiens)(b37) hg19; Paired-end two files for each sample. There were two output files: a BAM file contained alignment results; an alignment quality summary file.

3.2 Quality control on the alignments

FastQC (Version 0.72+galaxy1) on galaxy server was applied to perform the quality control. Number of reads were in the range of 21,450,348 to 68,026,190. Number of sequences were in the range of 45,322,851 to 147,128,996. Percentage of GC were in range 46 to 52. All the 6 alignment rates were close to 99.8%, average quality per read was 37, which indicated the alignment results were good and the quality of reads were good.
After, I want to compare the mapping rates between fetal and adult. Firstly, I gathered information for each sample from `https://www.ncbi.nlm.nih.gov/sra?linkname=bioproject_sra_all&from_uid=245228` by clicking "Send to - File - Download Full XML". Then, I converted the xml file into a csv file named "Week5QCdata.csv" by Python programming.

Then, I used R language to further investigate.

```
f = read.csv("Week5QCdata.csv")
phenotype_table = f
rownames(phenotype_table) = phenotype_table[,1]
phenotype_table[,1] = NULL
head(phenotype_table)
```

The code above outputs:

```
                                             Age age.group RIN sex race
SRR1554536 R3098_DLPFC_polyA_RNAseq_total 44.1700     adult 5.3 female   AA
SRR1554556 R3969_DLPFC_polyA_RNAseq_total 36.9800     adult 8.5   male   AA
SRR1554561 R4166_DLPFC_polyA_RNAseq_total 43.8800     adult 8.7   male   AA
SRR1554538 R3462_DLPFC_polyA_RNAseq_total -0.4027     fetal 6.4 female   AA
SRR1554566 R4706_DLPFC_polyA_RNAseq_total -0.4986     fetal 8.3   male HISP
SRR1554568 R4708_DLPFC_polyA_RNAseq_total -0.4986     fetal 8.0   male   AA
                                           Total.sequences alignment.rate
```

```
SRR1554536 R3098_DLPFC_polyA_RNAseq_total    45322851    99.87
SRR1554556 R3969_DLPFC_polyA_RNAseq_total   104578088    99.80
SRR1554561 R4166_DLPFC_polyA_RNAseq_total    83459762    99.70
SRR1554538 R3462_DLPFC_polyA_RNAseq_total   147128996    99.80
SRR1554566 R4706_DLPFC_polyA_RNAseq_total   116523909    99.79
SRR1554568 R4708_DLPFC_polyA_RNAseq_total   104269009    99.78
                                  Average.Quality.per.read X.GC
SRR1554536 R3098_DLPFC_polyA_RNAseq_total                37 46
SRR1554556 R3969_DLPFC_polyA_RNAseq_total                37 48
SRR1554561 R4166_DLPFC_polyA_RNAseq_total                37 52
SRR1554538 R3462_DLPFC_polyA_RNAseq_total                37 47
SRR1554566 R4706_DLPFC_polyA_RNAseq_total                37 49
SRR1554568 R4708_DLPFC_polyA_RNAseq_total                37 47
```

