GSE analysis in R

Model : *Mus Musculus* (+ others)

library installation : Bioconductor

```
if (!requireNamespace("BiocManager", quietly = TRUE))
  install.packages("BiocManager")
BiocManager::install()

library(BiocManger)

# BiocManager::install("library_name") or BiocManager::install(c("library1", "library2"))
BiocManager::install("GEOquery")
BiocManager::install("affy")
BiocManager::install("gcrma")
BiocManager::install("Biobase")
BiocManager::install("marray")
BiocManager::install("limma")
BiocManager::install("gplots")
BiocManager::install("oligo")
BiocManager::install("clusterProfiler")
BiocManager::install("enrichplot")
BiocManager::install("ggplot2")
BiocManager::install("org.Mm.eg.db")
                      # -> if you use Human data, then install "org.Hs.eg.db"

library(GEOquery)
library(affy)
library(gcrma)

library(Biobase)
library(marray)
library(limma)
library(gplots)
library(oligo)

library(clusterProfiler)
library(enrichplot)
library(ggplot2)

library(org.Mm.eg.db)
```

Download GSE data

```
setwd("/home/user/GSE") # for Ubuntu 18.04
getwd() # check

# check your number of GSE data (GSE21392 is just a example)
getGEOSuppFiles("GSE21392")
setwd("/home/user/GSE/GSE21392")
```

Read & Normalize GSE data

▼ **Affy** : `read.celfiles()` & `rma()`

```
untar("GSE31106/GSE31106_RAW.tar", exdir = "data")
cels = list.files("data/", pattern = "[gz]")
sapply(paste("data", cels, sep = "/"), gunzip)

# read files
exp.cel_31106<-read.celfiles(list.celfiles("data/", full.names = T))

GSE31106<- getGEO("GSE31106")
```

```
GSE31106 <- GSE31106[[1]]

  # in this case, I don't want to use all files (just options)
exp.cel_31106 <- exp.cel_31106[, -c(4:15)]


# normalization
exp.rma_31106 <- rma(exp.cel_31106)
mds_31106<-exprs(exp.rma_31106)

  # distinguish control vs. test sets (options)
colnames(mds_31106)<-c(rep("ctrl", 3), rep("test", 3))
```

▼ **Agilent** : `read.maimages()` & `normalizeBetweenArrays()`

```
untar("GSE46200_RAW.tar", exdir="data")
t<-list.files("data/", pattern="gz")
sapply(paste('data', t, sep='/'), gunzip)

FileName <- list.files("data/", pattern = "txt")

# read files
project <- read.maimages(FileName, source="agilent.median",
                         columns=list(G="gMedianSignal", Gb="gBGMedianSignal"),
                         path="/home/user/GSE/GSE46200/data")

GSE<-getGEO('GSE46200', GSEMatrix=TRUE)
GSE<-GSE[[2]]
eset<-exprs(GSE)

  # in this case, I don't want to use all files (just options)
exp.cel_46200 <- project[,-c(2,4,6,8,10,12,14,16,18,20,22,24,26,28,30)]


# normalization
exp.rma_46200 <- normalizeBetweenArrays(exp.cel_46200)
mds_46200 <- exp.rma_46200$E

  # distinguish control vs. test sets (options)
colnames(mds_46200) <- c(rep("ctrl",7), rep("test", 8))
```

▼ **Illumina** : `read.ilmn()` & `neqc()`

```
gunzip("GSE21392_non_normalized.txt.gz")

# read files
raw21392<-read.ilmn("GSE21392_non_normalized.txt", probeid = 'ID_REF', expr = 'SAMPLE')

  # in this case, I don't want to use all files (just options)
raw21392<-raw21392[, c(1,2,7,8)]


# normalization
norm21392<-neqc(raw21392, offset=16, robust=FALSE)

  # distinguish control vs. test sets (options)
colnames(norm21392)<-c(rep("ctrl",2), rep("test", 2))
```

