Anàlisi de dades òmiques. PAC1

Marina Gay

2024-11-06

Contents

1. Objectius

L'objectiu final és trobar aquells fosfopèptids que ens permeten diferenciar entre dos grups de tumor: PD i MSS. En aquest treball ens centrarem en avaluar la qualitat tècnica de les dades abans de procedir amb l'anàlisi diferencial.

2. Resum executiu

- Identifiquem 1438 fosfopèptdis de 718 grups de proteïnes. 1180 fosfopèptids tenen la fosforilació en serina o treonina i 258 en tirosina.
- El nombre de fosfopèptids i la distribució de les seves abundàncies és homogeni al llarg de totes les mostres.

- La correlació entre els fenotips MSS i PD és alta i les dades semblen ben normalitzades.
- De manera qualitativa no sembla que hi hagi grans diferències entre els dos fenotips. Tot i així hi ha dos fosfopètids que prodrien tenir menys abundància en PD que en MSS: GEPNVSYICSR[7] Phospho|[9] Carbamidomethyl_Y (P49840) (log2FC = -1.9, mean_abundance = 24.5) i LSLEGDHSTPPSAYGSVK[14] Phospho (P07355) (log2FC = -3.9, mean_abundance = 22.5).
- En global, les mostres semblen tenir una qualitat tècnica suficient com per a procedir amb l'anàlisi diferencial. No obstant hi ha uns quants punts que s'hauran d'abordar:
  - Identificar possibles efectes batch. Les mostres no s'acaben d'agrupar per fenotip. Caldrà parlar amb els investigadors per estudiar bé el procés d'obtenció de la mostra.
  - Gestió dels valors *missing*. Hem vist que aquest valors capturen part de la variància.
  - Possiblitat de treballar a nivell de *p-site* en comptes de fosfopèptid.

3. Materials i mètodes

El dataset escollit és de fosfoproteòmica, pel que tindrem en compte diversos punts clau a l'hora d'analitzar les dades, tenint en compte que es tracta de dades que provenen d'espectrometria de masses:

- Transformació: les haurem de passar a escala logarítmica per a que compleixin la normalitat.
- Exploració: visualització i avaluació de possibles efectes batch. Avaluació dels valors *missing*.

3.1. Selecció d'un dataset.

Descarrego el repositori de GitHub https://github.com/nutrimetabolomics/metaboData/ des de R studio. Per fer-ho vaig a la pentanya "Terminal" i clono el repositori a la meva carpeta d'inetrès emprant la següent comanda:

git clone https://github.com/nutrimetabolomics/metaboData/

Avaluo els datasets i escullo el dataset 2018-Phosphoproteomics. Es tracta d'unes dades de fosfoproteomòmica. Hi ha dos arxius:

- **description.md**: conté la descripció de l'experiment i de les mostres (*metadata*).
- **TIO2+PTYR-human-MSS+MSIvsPD.XLSX**: excel amb dues pestanyes:
  - **originalData**: conté la informació de quantificació a nivell de fosfopèptid (*expression data*) i informació adicional de com l'*Accession* i la descripció de la proteïna, etc.. (*feature data*).
  - **targets**: conté infroamció de les mostres i el seu fenotip (*pheno data*).

3.2. Creació d'un contenidor del tipus SummarizedExperiment.

Per crear el contenidor, en primet lloc, carrego totes les dades en format data frame o text. Incloc el *metadata* amb la descripció de l'experiment, l'*expression data*, que transformaré a matriu, el *feature data* amb informació extra de cada fosfopèptid i el *pheno data* amb la informació de cada mostra i el seu fenotip.

3.3. Pre processat i filtratge de les dades.

- Hem eliminat els fosfopèptids amb valors de quantificació igual a zero en totes les mostres.
- Hi ha una entrada duplicada ("GEPNVSYICSR[7] Phospho|[9] Carbamidomethyl") amb el *phopho site* o *p-site* en S/T o Y. Hem afegit l'etiqueta de S/T o Y per no tenir duplicitats.
- Hem substituït els zeros per NA. En dades obtingudes a partir d'espectrometria de mases no podem dir que un valor sigui zero, sinó que aquella identificació està per sota el nostre límit de detecció.
- Hem transformat les dades a logaritme emprant log2. En dades provinents d'espectrometria de masses sempre és necessària aquesta transformació per a tenir distribucions normals.
- Hem fet un estudi dels valors *missing* comptant quants *missing* hi ha en cadascuna de les mostres i el seu percetatge. Hem afegit aquesta informació a la *pheno data*.
- Hem calculat la intensitat total en cada mostra i hem afegit aquesta informació al *pheno data*.

3.4. Exploració del dataset.

Realizo la visualització de les dades mitjançant diferents gràfics, emprant les llibreries *ggplot2* i *MixOmics*.

