PEC1 Informe

Esperanza López Merino

2025-04-01

Contents

Resumen

En esta PEC hemos trabajado con unos datos de fosfoproteómica que, una vez nos hemos familiarizado brevemente con su estructura, hemos convertido en un objeto del tipo SummarizedExperiment. Una vez creado el objeto hemos hecho una exploración inicial de los datos identificando algunos puntos importantes a tener en cuenta durante el preprocesado. La mitad de las réplicas técnicas tienen una correlación superior al 90%, no obstante, la otra mitad muestra una correlación inferior al 75%. En esta línea, parece haber también valores iguales al 0 exacto (aunque en una pequeña proporción), lo cual habría que revisar pues probablemente se deba a la técnica. Atajar estas cuestiones estableciendo criterios objetivos de filtrado es importante para asegurarnos de la fiabilidad de los resultados que se obtengan en análisis posteriores.

Objetivos

El objetivo principal de este trabajo es hacer una exploración inicial de los datos de fosfoproteómica proporcionados para familiarizarse con Bioconductor, git y el manejo de este tipo de datos en R. No obstante, el objetivo principal del estudio era determinar si existen diferencias en los patrones de fosforilacion de dos modelos PDX (patient-derived xenograft) derivados de tumores humanos, los cuales pueden ser una nueva aproximación en la terapia del cancer (Liu et al. 2023). Con esto, planteamos los siguientes objetivos especificos:

- Obtener los ficheros y cargarlos en R.
- Generar un objeto Summarized Experiment (Bioconductor).
- Hacer una exploración inicial de los datos para detectar posibles puntos a tratar durante el preprocesado de los datos y evaluar las posibles diferencias entre los PDX.
- Compartir el código y los resultados en https://github.com/espelm/Lopez-Merino-Esperanza-PEC1

Métodos

Para realizar este trabajo hemos seleccionado el Dataset 2018-Phosphotroteomics del repositorio de github nutrimetabolomics/metaboData disponible en https://github.com/nutrimetabolomics/metaboData/tree/main/Datasets/2018-Phosphoproteomics. Se han seleccionado estos datos por un interés personal, ya que posiblemente en el futuro tendré que enfrentarme a estudios de este tipo.

El análisis se ha llevado a cabo usando RStudio RStudio 2023.12.1+402. Para más información acerca del código y las librerias usadas, el código completo se encuentra disponible en el repositirio de github espelm/Lopez-Merino-Esperanza-PEC1. No obstante, destacar el uso de la librería SummarizedExperiment para estructurar los datos de fosfoproteómica.

Inicialmente se intentaron descargar los datos del repositorio al proyecto creado con git directamente desde R, no obstante, el fichero xlsx no se podía abrir. Para bypassear el problema decicidimos descargar manualmente los archivos.

```
#data_url <- "https://github.com/nutrimetabolomics/metaboData/blob/main/Datasets/2018-Phosphoproteomics/
#download.file(data_url, "phosphoproteomics_data.XLSX", mode = "wb", method = "wininet")
#description_url <- "https://github.com/nutrimetabolomics/metaboData/blob/main/Datasets/2018-Phosphopro
#download.file(description_url, "description.md", mode = "wb")
```

Como indicábamos, los datos proporcionados se encontraban en formato xlsx, correspondiéndose la primera pestaña a los datos en sí y la segunda a la información relativa a las muestras. Además venian acompañados de un pequeño archivo de descripción de los mismos (.md), donde se explica con más detalle su origen. Una vez cargados los datos se exploraron brevemente y se pasó a crear el objeto Summarized Experiment.

Finalmente se hizo una breve exploración de los datos, identificando valores no asignados, 0 exactos y explorando brevemente las diferencias entre ambos grupos de PDX.

