#stringApp\_02\_GO\_KEGG\_top\_proteinlist.R

#5\_GO\_KEGG\_topに含まれるDEP list作成

#置換

#https://id.fnshr.info/2017/10/16/trim-string/

##############################################################

rm(list = ls(all.names = TRUE))

setwd("/Users/akira\_yoshimi/Dropbox/0\_Work/R/Perseus\_Like\_Analysis/AMY/2\_Cytoscape") #作業ディレクトリ設定

#setwd("/Users/akira\_yoshimi/Dropbox/0\_Work/R/Perseus\_Like\_Analysis/HIP/2\_Cytoscape") #作業ディレクトリ設定

#setwd("/Users/akira\_yoshimi/Dropbox/0\_Work/R/Perseus\_Like\_Analysis/NAc/2\_Cytoscape") #作業ディレクトリ設定

#setwd("/Users/akira\_yoshimi/Dropbox/0\_Work/R/Perseus\_Like\_Analysis/PFC/2\_Cytoscape") #作業ディレクトリ設定

#setwd("/Users/akira\_yoshimi/Dropbox/0\_Work/R/Perseus\_Like\_Analysis/STR/2\_Cytoscape") #作業ディレクトリ設定

getwd()#作業ディレクトリ確認

dir() #作業ディレクトリ内のファイル表示

options(digits=2) #change digit2桁表示指定

#############################################################

library(EnhancedVolcano)

library(magrittr)

library(tidyverse) #ライブラリtidyverse(ggplot2,dplyr),gcookbook読み込み

library(dplyr)

library(scales) #muted()関数使用のため

library(gridExtra) #svg出力のため

library(rJava)

library(readxl) #エクセル読み込み

library(openxlsx) #JAVA不使用で大きなデータも読み込める

library(cowplot)

##############################################################

### #5\_GO\_KEGG\_topに含まれるDEP list作成 ###

##############################################################

#xlsx入力

data1 <- read\_excel("GO\_KEGG.xlsx", 1) #シート1(GO)入力

data2 <- read\_excel("GO\_KEGG.xlsx", 5) #シート4(KEGG)入力

#置換

#grep("^genes$|nodes.SUID", colnames(data1))

data1$genes <- gsub("\\|", ",", data1$genes)

data1$nodes.SUID <- gsub("\\|", ",", data1$nodes.SUID)

data2$genes <- gsub("\\|", ",", data2$genes)

data2$nodes.SUID <- gsub("\\|", ",", data2$nodes.SUID)

##############################################################

#各群のデータフレーム作成

dat7 <- filter(data1, group == "PCP") #PCPのtop,NA除外：!is.na(`genes`)

dat8 <- filter(data1, group == "CLZ") #CLZのtop

dat9 <- filter(data1, group == "Interaction") #Interactionのtop

dat10 <- filter(data2, group == "PCP") #PCPのtop

dat11 <- filter(data2, group == "CLZ") #CLZのtop

dat12 <- filter(data2, group == "Interaction") #Interactionのtop

##############################################################

#行の抽出(CNS, Signaling,EndcrineMetabolism)

dat7[,grep("description", colnames(dat7))] #5列目(description)表示

dat8[,grep("description", colnames(dat8))] #5列目(description)表示

dat9[,grep("description", colnames(dat9))] #5列目(description)表示

dat10[,grep("description", colnames(dat10))] #5列目(description)表示

dat11[,grep("description", colnames(dat11))] #5列目(description)表示

dat12[,grep("description", colnames(dat12))] #5列目(description)表示

#GOindex

GO\_cns <- grep("myelin|synap|axon|cortex|Parkinson|addiction|Amyotrophic",dat7$description) #CNS-related

GO\_em <- grep("metabol|energy|Glycolysis|Citrate|synthesis|leucine|Fatty|Insulin",dat7$description) #Metabolic,energy-related

GO\_cv <- grep("cvdiac|Cardi|cardi|Cardio|cardio",dat7$description) #Cardio-vascular-related

GO\_sgn <- grep("signaling",dat7$description) #Signaling-related

#KEGGindex

KEGG\_cns <- grep("myelin|synap|axon|cortex|Parkinson|addiction|Amyotrophic",dat10$description) #CNS-related

KEGG\_em <- grep("metabol|energy|Glycolysis|Citrate|synthesis|leucine|Fatty|Insulin",dat10$description) #Energy,metabolic-related

