Lab9.R

daniellesenechal

2019-11-17

##############################################################  
# Name: Danielle Senechal  
# CSC-315  
# Lab #9: Limma, heatmaps, and analyzing processed GEO data  
#############################################################  
  
##########################################################################  
# Add R code to the script below and create a Notebook to complete  
# the steps below and to explicitly answer the following questions  
##########################################################################  
  
library(limma)  
library(GEOquery)

## Loading required package: Biobase

## 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 object is masked from 'package:limma':  
##   
## plotMA

## 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

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

## Setting options('download.file.method.GEOquery'='auto')

## Setting options('GEOquery.inmemory.gpl'=FALSE)

library(ggplot2)  
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

##################################################################################  
# 1.The code below reads in our class survey data and performs a   
# 2-sample t-test to evaluate whether there is a statistically  
# significant difference in Hours of Sleep between 'Cat' vs. 'Dog'  
# people. Based on the code below, (a) find the p-value and state   
# your conclusion regarding the null hypothesis of H0: mu\_cat - mu\_dog = 0;  
# and (b) calculate the difference in mean Alcohol consumption between  
# groups, using the formula:   
# mean hours of sleep for dog people - mean hours of sleep for cat people  
##################################################################################  
survey <- read.csv("https://gdancik.github.io/CSC-315/data/datasets/CSC-315\_survey.csv")  
s <- split(survey$Sleep, survey$CatOrDogPerson)  
res <- t.test(s$Cat, s$Dog, var.equal = TRUE)  
  
# a)   
   
 res$p.value

## [1] 0.7970757

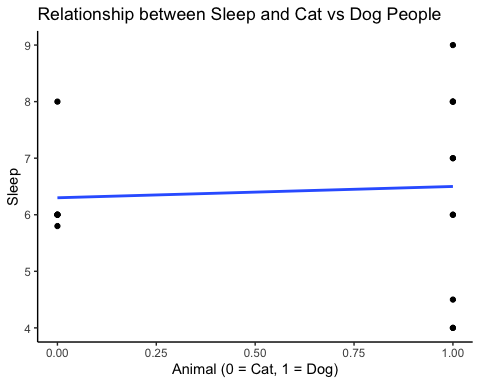
# The p-value is 0.7970757. Since the p-value does not fall below 0.05, we  
 # fail to reject the null hypothesis. This means that there not a statistically  
 # significant difference in Hours of Sleep between 'Cat' vs. 'Dog'  
 # people.  
   
# b)  
  
 diff(res$estimate)

## mean of y   
## 0.2

mean(s$Dog) - mean(s$Cat)

## [1] 0.2

# The difference of the means is 0.2.  
   
##################################################################################  
# 2.Fit a linear model that predicts Hours of Sleep based on   
# whether an individual is a cat or a dog person. You should use  
# the treatment contrast where 'cat person' is the reference (x = 0) and   
# 'dog person' is the treatment (x = +1)  
   
 animal <- as.integer(survey$CatOrDogPerson == "Dog")  
 df <- mutate(survey, animal)  
 # View(select(df, CatOrDogPerson, animal))  
  
 ggplot(df, aes(animal, Sleep)) +  
 geom\_point() + geom\_smooth(method = "lm", se = FALSE) +  
 theme\_classic() +   
 xlab("Animal (0 = Cat, 1 = Dog)") +  
 ggtitle("Relationship between Sleep and Cat vs Dog People")



# (a) Find and interpret the y-intercept of the regression line in the  
# context of this problem.  
   
 fit <- lm(Sleep ~ animal, data = survey)  
 fit

##   
## Call:  
## lm(formula = Sleep ~ animal, data = survey)  
##   
## Coefficients:  
## (Intercept) animal   
## 6.3 0.2

# The y-intercept is 6.3 hours of sleep. This means that the mean amount of sleep   
 # that cat lovers get is 6.3.  
  
# (b) Find and interpret the slope of the regression line in the context of   
# this problem  
   
 # As seen in the summary generated above, the slope of the regression line is 0.2,   
 # which shows the difference in the means of the hours of sleep for cat vs dog people.   
 # This slope shows that the mean average of sleep for cat lovers is 0.2 less than the  
 # mean average of sleep for dog lovers.  
  