3.5. Repository de GitHub

Per crear i gestionar el repositori de GitHub he emprat l'aplicatiu GitHub Desktop. En primer lloc he definit el compte de GitHub on vull crear el repository. Després he definit la carpeta on guardar el repositori. Finalment he creat el repositori i he fet un commit amb la informació i els resultats de l'estudi.

Enllaç al repositori de GitHub: https://github.com/marinaUOC?tab=repositories

Aquest repositori conté el present informe, les dades de l'estudi que s'han utilitzat així com el codi d'R emprat per a dur a terme l'anàlisi.

4. Resultats i discussió

El dataset d'estudi prové d'un experiment de fosfoproteòmica en el què volem determinar les diferències entre dos tipus de tumor (MSS i PD) en models PDX. Partim de 3 rèpliques biològiques de cada fenotip i dues rèpliques tècniques de cada mostra.

En aquesta anàlisi realizem una exploració qualitativa de les dades, que és el pas previ a l'anàlisi diferencial.

El dataset de partida conté informació quantitativa a nivell de fosfopèptid. Les dades provenen d'un experiment d'enriquiment en fosfopèptids i la seva anàlisi posterior per cormatografia líquida acoplada a espectrometria de masses (LC-MS).

En global, identifiquem 1438 fosfopèptdis que pertanyen a 718 grups de proteïnes diferents, dels quals 1180 tenen la fosforilació en serina o treonina i 258 en tirosina.

El nombre de fosfopèptids identificats és molt igual en els replicats tècnics i difereix una mica en les mostres biològiques (**Figure 1**). La mostra M43 és la que té un major nombre de fosfopèptids identificats.

La distribució de les intensitats, tot i no ser perfecte, és força homgènia al llarg de totes les mostes (**Figure 2**). Observem la presència d'outliers que ens crida l'atenció. Alguns valors log2Int són negatius, pel que la seva intensitat en escala linial ha d'estar per sota de la unitat. S'haurà d'estudiar d'on venen aquests outliers i si val la pena eliminar-los. És molt estrany tenir valors d'intensitat provinents espectrometria de masses

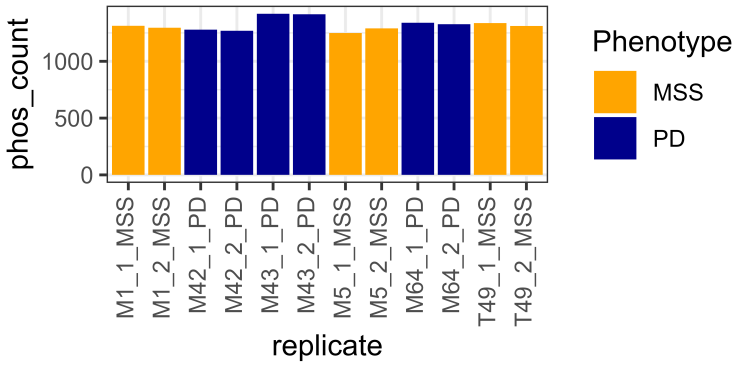

Figure 1: Barplot amb el número de fosfopèptids identificats (eix y) per a cada mostra (eix x), color = fenotip.

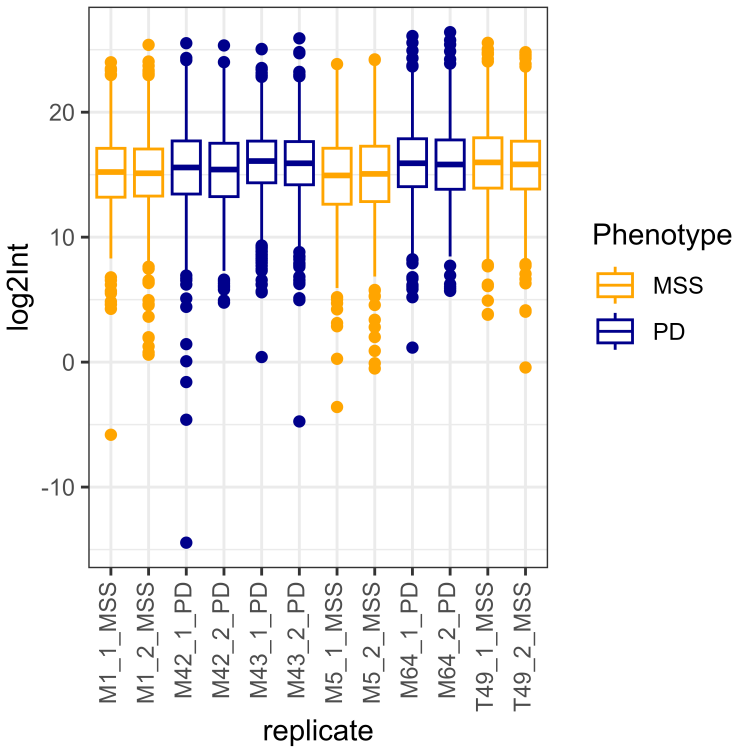

Figure 2: Boxplot del log2(intensity) (eix y) per a cada mostra (eix x), color = fenotip.

tan baixos ja que acostumen a estar al nivell del soroll. Hauríem de parlar amb els investigadors que han generat les dades per a tenir més informació i prendre una decisió més fonamentada.

