11DEGs.r

cris

2022-04-18

library(edgeR)

## Loading required package: limma

library(pheatmap)  
library(ggplot2)  
library(tidyverse)

## ── Attaching packages ─────────────────────────────────────── tidyverse 1.3.1 ──

## ✓ tibble 3.1.6 ✓ dplyr 1.0.8  
## ✓ tidyr 1.2.0 ✓ stringr 1.4.0  
## ✓ readr 2.1.2 ✓ forcats 0.5.1  
## ✓ purrr 0.3.4

## ── Conflicts ────────────────────────────────────────── tidyverse\_conflicts() ──  
## x dplyr::filter() masks stats::filter()  
## x dplyr::lag() masks stats::lag()

source("/home/cris/Documentos/EpiDiso/Disocactus\_transcriptome/bin/dif\_exp\_functions.R")  
  
#set workdir  
setwd("~/Documentos/Prosopis\_project/bin/")  
  
#Charge data  
count\_matrix<-read.table("../out/count\_matrix/prosopis\_count\_matrix.txt", header = TRUE,   
 stringsAsFactors = FALSE)  
#load metadata  
meta <- read.table("../metadata/meta.txt", header = T)  
  
#change colnames  
colnames(count\_matrix) <- c("Gene\_id","Chr","Start","End","Strand","Length",  
 "Ghaf12DT\_002\_CGATGT\_L007","Ghaf2DT\_005\_ACAGTG\_L007",  
 "Ghaf4DT\_006\_GCCAAT\_L007","Ghaf6DT\_007\_CAGATC\_L007",  
 "Ghaf8DT\_009\_GATCAG\_L007","PCDT3.10b","PCDT3.12","PCDT3.2b",  
 "PCDT3.4b","PCDT3.6","PCDT3.8","PCDT4.10b","PCDT5.4b",  
 "PCDT5.8b","PDT5\_10","PDT5\_2","PDT5\_6")  
  
#convert gen\_id in the rownames   
rownames(count\_matrix)<-count\_matrix[,1]  
  
count\_matrix<-count\_matrix[,-1]  
#select only colums whith counts  
count\_matrix<-count\_matrix%>%  
 select(., (6:22))  
  
class(count\_matrix)

## [1] "data.frame"

#explore matrix  
names(count\_matrix)

## [1] "Ghaf12DT\_002\_CGATGT\_L007" "Ghaf2DT\_005\_ACAGTG\_L007"   
## [3] "Ghaf4DT\_006\_GCCAAT\_L007" "Ghaf6DT\_007\_CAGATC\_L007"   
## [5] "Ghaf8DT\_009\_GATCAG\_L007" "PCDT3.10b"   
## [7] "PCDT3.12" "PCDT3.2b"   
## [9] "PCDT3.4b" "PCDT3.6"   
## [11] "PCDT3.8" "PCDT4.10b"   
## [13] "PCDT5.4b" "PCDT5.8b"   
## [15] "PDT5\_10" "PDT5\_2"   
## [17] "PDT5\_6"

#reoder columns as meta data is   
col\_order <- c("PCDT3.2b","PCDT3.4b","PCDT3.6", "PCDT3.8",   
 "PCDT3.10b","PCDT3.12","Ghaf2DT\_005\_ACAGTG\_L007", "Ghaf4DT\_006\_GCCAAT\_L007" ,   
 "Ghaf6DT\_007\_CAGATC\_L007","Ghaf8DT\_009\_GATCAG\_L007","Ghaf12DT\_002\_CGATGT\_L007",   
 "PCDT4.10b","PCDT5.4b","PCDT5.8b","PDT5\_10", "PDT5\_2","PDT5\_6")  
  
count\_matrix<- count\_matrix[, col\_order]  
  
#filter data by cpm  
keep <- rowSums(cpm(count\_matrix) >= 5) >=2  
table(keep)

## keep  
## FALSE TRUE   
## 54864 22354

count\_matrix <- count\_matrix[keep, ]  
  
#explore count matrix  
colnames(count\_matrix)

## [1] "PCDT3.2b" "PCDT3.4b"   
## [3] "PCDT3.6" "PCDT3.8"   
## [5] "PCDT3.10b" "PCDT3.12"   
## [7] "Ghaf2DT\_005\_ACAGTG\_L007" "Ghaf4DT\_006\_GCCAAT\_L007"   
## [9] "Ghaf6DT\_007\_CAGATC\_L007" "Ghaf8DT\_009\_GATCAG\_L007"   
## [11] "Ghaf12DT\_002\_CGATGT\_L007" "PCDT4.10b"   
## [13] "PCDT5.4b" "PCDT5.8b"   
## [15] "PDT5\_10" "PDT5\_2"   
## [17] "PDT5\_6"

groups <- factor(colnames(count\_matrix))  
table(groups)