Next, I get an overview of data.

```
write.table(phenotype_table, file="phenotype.txt", col.names=TRUE, row.names=TRUE)
adult = f[1:3,]
fetal = f[4:6,]

summary(adult)

                                          SAMPLE        Age          age.group
 SRR1554536    R3098_DLPFC_polyA_RNAseq_total:1    Min.   :36.98    adult:3
 SRR1554538    R3462_DLPFC_polyA_RNAseq_total:0    1st Qu.:40.43    fetal:0
 SRR1554556    R3969_DLPFC_polyA_RNAseq_total:1    Median :43.88
 SRR1554561    R4166_DLPFC_polyA_RNAseq_total:1    Mean   :41.68
 SRR1554566    R4706_DLPFC_polyA_RNAseq_total:0    3rd Qu.:44.02
 SRR1554568    R4708_DLPFC_polyA_RNAseq_total:0    Max.   :44.17
      RIN           sex         race     Total.sequences      alignment.rate
 Min.   :5.3    female:1    AA  :3    Min.   : 45322851    Min.   :99.70
 1st Qu.:6.9    male  :2    HISP:0    1st Qu.: 64391306    1st Qu.:99.75
 Median :8.5                          Median : 83459762    Median :99.80
 Mean   :7.5                          Mean   : 77786900    Mean   :99.79
 3rd Qu.:8.6                          3rd Qu.: 94018925    3rd Qu.:99.83
 Max.   :8.7                          Max.   :104578088    Max.   :99.87
 Average.Quality.per.read      X.GC
 Min.   :37                Min.   :46.00
 1st Qu.:37                1st Qu.:47.00
 Median :37                Median :48.00
 Mean   :37                Mean   :48.67
 3rd Qu.:37                3rd Qu.:50.00
 Max.   :37                Max.   :52.00

summary(fetal)

                                          SAMPLE        Age          age.group
 SRR1554536    R3098_DLPFC_polyA_RNAseq_total:0    Min.   :-0.4986    adult:0
```

3

```
SRR1554538    R3462_DLPFC_polyA_RNAseq_total:1    1st Qu.:-0.4986    fetal:3
SRR1554556    R3969_DLPFC_polyA_RNAseq_total:0    Median :-0.4986
SRR1554561    R4166_DLPFC_polyA_RNAseq_total:0    Mean   :-0.4666
SRR1554566    R4706_DLPFC_polyA_RNAseq_total:1    3rd Qu.:-0.4506
SRR1554568    R4708_DLPFC_polyA_RNAseq_total:1    Max.   :-0.4027
     RIN             sex      race    Total.sequences     alignment.rate
Min.   :6.400    female:1   AA  :2   Min.   :104269009   Min.   :99.78
1st Qu.:7.200    male  :2   HISP:1   1st Qu.:110396459   1st Qu.:99.78
Median :8.000                        Median :116523909   Median :99.79
Mean   :7.567                        Mean   :122640638   Mean   :99.79
3rd Qu.:8.150                        3rd Qu.:131826452   3rd Qu.:99.80
Max.   :8.300                        Max.   :147128996   Max.   :99.80
Average.Quality.per.read      X.GC
Min.   :37              Min.   :47.00
1st Qu.:37              1st Qu.:47.00
Median :37              Median :47.00
Mean   :37              Mean   :47.67
3rd Qu.:37              3rd Qu.:48.00
Max.   :37              Max.   :49.00
```

To determine whether the mapping rates and average quality score were different between adult and fetal group, I performed the student's t-test:

```
t.test(fetal$alignment.rate, adult$alignment.rate)

    Welch Two Sample t-test

data:  fetal$alignment.rate and adult$alignment.rate
t = 0, df = 2.0548, p-value = 1
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.2083256  0.2083256
sample estimates:
mean of x mean of y
    99.79     99.79

t.test(fetal$Average.Quality.per.read, adult$Average.Quality.per.read)

    Welch Two Sample t-test

data:  fetal$Average.Quality.per.read and adult$Average.Quality.per.read
t = 0, df = 2, p-value = 1
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.2083256  0.2083256
sample estimates:
mean of x mean of y
    37       37
```

The p-values were 1 and 1 for mapping rate and average quality score between the two groups respectively, which indicates there was no significant different between the two groups.

3.3 Get feature counts

To calculate the abundance of every gene in every sample, I used featureCounts (Version 1.6.4+galaxy1) on galaxy server, the gene annotation genome was hg19. The results were tables that was formatted with one gene per row with corresponding counts. After performed on each sample, I merged the all 6 tables into one table by their Geneid, and converted them to gene names.

```
library('tidyverse',quietly=TRUE)
library(org.Hs.eg.db,quietly=TRUE)
library(annotate,quietly=TRUE)
# read feature count files
tabular_files = list.files(path = "./Data/FeatureCount", pattern = "tabular\$", full.names =
tabular_list = lapply(tabular_files, read.table)

header.true <- function(df) {
  names(df) <- as.character(unlist(df[1,]))
  df[-1,]
}
tabular_list = lapply(tabular_list,header.true)

# merge the files by Geneid
feature_count_files = Reduce(function(x, y) merge(x, y, by="Geneid"), tabular_list)

# convert Geneid to gene_name
for (i in 1:nrow(feature_count_files)){
  feature_count_files[i,1] = lookUp(toString(feature_count_files[i,1]), 'org.Hs.eg', 'SYMBOL
}
rownames(feature_count_files) = make.names(feature_count_files[,1], unique=TRUE)
feature_count_files[,1] = NULL
feature_table = feature_count_files

head(feature_table)
```

