Assignment6

Sonali Joshi

4/22/2017

Exploratory Analysis

The Counts data and Phenotype data were loaded into R for analysis with Bioconductor. Rows with very low counts were filtered out to reduce the size of the counts data.

```
CountsData<-(read.delim('FeatureCountsData.txt', header = TRUE, row.names="Genes"))
CountsData = CountsData[rowMeans(CountsData)>5,]
CountsData = as.matrix(CountsData)
head(CountsData,5)
```

```
##        SRR1554537 SRR1554538 SRR1554541 SRR1554535 SRR1554536 SRR1554539
## 1MAR1         674       1520       1223        705        322        460
## 1MAR11       3250       5152       5550       1608        216       1934
## 2MAR1         523        947        418        846        277        828
## 2MAR2        1075        837       1058       3294       1582       2662
## 3MAR          469       1208        847        145         63         48
```

```
PhenoData<-as.matrix(read.delim('Sample_phenotypes.txt',header=TRUE))
sample_data <- as.matrix(PhenoData)
head(PhenoData)
```

```
##      Biosample      SRA         Run           Sex      Age       Race
## [1,] "SAMN02999520" "SRS686965" "SRR1554537" "Female" "-0.3836" "AA"
## [2,] "SAMN02999521" "SRS686966" "SRR1554538" "Female" "-0.4027" "AA"
## [3,] "SAMN02999524" "SRS686969" "SRR1554541" "Male"   "-0.3836" "AA"
## [4,] "SAMN02999518" "SRS686963" "SRR1554535" "Male"   "41.5800" "AA"
## [5,] "SAMN02999519" "SRS686964" "SRR1554536" "Female" "44.1700" "AA"
## [6,] "SAMN02999522" "SRS686967" "SRR1554539" "Female" "36.5000" "AA"
##      Tissue  Disease   RIN   ExpDesign
## [1,] "DLPFC" "Control" "9.6" "Fetal"
## [2,] "DLPFC" "Control" "6.4" "Fetal"
## [3,] "DLPFC" "Control" "5.7" "Fetal"
## [4,] "DLPFC" "Control" "8.7" "Adult"
## [5,] "DLPFC" "Control" "5.3" "Adult"
## [6,] "DLPFC" "Control" "9.0" "Adult"
```

The PCA done during exploratory analysis showed clustering by age. Hence, the hypothesis is that the genes are differentially expresses by age. The null hypothesis is that there is no difference in gene expression between the fetal and adult brain tissue.

Create DESeqDataSet object

The package DESeq2 was used for further analysis. A column called ExpDesign was added to the Phenotype table to label the fetal and adult samples. This is used by DESeq2 package to specify the experimental design for analysis. All samples have the same race, while there are four female and two male samples. Sex was adjusted as a covariate in the analysis.

Construct the DESeqDataSet object from the Counts data and the Phenotype data

```
data_dds <- DESeqDataSetFromMatrix(CountsData, PhenoData, ~ExpDesign + Sex)
head(data_dds)
```

```
## class: DESeqDataSet
## dim: 6 6
## metadata(1): version
## assays(1): counts
## rownames(6): 1MAR1 1MAR11 ... 3MAR 4MAR
## rowData names(0):
## colnames(6): SRR1554537 SRR1554538 ... SRR1554536 SRR1554539
## colData names(10): Biosample SRA ... RIN ExpDesign
```

Significance testing

```
data_ddsSE <- DESeq(data_dds)
```

```
## estimating size factors
## estimating dispersions
## gene-wise dispersion estimates
## mean-dispersion relationship
## final dispersion estimates
## fitting model and testing
```

```
head(data_ddsSE)
```

```
## class: DESeqDataSet
## dim: 6 6
## metadata(1): version
## assays(3): counts mu cooks
## rownames(6): 1MAR1 1MAR11 ... 3MAR 4MAR
## rowData names(25): baseMean baseVar ... deviance maxCooks
## colnames(6): SRR1554537 SRR1554538 ... SRR1554536 SRR1554539
## colData names(11): Biosample SRA ... ExpDesign sizeFactor
```