KEGG\_cv <- grep("Cardiac|Cardi|cardi|Cardio|cardio",dat10$description) #Cardio-vascular-related

KEGG\_sgn <- grep("signaling",dat10$description) #Signaling-related

##############################################################

t(names(dat7)) #行列の名前確認

t(names(dat8)) #行列の名前確認

t(names(dat9)) #行列の名前確認

t(names(dat10)) #行列の名前確認

t(names(dat11)) #行列の名前確認

t(names(dat12)) #行列の名前確認

#CNS-related

cns7 <- dat7 %>% slice(GO\_cns) %>% select(grep("^genes$",colnames(dat7)))

cns8 <- dat8 %>% slice(GO\_cns) %>% select(grep("^genes$",colnames(dat8)))

cns9 <- dat9 %>% slice(GO\_cns) %>% select(grep("^genes$",colnames(dat9)))

cns10 <- dat10 %>% slice(KEGG\_cns) %>% select(grep("^genes$",colnames(dat10)))

cns11 <- dat11 %>% slice(KEGG\_cns) %>% select(grep("^genes$",colnames(dat11)))

cns12 <- dat12 %>% slice(KEGG\_cns) %>% select(grep("^genes$",colnames(dat12)))

#Energy,metabolic-related

em7 <- dat7 %>% slice(GO\_em) %>% select(grep("^genes$",colnames(dat7)))

em8 <- dat8 %>% slice(GO\_em) %>% select(grep("^genes$",colnames(dat8)))

em9 <- dat9 %>% slice(GO\_em) %>% select(grep("^genes$",colnames(dat9)))

em10 <- dat10 %>% slice(KEGG\_em) %>% select(grep("^genes$",colnames(dat10)))

em11 <- dat11 %>% slice(KEGG\_em) %>% select(grep("^genes$",colnames(dat11)))

em12 <- dat12 %>% slice(KEGG\_em) %>% select(grep("^genes$",colnames(dat12)))

#Cardio-vascular-related

cv7 <- dat7 %>% slice(GO\_cv) %>% select(grep("^genes$",colnames(dat7)))

cv8 <- dat8 %>% slice(GO\_cv) %>% select(grep("^genes$",colnames(dat8)))

cv9 <- dat9 %>% slice(GO\_cv) %>% select(grep("^genes$",colnames(dat9)))

cv10 <- dat10 %>% slice(KEGG\_cv) %>% select(grep("^genes$",colnames(dat10)))

cv11 <- dat11 %>% slice(KEGG\_cv) %>% select(grep("^genes$",colnames(dat11)))

cv12 <- dat12 %>% slice(KEGG\_cv) %>% select(grep("^genes$",colnames(dat12)))

#Signalrelated

sgn7 <- dat7 %>% slice(GO\_sgn) %>% select(grep("^genes$",colnames(dat7)))

sgn8 <- dat8 %>% slice(GO\_sgn) %>% select(grep("^genes$",colnames(dat8)))

sgn9 <- dat9 %>% slice(GO\_sgn) %>% select(grep("^genes$",colnames(dat9)))

sgn10 <- dat10 %>% slice(KEGG\_sgn) %>% select(grep("^genes$",colnames(dat10)))

sgn11 <- dat11 %>% slice(KEGG\_sgn) %>% select(grep("^genes$",colnames(dat11)))

sgn12 <- dat12 %>% slice(KEGG\_sgn) %>% select(grep("^genes$",colnames(dat12)))

#列抽出

t(names(dat7)) #行列の名前確認

dat7 <- dat7 %>% select(grep("^genes$",colnames(dat7)))

dat8 <- dat8 %>% select(grep("^genes$",colnames(dat8)))

dat9 <- dat9 %>% select(grep("^genes$",colnames(dat9)))

dat10 <- dat10 %>% select(grep("^genes$",colnames(dat10)))

dat11 <- dat11 %>% select(grep("^genes$",colnames(dat11)))

dat12 <- dat12 %>% select(grep("^genes$",colnames(dat12)))

#縦に結合

GOcns <- rbind(cns7, cns8, cns9)

GOem <- rbind(em7, em8, em9)

GOcv <- rbind(cv7, cv8, cv9)

GOsgn <- rbind(sgn7, sgn8, sgn9)

KEGGcns <- rbind(cns10, cns11, cns12)

KEGGem <- rbind(em10, em11, em12)

KEGGcv <- rbind(cv10, cv11, cv12)

KEGGsgn<- rbind(sgn10, sgn11, sgn12)