# (c) What is the p-value for the hypothesis test that there is a  
# significant difference in Hours of Sleep between the two groups?  
# (show this result in R, based on the linear model) Note: the p-value   
# from the linear model should match the p-value from the two-sample   
# t-test from problem 1(a) above.  
   
 summary(fit)

##   
## Call:  
## lm(formula = Sleep ~ animal, data = survey)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -2.5 -0.5 -0.3 1.5 2.5   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 6.3000 0.6146 10.250 3.61e-08 \*\*\*  
## animal 0.2000 0.7641 0.262 0.797   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 1.506 on 15 degrees of freedom  
## Multiple R-squared: 0.004547, Adjusted R-squared: -0.06182   
## F-statistic: 0.06851 on 1 and 15 DF, p-value: 0.7971

# As seen in the summary generate above, the p-value is 0.7971, which does match  
 # the p-value generated in 1a. There is not a significant difference in hours of   
 # sleep between the two groups.   
   
##################################################################################  
  
  
###############################################################  
# 3. Get the processed data for GSE19143 and pull out the   
# expression data and phenotype data. Note that this  
# dataset contains gene expression samples from children  
# with Acute Lymphoblastic Leukemia (ALL), a cancer of  
# the bone marrow. Tumor samples were treated with  
# the anti-inflammatory drug prednisolone, and determined   
# to be either sensitive (responsive) or resistant   
# (non-responsive) to this drug.   
###############################################################  
   
 GSE19143 <- getGEO("GSE19143")

## Found 1 file(s)

## GSE19143\_series\_matrix.txt.gz

## Parsed with column specification:  
## cols(  
## .default = col\_double(),  
## ID\_REF = col\_character()  
## )

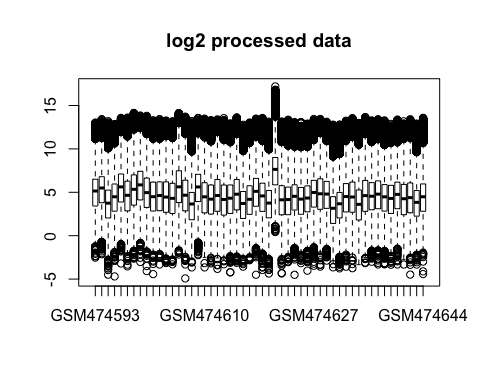
## See spec(...) for full column specifications.

## File stored at:

## /var/folders/dr/8c7c334x2yg5tc7lxlnj07kr0000gn/T//RtmpmBCDs7/GPL96.soft

## Warning: 68 parsing failures.  
## row col expected actual file  
## 22216 SPOT\_ID 1/0/T/F/TRUE/FALSE --Control literal data  
## 22217 SPOT\_ID 1/0/T/F/TRUE/FALSE --Control literal data  
## 22218 SPOT\_ID 1/0/T/F/TRUE/FALSE --Control literal data  
## 22219 SPOT\_ID 1/0/T/F/TRUE/FALSE --Control literal data  
## 22220 SPOT\_ID 1/0/T/F/TRUE/FALSE --Control literal data  
## ..... ....... .................. ......... ............  
## See problems(...) for more details.

GSE19143.expr <- exprs(GSE19143[[1]])   
 GSE19143.p <- pData(GSE19143[[1]])  
   
# (a) How many samples had their gene expression values profiled?  
   
 # There are 52 samples  
   
# (b) How many probes are on the array?  
  
 # There are 22283 probes.  
   
# (c) Take the log2 of the expression data, and generate a boxplot  
# to show that the samples are properly processed and normalized.  
# The analysis beginning with question 5 must use the log2 data;   
# otherwise the results will not be correct.  
  
