PEC1: ANÀLISI DE DADES ÒMIQUES

Arnau Mena Molina

6 de Novembre del 2024

Contents

1. Introducció i objectius, materials i mètodes

Aquesta primera PEC de l'assignatura d'Anàlisi de Dades Òmiques es realitza amb la finalitat de completar la introducció a les òmiques fent ús d'un exercici de repàs i ampliació que ens permet treballar amb Bioconductor i l'exploració multivariant de dades, entre d'altres.

Addicionalment, s'inclou l'ús i exploració de la plataforma github. Tornant a Bioconductor, en aquest cas, en comptes dels *expressionSets*, farem ús de l'objecte S4 *SummarizedExperiment*.

Per últim, el material utilitzat serà, per una banda, RStudio amb la versió d'R 4.4.2 i, pel que fa a les dades, s'han extret del repositori facilitat pel professor a github; concretament, s'ha utilitzat l'arxiu *TIO2+PTYR-human-MSS+MSIvsPD.xlsx*, ubicat dins la carpeta 2018-Phosphoproteomics.

2. *Scientific Background*

L'arxiu de dades escollit obté dades a partir d'un experiment de fosfoproteòmica realitzat per analitzar models de xenografies derivades de pacients (PDX) de dos subtipus tumorals diferents. Concretament, l'experiment es va dur a terme amb 3 models de cada subtipus, utilitzant mostres enriquides en fosfopèptids per aconseguir una anàlisi més precisa de les modificacions post-traduccionals. En cadascuna de les mostres, s'ha realitzat una anàlisi per cromatografia líquida acoblada a espectrometria de masses (LC-MS) amb 2 rèpliques tècniques per garantir la consistència i robustesa dels resultats.

El conjunt de dades resultant conté les abundàncies *normalitzades* dels senyals obtinguts per espectrometria de masses de prop de 1400 fosfopèptids. Aquestes dades proporcionen una visió detallada de les variacions en la fosforilació de proteïnes entre els dos subtipus tumorals, una modificació crucial que pot influir en les funcions de les proteïnes implicades en el creixement i la progressió del tumor.

L'objectiu principal de l'anàlisi és identificar els fosfopèptids que permetin diferenciar de manera significativa els dos grups tumorals estudiats.

Els grups estudiats es defineixen de la manera següent:

Grup MSS: Inclou les mostres M1, M5 i T49. Grup PD: Inclou les mostres M42, M43 i M64.

3. Resultats

3.1 Primera visualització de les dades

Com a primer objectiu, el que volem és familiaritzar-nos amb les dades, i veure quina estructura tenen, per tal d'optimitzar i generar, de la millor manera, el objecte SummariezExperiment amb el qual treballarem durant tota la PEC. Per tant, ens limitarem a veure les primeres línies de l'arxiu excel que tenim, per veure el format que obtenim:

Figura 1: Resultats de fer head sobre l'excel de dades; *chunk primera_visualització*

```
## # A tibble: 5 x 18
##   SequenceModifications   Accession Description Score M1_1_MSS M1_2_MSS M5_1_MSS
##   <chr>                   <chr>     <chr>       <dbl>    <dbl>    <dbl>    <dbl>
## 1 LYPELSQYMGLSLNEEEIR[2]~ O00560    Syntenin-1~  48.1     24.3   44476.       0
## 2 VDKVIQAQTAFSANPANPAILS~ O00560    Syntenin-1~  67.0      0     43139.    2102.
## 3 VIQAQTAFSANPANPAILSEAS~ O00560    Syntenin-1~  77.7   3413.   172143.   77323.
## 4 HADAEMTGYVVTR[6] Oxida~ O15264    Mitogen-ac~  44.9 220431.   145657.  104288.
## 5 HADAEMTGYVVTR[9] Phosp~ O15264    Mitogen-ac~  67.4  18255.     8530.   35956.
## # i 11 more variables: M5_2_MSS <dbl>, T49_1_MSS <dbl>, T49_2_MSS <dbl>,
## #   M42_1_PD <dbl>, M42_2_PD <dbl>, M43_1_PD <dbl>, M43_2_PD <dbl>,
## #   M64_1_PD <dbl>, M64_2_PD <dbl>, CLASS <chr>, PHOSPHO <chr>
```

Com podem veure obtenim sobre pantalla una primera línea que ens diu que ens trobem amb un dataset temporal de 5 files (les que hem demanat amb el codi) i un total de 18 columnes. D'aquestes 18 columnes veiem que, realment, no totes són necessàries. Mirant el nom de les columnes, sabem que ens interessen a nosaltres, com estadístics, la primera columna *SequenceModifications*, que ens dona informació sobre el fosfopèptid i la modificació que ha rebut i, d'altra banda, les columnes que pertanyen a la mesura per a cada mostra per duplicat (*M1_1_MSS, M1_2_MSS, M5_1_MSS, [...], M64_2_PD*). De tal manera, de les 18 columnes inicials, a nosaltres ens interessen només 13 columnes. Per tant, gràcies a aquesta primera observació de les dades originals, ja podem fer un filtre per excluir algunes columnes que no ens són d'interès per aquest context i, ergo, generar un SummarizedExperiment molt menys dens i net. A més a més, les columnes també ens donen informació sobre les mostres i quin fenotip presenten aquestes.