L'abundància dels fosfopèptids en els dos fenotips (PD i MSS) té una bona correlació (R = 0.86) i està centrada a la identitat (**Figure 3**). Observem també que el *fold change* està centrat a zero i que la dispersió de les dades és menor per valors més grans d'abundància, que és el que esperem en aquest tipus d'experiments (**Figure 4**). Hi ha dos punts que ens criden l'atenció, són els dos que tenen una abundància alta i un *fold change* negatiu gran. Corresponen als pèptids GEPNVSYICSR[7] Phospho|[9] Carbamidomethyl_Y (P49840) (log2FC = -1.9, mean_abundance = 24.5) i
LSLEGDHSTPPSAYGSVK[14] Phospho (P07355) (log2FC = -3.9, mean_abundance = 22.5). Aquests dos pèptids podrien ser rellevants.

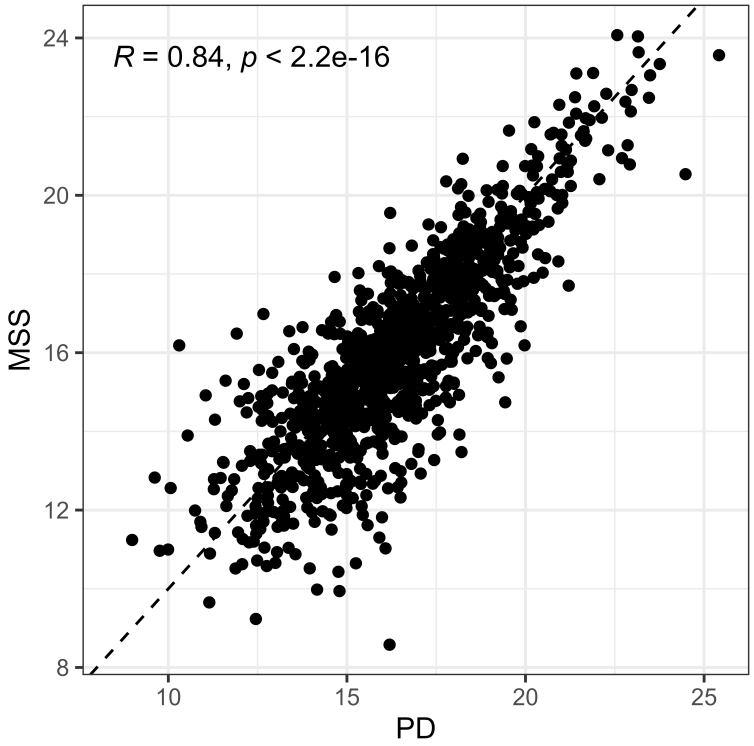

Figure 3: Scatter plot del log2(intensity) de MSS (eix y) vs PD (eix x).

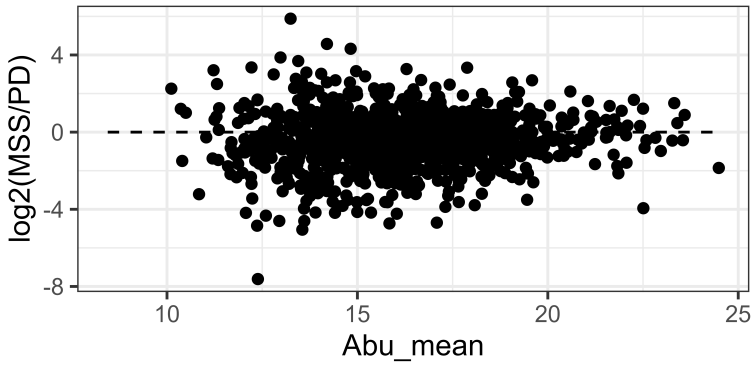

Figure 4: MA plot del log2 fold change de MMS/PD (eix y) vs la mitjana de les abundàncies (eix x).

El percentatge de valors *missing* no és del tot homogeni al llarg de les mostres (**Figure 5**). També és important avaluar quin tipus de *missing* generem per saber, en cas de que sigui necessari, el tipus d'imputació que farem. Aquest aspecte no l'abordarem en el present estudi.