	SRR1554536	SRR1554538	SRR1554556	SRR1554561	SRR1554566	SRR1554568
A1BG	48	136	328	315	304	149
X10	0	6	2	0	4	2
X100	135	189	619	113	186	92
X1000	1483	22765	6370	5105	24421	20543
X10000	457	24149	5411	6533	36168	44187
X100008586	0	0	0	0	0	0

```
write.table(feature_table, file="./feature_counts.txt", sep='\t', row.names=TRUE, col.names=
```

The whole workflow in my galaxy is here.

3.4   Exploratory analysis

I will use R to make analysis.

```
library(GenomicRanges)
library(SummarizedExperiment)
library(edgeR)
library(ggplot2)
feature_table = read.table(file="./feature_counts.txt", sep='\t', row.names=TRUE, col.names=
# remove low expression data
feature_table = feature_table[rowMeans(feature_table) > 10, ]

# create a SummarizedExperiment data
col_data = phenotype_table
row_data = relist(GRanges(), vector("list", length=nrow(feature_table)))
se = SummarizedExperiment(assays = list(counts = feature_table), rowRanges = row_data, colDa
print(se)

class: RangedSummarizedExperiment
dim: 18660 6
metadata(0):
assays(1): counts
rownames(18660): X100 X1000 ... X9994 X9997
rowData names(0):
colnames(6): SRR1554536 R3098_DLPFC_polyA_RNAseq_total SRR1554556
  R3969_DLPFC_polyA_RNAseq_total ... SRR1554566
  R4706_DLPFC_polyA_RNAseq_total SRR1554568
  R4708_DLPFC_polyA_RNAseq_total
colData names(9): Age age.group ... Average.Quality.per.read X.GC

# make a boxplot of the expression levels for each sample
dge <- DGEList(counts = assay(se, "counts"), group = phenotype_table$age.group )
dge$samples <- merge(dge$samples, as.data.frame(colData(se)), by = 0)
png("dgecount.png", width = 350, height = 350)
boxplot(dge$counts)
dev.off()
```
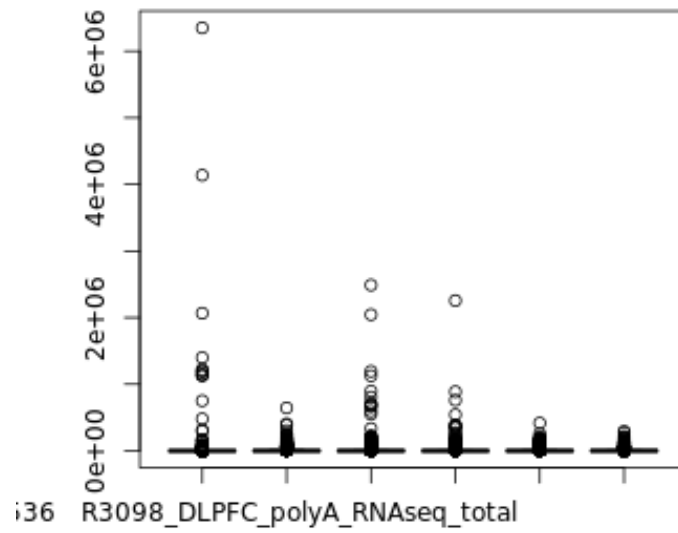
Figure 1: Boxplot of the expression levels for each sample