3.2 Generació del *SummarizedExperiment* i preprocessing

Per tant, podem passar a la generació del *SummarizedExperiment* (SE), el qual, com bé sabem, és un objecte S4 que està format per 3 components: Assay, colData i rowData. De manera resumida, el que volem emmagatzemar a cada component és:

- **Assay:** volem guardar les dades experimentals, les mesures generades per la màquina.
- **colData:** emmagatzemarem la informació sobre la mostra, és a dir, sobre els PDXs.
- **rowData:** emmagatzemarem la informació sobre les condicions d'anàlisi, és a dir, emmagatzemarem el fosfopèptid al qual pertany a cada fila.

Per tant, generem el codi per crear el SummarizedExperiment, tenint en compte les següents consideracions, gràcies a la primera visualització de les dades que hem fet anteriorment:

- **Per a Assay**, només guardarem les 13 columnes que ens interessen i que, per tant, contenen les mesures. Per tant, a partir de les dades originals (*original_data*), farem una primera neteja de columnes, generant *selected_columns*, que serà el que emmagatzemarem dins de Assay.
- **Per a colData**, redactarem manualment les mostres i els fenotips que ens podrem trobar, per així assegurar la coincidència amb Assay.
- **Per a rowData**, afegirem totes les metadades que ens aporten totes les files de la primera columna *SequenceModifications*.

3.2.1 Comprovació de la generació del SE

Un cop generat el *SummarizedExperiment*, fem un print sobre aquest per comprovar la seva correcta generació:

Figura 2: Resum de l'objecte *SummarizedExperiment* generat i print de les 5 primeres files d'aquest; *chunk Generació_SE_preprocessing_creació*

```
## class: SummarizedExperiment
## dim: 1438 12
## metadata(0):
## assays(1): counts
## rownames(1438): LYPELSQYMGLSLNEEEIR[2] Phospho|[9] Oxidation
##   VDKVIQAQTAFSANPANPAILSEASAPIPHDGNLYPR[35] Phospho ...
##   YQDEVFGGFVTEPQEESEEEVEEPEER[17] Phospho YSPSQNSPIHHIPSRR[1]
##   Phospho|[7] Phospho
## rowData names(1): SequenceModifications
## colnames(12): M1_1_MSS M1_2_MSS ... M64_1_PD M64_2_PD
## colData names(2): Sample Phenotype
```

```
##                                                  M1_1_MSS   M1_2_MSS
## LYPELSQYMGLSLNEEEIR[2] Phospho|[9] Oxidation      24.29438  44475.964
## VDKVIQAQTAFSANPANPAILSEASAPIPHDGNLYPR[35] Phospho   0.00000  43138.904
## VIQAQTAFSANPANPAILSEASAPIPHDGNLYPR[32] Phospho    3412.60332 172143.040
## HADAEMTGYVVTR[6] Oxidation|[9] Phospho          220431.17880 145656.887
## HADAEMTGYVVTR[9] Phospho                         18254.77813   8529.755
##                                                  M5_1_MSS   M5_2_MSS
## LYPELSQYMGLSLNEEEIR[2] Phospho|[9] Oxidation       0.000   6269.141
## VDKVIQAQTAFSANPANPAILSEASAPIPHDGNLYPR[35] Phospho  2102.056  50355.051
## VIQAQTAFSANPANPAILSEASAPIPHDGNLYPR[32] Phospho    77323.019 307637.429
## HADAEMTGYVVTR[6] Oxidation|[9] Phospho          104287.815  75887.365
## HADAEMTGYVVTR[9] Phospho                         35955.901  44102.316
##                                                  T49_1_MSS T49_2_MSS
## LYPELSQYMGLSLNEEEIR[2] Phospho|[9] Oxidation      1135.8169  21933.90
## VDKVIQAQTAFSANPANPAILSEASAPIPHDGNLYPR[35] Phospho   248.9275   3239.16
## VIQAQTAFSANPANPAILSEASAPIPHDGNLYPR[32] Phospho    98442.2773 192982.37
## HADAEMTGYVVTR[6] Oxidation|[9] Phospho          773377.4981 481165.54
## HADAEMTGYVVTR[9] Phospho                         57145.1682  34638.01
##                                                    M42_1_PD   M42_2_PD
## LYPELSQYMGLSLNEEEIR[2] Phospho|[9] Oxidation          0.000       0.00
## VDKVIQAQTAFSANPANPAILSEASAPIPHDGNLYPR[35] Phospho   1315.904       0.00
## VIQAQTAFSANPANPAILSEASAPIPHDGNLYPR[32] Phospho     24851.344   16547.95
## HADAEMTGYVVTR[6] Oxidation|[9] Phospho           1027196.292 1163747.38
## HADAEMTGYVVTR[9] Phospho                          21231.256   49499.70
##                                                    M43_1_PD    M43_2_PD
## LYPELSQYMGLSLNEEEIR[2] Phospho|[9] Oxidation        772.9056    2136.746
## VDKVIQAQTAFSANPANPAILSEASAPIPHDGNLYPR[35] Phospho      0.0000       0.000
## VIQAQTAFSANPANPAILSEASAPIPHDGNLYPR[32] Phospho      5565.2821       0.000
```

```
## HADAEMTGYVVTR[6] Oxidation|[9] Phospho            4080239.1820 4885818.113
## HADAEMTGYVVTR[9] Phospho                          666107.0448  379313.615
##                                                     M64_1_PD     M64_2_PD
## LYPELSQYMGLSLNEEEIR[2] Phospho|[9] Oxidation        1820.724    1727.9098
## VDKVIQAQTAFSANPANPAILSEASAPIPHDGNLYPR[35] Phospho       0.000     892.3565
## VIQAQTAFSANPANPAILSEASAPIPHDGNLYPR[32] Phospho      3264.563    5901.9577
## HADAEMTGYVVTR[6] Oxidation|[9] Phospho           3093786.793 2759104.5440
## HADAEMTGYVVTR[9] Phospho                          255792.117  579765.0018
```