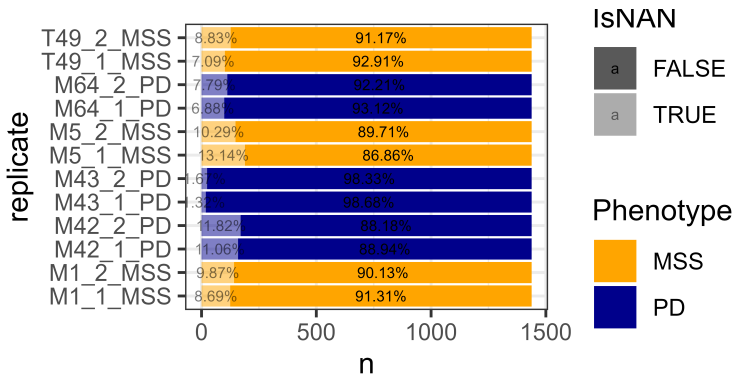

Figure 5: Barplot representant el percentatge de valors missing en les mostres (eix y) vs el número de missing (eix x). El color representa el fenotip i la intensitat de color si es missing o no.

L'anàlisi de components principals (PCA) revela que les dues primeres components s'enduen el 57.7 de variància explicada (**Figure 6**). La primera component (PC1) discrimina parcialment els dos fenotips, però no s'acaben de separar del tot. si veiem com s'agrupen les rèpliques tècniques. Les mostres T49 i M42 cauen al centre del gràfic (**Figure 7**). Podria ser que hi hagués algun efecte batch a l'hora de preparar les mostres o durant la injecció a l'espectrometre de masses, és una informació que no tenim i que hauríem de preguntar al investigadors per tenir en compte per procedir amb l'anàlisi.

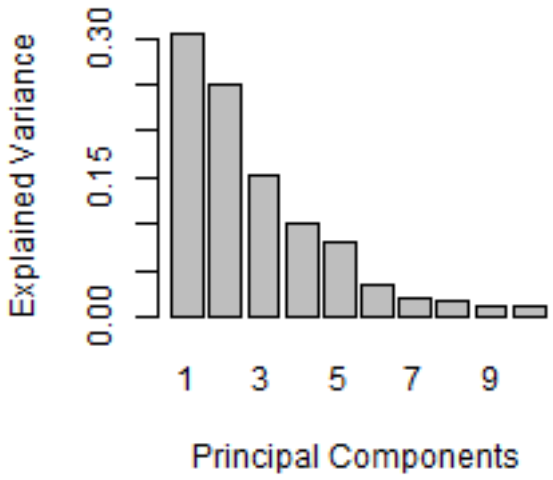

Figure 6: Barplot de la variància explicada per a cada component principal.

Avaluem la correlació entre la PC1 i el nombre de valors missing en cadascuna de les mostres. Veiem que que aquesta correlació és elevada (-0.81). És a dir la PC1 no només està separant per fenotip, sinó que també ho fa en funció del nombre de *missing* en cada mostra. A més a més, el nombre de *missing* sembla estar parcialment confós amb el fenotip, la qual cosa pot arribar a complicar l'anàlisi diferencial (**Figure 8**).

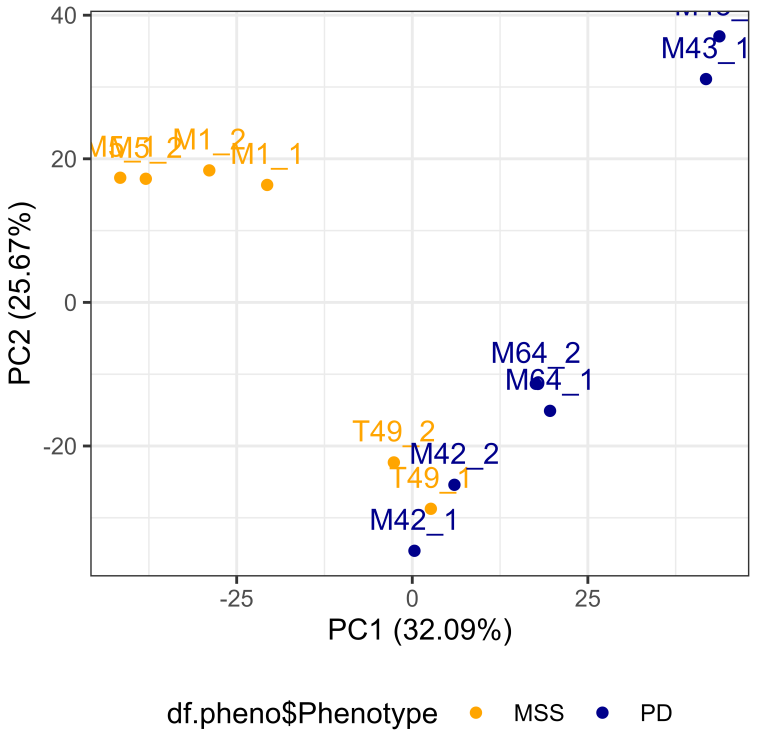

Figure 7: Anàlisi de components pprincipals (PCA). Segona component (PC2) (eix y) vs la primera component (PC1) (eix x), color = fenotip.