Resultados

El primer paso fue cargar los datos en RStudio. Para ello cargamos por un lado los datos de abundancia de cada phosphopéptido con su identificación. Separadamente cargamos la información de las muestras con sus fenótipos y sus réplicas técnicas. Para terminar cargamos la descripción del estudio.

```
phospho_data <- read_excel("phosphoproteomics_data.xlsx", sheet = "originalData")
head(phospho_data)
str(phospho_data)
samples <- read_excel("phosphoproteomics_data.xlsx", sheet = "targets")
head(samples)
str(samples)
description <- readLines("description.md")
```

Así, los datos muestran la abundancia de 1438 péptidos en 12 muestras. Estas muestras se separan en 3 muestras por línea PDX con 2 réplicas técnicas cada una.

A continuación, pasamos a crear el objeto de la clase SummarizedExperiment. Para ello debemos quedarnos con los datos de los niveles de fosfopéptidos de las muestras (para filtrarlos, nos beneficiamos de que son las únicas columnas que empiezan por M o T). Además, creamos los metadatos de las muestras (nombre, grupo de tratamiento y réplica) y los metadatos de los péptidos identificados(péptido, accession number, descripción y score de la identificación del peptido). Finalmente, tomamos la descripción del experimento como metadatos del objeto.

```
abundance_colums <- grep("^M|^T", colnames(phospho_data), value = TRUE)  #muestras
abundance_data <- phospho_data[, abundance_colums]  #datos

samples_metadata <- data.frame(
  Sample = colnames(abundance_data),
  Group = ifelse(grepl("^M1|^M5|^T49", colnames(abundance_data)), "MSS", "PD"),
  Replicate = rep(c(1,2), times = 6),  # Dos réplicas por muestra
  stringsAsFactors = FALSE)

peptides_metadata <- data.frame(  #Metadatos de los peptidos identificados
  Peptide = phospho_data$SequenceModifications,
  Accession = phospho_data$Accession,
  Description = phospho_data$Description,
  Score = phospho_data$Score,
  stringsAsFactors = FALSE)

phospho_se <- SummarizedExperiment(
  assays = list(counts = as.matrix(abundance_data)),  # Datos de abundancia
  colData = DataFrame(samples_metadata),  #Metadatos de las muestras
  rowData = DataFrame(peptides_metadata),  # Metadatos de los peptidos identificados
  metadata = description) #Metadatos del experimento
```

Una vez creado el objeto, lo exploramos brevemente. Por razones de longitud y claridad no imprimiremos todos los resultados del código.

```
summary(phospho_se)
```

```
## [1] "SummarizedExperiment object of length 1438 with 4 metadata columns"
```

```
dim(phospho_se)
```

```
## [1] 1438   12
```

```
colData(phospho_se)
```

```
## DataFrame with 12 rows and 3 columns
##                Sample       Group Replicate
##           <character> <character> <numeric>
## M1_1_MSS     M1_1_MSS         MSS         1
## M1_2_MSS     M1_2_MSS         MSS         2
## M5_1_MSS     M5_1_MSS         MSS         1
## M5_2_MSS     M5_2_MSS         MSS         2
```

```
## T49_1_MSS   T49_1_MSS         MSS         1
## ...               ...         ...       ...
## M42_2_PD     M42_2_PD          PD         2
## M43_1_PD     M43_1_PD          PD         1
## M43_2_PD     M43_2_PD          PD         2
## M64_1_PD     M64_1_PD          PD         1
## M64_2_PD     M64_2_PD          PD         2
```

