GSE22544 Compression

This R script uses a uniqued gene symbol vector to loop through a expression dataframe to subset and sort out gene symbols in a list. This list is then changed in a dataframe/matrix, meaned, and eventually contains indiv gene symbols on the rownames and sample names on colnames. Contains meaned expression data.

setwd("D:/Code/RE/My R scripts")

## (1) Loading up the combined dataset and gene symbol vector

bound\_GSE22544 <- read.csv("bound\_GSE22544.csv", header = T, sep = ",", stringsAsFactors = FALSE)  
bound\_GSE22544 <- bound\_GSE22544[,-1] # removes 1st column of numbers  
bound\_GSE22544 <- bound\_GSE22544[!(is.na(bound\_GSE22544$Gene.Symbol) | bound\_GSE22544$Gene.Symbol=="")  
   
gene\_symbol\_GSE22544 <- read.csv("gene\_symbol\_GSE22544.csv", header = T, sep = ",", stringsAsFactors = FALSE)  
gene\_symbol\_GSE22544 <- gene\_symbol\_GSE22544[,2] # chooses only the second column  
gene\_symbol\_GSE22544 <- na.omit(gene\_symbol\_GSE22544) # removes rows that have NAs  
# View(gene\_symbol\_GSE22544) # 16,072 entries

## (2) Using unique() function to remove duplicate genes

uniq\_genesymbol <- unique(gene\_symbol\_GSE22544) #16,972 entries before unique function  
# View(uniq\_genesymbol)# 9,995 entries after unique function

## (3) Making an Empty List

mat\_list <- list()

## (4) Using a ‘for’ loop to loop through bound\_GSE22544 using a gene symbol vector

for (i in uniq\_genesymbol){  
 df\_subset <- subset(bound\_GSE22544, Gene.Symbol == i, select = -c(1))  
 mat\_subset <- as.matrix(df\_subset)  
 mat\_list[[i]] <- mat\_subset  
}

## (5) Writing a function to find the mean for each sublist

mean\_fun <- function(x){  
 mean\_value <- colMeans(x)  
 return(mean\_value)  
}

## (6) Using lapply on the list of gene symbols and exprs data using the function i created (mean\_fun)

mean\_list <- lapply(mat\_list, mean\_fun)  
# View(mean\_list)

## (7) Exprs data compressed into mean exprs data

# new\_GSE22544 <- matrix(unlist(mean\_list), ncol = 19, byrow = TRUE) # rejected cos no row and column names; only mean exprs data  
# new\_GSE22544 <- data.frame(matrix(unlist(mean\_list), ncol = 19, byrow = TRUE)) # rejected cos no column and row names retained and needs extra steps to take the colnames from bound\_22544 df  
new\_df\_GSE22544 <- data.frame(t(sapply(mean\_list,c))) # has gene symbols as rownames and sample numbers as column names

## Writing files

write.csv(new\_df\_GSE22544, file = "D:/Code/RE/My R scripts/new\_df\_GSE22544.csv", row.names = T)