DEG selection

💡 Before selecting DEG, you must check PCA or MDS plot (refered later at ***Plot*** part)

```
ct<-factor(colnames(norm21392))
design<-model.matrix(~0+ct)
colnames(design)<-levels(ct)

fit<-lmFit(norm21392, design)
cont<-makeContrasts(test-ctrl, levels=design)
```

```
fit<-contrasts.fit(fit, cont)
fit<-eBayes(fit, trend = TRUE)
summary(decideTests(fit, method="global"))

topTable(fit, coef=1)
res2<-topTable(fit,number=Inf,lfc=1,adjust="BH")

  # up & down
up_res21392 <-res2[res2$P.Value < 0.05 & res2$logFC > 2,]
down_res21392 <-res2[res2$P.Value < 0.05 & res2$logFC < -2,]

  # save results as csv.file
write.csv(up_res21392, file='/home/user/GSE/GSE21392/up_res21392.csv')
write.csv(down_res21392, file='/home/user/GSE/GSE21392/down_res21392.csv')
```

Plot

```
# box plot
boxplot(log2(norm21392$E), ylab="log2 intensity")
# histogram
hist(norm21392$E, main="Histogram of GSE21392 after normalization")
```

▼ MA plot

```
limma::plotMA(fit, main="|logFC|>1.2")
abline(h=1.2, col="red", lwd=2)
abline(h=-1.2, col="red", lwd=2)
abline(h=0, col="blue")

mva.pairs(norm21392$E, log.it = TRUE)
```

▼ Dimension reduction

```
# MDS plot
plotMDS(norm21392, col=c(rep("blue", 2), rep("red", 2)), pch=c(rep(16, 2), rep(15, 2)))
legend("right", c("control", "test"), col = c("blue", "red"), pch=c(16, 15))

# PCA plot
pca_21392_b<-norm21392$E
pca_21392_b<-t(pca_21392_b)
pca_21392<-prcomp(pca_21392_b, center=T, scale. = T)
summary(pca_21392)

plot(pca_21392, type="l")
  # 2 dim
data2 <-data.frame("Samples"=rownames(pca_21392_b), pca_21392$x[, 1:2])
ggplot(data=data2) + geom_point(aes(x=PC1, y=PC2, col=Samples))+theme_minimal()
  # 3 dim
colors <- c("blue", "red")
colors<-colors[as.numeric(as.factor(rownames(pca_21392_b)))]

pca3_21392<-scatterplot3d(pca_21392$x[,1:3], pch = 16, color=colors)
legend("right", legend=levels(as.factor(rownames(pca_21392_b))),
       col=c("blue", "red"), pch = 16)
```

▼ Volcano plot

```
library(EnhancedVolcano)
EnhancedVolcano(res2,
                lab = rownames(res2),
                x = 'logFC',
                y = 'P.Value',
                xlab = bquote(~Log[2]~ 'fold change'),
                pCutoff = 0.05,
                FCcutoff = 2.0,
                pointSize = 2.0,
                labSize = 5.0,
```

```
                colAlpha = 1,
                legendPosition = 'right',
                legendLabSize = 14,
                legendIconSize = 5.0)
```

Convert to ENTREZ ID

(Important) check library and then apply correct `envir`

▼ Affy

```
library(mouse4302.db)

sym_up_31106<-data.frame(Gene=unlist(mget(x=rownames(up_res31106), envir=mouse4302ENTREZID, ifnotfound = NA)))
sym_down_31106<-data.frame(Gene=unlist(mget(x=rownames(down_res31106), envir=mouse4302ENTREZID, ifnotfound = NA)))

  # save results as csv.file
write.csv(sym_up_31106,file='/home/user/GSE/GSE31106/up_31106_id.csv')
write.csv(sym_down_31106,file='/home/user/GSE/GSE31106/down_31106_id.csv')
```