 GSE19143.expr <- log2(GSE19143.expr)  
 boxplot(GSE19143.expr, main = "log2 processed data")



#####################################################################  
# 4.How many individuals are resistant to prednisolone and  
# how many are sensitive?   
#####################################################################  
   
 indv <- as.character(GSE19143.p$characteristics\_ch1.5)  
 table(indv)

## indv  
## clinical response factor: prednisolone resistant   
## 25   
## clinical response factor: prednisolone sensitive   
## 27

# 27 individuals are sensitive and 25 are resistant.  
  
#####################################################################  
# 5. Find the top differentially expressed probes, with a FDR of 10%,  
# between individuals that are resistant vs. sensitive to prednisolone.  
# Note: there should be 16 probes total. How many of these probes   
# are up-regulated (i.e., have higher expression) in resistant   
# individuals and how many are down-regulated (i.e., have lower   
# expression) in resistant individuals.   
#####################################################################  
  
 designs <- model.matrix(~0+indv)  
 head(designs)

## indvclinical response factor: prednisolone resistant  
## 1 0  
## 2 0  
## 3 0  
## 4 0  
## 5 0  
## 6 0  
## indvclinical response factor: prednisolone sensitive  
## 1 1  
## 2 1  
## 3 1  
## 4 1  
## 5 1  
## 6 1

colnames(designs) <- c("Resistant", "Sensitive")  
 head(designs)

## Resistant Sensitive  
## 1 0 1  
## 2 0 1  
## 3 0 1  
## 4 0 1  
## 5 0 1  
## 6 0 1

fit <- lmFit(GSE19143.expr, designs)  
 contrast.matrix <- makeContrasts(Sensitive - Resistant, levels=designs)  
 contrast.matrix

## Contrasts  
## Levels Sensitive - Resistant  
## Resistant -1  
## Sensitive 1

fit2 <- contrasts.fit(fit, contrast.matrix)  
 fit2 <- eBayes(fit2)  
 tt <- topTable(fit2, sort.by = "p")  
 tt

## logFC AveExpr t P.Value adj.P.Val B  
## 209813\_x\_at -2.442641 4.908954 -4.849884 9.108073e-06 0.08778941 3.101527  
## 209374\_s\_at 2.005963 9.446631 4.720801 1.449816e-05 0.08778941 2.713316  
## 212827\_at 1.877708 9.542196 4.699004 1.567465e-05 0.08778941 2.648150  
## 207399\_at 1.389775 2.323241 4.697503 1.575899e-05 0.08778941 2.643668  
## 205297\_s\_at 1.442385 7.450054 4.485168 3.344075e-05 0.09738254 2.015267  
## 208068\_x\_at 1.648130 5.117845 4.462585 3.619610e-05 0.09738254 1.949144  
## 221976\_s\_at 1.497882 2.684540 4.457906 3.679391e-05 0.09738254 1.935464  
## 209153\_s\_at 1.543197 8.526154 4.431104 4.040831e-05 0.09738254 1.857213  
## 204928\_s\_at -1.595484 4.455320 -4.430349 4.051502e-05 0.09738254 1.855011  
## 215592\_at 1.895254 5.436788 4.315463 6.037197e-05 0.09738254 1.522004

tt.1 <- topTable(fit2, sort.by = "p", p.value = 0.1, number = nrow(GSE19143.expr))  
 nrow(tt.1)

## [1] 16

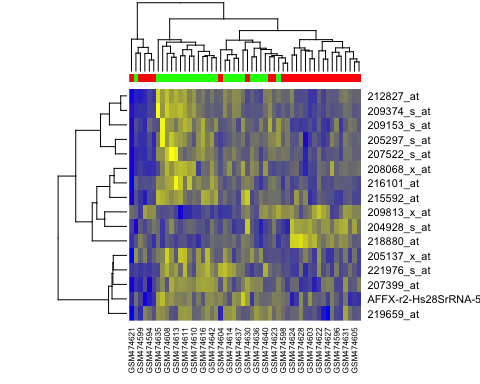
probe <- rownames(tt)[1]  
 m <- match(probe, rownames(GSE19143.expr))  
 df <- data.frame(expr = GSE19143.expr[m,], indv = indv)  
 means <- df %>% group\_by(indv) %>% summarize(mean = mean(expr))  
 means

## # A tibble: 2 x 2  
## indv mean  
## <fct> <dbl>  
## 1 clinical response factor: prednisolone resistant 6.18  
## 2 clinical response factor: prednisolone sensitive 3.73

diff(means$mean) #-2.442641

## [1] -2.442641

# Since the difference in means is less than zero, the expression is   
 # higher in those who are prednisolone resistant.   
   