## groups  
## Ghaf12DT\_002\_CGATGT\_L007 Ghaf2DT\_005\_ACAGTG\_L007 Ghaf4DT\_006\_GCCAAT\_L007   
## 1 1 1   
## Ghaf6DT\_007\_CAGATC\_L007 Ghaf8DT\_009\_GATCAG\_L007 PCDT3.10b   
## 1 1 1   
## PCDT3.12 PCDT3.2b PCDT3.4b   
## 1 1 1   
## PCDT3.6 PCDT3.8 PCDT4.10b   
## 1 1 1   
## PCDT5.4b PCDT5.8b PDT5\_10   
## 1 1 1   
## PDT5\_2 PDT5\_6   
## 1 1

#create edgeR list  
edgeRlist <- DGEList(counts = count\_matrix,  
 group = meta$month.Treatment,   
 genes = rownames(count\_matrix))  
str(edgeRlist)

## Formal class 'DGEList' [package "edgeR"] with 1 slot  
## ..@ .Data:List of 3  
## .. ..$ : int [1:22354, 1:17] 229 1920 2693 1275 1520 79 1096 755 154 143 ...  
## .. .. ..- attr(\*, "dimnames")=List of 2  
## .. .. .. ..$ : chr [1:22354] "KCPC\_00000005-RA" "KCPC\_00000006-RA" "KCPC\_00000013-RA" "KCPC\_00000015-RA" ...  
## .. .. .. ..$ : chr [1:17] "PCDT3.2b" "PCDT3.4b" "PCDT3.6" "PCDT3.8" ...  
## .. ..$ :'data.frame': 17 obs. of 3 variables:  
## .. .. ..$ group : Factor w/ 6 levels "month\_10","month\_12",..: 3 4 5 6 1 2 3 4 5 6 ...  
## .. .. ..$ lib.size : num [1:17] 41879743 43124317 40559418 32842486 34102995 ...  
## .. .. ..$ norm.factors: num [1:17] 1 1 1 1 1 1 1 1 1 1 ...  
## .. ..$ :'data.frame': 22354 obs. of 1 variable:  
## .. .. ..$ genes: chr [1:22354] "KCPC\_00000005-RA" "KCPC\_00000006-RA" "KCPC\_00000013-RA" "KCPC\_00000015-RA" ...  
## ..$ names: chr [1:3] "counts" "samples" "genes"

#Normalized count by TMM  
edgeRlist <- calcNormFactors(edgeRlist, method = "TMM")  
edgeRlist$samples

## group lib.size norm.factors  
## PCDT3.2b month\_2 41879743 0.8944485  
## PCDT3.4b month\_4 43124317 0.8330293  
## PCDT3.6 month\_6 40559418 1.2431947  
## PCDT3.8 month\_8 32842486 0.7856211  
## PCDT3.10b month\_10 34102995 0.7384695  
## PCDT3.12 month\_12 34887090 0.8875539  
## Ghaf2DT\_005\_ACAGTG\_L007 month\_2 30373184 1.2821889  
## Ghaf4DT\_006\_GCCAAT\_L007 month\_4 25640463 1.2275675  
## Ghaf6DT\_007\_CAGATC\_L007 month\_6 31010380 1.3678322  
## Ghaf8DT\_009\_GATCAG\_L007 month\_8 27172366 1.2617610  
## Ghaf12DT\_002\_CGATGT\_L007 month\_10 29775513 0.5684423  
## PCDT4.10b month\_12 38695465 1.0147384  
## PCDT5.4b month\_2 44846973 1.0151217  
## PCDT5.8b month\_4 37983556 1.1336470  
## PDT5\_10 month\_6 35384768 1.0500525  
## PDT5\_2 month\_8 33239856 0.8629483  
## PDT5\_6 month\_10 40184553 1.2831142

## Plot to evaluate the correct data normalization   
## Plot the results using absolute vs relative expression in each sample to check the correct normalization  
  
pdf("../figures/MD\_plots.pdf", height = 7, width = 10)  
par(mfrow = c(2, 3)) ##Generate a frame to store 6 plots in 2 rows and 3 columns  
for (i in c(1:17)) {  
 print(plotMD(cpm(edgeRlist, log = T), column = i))  
 grid(col = "blue")  
 abline(h = 0, col = "red", lty = 2, lwd = 2)  
}

