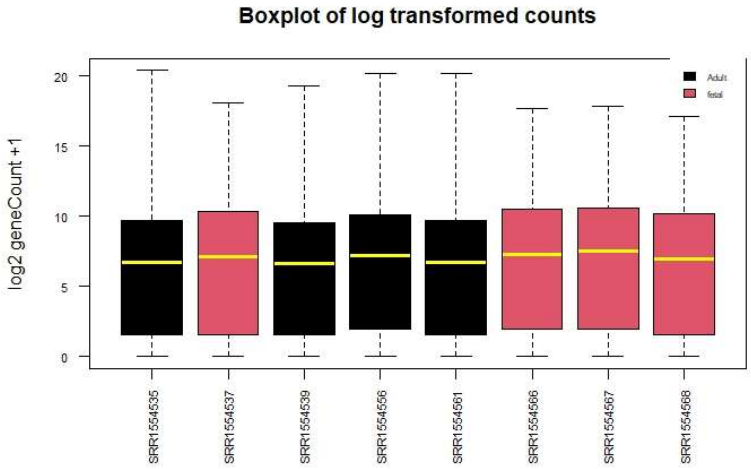
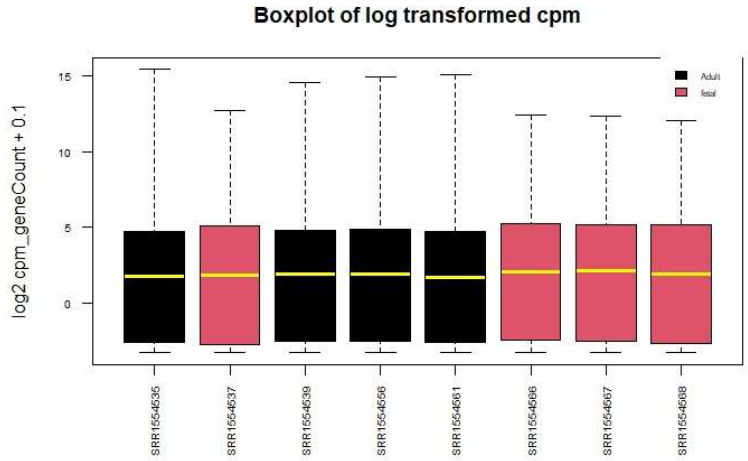
Genomic data Science Project – Week 7

The Gene expression table from week 6 and Phenotype table from week 4 were loaded into R studio. A summarized experiment object was created. For the exploratory data analysis-

1) Boxplot of log2 transformed gene counts was made.



2) To account for differences in library sizes, the CPM values were calculated using cpm function from edgeR package.

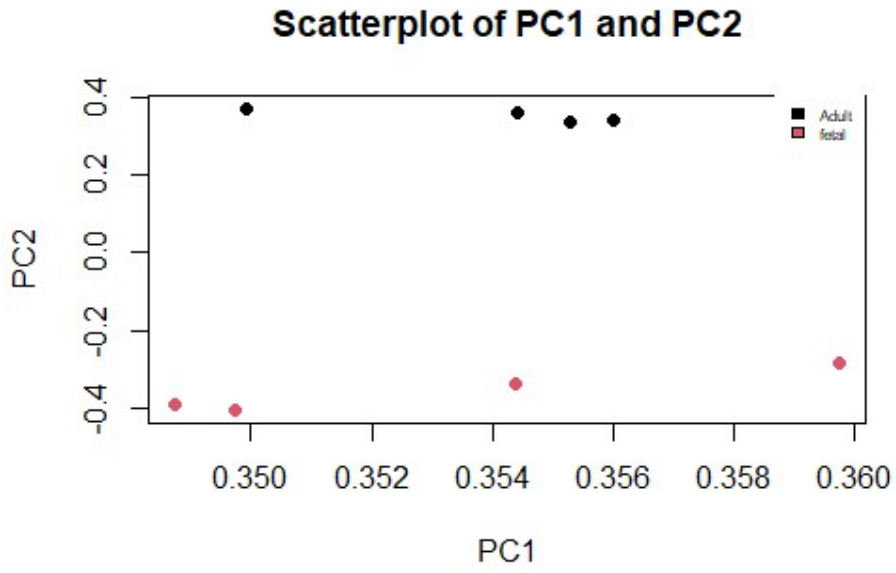


It can be seen that the counts appear to be more evenly distributed after calculating CPM.

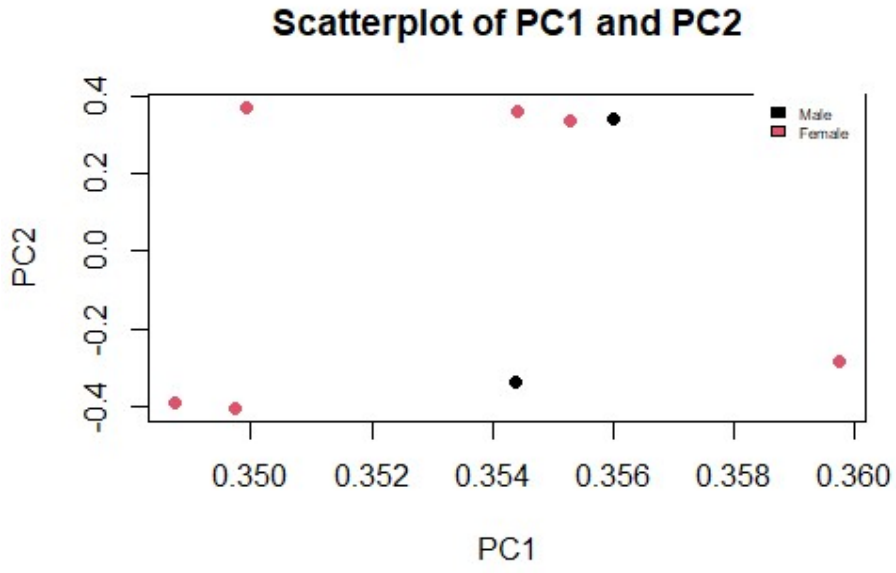
3) To remove lowly expressed genes, only those genes were retained where atleast one sample showed CPM values>0.5. Out of 25702 genes present in the gene expression data, 18237 were retained.

4) Principal component analysis was performed on the log transformed CPM values.

5) Plot of PC1 and PC2 coloring the plots by ‘Age Group” of sample, here ‘Adult’ or ‘fetal’.



6) Plot of PC1 and PC2 coloring the plots by ‘sex’ of sample, here ‘Male’ or ‘Female’.



From the principal component analysis, it can be interpreted that the maximum variability can be explained by the ‘age’ of the samples. The ‘sex’ of the sample does not seem to be correlated with the Principal components.