Lab8.R

daniellesenechal

2019-11-10

################################################################  
# Lab 8: GEO Lab - Raw Data  
# Danielle Senechal  
# In this lab you will analyze two probes from a gene expression  
# study of Alzheimer's Disease (AD). The dataset is  
# available from:   
# http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE1297  
################################################################  
  
library(affy)

## Loading required package: BiocGenerics

## Loading required package: parallel

##   
## Attaching package: 'BiocGenerics'

## The following objects are masked from 'package:parallel':  
##   
## clusterApply, clusterApplyLB, clusterCall, clusterEvalQ,  
## clusterExport, clusterMap, parApply, parCapply, parLapply,  
## parLapplyLB, parRapply, parSapply, parSapplyLB

## The following objects are masked from 'package:stats':  
##   
## IQR, mad, sd, var, xtabs

## The following objects are masked from 'package:base':  
##   
## anyDuplicated, append, as.data.frame, basename, cbind,  
## colnames, dirname, do.call, duplicated, eval, evalq, Filter,  
## Find, get, grep, grepl, intersect, is.unsorted, lapply, Map,  
## mapply, match, mget, order, paste, pmax, pmax.int, pmin,  
## pmin.int, Position, rank, rbind, Reduce, rownames, sapply,  
## setdiff, sort, table, tapply, union, unique, unsplit, which,  
## which.max, which.min

## Loading required package: Biobase

## Welcome to Bioconductor  
##   
## Vignettes contain introductory material; view with  
## 'browseVignettes()'. To cite Bioconductor, see  
## 'citation("Biobase")', and for packages 'citation("pkgname")'.

library(dplyr)

##   
## Attaching package: 'dplyr'

## The following object is masked from 'package:Biobase':  
##   
## combine

## The following objects are masked from 'package:BiocGenerics':  
##   
## combine, intersect, setdiff, union

## The following objects are masked from 'package:stats':  
##   
## filter, lag

## The following objects are masked from 'package:base':  
##   
## intersect, setdiff, setequal, union

library(ggplot2)  
  
#####################################################################  
# Download the raw data (GSE1297\_RAW.tar file) and extract the files   
# Read in the data using theReadAffy() function and the   
# celfile.path argument which must be set appropriately below  
################################################################  
  
GSE1297 <- ReadAffy(celfile.path =   
 "/Users/daniellesenechal/Documents/ECSU/Fall 2019/CSC 315 Bioinformatics/Lab 8/GSE1297\_RAW")  
  
################################################################  
# Process the gene expression data using the Robust Multi-Array   
# Average (RMA) method and extract the expression data.  
# How many probes and samples does this dataset contain?  
# Generate a boxplot of the expression values of each sample  
# to confirm that the data has been normalized  
################################################################  
  
GSE1297.rma <- rma(GSE1297)

## Warning: replacing previous import 'AnnotationDbi::tail' by 'utils::tail'  
## when loading 'hgu133acdf'

## Warning: replacing previous import 'AnnotationDbi::head' by 'utils::head'  
## when loading 'hgu133acdf'

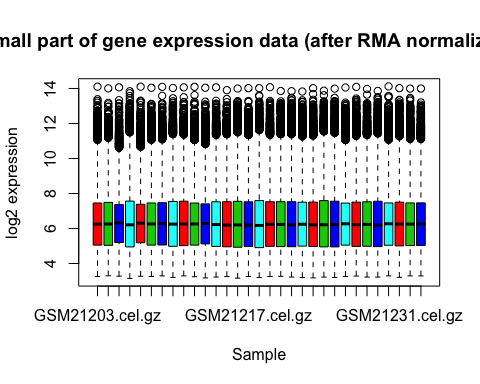
##

## Background correcting  
## Normalizing  
## Calculating Expression

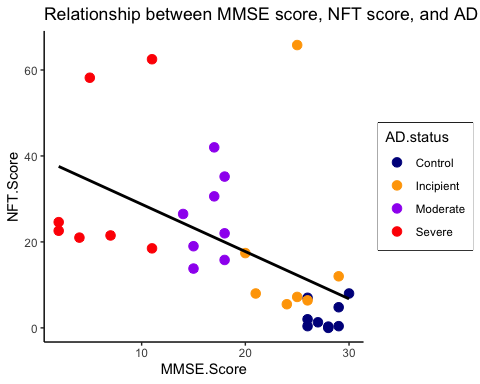
GSE1297.expr <- exprs(GSE1297.rma)   
  