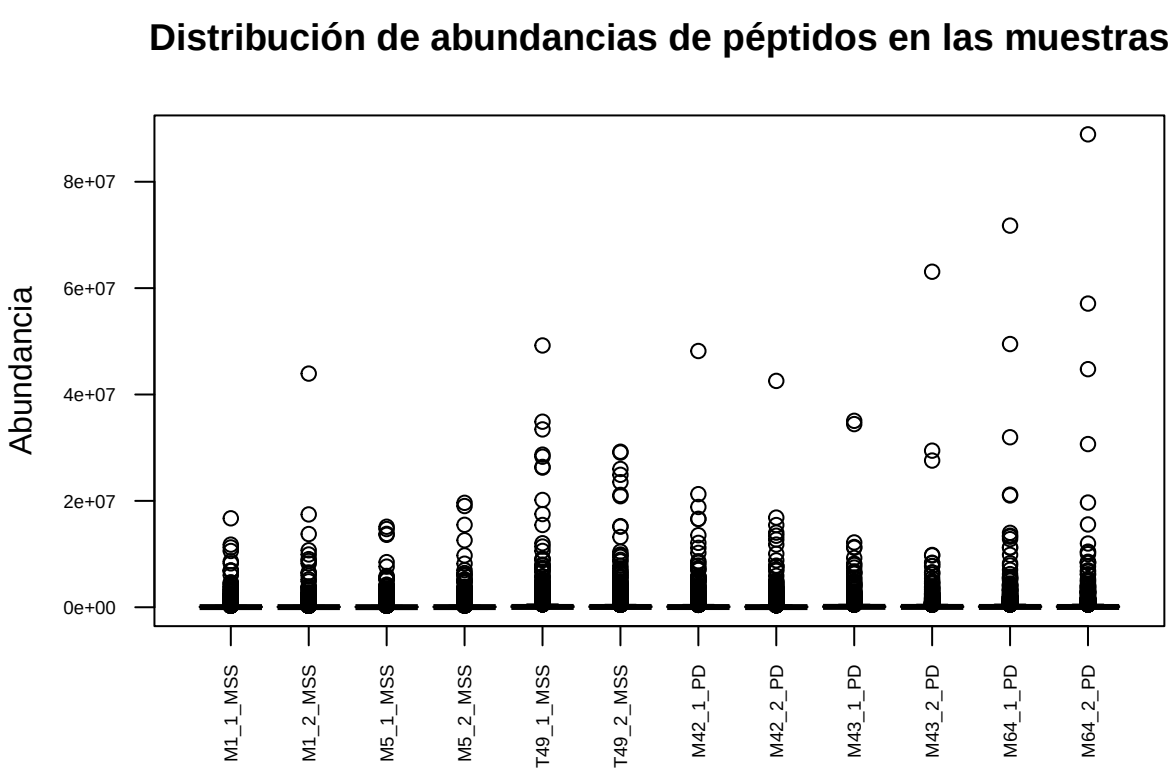
```
rowData(phospho_se)
```

```
## DataFrame with 1438 rows and 4 columns
##                     Peptide   Accession            Description     Score
##                 <character> <character>            <character> <numeric>
## 1    LYPELSQYMGLSLNEEEIR[..      O00560 Syntenin-1 OS=Homo s..     48.07
## 2    VDKVIQAQTAFSANPANPAI..      O00560 Syntenin-1 OS=Homo s..     67.05
## 3    VIQAQTAFSANPANPAILSE..      O00560 Syntenin-1 OS=Homo s..     77.71
## 4    HADAEMTGYVVTR[6] Oxi..      O15264 Mitogen-activated pr..     44.87
## 5    HADAEMTGYVVTR[9] Pho..      O15264 Mitogen-activated pr..     67.42
## ...                     ...         ...                    ...       ...
## 1434 YLLSQSSPAPLTAAEEELR[..      Q12792 Twinfilin-1 OS=Homo ..     56.19
## 1435   YLSFTPPEK[3] Phospho      Q13177 Serine/threonine-pro..     39.14
## 1436 YNLDASEEEDSNK[6] Pho..      O95218 Zinc finger Ran-bind..     80.66
## 1437 YQDEVFGGFVTEPQEESEEE..      Q13283 Ras GTPase-activatin..     40.01
## 1438 YSPSQNSPIHHIPSRR[1] ..      Q9NYF8 Bcl-2-associated tra..     36.71
```

```
metadata(phospho_se)
```

