GSE22544 Extraction Procedure

## (1) Set the working directory

The working directory is always set at the start of any R script to ensure that RStudio can read the files in the working directory.

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

## (2) Obtain GEO GSE file

Firstly, the data is extracted from a GEO GSE file using the GEOquery package:

library(GEOquery) # always use the library() function first in order to use a certain package  
   
GSE22544 <- getGEO(filename = 'GSE22544\_series\_matrix.txt.gz')  
GSE22544

The code has been commented to prevent the code from running.

## (3) Putting pheno data and feature data into respective vectors

Then I assigned pheno data and feature data of the GSE into separate variables as such:

pheno\_GSE22544 <- phenoData(GSE22544) # pheno data  
feature\_GSE22544 <- fData(GSE22544) # feature data

## (4) Obtaining expression data from feature data

exprs\_GSE22544 <- exprs(GSE22544) # exprs() function applied to GSE22544  
 class(exprs\_GSE22544) # shows as matrix  
 exprs\_GSE22544 <- as.data.frame(exprs\_GSE22544) # change matrix to data.frame  
 View(exprs\_GSE22544) # GSM on the columns, 'at' on the rows

## (5) Subsetting the gene Symbols from feature data

GENE\_Symbols <- (feature\_GSE22544 [11])  
 rownames(GENE\_Symbols) <- NULL # removes probe IDs as rownames  
 class(GENE\_Symbols)

## (6) Combine GENE\_Symbols df with expression df

bound\_GSE22544 <- cbind(GENE\_Symbols, exprs\_GSE22544)  
 rownames(bound\_GSE22544) <- NULL  
   
 View(bound\_GSE22544)  
 ncol(bound\_GSE22544)

## (7) Creating a class information vector and factor

# using gsub to change the class labels  
pheno\_GSE22544@data$characteristics\_ch1 <- gsub("tissue:", "", pheno\_GSE22544@data$characteristics\_ch1)  
pheno\_GSE22544@data$characteristics\_ch1 <- gsub("breast cancer", "IDC", pheno\_GSE22544@data$characteristics\_ch1)  
pheno\_GSE22544@data$characteristics\_ch1 <- gsub("normal breast", "normal", pheno\_GSE22544@data$characteristics\_ch1)  
   
 # making a factor for use later in bootstrap step  
 class\_factor <- as.factor(pheno\_GSE22544@data$characteristics\_ch1)  
 levels(class\_factor)  
 levels(class\_factor) [1]  
 levels(class\_factor) [2]  
   
 class\_factor == levels(class\_factor) [1] # [1] TRUE TRUE TRUE FALSE TRUE TRUE TRUE FALSE FALSE TRUE TRUE TRUE FALSE TRUE TRUE TRUE TRUE TRUE TRUE TRUE  
   
   
 class\_factor == levels(class\_factor) [2] # [1] FALSE FALSE FALSE TRUE FALSE FALSE FALSE TRUE TRUE FALSE FALSE FALSE TRUE FALSE FALSE FALSE FALSE FALSE FALSE FALSE  
   
   
 # making a vector for insertion into exprs dataframe later in bootstrap step  
 class\_vector <- c(pheno\_GSE22544@data$characteristics\_ch1)

## (8) Creating a gene symbols vector

gene\_symbol\_GSE22544 <- (feature\_GSE22544 [11])  
 rownames(gene\_symbol\_GSE22544) <- NULL  
 # View(gene\_symbol\_GSE22544)  
   
 gene\_symbol\_GSE22544 <- as.vector(gene\_symbol\_GSE22544[[1]])  
 is.vector(gene\_symbol\_GSE22544)  
   
 # writing into a csv file  
 # write.csv()

## Writing files

write.csv(bound\_GSE22544, file = "D:/Code/RE/My R scripts/bound\_GSE22544.csv") # writes bound df with gene symbol and expression data  
write.csv(gene\_symbol\_GSE22544, file = "D:/Code/RE/My R scripts/gene\_symbol\_GSE22544.csv") # writes gene symbol vector  
write.csv(class\_factor, file = "D:/Code/RE/My R scripts/class\_factor.csv") # writes class labels factor  
write.csv(class\_vector, file = "D:/Code/RE/My R scripts/class\_vector.csv")