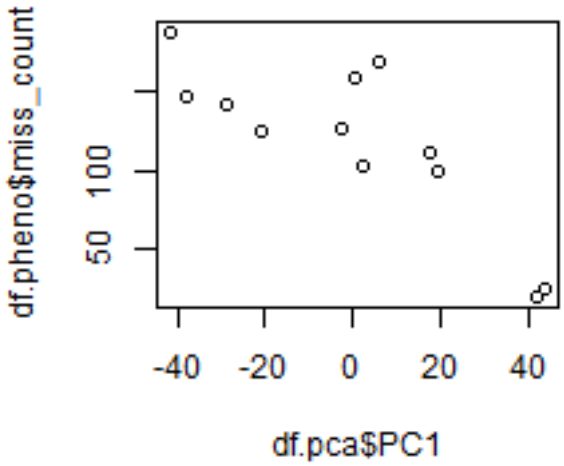

Figure 8: Scatter plot del nombre de missing (eix y) vs la PC1 (eix x).

El mapa de correlacions i el dendogrma ens confirmen el que ja intuíem a la PCA. Tal i com esperàvem, la correlació entre les rèpliques tècniques és superior a la correlació entre rèpliques biològiques. També observem que les mostres no s'acaben de separar per fenotip (PD i MSS). La mostra T49_MSS queda més a prop de les mostres PD i la mostra M43_PD va una mica a part de la resta (**Figure 9** i **Figure 10**).

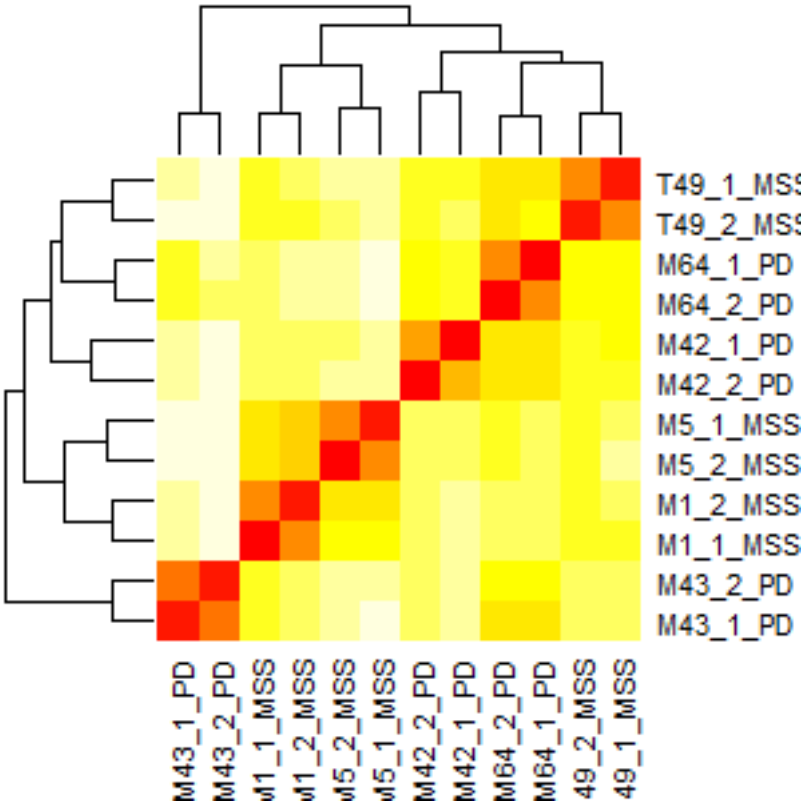

Figure 9: Mapa de colors de la correlació entre les diferents mostres.

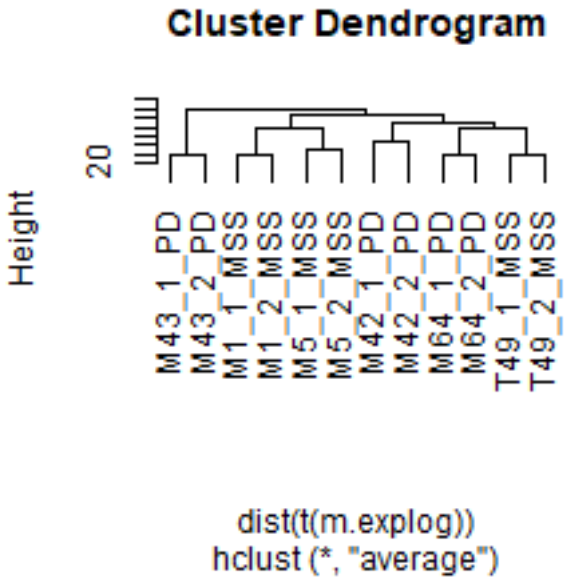

Figure 10: Mapa de colors de la correlació entre les diferents mostres.