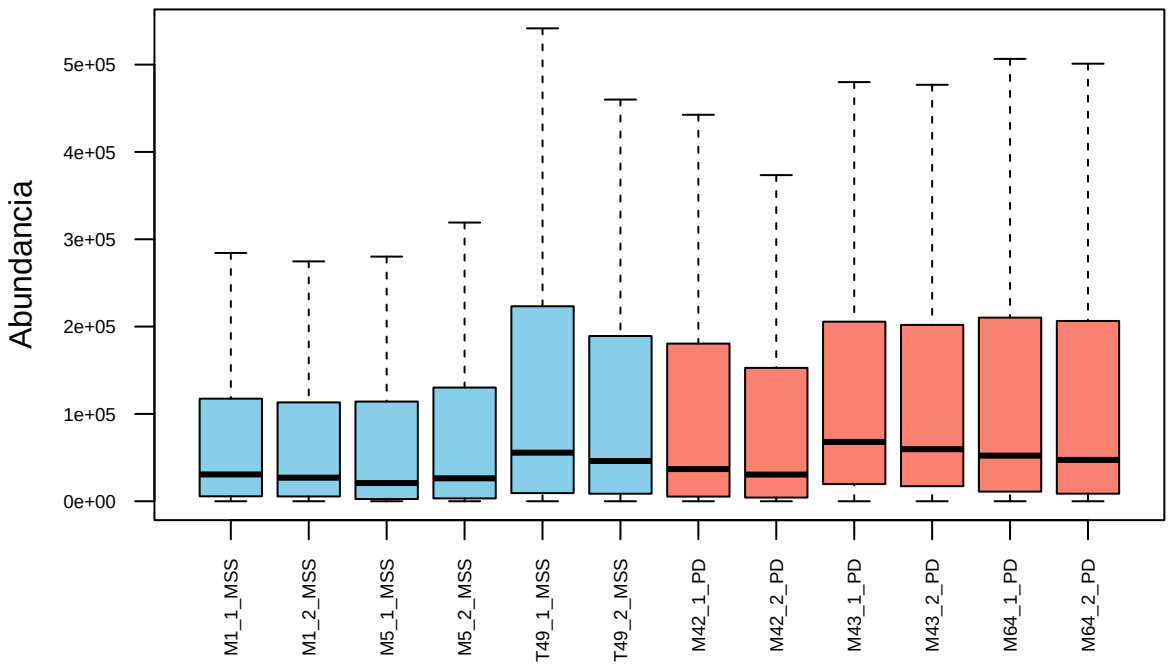
Una vez comprobado que el objeto SummarizedExperiment se ha creado correctamente, empezamos a explorar los datos para ver si es necesario hacer algún preprocesado antes de hacer el análisis estadístico (fuera de los objetivos de esta PEC).

Empezamos por hacer una representacion rápida de la abundancia de fosfopéptidos por muestra con un boxplot.