Most of the data push to the bottom in the boxplot, so I perform log2 transformation on the data.

```
log2_dge_count = log2(dge$counts + 1)
boxplot(log2_dge_count)
```
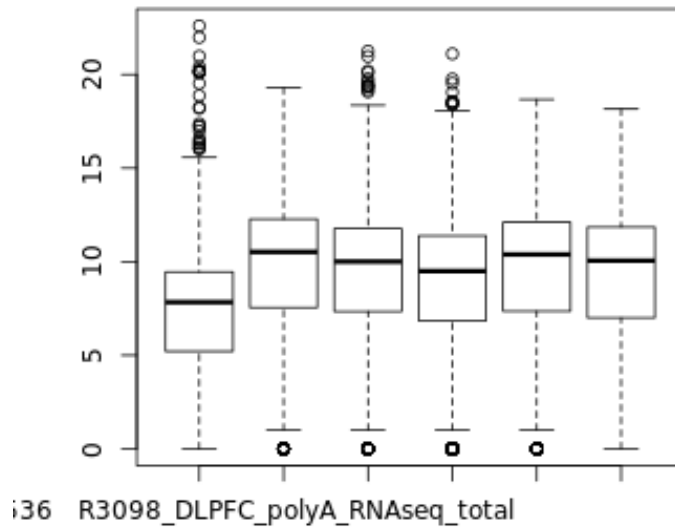
Figure 2: Boxplot of the expression levels for each sample with log

Now the boxplot looked much better. It seems many outliers with extremly high expression in adult data but not in fetal data.

Next I performed a principal component analysis.

```
library(ggfortify)
# perform PCA
count_pca = prcomp(log2_dge_count, center=TRUE, scale=TRUE)
dat = data.frame(X=count_pca$rotation[,1], Y=count_pca$rotation[,2], age_group=phenotype_tab

# scatterplot using PC1 and PC2, colored by RIN, shaped by age.group
ggplot(dat, aes(x=X, y=Y, shape=age_group, color=RIN)) + geom_point(size=5) + xlab("PC1") +
```
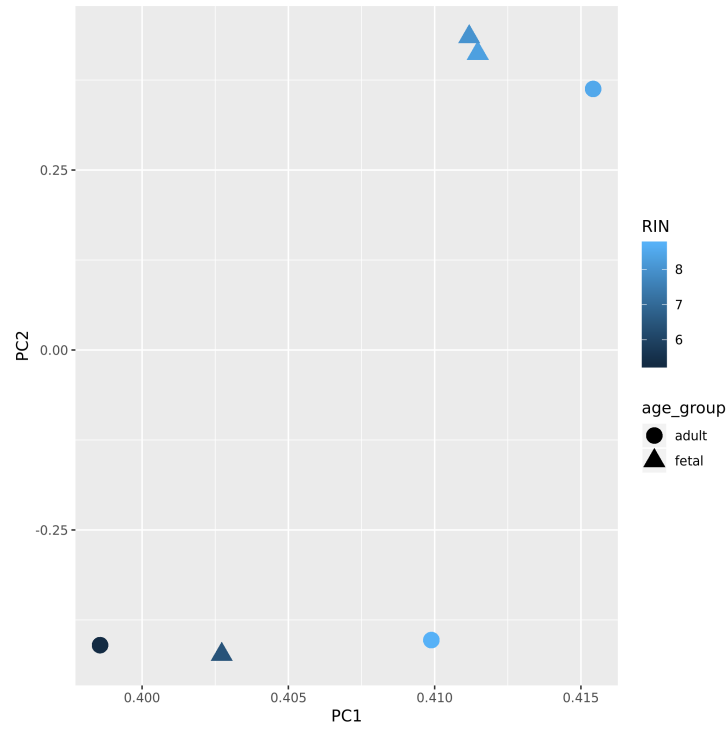
8

Figure 3:

Adult gene expression and fetal gene expression data were hardly differenti-
ate by PC1 and PC2. If we only use RIN, we also cannot distinguish adult and
fetal group.

