```
################################################################################
##### RNA-SEQ DATA ANALYSIS - PSEUDOMONAS S5 - MAIN FIGURES OF THE REPORT
################################################################################
##### Spring 2016 - MLS - UNIL - Marie Zufferey
##### !!! some hard-coded parameters, file shape and formats not checked
rm(list=ls())
setwd("PATH_TO_DIRECTORY")
outfolder = "YOUR_OUTFOLDER"
system(paste("rm -rf", outfolder))
system(paste("mkdir", outfolder)) #not overwritten if already existing
source("functions_4.R")
library(edgeR)
library(readr)
library(ggplot2)
library(pheatmap)
library(reshape2)
library(rtracklayer)
library(magrittr)
library(dplyr)
library(VennDiagram)

#**********************************************
# DATA PREPARATION
#**********************************************

annot <- read.csv("../data/annot_mot.csv", sep=",")
annot$Gene_position!="S5_genome_1619"
annot <- annot[-which(annot$Gene_position=="S5_genome_4011"),]
annot <- annot[-which(annot$Gene_position=="S5_genome_1619"),]
rawannot <- read.csv("../data/annot_mot.csv", sep=",")
rawannot <- annot[-which(rawannot$Gene_position=="S5_genome_1619"),]
S5_stat <- read.csv("../data/Pseud_S5_stat.txt", sep="\t")
gbkData <- read.csv("../data/S5_gbk_short.csv", sep=",")
abd_fld <- "../data/abundances/"
dt <- getDGE(abd_fld)  # compute it once here # normalized for dispersion !!!
# dt after  estimateTagwiseDisp(dt)  !!!!


##### Manual curation motility genes
a <- as.character(gbkData$Locus_tag[which(
  regexpr("pilus|motility|mobility|flagella|swarming|flagellum|pili", gbkData$Function)>0)])
all(a %in% annot$Gene_position) # TRUE -> ok
b <- as.character(gbkData$Locus_tag[which(
  regexpr("pilus|motility|mobility|flagella|swarming|flagellum|pili", gbkData$Product)>0)])
all(b %in% annot$Gene_position) # F
b[which(! b %in% annot$Gene_position)]

gbkData[gbkData$Locus_tag %in% b[which(! b %in% annot$Gene_position)],]

# Type Strand   Start     End      Locus_tag Gene_id                                Product Function
# 445   CDS      +  477446  478882  S5_genome_522       0            pilus assembly protein
PilQ        0
# 1042  CDS      - 1145050 1145361 S5_genome_1109       0 motility quorum-sensing regulator
MqsR        0
# 2060  CDS      - 2236727 2237314 S5_genome_2116       0            pilus assembly protein
PilZ        0
# 2071  CDS      - 2246411 2246710 S5_genome_2127       0                 pilus assembly
protein        0
# 3974  CDS      - 4407774 4408310 S5_genome_4013       0            type I pilus protein CsuA/
B        0
# 4334  CDS      + 4831767 4832201 S5_genome_4365       0            pilus assembly protein
PilZ        0
# 4759  CDS      - 5275681 5276040 S5_genome_4781       0            pilus assembly protein
PilZ        0

##### Manual curation chemotaxis
c <- grep("che", gbkData$Gene_id) # 5
c[which(! gbkData$Locus_tag[c] %in% annot$Gene_position)]   #5
gbkData[c,]
# Type Strand   Start     End      Locus_tag Gene_id
Product                                 Function
# 1123  CDS      + 1242569 1243579 S5_genome_1190   cheB2 Chemotaxis response regulator protein-
glutamate Involved in the modulation of the chemotaxis
# 1761  CDS      + 1915176 1915547 S5_genome_1824    cheY                         Chemotaxis protein
```

```
CheY      Involved in the transmission of sensory
# 1762  CDS      + 1915578 1916366 S5_genome_1825    cheZ                        Protein phosphatase
CheZ         Plays an important role in bacterial