GSE1297.rma

## ExpressionSet (storageMode: lockedEnvironment)  
## assayData: 22283 features, 31 samples   
## element names: exprs   
## protocolData  
## sampleNames: GSM21203.cel.gz GSM21204.cel.gz ... GSM21233.cel.gz  
## (31 total)  
## varLabels: ScanDate  
## varMetadata: labelDescription  
## phenoData  
## sampleNames: GSM21203.cel.gz GSM21204.cel.gz ... GSM21233.cel.gz  
## (31 total)  
## varLabels: sample  
## varMetadata: labelDescription  
## featureData: none  
## experimentData: use 'experimentData(object)'  
## Annotation: hgu133a

# The data set contains 22283 probes and 31 samples.  
  
boxplot(GSE1297.expr, col = 2:5, ylab = "log2 expression",  
 main = "Small part of gene expression data (after RMA normalization)",  
 xlab = "Sample")



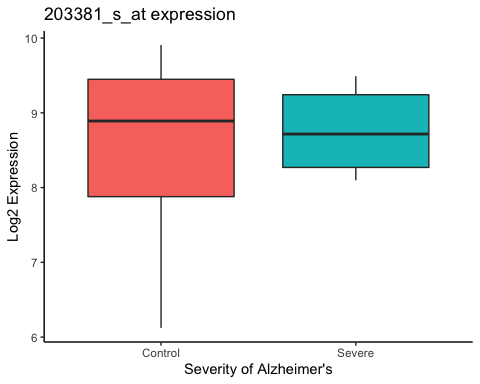
###################################################################  
# We will see how to get the phenotype data from GEO in a later  
# class; for this lab, the data has already been processed  
# and can be read in using the statement below. The data includes  
# MMSE.Score = mini–mental state examination score for   
# cognitive impairment (low scores indicate impairment)  
# NFT.Score = protein markers for AD  
################################################################  
  
GSE1297.p <- read.delim("http://bioinformatics.easternct.edu/BCBET2/GSE1297.p.xlsx")  
  
##################################################################  
# The code below gets the group names from the sample names of the   
# pheno table and constructs a scatterplot of MMSE and NFT  
# scores with points color-coded by AD severity.   
# (note: code assumes that the pheno table is stored in GSE1297.p)  
###################################################################  
  
##################################################  
# get group names from sample names, which have  
# format "Group SampleNumber"  
##################################################  
sample.names <- as.character(GSE1297.p$Sample)  
groups <- gsub(" .\*", "", sample.names)  
  
# update the phenotype data with the group names  
GSE1297.p <- mutate(GSE1297.p, AD.status = groups)  
  
ggplot(GSE1297.p, aes(MMSE.Score, NFT.Score)) +  
 geom\_point(aes(color = AD.status), size = 3) +   
 geom\_smooth(method = "lm", color = "black", se = FALSE) +  
 theme\_classic() +  
 ggtitle("Relationship between MMSE score, NFT score, and AD severity") +  
 theme(legend.box.background = element\_rect(color = "black")) +  
 scale\_color\_manual(values = c("darkblue", "orange", "purple", "red"))



cor(GSE1297.p$MMSE.Score, GSE1297.p$NFT.Score)

## [1] -0.5304362

########################################################  
# Describe the relationship between MMSE and NFT score.  
# Would you expect a person with a high MMSE score to  
# have Alzheimer's Disease?  
#####################################################  
  
# The relationship between MMSE and NFT has moderately weak association and is decreasing.   
# The correlation is -0.5304362. It can be expected that a person with a high MMSE score   
# does not have Alzheimer's Disease, because the Disease Status for those with the high   
# MMSE scores are control or incipient. The people with high MMSE scores also have low   
# protien markers for Alzheimer's.   
  
#####################################################  
# A gene called APOE is associated with late onset  
# Alzheimer's disease. One of the probes for   
# APOE is 203381\_s\_at  
#####################################################  
  
############################################################  
# Construct side-by-side boxplots showing the expression  
# of the probe 203381\_s\_at for CONTROL patients and   
# patients with SEVERE AD. (The boxplot must be constructed  
# using ggplot -- see notes for an example)  
############################################################  
  
m <- match("203381\_s\_at", rownames(GSE1297.expr))  
df <- data.frame(expr = GSE1297.expr[m,], type = GSE1297.p$AD.status)  
df <- filter(df, type == "Control" | type == "Severe")  
ggplot(df,aes(type, expr, fill = type)) + geom\_boxplot() +  
 theme\_classic() + theme(legend.position = "none") +   
 labs(x = "Severity of Alzheimer's", y = "Log2 Expression") +  
 ggtitle("203381\_s\_at expression")