########################################################################  
# 6. Construct a heatmap of these top 16 probes, with individuals   
# color-coded by response to prednisolone (with green=sensitive and   
# red = resistant). (Note: if you are unable to complete question 5),   
# you may do this with the first 16 probes in the expression matrix).  
########################################################################  
  
 m <- match(rownames(tt.1), rownames(GSE19143.expr))  
 X <- GSE19143.expr[m,]  
 col.heat <- colorRampPalette(c("yellow", "blue"))(200)  
 col.indv <- as.integer(as.factor(indv))  
 col.indv <- c("green", "red")[col.indv]  
  
 heatmap(X, ColSideColors = col.indv, col = col.heat)



########################################################################  
# 7. If you answered question 5 correctly, the SECOND hit   
# should be for the probe 209374\_s\_at. Show that this probe  
# corresponds to the gene IGHM, by first downloading the   
# correct platform data from GEO, and then finding the gene  
# associated with this probe.   
#######################################################################  
  
 platform <- annotation(GSE19143[[1]])   
 platform

## [1] "GPL96"

pl <- getGEO(platform)

## Using locally cached version of GPL96 found here:  
## /var/folders/dr/8c7c334x2yg5tc7lxlnj07kr0000gn/T//RtmpmBCDs7/GPL96.soft

## Warning: 68 parsing failures.  
## row col expected actual file  
## 22216 SPOT\_ID 1/0/T/F/TRUE/FALSE --Control literal data  
## 22217 SPOT\_ID 1/0/T/F/TRUE/FALSE --Control literal data  
## 22218 SPOT\_ID 1/0/T/F/TRUE/FALSE --Control literal data  
## 22219 SPOT\_ID 1/0/T/F/TRUE/FALSE --Control literal data  
## 22220 SPOT\_ID 1/0/T/F/TRUE/FALSE --Control literal data  
## ..... ....... .................. ......... ............  
## See problems(...) for more details.

pl <- Table(pl)  
 probe <- rownames(tt.1)[2]   
 m <- match(probe, pl$ID)  
 pl$`Gene Symbol`[m]

## [1] "IGHM"

#####################################################################  
# 8. How many probes are there for the gene IGHM on the platform  
# in this study? Note: you must search for this gene using the  
# regular expressions covered in the GEO-and-limma.R script. Your   
# code must also output the number of probes.   
####################################################################  
  
 m <- match(rownames(tt.1), pl$ID)  
 pl$`Gene Symbol`[m]

## [1] "TARP /// TRGC2 /// TRGV9" "IGHM"   
## [3] "IGHM" "BFSP2"   
## [5] "CD79B" "CSH1 /// CSHL1 /// GH1 /// GH2"  
## [7] "HDGFRP3" "TCF3"   
## [9] "SLC10A3" ""   
## [11] "ATP2A3" ""   
## [13] "USH1C" "FOSL2"   
## [15] "" "ATP8A2"

length(pl$`Gene Symbol`[m])

## [1] 16

########################################################################  
# Final Notes: the heatmap in question 6 provides a candidate list  
# of probes associated with prednisolone response in children with   
# leukemia. Although much additional work and testing would need to be   
# done, this kind of gene signature could ultimately be used to   
# determine whether a child with leukemia would benefit from   
# prednisolone treatment, or whether an alternative treatment might be   
# more effective.  
  
# The IGHM finding is also interesting. IGHM is a gene that codes  
# for an antibody protein involved in the immune reponse; the   
# fact that this gene is differentially expressed beween responders and   
# non-respnoders suggests that a patient's immune system may play a  
# role in how they respond to prednisolone)  
########################################################################