Sabem que el SE s'ha generat correctament per diversos motius:

1. Al resum del SE, veiem que les dimensions són de 1438 files i 12 columnes. Sabem que són 13 columnes les que ens interessen, pero la que falta és, essencialment, rowData, que conté la columna *SequenceModifications*.
2. Observem 12 colnames diferents, amb 2 coldata names, indicant així les mostres (6) i els diferents fenotips (2).
3. Quan fem un print de les cinc primeres files de l'assay, veiem com, per una banda, la columna de *SequenceModifications*, emmagatzemada a rowData, s'ha afegit correctament, i que les columnes pertanyents a les mostres i els fenotips, emmagatzemat a colData, també s'han afegit correctament, amb els seus respectius valors, emmagatzemats a Assay.

3.3 Data visualization

Després de fer una primera ullada a les dades per així poder crear el SE, i en la infomació que ens dona el SE *per se*, podem veure que estem davant un conjunt de dades relativament senzill, on les metadades son caracters i les mesures son numèriques, precises. Per tant, podem trucar un *summary* d'aquestes mesures:

Figura 3: *summary* de les mesures guardades al Assay del *SummarizedExperiment*; *chunk Data_Visualization*

```
##     M1_1_MSS           M1_2_MSS           M5_1_MSS           M5_2_MSS
##  Min.   :       0   Min.   :       0   Min.   :       0   Min.   :       0
##  1st Qu.:    5653   1st Qu.:    5497   1st Qu.:    2573   1st Qu.:    3273
##  Median :   30682   Median :   26980   Median :   20801   Median :   26241
##  Mean   :  229841   Mean   :  253151   Mean   :  232967   Mean   :  261067
##  3rd Qu.:  117373   3rd Qu.:  113004   3rd Qu.:  113958   3rd Qu.:  130132
##  Max.   :16719906   Max.   :43928481   Max.   :15135169   Max.   :19631820
##    T49_1_MSS          T49_2_MSS          M42_1_PD           M42_2_PD
##  Min.   :       0   Min.   :       0   Min.   :       0   Min.   :       0
##  1st Qu.:    9306   1st Qu.:    8611   1st Qu.:    5341   1st Qu.:    4216
##  Median :   55641   Median :   46110   Median :   36854   Median :   30533
##  Mean   :  542449   Mean   :  462616   Mean   :  388424   Mean   :  333587
##  3rd Qu.:  223103   3rd Qu.:  189141   3rd Qu.:  180252   3rd Qu.:  152088
##  Max.   :49218872   Max.   :29240206   Max.   :48177680   Max.   :42558111
##     M43_1_PD           M43_2_PD           M64_1_PD           M64_2_PD
##  Min.   :       0   Min.   :       0   Min.   :       0   Min.   :       0
##  1st Qu.:   19641   1st Qu.:   17299   1st Qu.:   11038   1st Qu.:    8660
##  Median :   67945   Median :   59607   Median :   52249   Median :   47330
##  Mean   :  349020   Mean   :  358822   Mean   :  470655   Mean   :  484712
##  3rd Qu.:  205471   3rd Qu.:  201924   3rd Qu.:  209896   3rd Qu.:  206036
##  Max.   :35049402   Max.   :63082982   Max.   :71750330   Max.   :88912734
```

