Gene Expression Data Analysis and Visualization
Final Project

For a final project, you will be conducting an analysis pipeline in attempt to answer some questions about the data set being used. The data selected should include some sort of class structure with different levels (e.g. disease vs. normal, treated vs. non-treated, age<65 vs. age≥65, low dose vs. high dose vs. control etc.).

Each project will include analysis of a publicly available data set from some expression database (e.g., Stanford MicroArray Database, Gene Expression Omnibus, EMBL ArrayExpress). The student should first test for outlier samples and provide visual proof. Remove these outliers. Then, filter out genes that have low expression values using some criterion. Next, conduct some method of feature selection with a statistical test or other machine learning method. The type of test will depend upon how many factor levels are included in your data set. For example. two conditions would require a two-sample test, while greater than two conditions would require other tests. Adjust for multiplicity, then provide the number of genes retained with the associated score (p-value, weight, test statistic, etc.) and threshold value that you used. Plot the scores of those genes retained in a histogram. Next, subset your data by the genes that you determined and use one of the clustering or dimensionality reduction methods discussed in class to visualize the samples in two-dimensional space (xy scatter plot, dendrogram, etc.).

Using these linear projections of the original data (i.e. cluster centroids, latent variables, etc.), use a classification method to classify the samples into their respective classes. Make sure to color the samples appropriately by their predicted class membership and use different symbols for the actual class memberships.

Finally, using the top 5 discriminant genes (positive and negative direction) from your analysis, go to NCBI’s DAVID and look up the gene information. Provide the gene name and functional information (associated pathways, GO terms, etc) for these 10 genes.

Below is a function with the function call to a 1-factor ANOVA with 3 levels. The commented code can be adjusted to return either a p-value or an F-statistic.

```
aov.all.genes <- function(x,s1,s2,s3) {
        x1 <- as.numeric(x[s1])
        x2 <- as.numeric(x[s2])
        x3 <- as.numeric(x[s3])
        fac <- c(rep(“A”,length(x1)), rep(“B”,length(x2)), rep(“C”,length(x3)))
        a.dat <- data.frame(as.factor(fac),c(x1,x2,x3))
        names(a.dat) <- c(“factor”,”express”)
        p.out <- summary(aov(express~factor, a.dat))[[1]][1,5]
        #p.out <- summary(aov(express~factor, a.dat))[[1]][1,4]        # use to get F-statistic
        return(p.out)
}
aov.run <- apply(dat,1, aov.all.genes,s1=bcell,s2=tcell,s3=aml)
```

Put together at least 15 slides in PowerPoint that explain your methodology and findings. Be sure to include things such as an introduction, background, conclusions, and all tables and figures. Also include your code in a text file and turn this in.