5. Conclusions

La qualitat tècnica de les mostres de l'estudi no és excel · lent i hi ha alguns punts que s'haurien d'abordar abans de procedir amb l'anàlisi d'abundància diferencial:

- Gestió dels valors *missing*: Hem d'acabar d'estudiar la influència d'aquests valors en l'agrupació de les mostres. També hem de valorar si cal filtrar i/o imputar aquests valors.
- Efecte batch: No acabem de diferenciar per fenotip i podria ser que no estiguessim tenint en compte algun possible efecte batch. Caldria parlar amb els investigadors per entendre ben bé com s'han processat les mostres (dia, personal, ordre d'injecció l'espectrometre de masses etc...).
- L'estudi està basat amb l'anàlisi dels fosfopèptids. Aquest estudi és informatiu, però suggeriríem fer l'estudi a nivell de *phospho site* (*p-site*) i no de fosfopèptid. Per això hauríem d'anotar el lloc de fosforilació dins la proteïna i fer un resum dels diferents pèptids que apunten al mateix *p-site*. Aquest nou enfoc pot ser més informatiu ja que des d'un punt de vista biològic té més sentit parlar de llocs de fosforilació que no pas de fosfopèptids.

6. Annex I: Codi R

Script en R per explorar les dades i generar els plots presents a l'informe.

```
# Analisi de dades omiques. PAC 1
# Marina Gay
# Exploratory data analysis of a phosphoproteomics experiment


# Libraries and folders -----------------------------------------------------
# Load libraries
library(tidyverse)
library(ggplot2)
library(xlsx)
library(SummarizedExperiment)
library(mixOmics)
library(ggpubr)
library(plotly)
library(htmlwidgets)

# Getting the dir where this R script is located: mainDir
p_mainDir <- dirname(rstudioapi::getSourceEditorContext()$path)

# Setting mainDir as the current working dir
setwd(p_mainDir)

# Define folders
p_mainDir <-  paste0(p_mainDir,"/")
p_figures <-  paste0(p_mainDir,"figures/")
p_tables <-  paste0(p_mainDir,"tables/")

# Create figures and table folders
dir.create(file.path(p_figures), showWarnings=F)
dir.create(file.path(p_tables), showWarnings=F)

# Load data and prepare for the SummarizedExperiment ------------------------
# Load data expression and feature data
df.raw <- read.xlsx(file=paste0(p_mainDir,"metaboData/Datasets/2018-Phosphoproteomics/TIO2+PTYR-human-MS
# df.raw <- read.delim(file=paste0(p_mainDir,"phosphopep.txt"), sep = "\t", stringsAsFactors = FALSE)

# There is a phosphopeptide entry duplicate (GEPNVSYICSR[7] Phospho|[9] Carbamidomethyl):
# One is Y and the other is S/T
# I will add this information to distinguish them
df.raw$SequenceModifications <- with(df.raw, ifelse(SequenceModifications == "GEPNVSYICSR[7] Phospho|[9]
                                                    paste0(SequenceModifications, "_", PHOSPHO),
                                                    SequenceModifications))

# Select only expression data
df.expr <- df.raw %>%
  dplyr::select(SequenceModifications, ends_with("_MSS"), ends_with("_PD"))

# Convert to matrix
m.expr <- as.matrix(df.expr %>% dplyr::select(-SequenceModifications))
```

```r
# Add phosphopeptides as rownames
rownames(m.expr) <- df.expr$SequenceModifications

# Select feature data and add column with number of phospho in the sequence
df.fd <- df.raw %>%
  dplyr::select(-ends_with("_MSS"), -ends_with("_PD")) %>%
  mutate(phos = str_count(SequenceModifications, "Phospho") # Count number of phospho
  )

# Load pheno data
df.pheno <- read.xlsx(file=paste0(p_mainDir, "TIO2+PTYR-human-MSS+MSIvsPD.XLSX"), sheetIndex=2)
# df.pheno <- read.delim(file=paste0(p_mainDir, "targets.txt"), sep = "\t", stringsAsFactors = TRUE)

# load metadata
meta <- readLines(paste0(p_mainDir, "description.md"))

# Create the summarized experiment container ----------------------------------
se <- SummarizedExperiment(assays = list(counts = m.expr),
                           colData = df.pheno,
                           rowData = df.fd,
                           metadata = meta)