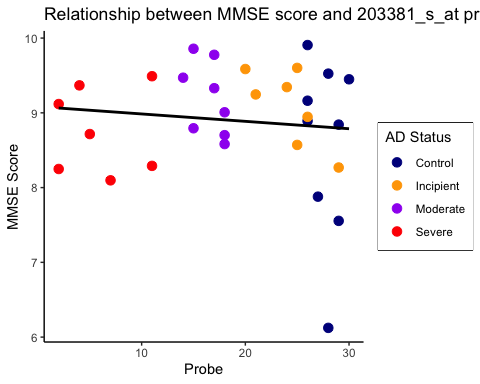
###########################################################  
# Perform a two sample t-test to evaluate whether or not  
# expression is significantly different between CONTROL  
# patients and patients with SEVERE AD. Report the   
# fold change and the p-value and state your conclusion.  
###########################################################  
  
s <- split(df$expr, df$type, drop = TRUE)   
  
l <- lapply(s, mean)  
logFC <- l$Severe - l$Control  
FC <- 2\*\*logFC   
FC

## [1] 1.123929

# The fold change is 1.123929.  
  
res <- t.test(s$Control, s$Severe)  
res$p.value

## [1] 0.7168644

# The p-value is 0.7168644.  
# Since the p-value of 0.7168644 is above 0.05, we fail to reject the null hyopthesis.  
# This means that expression for the gene APOE is not significantly different between   
# CONTROL patients and patients with SEVERE AD.  
  
#####################################################  
# Construct a scatterplot of gene expression of  
# the probe 203381\_s\_at on the x-axis and MMSE score  
# on the y-axis, coloring the points by AD status as  
# was done for the above scatterplot. Give the graph   
# an appropriate title, axis labels, and legend,   
# and also add the regression line, as was done above.   
# What is the correlation between MMSE score and   
# expression?   
#####################################################  
  
ggplot(NULL, aes(GSE1297.p$MMSE.Score, GSE1297.expr[m,])) +  
 geom\_point(aes(color = GSE1297.p$AD.status), size = 3) +   
 geom\_smooth(method = "lm", color = "black", se = FALSE) +  
 theme\_classic() +  
 labs(x = "Probe", y = "MMSE Score", color = "AD Status") +  
 ggtitle("Relationship between MMSE score and 203381\_s\_at probe") +  
 theme(legend.box.background = element\_rect(color = "black")) +  
 scale\_color\_manual(values = c("darkblue", "orange", "purple", "red"))



cor(GSE1297.p$MMSE.Score, GSE1297.expr[m,])

## [1] -0.1096518

# The correlation between MMSE score and expression is -0.1096518. The correlation is weak.  
  
##########################################################  
# The cor.test function can be used to evaluate the  
# following hypotheses:  
  
# H0: r = 0, where r is the correlation between x and y  
# HA: r != 0  
  
# The function is called using cor.test(x,y), where x   
# and y are the vectors of observations. Find the p-value,   
# report the correlation, and state whether or not the   
# correlation between expression and MMSE score is   
# statistically significant   
##########################################################  
  
res <- cor.test(GSE1297.p$MMSE.Score, GSE1297.expr[m,])  
res

##   
## Pearson's product-moment correlation  
##   
## data: GSE1297.p$MMSE.Score and GSE1297.expr[m, ]  
## t = -0.59408, df = 29, p-value = 0.5571  
## alternative hypothesis: true correlation is not equal to 0  
## 95 percent confidence interval:  
## -0.4466383 0.2545797  
## sample estimates:  
## cor   
## -0.1096518

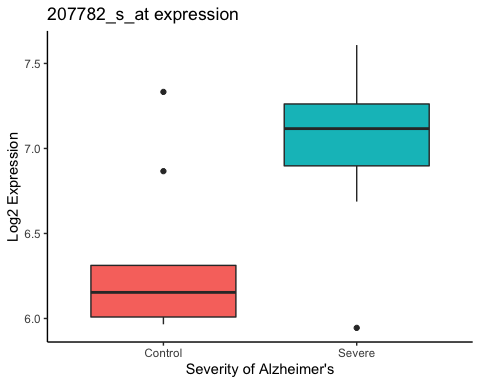
res$p.value

## [1] 0.557068

cor(GSE1297.p$MMSE.Score, GSE1297.expr[m,])

## [1] -0.1096518

# The correlation is -0.1096518. It is weak. The p-value is 0.557068. Since the p-value  
# does not fall below, we fail to reject the null hypothesis. This means there is not   
# significant evidence that the correlation between expression and MMSE score is   
# statistically significant.   
  