GO <- rbind(dat7, dat8, dat9)

KEGG <- rbind(dat10, dat11, dat12)

GO\_KEGG <- rbind(GO, KEGG)

##############################################################

#関数作成

split\_into\_multiple <- function(column, pattern = ",", into\_prefix){

cols <- str\_split\_fixed(column, pattern, n = Inf)

# Sub out the ""'s returned by filling the matrix to the right, with NAs which are useful

cols[which(cols == "")] <- NA

cols <- as.tibble(cols)

# name the 'cols' tibble as 'into\_prefix\_1', 'into\_prefix\_2', ..., 'into\_prefix\_m'

# where m = # columns of 'cols'

m <- dim(cols)[2]

names(cols) <- paste(into\_prefix, 1:m, sep = "\_")

return(cols)

}

##############################################################

#gene list抽出

dat7 <- dat7 %>%

bind\_cols(split\_into\_multiple(.$genes, ",", "gene1")) %>%

# selecting those that start with 'gene\_' will remove the original 'AssociatedGenesFound' column

# select(AssociatedGenesFound, starts\_with("type\_"))

gather(key, value = gene1, -genes, na.rm = T) %>% #geneを1列にまとめる

#select(gene) %>% #gene listのみ抽出, 省略可能

distinct(gene1,.keep\_all=FALSE) #gene重複削除しgeneのみ返す

dat8 <- dat8 %>%

bind\_cols(split\_into\_multiple(.$genes, ",", "gene2")) %>%

gather(key, value = gene2, -genes, na.rm = T) %>%

distinct(gene2,.keep\_all=FALSE)

dat9 <- dat9 %>%

bind\_cols(split\_into\_multiple(.$genes, ",", "gene3")) %>%

gather(key, value = gene3, -genes, na.rm = T) %>%

distinct(gene3,.keep\_all=FALSE)

dat10 <- dat10 %>%

bind\_cols(split\_into\_multiple(.$genes, ",", "gene4")) %>%

gather(key, value = gene4, -genes, na.rm = T) %>%

distinct(gene4,.keep\_all=FALSE)

dat11 <- dat11 %>%

bind\_cols(split\_into\_multiple(.$genes, ",", "gene5")) %>%

gather(key, value = gene5, -genes, na.rm = T) %>%

distinct(gene5,.keep\_all=FALSE)

dat12 <- dat12 %>%

bind\_cols(split\_into\_multiple(.$genes, ",", "gene6")) %>%

gather(key, value = gene6, -genes, na.rm = T) %>%

distinct(gene6,.keep\_all=FALSE)

##############################################################

GOcns <- GOcns %>%

bind\_cols(split\_into\_multiple(.$genes, ",", "GOcnsgene")) %>%

gather(key, value = GOcnsgene, -genes, na.rm = T) %>%

distinct(GOcnsgene,.keep\_all=FALSE)

GOcv <- GOcv %>%

bind\_cols(split\_into\_multiple(.$genes, ",", "GOcvgene")) %>%

gather(key, value = GOcvgene, -genes, na.rm = T) %>%

distinct(GOcvgene,.keep\_all=FALSE)

GOem <- GOem %>%

bind\_cols(split\_into\_multiple(.$genes, ",", "GOemgene")) %>%

gather(key, value = GOemgene, -genes, na.rm = T) %>%

distinct(GOemgene,.keep\_all=FALSE)

GOsgn <- GOsgn %>%

bind\_cols(split\_into\_multiple(.$genes, ",", "GOsgngene")) %>%

gather(key, value = GOsgngene, -genes, na.rm = T) %>%

distinct(GOsgngene,.keep\_all=FALSE)

KEGGcns <- KEGGcns %>%

bind\_cols(split\_into\_multiple(.$genes, ",", "KEGGcnsgene")) %>%

gather(key, value = KEGGcnsgene, -genes, na.rm = T) %>%

distinct(KEGGcnsgene,.keep\_all=FALSE)

KEGGcv <- KEGGcv %>%

bind\_cols(split\_into\_multiple(.$genes, ",", "KEGGcvgene")) %>%

gather(key, value = KEGGcvgene, -genes, na.rm = T) %>%

distinct(KEGGcvgene,.keep\_all=FALSE)

KEGGem <- KEGGem %>%

bind\_cols(split\_into\_multiple(.$genes, ",", "KEGGemgene")) %>%

gather(key, value = KEGGemgene, -genes, na.rm = T) %>%

distinct(KEGGemgene,.keep\_all=FALSE)