3.5 Statistical analysis

Next I will perform a statistical analysis to detect genes that are differentially expressed. The null and alternative hypothesis is:

H0: the mean gene expression level of each gene is equal between adult and fetal samples.

H1: the mean gene expression level of each gene is not equal between adult and fetal samples. To test the hypothesis, I used limma package to make linear models for the assessment of differential expression.

```
library(limma)
library(edge)

# make log2 transformation and remove low expression
edata = assay(se)
edata = log2(as.matrix(edata) + 1)
edata = edata[rowMeans(edata) > 10, ]

# fit the model by age.group and write results to a tab-delimited file with gene name, log2
mod = model.matrix(~ se\$age.group)
fit_limma = lmFit(edata,mod)
ebayes_limma = eBayes(fit_limma)
limma_toptable = topTable(ebayes_limma,number=dim(edata)[1])
limma_table_output = limma_toptable[,c(1,4,5)]
write.table(limma_table_output, file="dif_exp_genes.txt", sep='\t', row.names=TRUE, col.name

head(limma_table_output)

Removing intercept from test coefficients
           logFC      P.Value adj.P.Val
X361    -4.886335 4.046429e-05 0.3459697
X1410   -4.266573 3.380089e-04 0.9784850
X4155   -4.163006 4.700087e-04 0.9784850
X10716   3.556653 2.808341e-03 0.9784850
X4645    3.550112 2.859240e-03 0.9784850
X2246   -3.544767 2.901458e-03 0.9784850
```

Then I made a volano plot, which is a plot of the fold change for age in each linear model versus log10 p-value.

```
# make the volcano plot, mark those gene with p-value less than 0.05 as red
with(limma_toptable, plot(logFC, -log10(adj.P.Val), pch=20, main="Volcano plot"))
with(subset(limma_toptable, adj.P.Val < 0.05), points(logFC, -log10(adj.P.Val), pch=20, col=
```
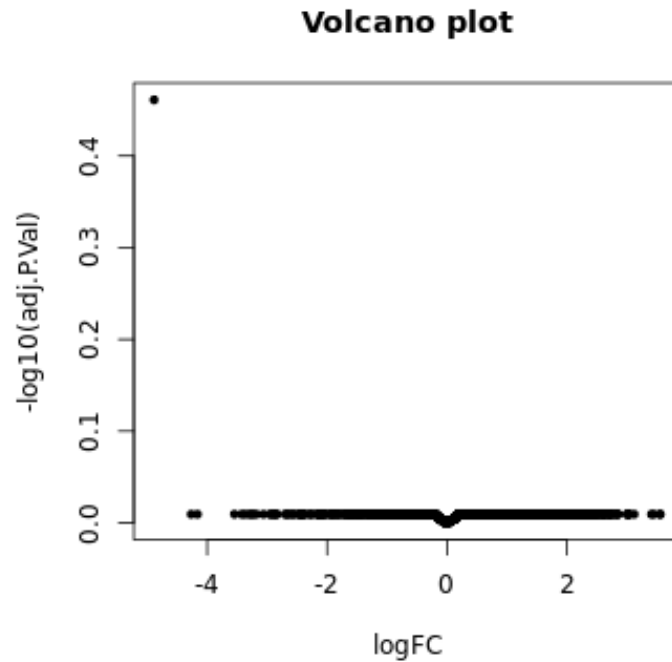
Volcano plot

Figure 4:

Those genes show in red dots were considered as differentially expressed.
Next I summarized those genes:

```
print(paste0("Genes differentially expressed: ", sum(limma_table_output$adj.P.Val < 0.05)))
print(paste0("Genes differentially expressed and down-regulated from fetal to adult: ", sum(
print(paste0("Genes differentially expressed and up-regulated from fetal to adult: ", sum(li

[1] "Genes differentially expressed: 0"
[1] "Genes differentially expressed and down-regulated from fetal to adult: 0"
[1] "Genes differentially expressed and up-regulated from fetal to adult: 0"
```

4   SessionInfo

5   Reference

http://binf.gmu.edu/swang36/NGS/Genomic_Capstone_Report.html