#####################################################  
## Repeat the boxplot, t.test, scatterplot, and   
## cor.test for the gene PSEN1 using the probe   
## 207782\_s\_at  
#####################################################  
  
# boxplot  
m <- match("207782\_s\_at", rownames(GSE1297.expr))  
df <- data.frame(expr = GSE1297.expr[m,], type = GSE1297.p$AD.status)  
df <- filter(df, type == "Control" | type == "Severe")  
ggplot(df,aes(type, expr, fill = type)) + geom\_boxplot() +  
 theme\_classic() + theme(legend.position = "none") +   
 labs(x = "Severity of Alzheimer's", y = "Log2 Expression") +  
 ggtitle("207782\_s\_at expression")



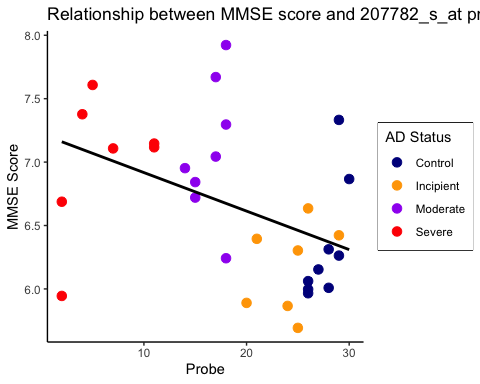
# t.test  
s <- split(df$expr, df$type, drop = TRUE)   
  
l <- lapply(s, mean)  
logFC <- l$Severe - l$Control  
FC <- 2\*\*logFC   
FC

## [1] 1.590322

# The fold change is 1.590322.  
  
res <- t.test(s$Control, s$Severe)  
res$p.value

## [1] 0.02343402

# The p-value is 0.02343402.  
# Since the p-value of 0.02343402 falls below 0.05, we reject the null hyopthesis in  
# favor of the alternatice.This means that expression for the gene PSEN1 is   
# significantly different between CONTROL patients and patients with SEVERE AD.  
  
# scatterplot  
ggplot(NULL, aes(GSE1297.p$MMSE.Score, GSE1297.expr[m,])) +  
 geom\_point(aes(color = GSE1297.p$AD.status), size = 3) +   
 geom\_smooth(method = "lm", color = "black", se = FALSE) +  
 theme\_classic() +  
 labs(x = "Probe", y = "MMSE Score", color = "AD Status") +  
 ggtitle("Relationship between MMSE score and 207782\_s\_at probe") +  
 theme(legend.box.background = element\_rect(color = "black")) +  
 scale\_color\_manual(values = c("darkblue", "orange", "purple", "red"))



cor(GSE1297.p$MMSE.Score, GSE1297.expr[m,])

## [1] -0.4342109

# cor.test  
res <- cor.test(GSE1297.p$MMSE.Score, GSE1297.expr[m,])  
res

##   
## Pearson's product-moment correlation  
##   
## data: GSE1297.p$MMSE.Score and GSE1297.expr[m, ]  
## t = -2.5958, df = 29, p-value = 0.01466  
## alternative hypothesis: true correlation is not equal to 0  
## 95 percent confidence interval:  
## -0.68340363 -0.09439404  
## sample estimates:  
## cor   
## -0.4342109

res$p.value

## [1] 0.01465959

cor(GSE1297.p$MMSE.Score, GSE1297.expr[m,])

## [1] -0.4342109

# The correlation is -0.4342109. It is moderately weak. The p-value is 0.01465959. Since   
# the p-value falls below 0.05, we reject the null hypothesis in favor of the alternative.  
# This means that there is evidence that the correlation between expression and MMSE score is   
# statistically significant.   
  
########################################################  
## Based on the above analyses, what is your conclusion  
## about the association between the genes APOE and  
## PSEN1 and Alzheimer's Disease / cognitive   
## impairment?  
###############################################s######  
  
# The association between the APOE gene and Alzheimer's Disease is weak, and there is no  
# association between cognitive impairment and the the gene.  
# The association between the PSEN1 gene and Alzheimer's Disease is more signifcant, and   
# it can be concluded that there may be some association between PSEN1 and cognitive impairment.   
  
###########################################################  
## If you are interested, more information about  
## Alzheimer's Disease and these genes can be found  
## at: http://ghr.nlm.nih.gov/condition/alzheimer-disease  
##########################################################