Direction of the fold change

```
data_results <- results(data_ddsSE,contrast=c("ExpDesign","Fetal","Adult"), alpha = 0.05)
mcols(data_results, use.names = T)
```

```
## DataFrame with 6 rows and 2 columns
##                         type
##                  <character>
## baseMean       intermediate
## log2FoldChange      results
## lfcSE               results
## stat                results
## pvalue              results
## padj                results
##                                                 description
##                                                 <character>
```

```
## baseMean              mean of normalized counts for all samples
## log2FoldChange log2 fold change (MLE): ExpDesign Fetal vs Adult
## lfcSE                  standard error: ExpDesign Fetal vs Adult
## stat                   Wald statistic: ExpDesign Fetal vs Adult
## pvalue              Wald test p-value: ExpDesign Fetal vs Adult
## padj                                     BH adjusted p-values
```

```
summary(data_results)
```

```
##
## out of 18091 with nonzero total read count
## adjusted p-value < 0.05
## LFC > 0 (up)     : 3749, 21%
## LFC < 0 (down)   : 4451, 25%
## outliers [1]     : 0, 0%
## low counts [2]   : 0, 0%
## (mean count < 3)
## [1] see 'cooksCutoff' argument of ?results
## [2] see 'independentFiltering' argument of ?results
```

Dataframe of p-values and fold change

Use the adjusted p-value as it is corrected for multiple comparisons. As the values are small take the -log10 of the adjust p-value to better visualize the magnitude.

```
data_values <- data.frame(gene = row.names(CountsData),
                          pvalue = data_results$pvalue,
                          padj = data_results$padj,
                          log10_adj_pvalue = -log10(data_results$padj),
                          logfc = data_results$log2FoldChange)

data_values <- na.omit(data_values)

sorted_data_values <- data_values[order(-data_values$log10_adj_pvalue),]
head(sorted_data_values,10)
```

```
##          gene        pvalue          padj log10_adj_pvalue      logfc
## 14854   SOX11 2.335330e-159 4.224846e-155        154.37419   8.488733
## 14114     SLA 1.897727e-113 1.716589e-109        108.76533   6.652278
## 5323     FBN3 4.754483e-101  2.867112e-97         96.54256   5.928111
## 11098  OPALIN 7.639214e-101  3.455025e-97         96.46155 -10.849468
## 14573    SNCG  4.310038e-98  1.559458e-94         93.80703  -7.798304
## 15131 ST8SIA2  5.141463e-87  1.550237e-83         82.80960   7.609435
## 4835     ERMN  4.365789e-83  1.128307e-79         78.94757  -9.218887
## 16939   VASH2  6.710405e-81  1.517474e-77         76.81888   5.764825
## 3455    CORO6  9.104068e-81  1.830019e-77         76.73754  -6.512475
## 7093    IGSF9  2.522837e-79  4.564065e-76         75.34065   7.488609
```

```
#generate the required tab delimited file
write.table(sorted_data_values, file = "sorted_data_values.txt", sep = "\t",
            row.names = F)

#Find number of differentially expressed genes with adjusted p value < 0.001
num_diffex_genes <- subset(sorted_data_values, padj <= 0.001)
dim(sorted_data_values)
```

```
## [1] 18091     5
```