## NULL

## NULL

## NULL

## NULL

## NULL

## NULL

## NULL

## NULL

## NULL

## NULL

## NULL

## NULL

## NULL

## NULL

## NULL

## NULL

## NULL

dev.off()

## png   
## 2

#Heatmap to explor data  
#calculating the correlation (Pearson) that exists between the samples  
pdf("../figures/corr\_rep\_plots.pdf", height = 7, width = 10)  
cormat <- cor(cpm(edgeRlist$counts, log = T))  
pheatmap(cormat, border\_color = NA, main = "P. cineraria correlation of replicates")  
dev.off()

## pdf   
## 3

#In order to check the correct normalization of the samples we repeat the boxplot  
# Get log2 counts per million  
jpeg("../figures/boxplot\_logCPMs\_norm.jpg")  
logcounts <- cpm(edgeRlist,log=TRUE)  
# Check distributions of samples using boxplots  
boxplot(logcounts, xlab="", ylab="Log2 counts per million",las=2)  
# Let's add a blue horizontal line that corresponds to the median logCPM  
abline(h=median(logcounts),col="blue")  
title("transformed logCPMs")  
dev.off()

## pdf   
## 3

#Experiimental matrix design  
design <- model.matrix(~0+edgeRlist$samples$group)  
design

## edgeRlist$samples$groupmonth\_10 edgeRlist$samples$groupmonth\_12  
## 1 0 0  
## 2 0 0  
## 3 0 0  
## 4 0 0  
## 5 1 0  
## 6 0 1  
## 7 0 0  
## 8 0 0  
## 9 0 0  
## 10 0 0  
## 11 1 0  
## 12 0 1  
## 13 0 0  
## 14 0 0  
## 15 0 0  
## 16 0 0  
## 17 1 0  
## edgeRlist$samples$groupmonth\_2 edgeRlist$samples$groupmonth\_4  
## 1 1 0  
## 2 0 1  
## 3 0 0  
## 4 0 0  
## 5 0 0  
## 6 0 0  
## 7 1 0  
## 8 0 1  
## 9 0 0  
## 10 0 0  
## 11 0 0  
## 12 0 0  
## 13 1 0  
## 14 0 1  
## 15 0 0  
## 16 0 0  
## 17 0 0  
## edgeRlist$samples$groupmonth\_6 edgeRlist$samples$groupmonth\_8  
## 1 0 0  
## 2 0 0  
## 3 1 0  
## 4 0 1  
## 5 0 0  
## 6 0 0  
## 7 0 0  
## 8 0 0  
## 9 1 0  
## 10 0 1  
## 11 0 0  
## 12 0 0  
## 13 0 0  
## 14 0 0  
## 15 1 0  
## 16 0 1  
## 17 0 0  
## attr(,"assign")  
## [1] 1 1 1 1 1 1  
## attr(,"contrasts")  
## attr(,"contrasts")$`edgeRlist$samples$group`  
## [1] "contr.treatment"

##the term ~0 tells the function not to include a column of intersections and only include as many columns as groups in our experimental design  
colnames(design) <- levels(edgeRlist$samples$group)  
  
#explore design  
design

## month\_10 month\_12 month\_2 month\_4 month\_6 month\_8  
## 1 0 0 1 0 0 0  
## 2 0 0 0 1 0 0  
## 3 0 0 0 0 1 0  
## 4 0 0 0 0 0 1  
## 5 1 0 0 0 0 0  
## 6 0 1 0 0 0 0  
## 7 0 0 1 0 0 0  
## 8 0 0 0 1 0 0  
## 9 0 0 0 0 1 0  
## 10 0 0 0 0 0 1  
## 11 1 0 0 0 0 0  
## 12 0 1 0 0 0 0  
## 13 0 0 1 0 0 0  
## 14 0 0 0 1 0 0  
## 15 0 0 0 0 1 0  
## 16 0 0 0 0 0 1  
## 17 1 0 0 0 0 0  
## attr(,"assign")  
## [1] 1 1 1 1 1 1  
## attr(,"contrasts")  
## attr(,"contrasts")$`edgeRlist$samples$group`  
## [1] "contr.treatment"

#calculate data dispesion  
jpeg("../figures/data\_dispersion.jpg")  
edgeRlist <- estimateDisp(edgeRlist, design = design, robust = T)  
plotBCV(edgeRlist)  
dev.off()