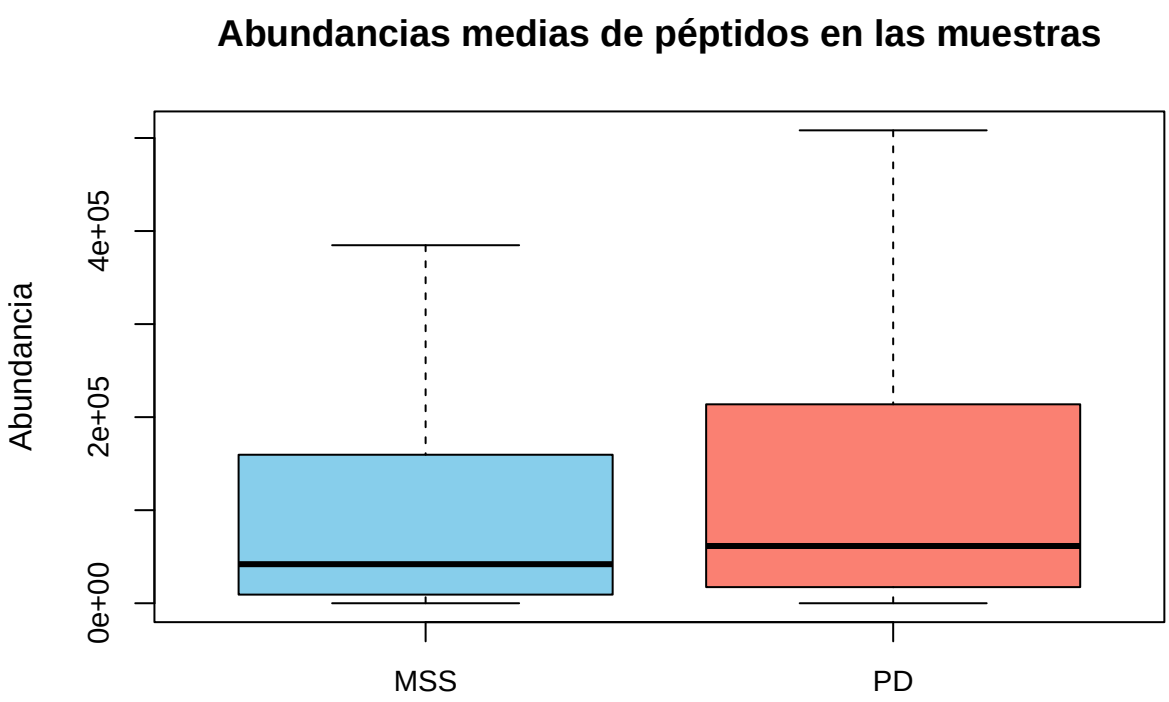
Pudimos ver que hay bastanes puntos fuera de los bigotes del boxplot (de hecho se ven ni las cajas ni los bigotes), lo que probablemente puedan ser outliers. Valdría la pena revisar si estos puntos se conservan en las replicas técnicas (en cuyo caso es más probable que representen variabilidad biológica) o bien puedan deberse a errores técnicos. Por ejemplo, en la primera muestra M1 del grupo MSS, el punto entorno a $4x10^7$ es sospechoso de ser un error técnico.

Aunque no es correcto, con fines exploratorios eliminamos los outliers para ver bien las cajas y los bigotes.



Así pudimos ver que en general las muestras del grupo PD parecen tener más variabilidad y tal vez unos mayores niveles de fosfoproteínas (aunque las dos réplicas técnicas T49 del grupo MSS tienen una variabilidad parecida, lo que llama la atención además de que tienen una codificación distinta al resto -T y no M y una numeración alta-).

Viendo estos datos decidimos representar la abundancia de los peptidos por tipo de PDX.



Comprobamos que la distribución de los datos no es normal (se ve que la distribución es asimétrica), pero aunque no es correcto hacerlo antes del preprocesado, no pudimos resistirnos a ver que efectivamente parece que SÍ hay diferencias entre los niveles globales de phosphoproteínas. IMPORTANTE, este análisis global no es concluyente hasta que no se repita con los datos filtrados.

```
shapiro.test(mean_abundance$MSS)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  mean_abundance$MSS
## W = 0.22299, p-value < 2.2e-16
```

```
shapiro.test(mean_abundance$PD)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  mean_abundance$PD
## W = 0.15894, p-value < 2.2e-16
```

```
wilcox.test(mean_abundance$MSS, mean_abundance$PD)
```

```
##
```

```
##  Wilcoxon rank sum test with continuity correction
##
## data:  mean_abundance$MSS and mean_abundance$PD
## W = 908745, p-value = 1.888e-08
## alternative hypothesis: true location shift is not equal to 0
```

Continuando con nuestra evaluación inicial de los datos, comprobamos que había 0 *missing values*, pero 1416 ceros exactos (de 6.2801931×10^9 medidas). Esto puede ser tecnicamente incorrecto, ya que debemos poder diferenciar entre péptidos que se encuentran fosforilados a muy bajos niveles y péptidos que simplemente no hemos sido capaces de identificar en nuestras muestras.

