<http://afodor.github.io/classes/stats2015/prePostPhylum.txt>

1. Download the dataset. Perform PCA ordination.

(For example:

rm(list=ls())

setwd("C:\\Users\\afodor\\git\\afodor.github.io\\classes\\stats2015\\")

inFileName <- paste("prePostPhylum.txt", sep ="")

myT <-read.table(inFileName,header=TRUE,sep="\t")

numCols <- ncol(myT)

myColClasses <- c(rep("character",4), rep("numeric", numCols-4))

myT <-read.table(inFileName,header=TRUE,sep="\t",colClasses=myColClasses)

myTData<-myT[,5:10]

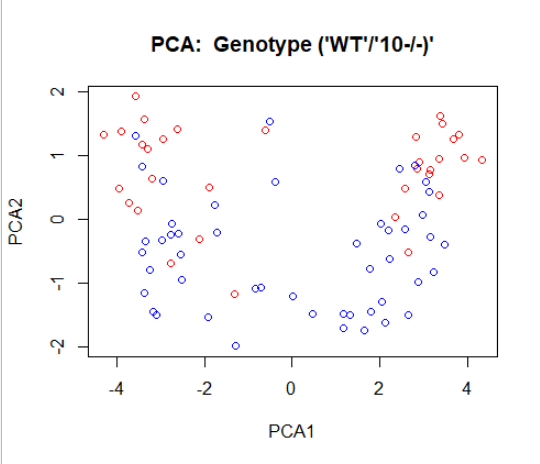
myPCOA <- princomp(myTData)

1. Graph PCA1 vs. PCA2. Make three versions of the graph. One colored by genotype,

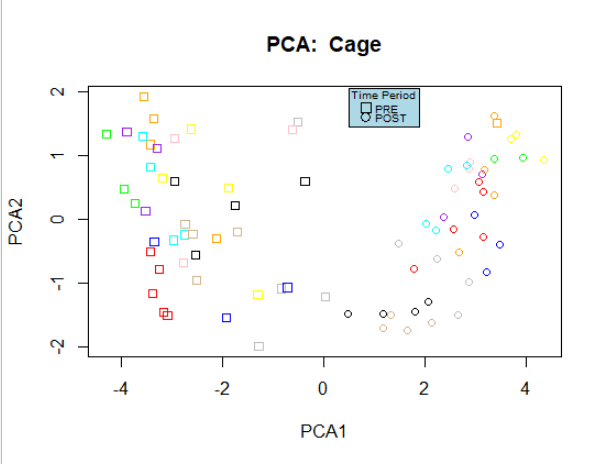
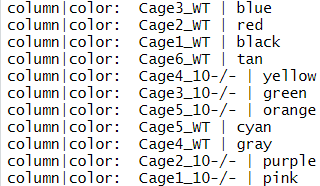
one colored by cage and one colored by timepoint (pre-vs-post)

The following three R scripts were utilized to generate three plots for genotype, cage and timepoint at figures 1, 2 & 3, respectively.

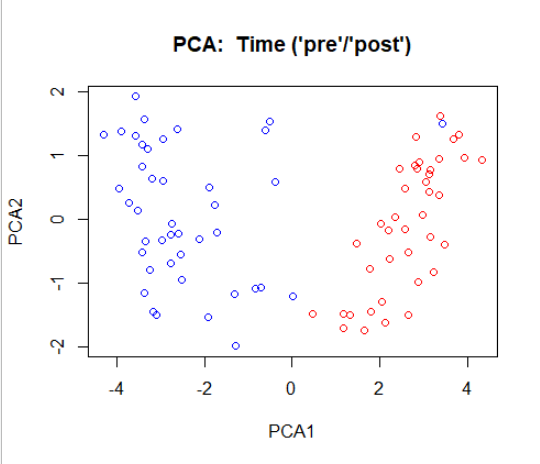
* <https://github.com/jyoung67/advstatistics-labs/blob/master/labs/lab11/performPCA.R>
* <https://github.com/jyoung67/advstatistics-labs/blob/master/labs/lab11/getColorVector.R>
* <https://github.com/jyoung67/advstatistics-labs/blob/master/labs/lab11/getPValues.R>



**Figure 1: Genotype Plot**

**Figure 2: Cage Plot**



**Figure 3: Time Plot**

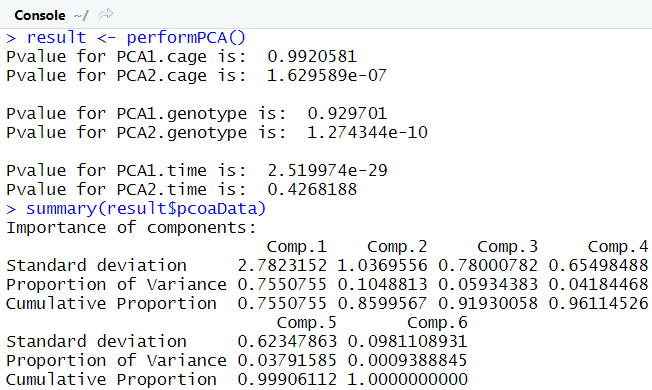
1. Fill in the following table for p-values testing the null hypothesis for PCA 1 and 2. For cage, use a way one-ANOVA. For genotype and timepoint (“pre” vs “post”) use a t-test

The following two scripts were utilized to calculate the p-values for PCA1 & PCA2, listed at table 1. Figure 4 contains the output after executing these scripts.

* <https://github.com/jyoung67/advstatistics-labs/blob/master/labs/lab11/performPCA.R>
* <https://github.com/jyoung67/advstatistics-labs/blob/master/labs/lab11/getPValues.R>

|  |  |  |
| --- | --- | --- |
|  | PCA1 | PCA2 |
| Cage | 0.9920581 | 1.629589e-07 |
| Genotype | 0.929701 | 1.274344e-10 |
| Time (pre vs. post) | 2.519974e-29 | 0.4268188 |

**Table 1: Calculated p-values for PCA1 & PCA2**



**Figure 4: Output of R-command scripts for calculating p-values**

Which variable seems to be most associated with the first PCA axis? Which variable is most associated with the second PCA axis? Does cage seem to be having an effect on these data?

* Time is most associated with the first PCA axis, because it has the most significant p-value of 2.519974e-29 among cage and genotype (listing 1).
* Genotype is most associated with the second PCA axis, because it has the most significant p-value of 1.274344e-10 among time and cage (listing 1).
* Yes, cage seems to have an effect on these data. First of all, each of the eleven cage names displays a notable degree of clustering within pre and post time points (figure 2). In addition, cage has a significant p-value of 1.629589e-07 for PCA2 (listing 1).