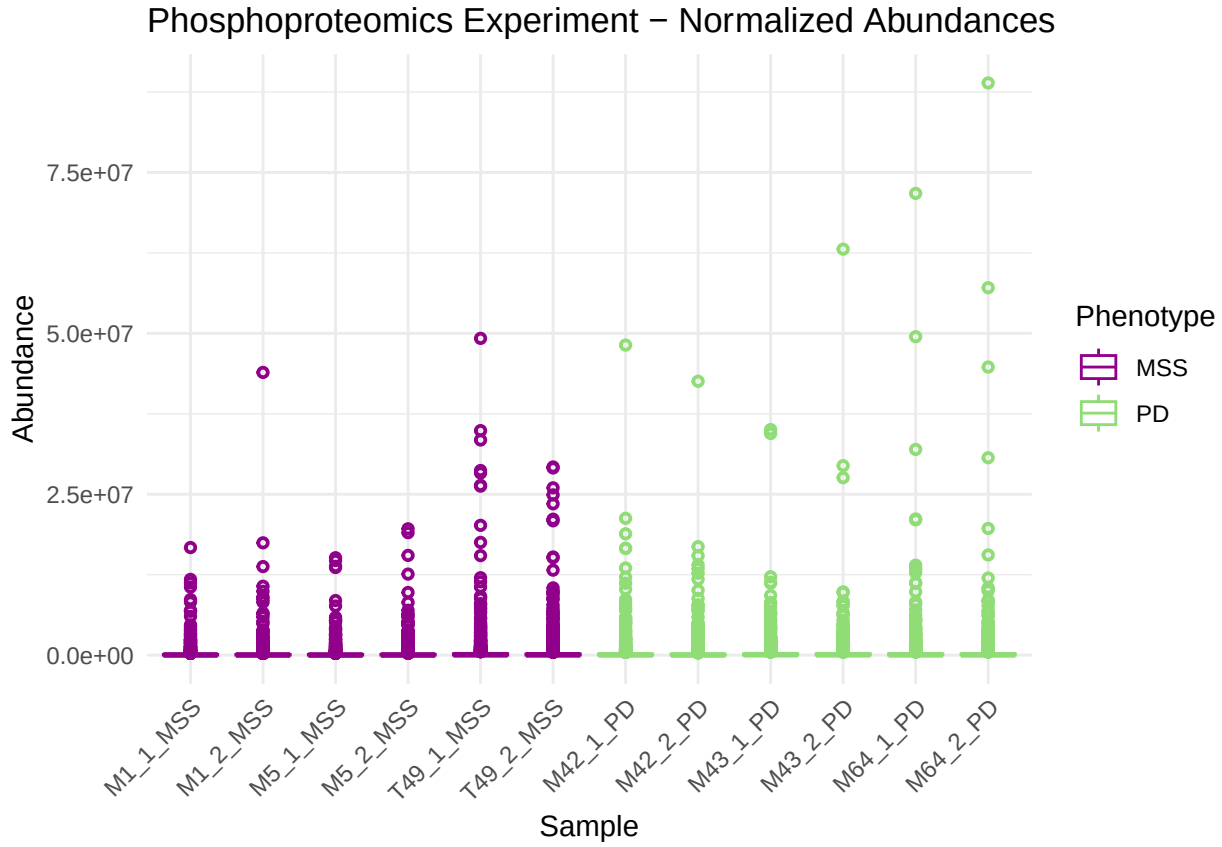
Com es pot observar, s'ha generat un *summary* de les dades que tenim grabades al SE. Com veiem, aquestes dades les hem analitzat per mostra i rèplica, i podem veure que aquestes mesures d'abundància són semblants entre si, fet que ens porta a pensar que té sentit realitzar **un boxplot on cada caixa seria cada rèplica**. No tindria sentit fer un boxplot del conjunt de dades sencer, sense tenir en compte a quina mostra pertanyen

aquestes mesures, ja que seriem incapaços d'obtenir dades sobre el nostre objectiu: identificar els fosfopèptids que permetin diferenciar de manera significativa els dos grups tumorals estudiats.

3.3.1 Boxplots

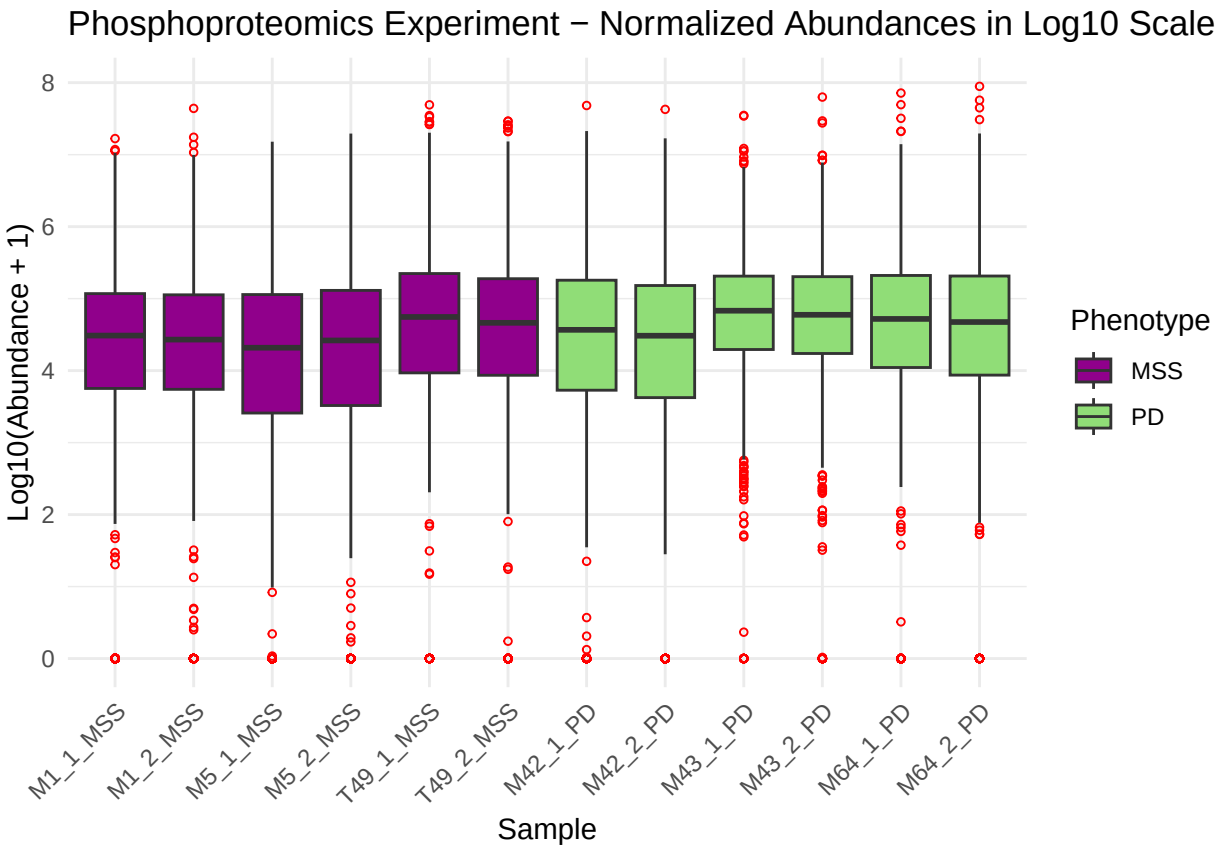
Abans de poder generar els boxplots, tenim un inconvenient, i és la naturalesa de les dades que utilitza *ggplot2*: aquest paquet treballa amb dades en *long format*. Al tenir diverses columnes, pertinents a diferents mostres, nosaltres tenim *wide format*. Per tant, hem de, primer, transferir les dades del nostre SE a un data frame per a poder treballar amb elles i, una vegada fet això, transformar aquestes dades a un format *long*, generant *boxplot_data_long* i *phenotype_map*, per poder agrupar i ordenar més tard per fenotips (*chunk Data_Visualization*). Un cop feta la transofrmació passem, ara sí, a generar els boxplots:

Figura 4: Boxplot no logarítmic de les abundàncies normalitzades; *chunk Data_Visualization*



Veient com es veu el boxplot anterior, pràcticament no es pot extreure cap conclusió ni cap anàlisi visual. De manera alternativa, podem provar a generar aquest mateix boxplot, però en escala logarítmica:

Figura 5: Boxplot logarítmic de les abundàncies normalitzades; *chunk Data_Visualization*

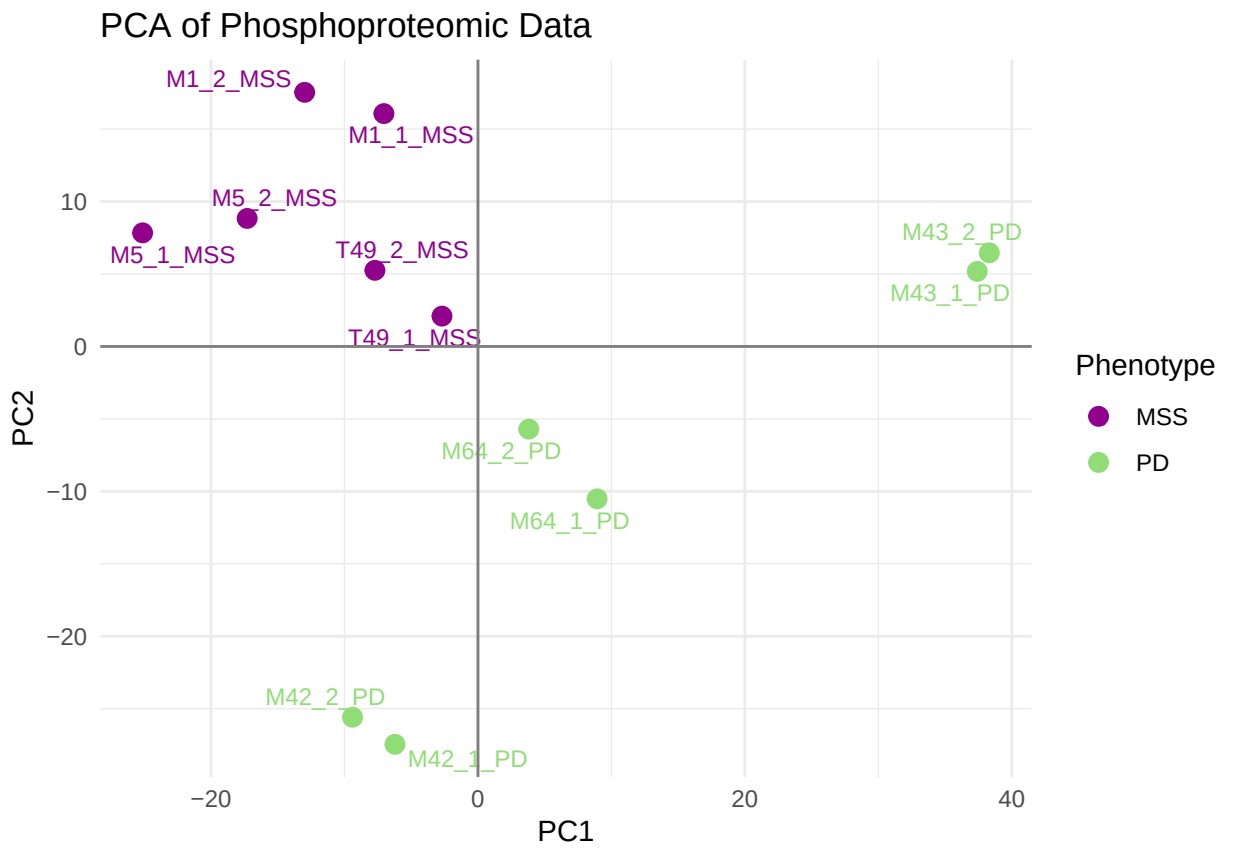


Com bé veniem dient, podem confirmar que obtenim un boxplot molt més "maco", útil i llegible si el realitzem amb les abundàncies en escala logarítmica. En aquest cas, ara podem analitzar que, per una banda, si ens fixem en les medianes segons fenotip, el fenotip MSS presenta una distribució més homogenia, notant-se, sobre tot, pel número d'outliers que presenta aquest fenotip, en comparació al PD. No obstant, no podem observar un patró diferencial clar entre els fenotips. Tenint en compte totes aquestes dades, estem en una situació ideal per comprovar si podem obtenir més informació al fer una visualització multivariant, concretament, un anàlisi de components principals (PCA).