KEGGsgn <- KEGGsgn %>%

bind\_cols(split\_into\_multiple(.$genes, ",", "KEGGsgngene")) %>%

gather(key, value = KEGGsgngene, -genes, na.rm = T) %>%

distinct(KEGGsgngene,.keep\_all=FALSE)

GO <- GO %>%

bind\_cols(split\_into\_multiple(.$genes, ",", "GOgene")) %>%

gather(key, value = GOgene, -genes, na.rm = T) %>%

distinct(GOgene,.keep\_all=FALSE)

KEGG <- KEGG %>%

bind\_cols(split\_into\_multiple(.$genes, ",", "KEGGgene")) %>%

gather(key, value = KEGGgene, -genes, na.rm = T) %>%

distinct(KEGGgene,.keep\_all=FALSE)

GO\_KEGG <- GO\_KEGG %>%

bind\_cols(split\_into\_multiple(.$genes, ",", "GO\_KEGGgene")) %>%

gather(key, value = GO\_KEGGgene, -genes, na.rm = T) %>%

distinct(GO\_KEGGgene,.keep\_all=FALSE)

##############################################################

#通し番号列追加

dat7 <- dat7 %>% mutate(num = row\_number())

dat8 <- dat8 %>% mutate(num = row\_number())

dat9 <- dat9 %>% mutate(num = row\_number())

dat10 <- dat10 %>% mutate(num = row\_number())

dat11 <- dat11 %>% mutate(num = row\_number())

dat12 <- dat12 %>% mutate(num = row\_number())

GOcns <- GOcns %>% mutate(num = row\_number())

GOcv <- GOcv %>% mutate(num = row\_number())

GOem <- GOem %>% mutate(num = row\_number())

GOsgn <- GOsgn %>% mutate(num = row\_number())

KEGGcns <- KEGGcns %>% mutate(num = row\_number())

KEGGcv <- KEGGcv %>% mutate(num = row\_number())

KEGGem <- KEGGem %>% mutate(num = row\_number())

KEGGsgn <- KEGGsgn %>% mutate(num = row\_number())

GO <- GO %>% mutate(num = row\_number())

KEGG <- KEGG %>% mutate(num = row\_number())

GO\_KEGG <- GO\_KEGG %>% mutate(num = row\_number())

##############################################################

#結合

m <- merge(dat7, dat8, by = "num", all = T)

m <- merge(m, dat9, by = "num", all = T)

m <- merge(m, dat10, by = "num", all = T)

m <- merge(m, dat11, by = "num", all = T)

m <- merge(m, dat12, by = "num", all = T)

m <- merge(m, GOcns, by = "num", all = T)

m <- merge(m, GOcv, by = "num", all = T)

m <- merge(m, GOem, by = "num", all = T)

m <- merge(m, GOsgn, by = "num", all = T)

m <- merge(m, KEGGcns, by = "num", all = T)

m <- merge(m, KEGGcv, by = "num", all = T)

m <- merge(m, KEGGsgn, by = "num", all = T)

m <- merge(m, KEGGem, by = "num", all = T)

m <- merge(m, GO, by = "num", all = T)

m <- merge(m, KEGG, by = "num", all = T)

m <- merge(m, GO\_KEGG, by = "num", all = T)

##############################################################

#積集合

t(names(m)) #行列の名前確認

GOint <- intersect(m$gene1, m$gene2)

GOint <- intersect(GOint, m$gene3)

KEGGint <- intersect(m$gene4, m$gene5)

KEGGint <- intersect(KEGGint, m$gene6)

#データフレーム変換

GOint <- data.frame(GOint=GOint)

GOint <- GOint %>% distinct(GOint,.keep\_all=FALSE) #gene重複削除しgeneのみ返す

GOint <- GOint %>% mutate(num = row\_number()) #通し番号列追加

KEGGint <- data.frame(KEGGint=KEGGint)

KEGGint <- KEGGint %>% distinct(KEGGint,.keep\_all=FALSE) #gene重複削除しgeneのみ返す

KEGGint <- KEGGint %>% mutate(num = row\_number()) #通し番号列追加

m <- merge(m, GOint, by = "num", all = T)

m <- merge(m, KEGGint, by = "num", all = T)

##############################################################

#csv出力

write.csv(m, file="GO\_KEGG\_top\_proteinlist.csv", na="", quote=F, row.names=F)

##############################################################