**Berk Yıldız**

**21502040**

**GEOquery Package PCA Worksheet Takehome**

**MBG326- Fall 2018**

Install GEOquery package

source("https://bioconductor.org/biocLite.R")

biocLite("GEOquery")

library(GEOquery)

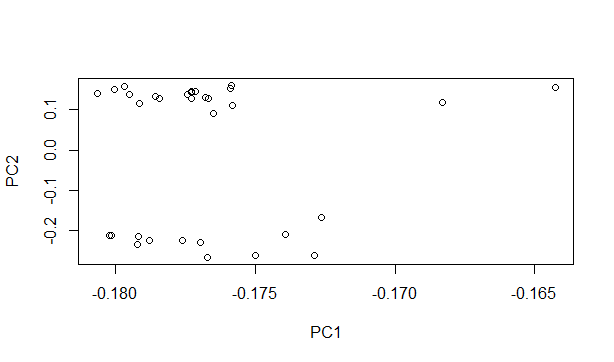
1. Download GDS2577 dataset using GEOQuery package (function: getGEO)

library(GEOquery)

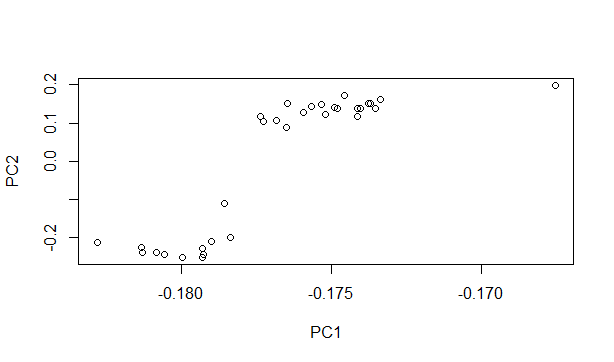
dset<- getGEO(“GDS2577”)

dset

1. Take log of data (function (GDS2eSet() with option do.log=TRUE))
2. Select randomly 50 rows from the dataset and perform a PCA analysis using this subset of genes and draw a PCA plot and label each sample with its time point and paste it here (prcomp function is used for PCA plot). Are regenerating liver samples distinctly clustered from the embryonic samples?



1. Select randomly 150 rows from the dataset and perform a PCA analysis using this subset of genes and draw a PCA plot and label each sample with its time point and paste it here. regenerating liver samples distinctly clustered from the embryonic samples? Compare with question 3.



1. Next, obtain the top 50 most variable (based on standard deviation of the gene across samples divided by the mean gene expression value, which is coefficient of variation) genes and perform a PCA analysis and draw a PCA plot and label each sample with its time point and paste it here. Are regenerating liver samples distinctly clustered from the embryonic samples? Compare with question 3 and 4.
2. Discuss the answers you have obtained to question 4, 5 and 6 in terms of their results.
3. Based on liver regenation/development paper and the existing bioinformatics literature on PCA for expression data on Pubmed, write a one paragraph discussion on how the features (genes) should be selected to perform PCA on an expression dataset to visualize samples? You need to cite three references in addition to the original dataset paper (GDS2577 paper on Moodle).