3.3.2 Visualització mulitvariant

Basant-nos en una publicació d'Alboukadel Kassambara a STHDA (Statistical Tools for High-throughput Data Analysis, https://sthda.com/english/articles/31-principal-component-methods-in-r-practical-guide/118-principal-component-analysis-in-r-prcomp-vs-princomp), procedim a generar el PCA:

Figura 6: plot PCA per a les dades en escala logarítmica; *chunk Data_Visualization*

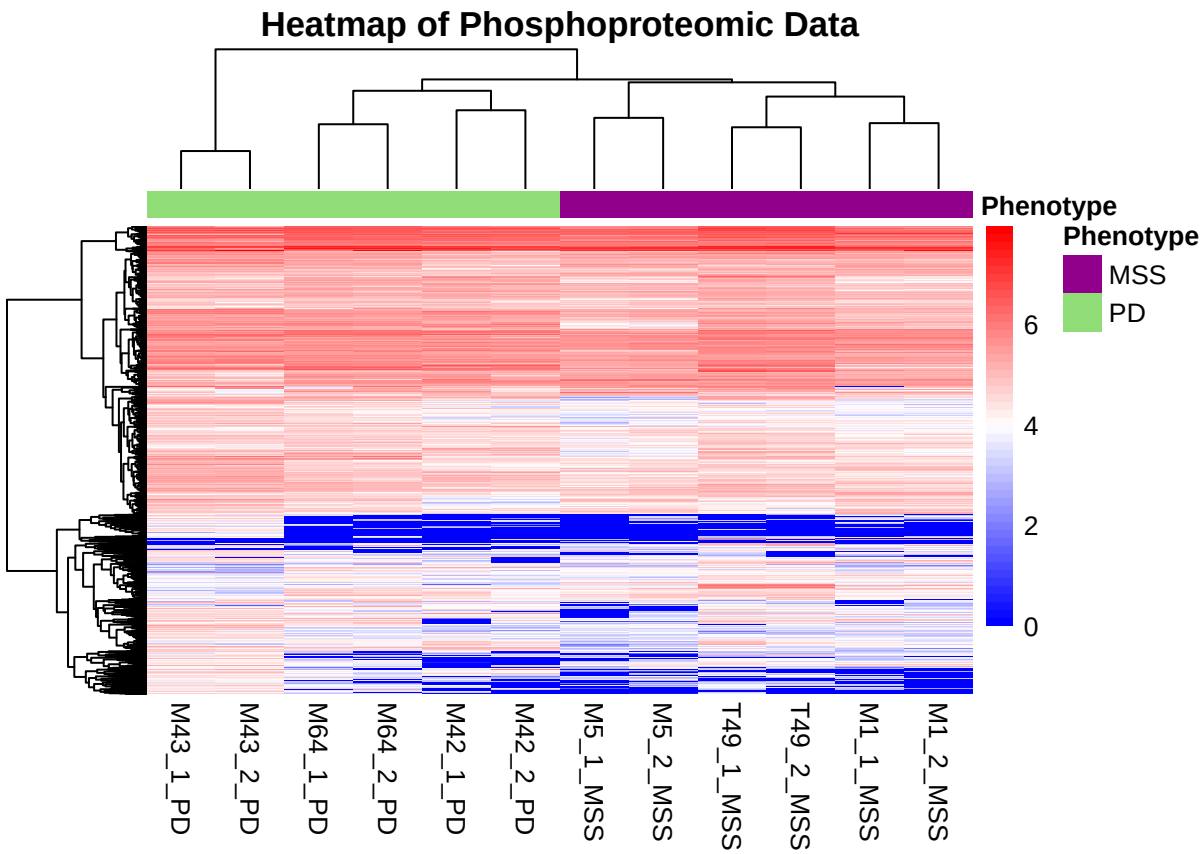


A partir d'aquest plot, podem extreure més informació, al veure que, primer de tot, els dos fenotips es troben clarament separats i diferenciats. En específic, el fenotip MSS s'agrupa en el quadrant superior exclusivament, mentre que les mostres del fenotip PD es distribueixen pels altres tres quadrants. Per tant, també podem afirmar que PD presenta una major heterogeneitat en comparació a MSS. Si ens fixem en els eixos, podem dir també que aquesta diferenciació es produeix, essencialment, degut a PC1 **Per tant, podriem afirmar, de manera prematura, que els fosfopèptids que contribueixen més a PC1, són els fosfopèptids més probables a ser diferencials entre els fenotips MSS i PD.** Igualment, no podem afirmar que cap fosfopèptid que contribueix en PC2 no pugui ser diferencial. Caldria, per tant, fer una anàlisi estadístic més complex i metòdic.

3.3.3 Heatmap

Una vegada confirmada l'heterogeneitat entre MSS i PD amb la visualització multivariant, podem fer una visualització més individualitzada del fosfopèptids per veure cap patró en específic mitjançant un *heatmap* amb dendogrames a cada eix:

Figura 7: Heatmap del fosfopèptids segon mostra i fenotip amb dendogrames als eixos; *chunk Data_Visualization*



Al observar el heatmap i els dendogrames, podem extreure diverses conclusions.

Si començem amb els dendogrames dels fenotips, veiem com MSS i PD s'agrupen en diferents branques, suggerint el que ja veniem imaginant des de fa un parell de plots: hi ha un patró de fosforilació distintiu entre aquests dos fenotips.