# Data pre processing ---------------------------------------------------------
# Remove phosphopeptides with no quantification values in any sample
# Check for rows with non-zero values in any sample
non.zero.pep <- rowSums(assay(se, "counts") != 0) > 0

# Subset the SummarizedExperiment object to keep only non-zero rows
se <- se[non.zero.pep, ]

# Replace 0 with NA
assays(se)$counts[assays(se)$counts == 0] <- NA

# log 2 transform
assays(se)$counts <- log2(assays(se)$counts)

# Tidy data for ggplot
df.t <- df.raw %>%
  pivot_longer(cols = ends_with("_MSS") | ends_with("_PD"),
               names_to = "replicate",
               values_to = "intensity") %>%
  mutate_at(c('intensity'), ~na_if(., 0)) %>% # replace 0 with NA
  mutate(log2Int = log2(intensity), # log transform
         phos = str_count(SequenceModifications, "Phospho"), # Count number of phospho
         Phenotype =  str_split_fixed(replicate, "_", 3), # Add pheno data
         sample = Phenotype[,1],
         Individual = Phenotype[,2],
         Phenotype = Phenotype[,3],
         IsNAN = is.na(log2Int)) # Find NA

# Average data by phenotype and calculate FC
df.av <- df.t %>%
  group_by(SequenceModifications, Accession, Description, Phenotype) %>%
```

```
  summarise(abundance = mean(log2Int)) %>%
  ungroup() %>%
  pivot_wider(names_from = Phenotype,
              values_from = abundance) %>%
  mutate(Abu_mean = rowMeans(dplyr::select(., MSS, PD), na.rm = TRUE),
         log2FC = MSS - PD)

# Prepare data form missigness plot
df.miss <- df.t %>%
  mutate(replicate = factor(replicate)) %>%
  group_by(replicate, Phenotype, IsNAN) %>%
  summarise(n = n(), .groups = "drop") %>%
  group_by(replicate, Phenotype) %>%
  mutate(Percentage = n / sum(n),
         Sample = str_split_fixed(replicate, "_", 3),
         Sample = paste0(Sample[,1], "_", Sample[,2]))


# Add technical variables to pheno data
# Filter missing data
df.miss.filt <- df.miss %>%
  filter(IsNAN == TRUE) %>%
  rename("miss_count" = "n",
         "miss_per" = "Percentage")

# Extract total intensity
df.totInt <- df.t %>%
  drop_na(log2Int) %>%
  group_by(replicate, Phenotype) %>%
  summarize(totalInt = sum(intensity, na.rm = TRUE),
            phos_count = n()) %>%
  mutate(Sample = str_split_fixed(replicate, "_", 3),
         Sample = paste0(Sample[,1], "_", Sample[,2]))

# Join data frames
df.pheno <- merge(df.pheno, df.miss.filt, by = c("Sample", "Phenotype"),
                  all.x = TRUE, sort = FALSE)

df.pheno <- merge(df.pheno, df.totInt, by = c("Sample", "Phenotype", "replicate"),
                  all.x = TRUE, sort = FALSE)

# Update pheno data
colData(se)$miss_count <- df.pheno$miss_count
colData(se)$totalInt <- df.pheno$totalInt

# Exploratory analysis ------------------------------------------------------
df.raw$PHOSPHO <- as.factor(df.raw$PHOSPHO)
summary(df.raw)

# number of phosphopeptides identified
length(unique(df.raw$SequenceModifications))

# number of proteins identified
```

```r
length(unique(df.raw$Accession))

# Data visualization ------------------------------------------------------------
# Define phenotype color
col <- c("PD" = "darkblue",
         "MSS" = "orange")

# Boxplot of log2 Intensities
plot.boxInt <- ggplot(data = df.t,
       aes(x = replicate, y = log2Int, color = Phenotype)) +
  theme_bw() +
  geom_boxplot() +
  theme(axis.text.x  = element_text(angle=90, vjust= 0.3, hjust= 1)) +
  scale_color_manual(values = col)
plot.boxInt

# Number of phosphopeptides identified by sample
plot.barNum <- ggplot(data = df.totInt,
                      aes(x = replicate, y = phos_count, fill = Phenotype)) +
  theme_bw() +
  geom_col() +
  theme(axis.text.x  = element_text(angle=90, vjust= 0.3, hjust= 1)) +
  scale_fill_manual(values = col)
plot.barNum

# Scatter plot of phenotype
plot.pt_groups <- ggplot(data = df.av,
                  aes(x = PD, y = MSS)) +
  theme_bw() +
  geom_point() +
  stat_cor(method="pearson") +
  geom_abline(slope = 1, intercept = 0, linetype = 2)
