Module 6 Assignment

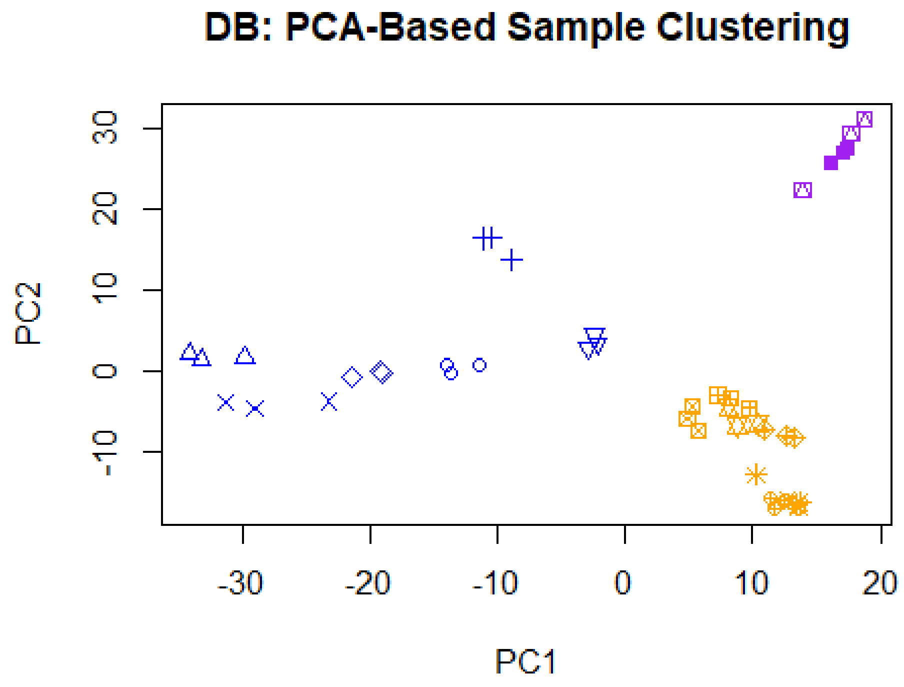
Daniel Bihnam

Load in Data

```r
load('C:/Users/m293215/Downloads/tobacco_clr.Rdata')
```

PCA-Based Clustering Analysis

```r
tobacco_c <- scale(tobacco_clr$data,center=T,scale=F)
#Centers data columns
sigmahat <- 1/150*t(tobacco_c)%*%tobacco_c
#Find sigma hat
sigmahat.pca <- eigen(sigmahat)
#Eigen decomposition
tobacco_eigenv <- sigmahat.pca$vectors
#Extract eigenvectors
pca.cord1 <- as.matrix(tobacco_c)%*%tobacco_eigenv[,1]
pca.cord2 <- as.matrix(tobacco_c)%*%tobacco_eigenv[,2]

plot(pca.cord1,pca.cord2,
     xlab='PC1',
     ylab='PC2',
     col=tobacco_clr$sample.color,
     pch=tobacco_clr$sample.pch,
     main='DB: PCA-Based Sample Clustering')
```

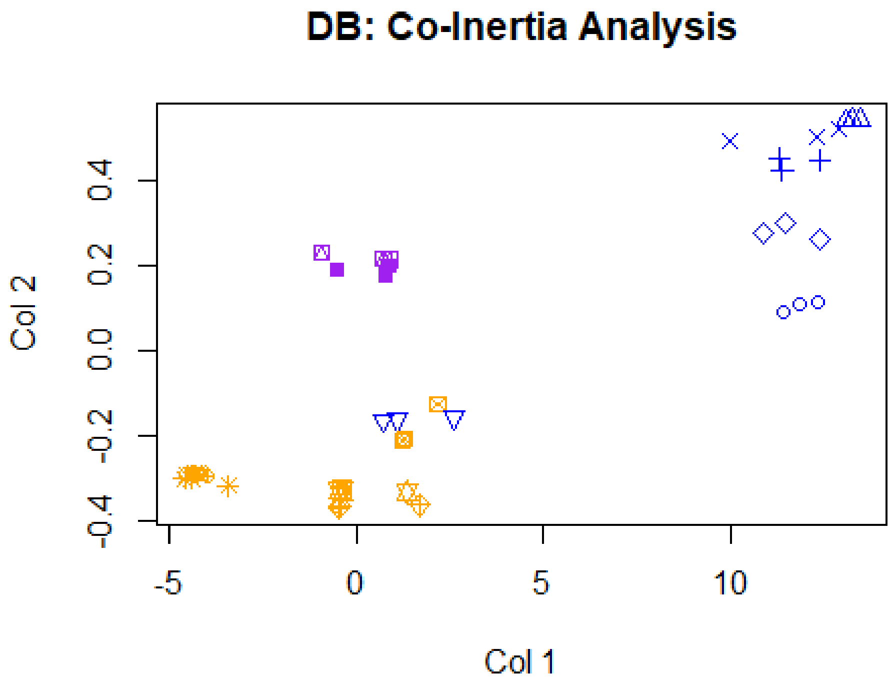


```
#Plot clusters
```

This PCA-based clustering algorithm produces a plot that shows us 3 main clusters. These are separate clusters made up of a heterogeneous group of microorganisms. It is likely that these microorganisms share some common ancestry or features with the other microorganisms in their cluster. This clustering method is only taking the data variable into account, and is just reducing the dimensionality of the dataset.

Co-Inertia Analysis

```
library(CCA)

x=tobacco_clr$data[,1:20]
y=tobacco_clr$H[,1:20]
#Assign Variables
cca.tobacco <- cc(x,y)

cca.x1 <- cca.tobacco$scores$xscores[,1]
cca.y1 <- cca.tobacco$scores$yscores[,1]

coIa <- svd(t(x)%*%y)

coIa.x1 <- x%*%coIa$u[,1]
coIa.y1 <- y%*%coIa$v[,1]

plot(coIa.x1,coIa.y1,
     xlab='CoI 1',
     ylab='CoI 2',
     col=tobacco_clr$sample.color,
     pch=tobacco_clr$sample.pch,
     main='DB: Co-Inertia Analysis')
```



```
#Plot Clusters
```

The co-inertia analysis of the data yields similar results to the PCA-based analysis. There are still 3 groups, and they are made up of the same microorganisms as before. However, an important difference here is that the microorganisms that are the same cluster together within their larger group. A good example is the yellow cluster. In the PCA analysis, the points were more spread out, and some species were overlapping. In the co-inertia analysis, species cluster closer together. It is able to do this because the co-inertia analysis is simultaneously reducing two sets of data, and using that to find similarities between the microorganisms.