Baixant ja a l'anàlisi del heatmap com a tal, podem veure com hi ha un patró de fosfopèptids marcat en blau, majoritàriament al fenotip MSS. Aquests fosfopèptids blaus, molt probablement siguin els que hagin provocat aquesta migració per part de PC1 al PCA anterior, confirmant que aquests serien els diferencials i essencials per poder distinguir entre els dos fenotips. D'altra banda, i com bé veniem avançant, també trobem fosfopèptids diferencials dins del fenotip PD, ja que aquest fenotip també presenta un nombre de fosfopèptids en tonalitat blava.

No obstant, sabem que estem tractant un total de 1438 fosfopèptids, i és impossible identificar-los així visualment. Caldria, per tant, un meta-anàlisi sobre el heatmap per acotar a les tonalitats més blavoses i, per tant, menys freqüents, per tal d'així poder acotar fins arribar als fosfopèptids diferencials.

4. Conclusions

Després de tota aquesta sèrie d'anàlisi i de generació de plots de diferent naturalesa, podem afirmar que, simplement fent una visualització de les dades, podem afirmar que, en efecte, tenim una sèrie de fosfopèptids que ens serveixen com a factor diferencial per poder identificar i diferenciar els fenotips MSS i PD.

Pel fet del tractament de les dades, en efecte l'objecte SummarizedExperiment ha sigut d'utilitat per entendre l'experiment, l'organització de les dades i, a més, per poder manipular les dades a l'hora d'escriure codi.

En definitiva, podem afirmar que hem realitzat un treball previ a l'anàlisi estadístic força complet i informatiu, el qual ja ens dona una tendència clara a els resultats que podem esperar de l'anàlisi en profunditat.

5. Repositori

Tot i que en aquest informe es troba la teoria, el codi i tots els plots generats a partir d'aquest codi, també s'ha creat un repositori en github on s'ha pujat:

- una còpia d'aquest informe
- l'arxiu Rmarkdown d'aquest informe, amb tot el codi
- el codi per separat, en un arxiu R
- una carpeta amb totes les dades generades en format text, a més de l'arxiu de dades original
- l'arxiu de format binari .RData
- una carpeta amb imatges dels plots generats.

L'enllaç al repositori de github és el següent: https://github.com/arnaumenamolina/Mena-Molina-Arnau-PEC1/tree/main

Annex I. Codi complet

1. Primera visualització prèvia a la generació del SummarizedExperiment

```
# Carreguem readxl per poder veure l'arxiu original
library(readxl)

# Definim la ruta de l'arxiu
file_path <-
↪  "C:/Users/Arnau/Desktop/PEC1_ADO/original_data/TIO2+PTYR-human-MSS+MSIvsPD.XLSX"

# Carreguem les dades com original_data
original_data <- read_excel(file_path, sheet = "originalData")

# Mostrar las primeras 5 filas
head(original_data, 5)
```

2. Generació del SummarizedExperiment i preprocessing de les dades

```
# Carreguem els paquets necessaris
library(SummarizedExperiment)
library(readxl)

# Un cop vist el format original de les dades, seleccionem les que ens interessen i
↪  generem data_filtered
selected_columns <- c("SequenceModifications",
                      "M1_1_MSS", "M1_2_MSS", "M5_1_MSS", "M5_2_MSS",
                      "T49_1_MSS", "T49_2_MSS",
                      "M42_1_PD", "M42_2_PD", "M43_1_PD", "M43_2_PD",
                      "M64_1_PD", "M64_2_PD")
data_filtered <- original_data[selected_columns]

# Generem la matriu assay amb les abundàncies de fosfopèptids
assay_data <- as.matrix(data_filtered[, -1]) # Treiem la columna SequenceModifications,
↪  només volem les dades a assay
rownames(assay_data) <- data_filtered$SequenceModifications

# Definim colData manualment per fer coincidir amb assay_data
samples <- c("M1", "M1", "M5", "M5", "T49", "T49", "M42", "M42", "M43", "M43", "M64",
↪  "M64")
```

```r
phenotypes <- c("MSS", "MSS", "MSS", "MSS", "MSS", "MSS", "PD", "PD", "PD", "PD", "PD",
↪  "PD")
col_data <- DataFrame(Sample = samples, Phenotype = phenotypes)
rownames(col_data) <- colnames(assay_data)

# Afegim a row_data les metadades de SequenceModifications com informació addicional dels
↪  fosfopèptids
row_data <- DataFrame(SequenceModifications = data_filtered$SequenceModifications)
rownames(row_data) <- data_filtered$SequenceModifications

# Un cop format Assay, colData i rowData, generem el SummarizedExperiment
se <- SummarizedExperiment(
  assays = list(counts = assay_data),
  rowData = row_data,
  colData = col_data
)

# Visualitzem assay del nostre SummarizedExperiment per veure que les dades hi són, igual
↪  que les metadades de les columnes i les files
se
head(assay(se), 5)
```