plot.pt_groups

# MA plot
plot.MA <- ggplot(data = df.av,
               aes(x = Abu_mean, y = log2FC,
                   label = paste(Accession, SequenceModifications))) +
  theme_bw() +
  geom_point() +
  ylab("log2(MSS/PD)") +
  geom_line(y = 0, linetype = 2)
plot.MA

# Barplot with the percentage of missing by sample
plot.miss <- ggplot(df.miss,
                  aes(x=n, y=replicate, alpha=IsNAN, fill=Phenotype, label=scales::percent(Percentage)))
  theme_bw() +
  geom_col() +
  geom_text(position = position_stack(0.5), size = 2) +
  scale_alpha_manual(values=c(1.0, 0.5)) +
  scale_fill_manual(values = col)
plot.miss
```

```r
# Extract expression matrix form the se object
m.explog <- assay(se)

# remove NA
m.explog <- m.explog[complete.cases(m.explog), ]

# Computing PCA
pca <- prcomp(t(m.explog))
pcvara <- pca$sdev^2/sum(pca$sdev^2)*100;

# Convert to dataframe
df.pca <- as.data.frame(pca$x)
df.pca$sample <- rownames(df.pca)

plot.pca <- ggplot(data=df.pca,
                       aes(x=PC1, y=PC2, color=df.pheno$Phenotype)) +
  geom_point()+
  geom_text(label=df.pheno$Sample, show.legend = FALSE, vjust = -1) +
  theme_bw() +
  xlab(paste0("PC1 (", round(pcvara[1], 2), "%)")) +
  ylab(paste0("PC2 (", round(pcvara[2], 2), "%)")) +
  theme(legend.position = "bottom") +
  scale_color_manual(values = col)
plot.pca

# PC1 correlation with technical variables
plot(df.pca$PC1, log2(df.pheno$totalInt))
cor(df.pca$PC1, log2(df.pheno$totalInt))

plot(df.pca$PC1, df.pheno$miss_count)
cor(df.pca$PC1, df.pheno$miss_count)


# heatmap sample correlation
manDist <- dist(t(m.explog))
heatmap (as.matrix(manDist), col=heat.colors(16))

# Dendogram
clust.euclid.average <- hclust(dist(t(m.explog)),method="average")
plot(clust.euclid.average, hang=-1)

# MixOmics plots -------
MyResult.pca <- pca(t(m.explog))  # 1 Run the method
plotIndiv(MyResult.pca) # 2 Plot the samples
plotVar(MyResult.pca)   # 3 Plot the variables

tune.pca.multi <- tune.pca(t(m.explog), ncomp = 10, scale = TRUE)
plot(tune.pca.multi)

final.pca.multi <- pca(t(m.explog), ncomp = 3, center = TRUE, scale = TRUE)

# Top variables on the first component only:
head(selectVar(final.pca.multi, comp = 1)$value)
```

```r
biplot(final.pca.multi)

# Save plots and data ----------------------------------------------------------
# Save plots in png
# List all objects in the environment that start with "plot."
plot_names <- ls(pattern = "^plot\\.")

# Initialize an empty list to store the plots
plot_list <- list()

# Populate the list with plots
for (plot_name in plot_names) {
  plot_list[[plot_name]] <- get(plot_name)
}

# Save the plots
lapply(names(plot_list), function(f) {
  ggsave(paste0(p_figures, f, '4x4.png'),
         plot_list[[f]], width=4, height=4, dpi=1200)
})

lapply(names(plot_list), function(f) {
  ggsave(paste0(p_figures, f, '2x4.png'),
         plot_list[[f]], width=4, height=2, dpi=1200)
})

png(filename = paste0(p_figures, "pca_comp.png"), width = 240, height = 240, units = "px")
plot(tune.pca.multi)
dev.off()

png(filename = paste0(p_figures, "heatmap.png"), width = 300, height = 300, units = "px")
heatmap (as.matrix(manDist), col=heat.colors(16))
dev.off()

png(filename = paste0(p_figures, "dendogram.png"), width = 240, height = 240, units = "px")
plot(clust.euclid.average, hang=-1)
dev.off()

png(filename = paste0(p_figures, "PCA_miss.png"), width = 240, height = 240, units = "px")
plot(df.pca$PC1, df.pheno$miss_count)
dev.off()

saveWidget(ggplotly(plot.MA), file = paste0(p_figures, "MAplot.html"))

# Save Rda object
save(se, df.av, plot_list, file=paste0(p_mainDir, "Phosphopeptide_SE", '.Rda'))
```

Referències

The R Project for Statistical Computing. [cited 4 Jun 2021]. Available: https://www.r-project.org/

https://aspteaching.github.io/AMVCasos/

https://aspteaching.github.io/Analisis_de_datos_omicos-Ejemplo_0-Microarrays/ExploreArrays.html
https://mixomicsteam.github.io/mixOmics-Vignette/index.html