Sonali Joshi - Assignment 5

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)
CountsData = CountsData[rowMeans(CountsData[,2:7])>5,]
head(CountsData,5)
```

```
##           Genes SRR1554537 SRR1554538 SRR1554541 SRR1554535 SRR1554536
## 2        WASH7P       1711        950       1230        849        257
## 6     LOC729737         25        280         62        134          2
## 9  LOC100133331        454        602        283        137         22
## 13 LOC100288069        127        277        144         27          4
## 14    LINC00115        261        459        212         94         12
##    SRR1554539
## 2         486
## 6         181
## 9         135
## 13         52
## 14         55
```

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

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

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. Construct the DESeqDataSet object from the Counts data and the phenotype data

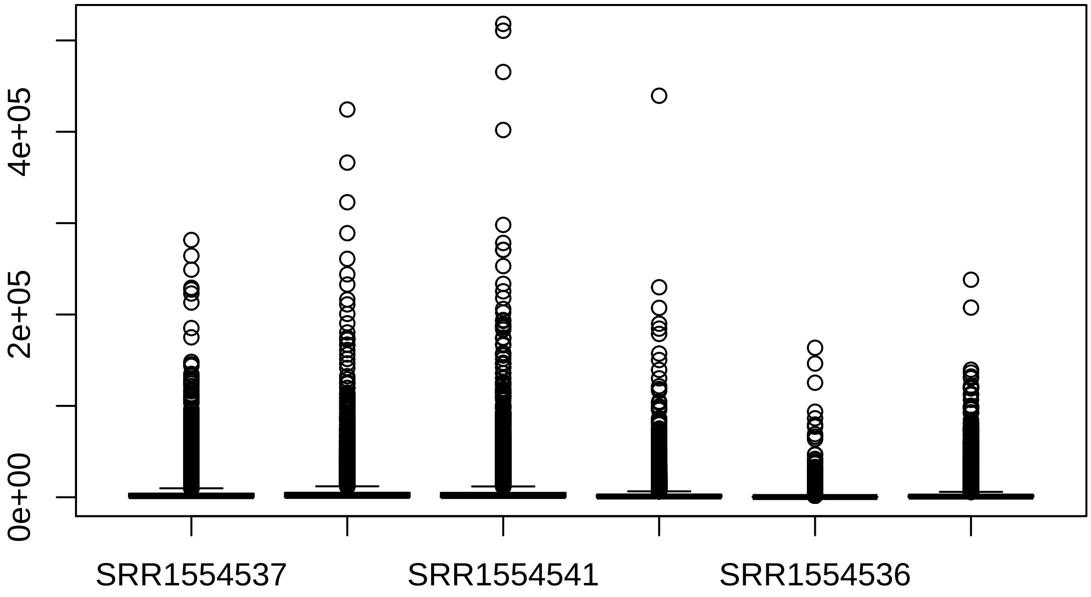
```
data_dds <- DESeqDataSetFromMatrix(CountsData[,2:7], PhenoData, ~ExpDesign)
```

The DESeq2 package recommends the use of raw data without normalizing for sequencing depth as it accounts for library size differences internally.

Visualize data

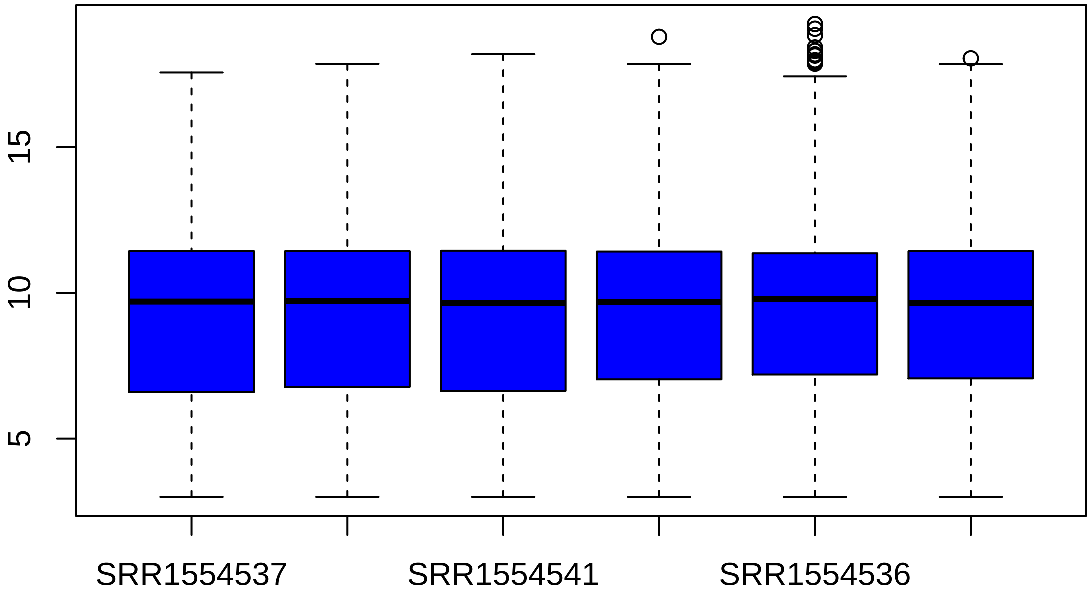
The boxplot on raw data indicates that data transformation is needed before PCA.