# 1764  CDS      + 1918694 1919809 S5_genome_1827   cheB1 Chemotaxis response regulator protein-
glutamate Involved in the modulation of the chemotaxis
# 4546  CDS      - 5056065 5056892 S5_genome_4573    cheR            Chemotaxis protein
methyltransferase            Methylation of the membrane-bound

## Done manually
# colnames(annot): Gene_position Gene_name Motility_type
# we do not add the S5_genome_1109 and S5_genome_4013
addAnnot <- read.table(textConnection("
S5_genome_522 pilQ  pili
S5_genome_2116  pilZ  pili
S5_genome_2127  no_name2127 pili
S5_genome_4365  pilZ  pili
S5_genome_4781  pilZ  pili"), header=F)
colnames(addAnnot) <- c("Gene_position", "Gene_name", "Motility_type")
annot <- read.csv("../data/annot_mot.csv", sep=",")
annot <- annot[-which(annot$Gene_position=="S5_genome_4011"),]
annot <- annot[-which(annot$Gene_position=="S5_genome_1619"),]

annot <- rbind(annot, addAnnot)

addAnnot_c <- read.table(textConnection("
S5_genome_1190  cheB2 chemotaxis
S5_genome_1824  cheY chemotaxis
S5_genome_1825  cheZ chemotaxis
S5_genome_1827  cheB1 chemotaxis
S5_genome_4573  cheR chemotaxis"), header=F)
colnames(addAnnot_c) <- c("Gene_position", "Gene_name", "Motility_type")
annot_chemo <- rbind(annot, addAnnot_c)

# Pairwise comparisons
# exact test for the 2 conditions passed in argument (last 2 arguments)
# for a given set of genes (2nd argument)
dataLMSA <- pairTestGenes(dt, annot$Gene_position , "LM", "SA")  #1
dataLMWL <- pairTestGenes(dt, annot$Gene_position , "LM", "WL")  #2
dataLMWR <- pairTestGenes(dt, annot$Gene_position , "LM", "WR")  #3
dataSAWL <- pairTestGenes(dt, annot$Gene_position , "SA", "WL")  #4
dataSAWR <- pairTestGenes(dt, annot$Gene_position , "SA", "WR")  #5
dataSALM <- pairTestGenes(dt, annot$Gene_position , "SA", "LM")  #1b
dataWLWR <- pairTestGenes(dt, annot$Gene_position , "WL", "WR")  #6
dataWLSA <- pairTestGenes(dt, annot$Gene_position , "WL", "SA")  #4b
dataWLLM <- pairTestGenes(dt, annot$Gene_position , "WL", "LM")  #4b


#**********************************************
# MATRIX OF PLOTS
#**********************************************
# First we do the matrix with all pairs of conditions
# it will allow us to justify which pairs we choose
# before merging, select only needed data
# (not mandatory)
subLMSA <- dataLMSA[,c("logFC", "FDR", "Transcript")]      #1
colnames(subLMSA)[1:2] %<>% paste0(., ".LMSA")
subLMWL <- dataLMWL[,c("logFC", "FDR", "Transcript")]      #2
colnames(subLMWL)[1:2] %<>% paste0(., ".LMWL")
subLMWR <- dataLMWR[,c("logFC", "FDR", "Transcript")]      #3
colnames(subLMWR)[1:2] %<>% paste0(., ".LMWR")
subSAWL <- dataSAWL[,c("logFC", "FDR", "Transcript")]      #4
colnames(subSAWL)[1:2] %<>% paste0(., ".SAWL")
subSAWR <- dataSAWR[,c("logFC", "FDR", "Transcript")]      #5
colnames(subSAWR)[1:2] %<>% paste0(., ".SAWR")
subWLWR <- dataWLWR[,c("logFC", "FDR", "Transcript")]      #6
colnames(subWLWR)[1:2] %<>% paste0(., ".WLWR")

# merge all in a single DF
allJoins <- full_join(subLMSA, subLMWL, by="Transcript") %>%   #1,2
  full_join(., subLMWR, by="Transcript")  %>% #3
  full_join(., subSAWL, by="Transcript")  %>% #4
  full_join(., subSAWR, by="Transcript")  %>% #5
  full_join(., subWLWR, by="Transcript")  #6
```

```
# convert into a matrix with only logFC values
matAllJoins <- allJoins
rownames(matAllJoins) <- matAllJoins$Transcript
matAllJoins <- matAllJoins[,grep("log", colnames(matAllJoins))]
# change the colnames for nicer titles in the matrix plot
colnames(matAllJoins) %<>% gsub("logFC.", "",.) %<>%
  gsub('(^.{2})(.{2}$)', '\\2 vs. \\1', .)