## pdf   
## 3

#estimation of QL dispersions  
#Data must fit a negative bi-nominal linear model. For this, the glmQLfit function will be used with which there is a more robust control of the error  
fit <- glmQLFit(edgeRlist, design, robust=TRUE)  
head(fit$coefficients)

## month\_10 month\_12 month\_2 month\_4 month\_6  
## KCPC\_00000005-RA -12.230858 -12.049787 -12.347226 -12.246639 -12.418340  
## KCPC\_00000006-RA -9.992375 -9.838976 -10.000756 -9.977719 -10.154735  
## KCPC\_00000013-RA -9.790987 -9.651588 -9.376311 -9.215017 -9.431225  
## KCPC\_00000015-RA -9.649078 -9.953982 -9.973272 -10.068398 -9.965221  
## KCPC\_00000019-RA -10.368387 -10.275671 -9.959031 -10.409000 -10.292389  
## KCPC\_00000020-RA -11.460259 -12.564344 -12.029589 -11.988838 -11.526421  
## month\_8  
## KCPC\_00000005-RA -12.313240  
## KCPC\_00000006-RA -10.006816  
## KCPC\_00000013-RA -9.679837  
## KCPC\_00000015-RA -9.804698  
## KCPC\_00000019-RA -10.282494  
## KCPC\_00000020-RA -11.507426

#Plot QL dispersion using fit object  
jpeg("../figures/QL\_disp.jpg")  
plotQLDisp(fit, main = " Quasi Likelihood dispersion in P.cinerase")  
dev.off()

## pdf   
## 3

#Contrast matrix for 3 comparisons   
#Since we are interested in differences between groups, we need to specify which comparisons we want to test.  
contrast\_matrix <- makeContrasts(  
 "PC\_2vsPC\_4" = "month\_2 - month\_4",  
 "PC\_4vsPC\_6" = "month\_4 - month\_6",   
 "PC\_6vsPC\_8" = "month\_6 - month\_8",  
 "PC\_8vsPC\_10" = "month\_8 - month\_10",levels=design)  
  
  
contrast\_matrix

## Contrasts  
## Levels PC\_2vsPC\_4 PC\_4vsPC\_6 PC\_6vsPC\_8 PC\_8vsPC\_10  
## month\_10 0 0 0 -1  
## month\_12 0 0 0 0  
## month\_2 1 0 0 0  
## month\_4 -1 1 0 0  
## month\_6 0 -1 1 0  
## month\_8 0 0 -1 1

#Create the object contr\_leves  
contr\_levels<-attributes(contrast\_matrix)$dimnames$Contrasts  
  
#Adjust data to Binomial (BN) method and generate volcano plots for every comparisson  
pdf("../figures/prosopis\_volcanos.pdf", height = 7, width = 10)  
par(mfrow = c(2, 3)) ##Generate a frame to store 3 plots in 2 rows and 3 columns  
dif\_exp\_results<-data.frame() #Empty data frame  
for (i in c(1:4)) {  
 qlf.BvsA.lfc1 <- glmTreat(fit, ##Object in list form with data fitted to a negative bi-nominal model  
 contrast = contrast\_matrix[, i],   
 lfc = 1)  
 ##We obtain the SDRs with an expression other than 1, p value less than 0.05, correcting the p value by the Benjamini-Hochberg method  
 deg.BvsA.lfc1 <- decideTestsDGE(qlf.BvsA.lfc1, p.value = 0.05, adjust.method = "BH", lfc = 1)  
 table(deg.BvsA.lfc1)  
 #select the genes that statistically have | lfc | > 1 and strengthen our results  
 DEG.BvsA.lfc1 <- DEGResults(qlf.BvsA.lfc1)  
 DEG.BvsA.lfc1 <- edgeResults(DEG.BvsA.lfc1, logfc = 1, padj = 0.05)  
 comparacion<-data.frame(comparacion = rep(contr\_levels[i], times = nrow(DEG.BvsA.lfc1)))  
 DEG.BvsA.lfc1 <-cbind(comparacion, DEG.BvsA.lfc1)  
 dif\_exp\_results<-rbind(dif\_exp\_results, DEG.BvsA.lfc1)  
 print(volcano\_edgeR(DEG.BvsA.lfc1, lfc = 1, padj = 0.05) + #print volcano\_plots  
 labs(title = contr\_levels[i]) + #labs of comparissons  
 xlim(-15, 15) +  
 ylim(0, 10))  
   
}  
dev.off()

## pdf   
## 3