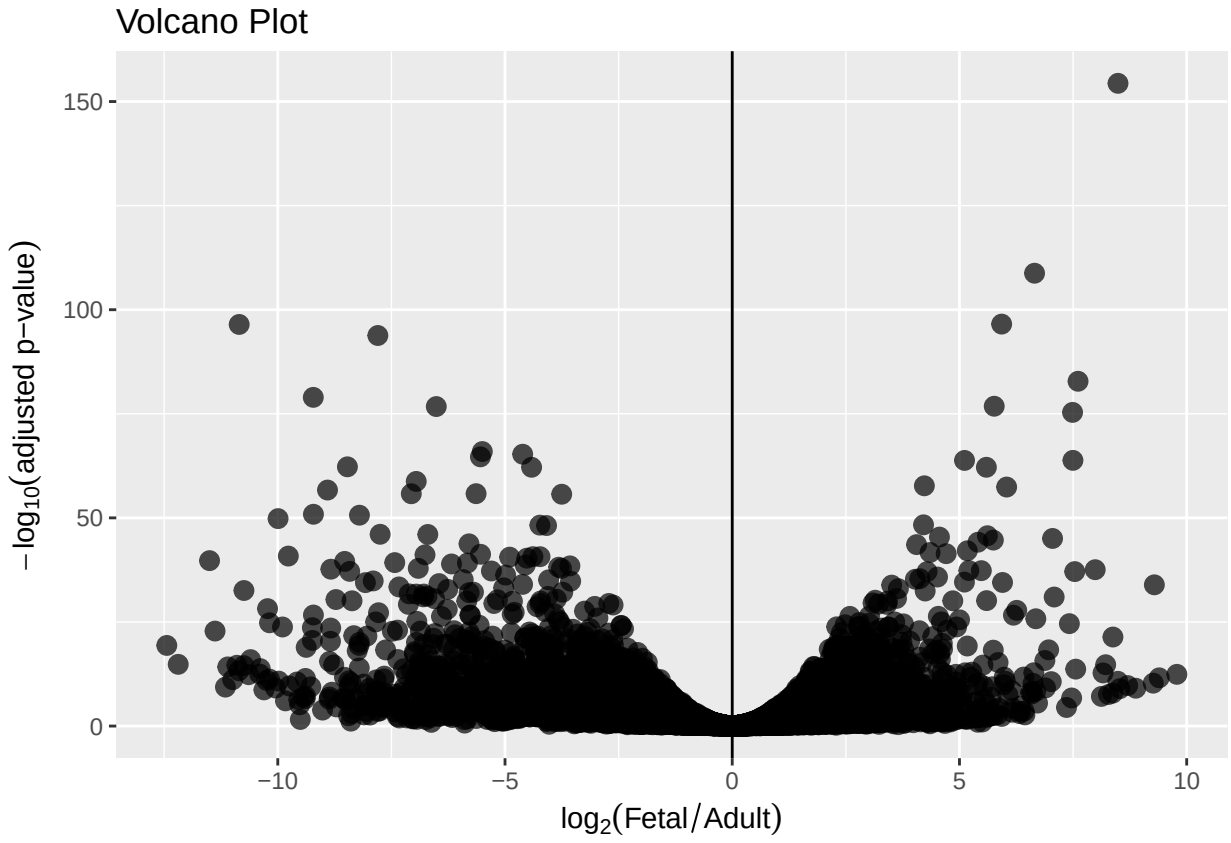
```
dim(num_diffex_genes)
```

```
## [1] 4482    5
```

Make a Volcano plot

```
Vplot1 <- ggplot(data_values, aes(x=logfc, y = log10_adj_pvalue)) +
  geom_point(size = 3, alpha = 0.7, na.rm = T) +
  ggtitle(label = "Volcano Plot") + # Add a title
  xlab(expression(log[2]("Fetal" / "Adult"))) + # x-axis label
  ylab(expression(-log[10]("adjusted p-value"))) + # y-axis label
  geom_vline(xintercept = 0, colour = "black")# + # Add 0 lines

Vplot1
```




```
#Genes up regulated
up_genes <- data_values %>% filter(logfc > 1) %>% arrange (padj)
head(up_genes)
```

```
##      gene        pvalue          padj log10_adj_pvalue    logfc
## 1   SOX11 2.335330e-159 4.224846e-155        154.37419 8.488733
## 2     SLA 1.897727e-113 1.716589e-109        108.76533 6.652278
## 3    FBN3 4.754483e-101  2.867112e-97         96.54256 5.928111
## 4 ST8SIA2  5.141463e-87  1.550237e-83         82.80960 7.609435
## 5   VASH2  6.710405e-81  1.517474e-77         76.81888 5.764825
```

```
## 6   IGSF9  2.522837e-79  4.564065e-76         75.34065 7.488609
```

```
up_gene_list <- (as.matrix(up_genes$gene))

#Genes down regulated
down_genes <- data_values %>% filter(logfc < -1 ) %>% arrange (padj)
down_gene_list <- (as.matrix(down_genes$gene))
head(down_genes)
```

```
##     gene        pvalue         padj log10_adj_pvalue      logfc
## 1 OPALIN 7.639214e-101 3.455025e-97         96.46155 -10.849468
## 2   SNCG  4.310038e-98 1.559458e-94         93.80703  -7.798304
## 3   ERMN  4.365789e-83 1.128307e-79         78.94757  -9.218887
## 4  CORO6  9.104068e-81 1.830019e-77         76.73754  -6.512475
## 5   LDB3  7.104478e-70 1.168428e-66         65.93240  -5.499103
## 6  SIRPA  3.440494e-69 5.186832e-66         65.28510  -4.610796
```