png(paste0(outfolder,"/scatterplotMatrix_all.png"))
pairs(matAllJoins,panel=panel.smooth, upper.panel=panel.cor,
      diag.panel=panel.hist)  # panel.hist defined in functions_4.R
title("Log2FC for motility associated genes - all pairs", line=3)
dev.off()

# => we choose cond1=LM, cond2=SA
# => and cond1=SA, cond2=WR

#*********************************************
# VOLCANO PLOTS WITH ALL DATA
#*********************************************
# with label for the top 5 genes
png(paste0(outfolder,"/volcanoAll_LMSA.png"))
volcanoAllPoints(dt, annot, "LM", "SA", plotAnnot=T, myT=3)  %>% plot
dev.off()
png(paste0(outfolder,"/volcanoAll_SAWR.png"))
volcanoAllPoints(dt, annot, "SA", "WR")  %>% plot
dev.off()

#*********************************************
# VOLCANO PLOTS WITH MOTILITY ASSOCIATED GENES
#*********************************************

svg(paste0(outfolder,"/volcanoMot_LMSA.svg"))
volcanoMotilityPointsAnnot(dt, annot, "LM", "SA")
dev.off()

svg(paste0(outfolder,"/volcanoMot_SAWR.svg"))
volcanoMotilityPointsAnnot(dt, annot, "SA", "WR")
dev.off()

#*********************************************
# BOX PLOT FOR MOTILITY ASSOCIATED GENES
#*********************************************
dt_raw <- getRawData(abd_fld)
all_data <- dt_raw$counts %>% as.data.frame  #6087
all_data$Tra <- rownames(all_data)

mot_data <-left_join(annot_chemo, all_data, by=c("Gene_position"="Tra")) %>%
  left_join(., S5_stat, by=c("Gene_position"="Seq_tag"))

# WITH OWN DEFINED "MYRPKM" ***********************
motRP <- myrpkm(mot_data[,grep("1|2|3|4", colnames(mot_data))], mot_data$Length)

# WITH edgeR "RPKM" ***********************
# get DGE object
dt <- getDGE(abd_fld)
dt <- calcNormFactors(dt)
temp <- dt$counts

# retrieve the length
getN <- temp
getN %<>% as.data.frame
getN$Tr <- rownames(getN)
getN <- left_join(getN, S5_stat, by=c("Tr"="Seq_tag"))

# rpkm
temp <- rpkm(temp, getN$Length)
temp %<>% as.data.frame
#temp$Tr <- rownames(temp)
temp <- temp[which(rownames(temp) %in% mot_data$Gene_position),]
temp <- temp[match(mot_data$Gene_position, rownames(temp)),]
motRP <- temp

# we want to compare across genes and across conditions -> RPKM
```

```
#
mot_data[,grep("1|2|3|4", colnames(mot_data))]  <- motRP

# take the mean of the replicates for all conditions
#mot_data2 <- cbind(mot_data[,1:3], getMeanData(mot_data))
mot_data2 <- mot_data[,1:(ncol(mot_data)-3)]

# select only motility genes (without chemotaxis)
data_mot <- mot_data2[which(mot_data2$Motility_type!="chemotaxis"),]

png(paste0(outfolder,"/boxplot_Mot.png"))
boxplotMotGenes(data_mot, annot, "Global expression mot. genes", chemo=F)
dev.off()

png(paste0(outfolder,"/boxplot_withChem.png"))
boxplotMotGenes(mot_data2, annot, "Global expression mot. genes (with chemo.)", chemo=T)
dev.off()

#********************************************
# LINE PLOTS WITH MOTILITY ASSOCIATED GENES
#********************************************
#************ cond1=SA, cond2=WL
# Draw it for SAWL, motility associated genes
dataSAWR <- pairTestGenes(dt, annot$Gene_position , "SA", "WR")  #1
tit <- "log 2 FC (WR vs. SA - motility associated genes)"
png(paste0(outfolder,"/2axis_SAWR_mot.png"))
fc_barAndCpm_line(dataSAWR, annot, gbkData, tit, pt=F) %>% grid.draw
dev.off()

dataLMSA <- pairTestGenes(dt, annot$Gene_position , "LM", "SA")  #1
tit <- "log 2 FC (SA vs. LM - motility associated genes)"
png(paste0(outfolder,"/2axis_LMSA_mot.png"))
fc_barAndCpm_line(dataLMSA, annot, gbkData, tit,pt=F) %>% grid.draw
dev.off()
```