3. Data Visualization: boxplots, PCA i heatmap + dendogrames

```r
library(ggplot2)
library(ggrepel)
library(tidyr)
library(pheatmap)

summary(assay(se))

# Definim els colors que volem que s'utilitzin per als plots
custom_colors <- c("MSS" = "#90008B", "PD" = "#90DD77")

# Preparem les dades pel boxplot
# Utilitzant tidyr, passem Assay a un data frame per poder generar els boxplots.
boxplot_data <- as.data.frame(assay_data)

# Degut a la naturalesa de ggplot2, paquet que utilitzarem per generar els boxplots, hem
↪  de passar les dades a un format "long", ja que així facilitem el mapeig de variables
↪  estétiques (color, fill, tema...), de mateixa manera que així també podrem agrupar
↪  les dades per grups (fenotips, en aquest cas)
boxplot_data_long <- pivot_longer(boxplot_data, cols = everything(), names_to = "Sample",
↪  values_to = "Abundance")

# Creem un vector de Phenotype que coincideixi amb cada mostra generant els dos grups
↪  segons fenotip.
phenotype_map <- setNames(phenotypes, colnames(assay_data))
boxplot_data_long$Phenotype <- phenotype_map[boxplot_data_long$Sample]

# Ordenem els fenotips per a que apareguin així als boxplots
sample_order <- c(colnames(assay_data)[phenotypes == "MSS"],
↪  colnames(assay_data)[phenotypes == "PD"])
boxplot_data_long$Sample <- factor(boxplot_data_long$Sample, levels = sample_order)
```

```r
# 1. Boxplot d'abundàncies sense transformació logarítmica
ggplot(boxplot_data_long, aes(x = Sample, y = Abundance, fill = Phenotype)) +
  geom_boxplot(aes(color = Phenotype), outlier.shape = 21, outlier.size = 1,
  ↪  outlier.stroke = 1, outlier.fill = NA) +
  scale_fill_manual(values = custom_colors) +
  scale_color_manual(values = custom_colors) +
  labs(title = "Phosphoproteomics Experiment - Normalized Abundances", x = "Sample", y =
  ↪  "Abundance") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1)) +
  guides(fill = "none")

# 2. Boxplot logarítmic
# Creem una nova columna para les abundàncies logarítmiques. Afegim el + 1 al final per
↪  evitar problemes amb els 0s, tenint en compte que treballem en escala log10.
boxplot_data_long$LogAbundance <- log10(boxplot_data_long$Abundance + 1)

ggplot(boxplot_data_long, aes(x = Sample, y = LogAbundance, fill = Phenotype)) +
  geom_boxplot(outlier.colour = "red", outlier.fill = NA, outlier.shape = 21,
  ↪  outlier.size = 1) +  # Outliers vermells
  scale_fill_manual(values = custom_colors) +
  labs(title = "Phosphoproteomics Experiment - Normalized Abundances in Log10 Scale", x =
  ↪  "Sample", y = "Log10(Abundance + 1)") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))

# 3. PCA
# Generem el PCA amb les dades logarítmiques
pca_res <- prcomp(t(log10(assay_data + 1)), scale. = FALSE)
pca_data <- data.frame(PC1 = pca_res$x[, 1], PC2 = pca_res$x[, 2], Phenotype =
↪  phenotypes)
pca_data$Sample <- colnames(assay_data)  # Cridem a assay_data per poder identificar els
↪  dots dins el PCA

# Creem el plot PCA
ggplot(pca_data, aes(x = PC1, y = PC2, color = Phenotype, label = Sample)) +
  geom_point(size = 3) +
  geom_text_repel(size = 3) +
  scale_color_manual(values = custom_colors) +
  labs(title = "PCA of Phosphoproteomic Data", x = "PC1", y = "PC2") +
  geom_hline(yintercept = 0, color = "grey50", linetype = "solid") +
  geom_vline(xintercept = 0, color = "grey50", linetype = "solid") +
  theme_minimal()

# 4. Heatmap + dendograma en els eixos
# Creem el log d'Assay i eliminem els NA, en cas de que n'hi hagin.
log_data <- log10(assay_data + 1)
log_data[is.na(log_data)] <- 0

# Configurem les columnes de les mostres
annotation_col <- data.frame(Phenotype = phenotypes)
rownames(annotation_col) <- colnames(log_data)
```

```r
# Assignem els colors que ja veniem utilitzant per a cada fenotip
annotation_colors <- list(Phenotype = custom_colors)

# Creem el heatmap
pheatmap(log_data,
         annotation_col = annotation_col,
         annotation_colors = list(Phenotype = custom_colors),
         color = colorRampPalette(c("blue", "white", "red"))(50),
         cluster_rows = TRUE, cluster_cols = TRUE,
         main = "Heatmap of Phosphoproteomic Data",
         show_rownames = FALSE)

# Exportació de totes les dades en format text
# Directori on guardar els CSV
output_dir <- "C:/Users/Arnau/Desktop/PEC1_ADO/data/"

# 1. Dades filtrades
write.csv(data_filtered, file.path(output_dir, "data_filtered.csv"), row.names = FALSE)

# 2. assay_data
write.csv(as.data.frame(assay_data), file.path(output_dir, "assay_data.csv"), row.names =
↪  TRUE)

# 3. colData
write.csv(as.data.frame(col_data), file.path(output_dir, "col_data.csv"), row.names =
↪  TRUE)

# 4. rowData
write.csv(as.data.frame(row_data), file.path(output_dir, "row_data.csv"), row.names =
↪  TRUE)

# 5. boxplot_data_long
write.csv(boxplot_data_long, file.path(output_dir, "boxplot_data_long.csv"), row.names =
↪  FALSE)

# 6. pca_data
write.csv(pca_data, file.path(output_dir, "pca_data.csv"), row.names = FALSE)

# 7. log_data
write.csv(as.data.frame(log_data), file.path(output_dir, "log_data.csv"), row.names =
↪  TRUE)

# 8. annotation_col
write.csv(annotation_col, file.path(output_dir, "annotation_col.csv"), row.names = TRUE)
```