▼ Agilent

```
sym_up_107139 <- bitr(up_res107139$SystematicName, fromType="REFSEQ", toType="ENTREZID", OrgDb="org.Mm.eg.db")
sym_down_107139 <- bitr(down_res107139$SystematicName, fromType="REFSEQ", toType="ENTREZID", OrgDb="org.Mm.eg.db")

  # save results as csv.file
write.csv(sym_up_107139,file='/home/user/GSE/GSE107139/up_107139_id.csv')
write.csv(sym_down_107139,file='/home/user/GSE/GSE107139/down_107139_id.csv')
```

▼ Illumina

```
library(illuminaMousev1.db)

sym_up_21392<-data.frame(Gene=unlist(mget(x=rownames(up_res21392), envir=illuminaMousev1ENTREZID, ifnotfound = NA)))
sym_down_21392<-data.frame(Gene=unlist(mget(x=rownames(down_res21392), envir=illuminaMousev1ENTREZID, ifnotfound = NA)))

  # save results as csv.file
write.csv(sym_up_21392,file="/home/user/GSE/GSE21392/up_21392_id.csv")
write.csv(sym_down_21392,file="/home/user/GSE/GSE21392/down_21392_id.csv")
```

Pathway analysis

If use another organism's data (not Mus Musculus), then convert database of organism.

▼ GO pathway

```
go_up_21392 <- enrichGO(sym_up_21392$Gene, OrgDb = "org.Mm.eg.db", ont="all")
head(go_up_21392) # check

go_down_21392 <- enrichGO(sym_down_21392$Gene, OrgDb = "org.Mm.eg.db", ont="all")
head(go_down_21392) # check

## plot
dotplot(go_up_21392, split="ONTOLOGY", showCategory=10) + facet_grid(ONTOLOGY~., scale="free")
dotplot(go_down_21392, split="ONTOLOGY", showCategory=10) + facet_grid(ONTOLOGY~., scale="free")
```

▼ KEGG pathway

```
kegg_up_21392 <- enrichKEGG(gene = sym_up_21392$Gene,
                            organism = 'mmu',
                            pvalueCutoff = 0.05)
kegg_down_21392 <- enrichKEGG(gene = sym_down_21392$Gene,
                              organism = 'mmu',
                              pvalueCutoff = 0.05)

## plot
dotplot(kegg_up_21392)
dotplot(kegg_down_21392)
```

Heat map

```
library(RColorBrewer)

row1<-rownames(norm21392)
heat_21392<-cbind(row1, norm21392$E)

name<-c(rownames(up_res21392), rownames(down_res21392))
name<-name[order(name)]

heat_21392<-subset(heat_21392, row1 %in% name)
heat_21392<-heat_21392[, -1] # erase name
ma<-matrix(as.numeric(heat_21392), ncol=4)
colnames(ma)<-c("ctrl1", "ctrl2", "test1", "test2")
rownames(ma)<-name

library(ComplexHeatmap)

# plot
Heatmap(ma, name="Expression",
        column_title="model",
        row_title="gene",
        row_names_gp=gpar(fontsize=1),
        width=unit(4, "cm"),
        height = unit(20, "cm"))
```

DEG venn diagram

```
g2<-c(sym_up_21392$Gene, sym_down_21392$Gene)
g2<-na.omit(g2)

g3<-c(sym_up_31106$Gene, sym_down_31106$Gene)
g3<-na.omit(g3)

g1<-c(sym_up_107139$ENTREZID, sym_down_107139$ENTREZID)
g1 <- na.omit(g1)

gse1<-list(GSE21392=g2, GSE31106=g3, GSE107139=g1)

setwd('/home/user')

library(VennDiagram)

venn.diagram(
  gse1,
  fill=c(3, 2, 7),
  alpha=c(0.5, 0.5, 0.5),
  lty=c(1, 2, 3),
  filename = "DEG_venn_diagram"
)
```

Reference : Bioconductor, limma, Heatmap, Volcano plots, Venn diagram

Connect : lifescience18@naver.com