```
boxplot(counts(data_dds))
```



Data Transformation DESeq2 offers two transformations for count data to stabilize variance. Transform the data using the VST - Variance Stabilizing Transformation and plot the data.

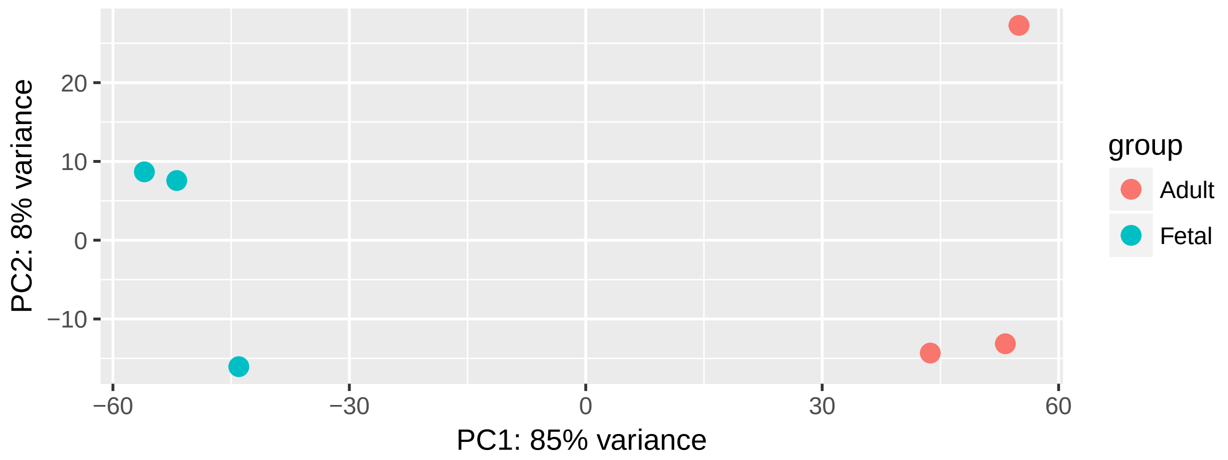
```
vdata_dds <- vst(data_dds, blind = FALSE)
boxplot(vst(counts(data_dds)), col="blue")
```



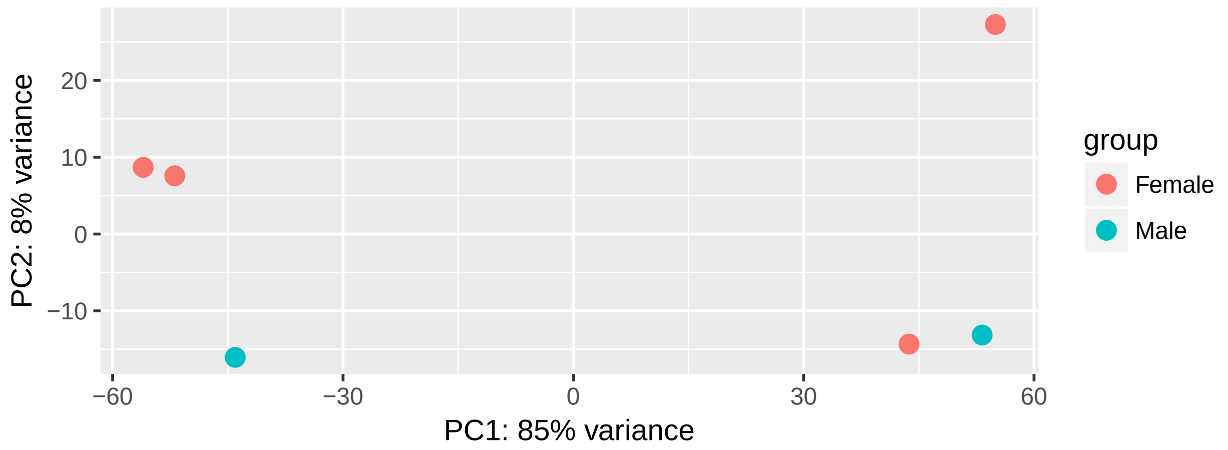
PCA

Explore the data further by doing PCA on the data, to check if the age (fetal, adults) or sex have correlations with the principal components.

```r
plotPCA(vdata_dds, c("ExpDesign"))
```



```r
plotPCA(vdata_dds, c("Sex"))
```



It is clear from the PCA that age (Fetal, Adult) explains the difference in the counts data as compared to the sex.