Para terminar nuestra exploración, retomamos el tema de las réplicas técnicas y construimos una matriz de correlación.

```
##            M1_1_MSS  M1_2_MSS  M5_1_MSS  M5_2_MSS T49_1_MSS T49_2_MSS  M42_1_PD
## M1_1_MSS  1.0000000 0.6879669 0.8254004 0.7922283 0.6957629 0.7301037 0.7051752
## M1_2_MSS  0.6879669 1.0000000 0.8266863 0.8554969 0.4586948 0.5822222 0.4536530
## M5_1_MSS  0.8254004 0.8266863 1.0000000 0.9877164 0.6137969 0.7210519 0.6142241
## M5_2_MSS  0.7922283 0.8554969 0.9877164 1.0000000 0.5953579 0.7093807 0.5902427
## T49_1_MSS 0.6957629 0.4586948 0.6137969 0.5953579 1.0000000 0.9158050 0.8017620
## T49_2_MSS 0.7301037 0.5822222 0.7210519 0.7093807 0.9158050 1.0000000 0.7815558
## M42_1_PD  0.7051752 0.4536530 0.6142241 0.5902427 0.8017620 0.7815558 1.0000000
## M42_2_PD  0.6447901 0.3640460 0.4801517 0.4387735 0.7802924 0.7228780 0.9122772
## M43_1_PD  0.5602113 0.4328218 0.4500716 0.4343694 0.6441791 0.6003283 0.7112070
## M43_2_PD  0.4171569 0.7939070 0.5377181 0.5732886 0.4331291 0.4956160 0.4588941
## M64_1_PD  0.6297437 0.3644654 0.4941531 0.4728644 0.7445057 0.6931224 0.8488577
## M64_2_PD  0.5179559 0.8077572 0.6410269 0.6739684 0.5706957 0.6221025 0.6361994
##            M42_2_PD  M43_1_PD  M43_2_PD  M64_1_PD  M64_2_PD
## M1_1_MSS  0.6447901 0.5602113 0.4171569 0.6297437 0.5179559
## M1_2_MSS  0.3640460 0.4328218 0.7939070 0.3644654 0.8077572
## M5_1_MSS  0.4801517 0.4500716 0.5377181 0.4941531 0.6410269
## M5_2_MSS  0.4387735 0.4343694 0.5732886 0.4728644 0.6739684
## T49_1_MSS 0.7802924 0.6441791 0.4331291 0.7445057 0.5706957
## T49_2_MSS 0.7228780 0.6003283 0.4956160 0.6931224 0.6221025
## M42_1_PD  0.9122772 0.7112070 0.4588941 0.8488577 0.6361994
## M42_2_PD  1.0000000 0.8254295 0.5158780 0.8622895 0.6102590
## M43_1_PD  0.8254295 1.0000000 0.7474376 0.8935105 0.7560768
## M43_2_PD  0.5158780 0.7474376 1.0000000 0.5470780 0.9300472
## M64_1_PD  0.8622895 0.8935105 0.5470780 1.0000000 0.7145727
## M64_2_PD  0.6102590 0.7560768 0.9300472 0.7145727 1.0000000
```

Así, es especialmente interesante fijarse en la correlación enre las réplicas técnicas. Dentro del grupo MSS, las réplicas técnicas de M1 muestran una correlación de 0.69, M5 0.99 y T49 0.91. Por su parte, los PDX PD muestran una correlación de 0.91 para M42, 0.74 para M43 y 0.71 para M64.

Discusión

En este trabajo hemos creado un objeto de la clase Summarized Experiment para los datos de un estudio de fosfoproteómica. El manejo del objeto en sí es bastante asequible. No obstante, en cuanto a la exploración y preprocesado de los datos debemos ser cautelosos.

Es especialmente preocupante la baja reproducibilidad entre las réplicas técnicas (baja correlación), que está pro debajo del 75% en 3 de las muestras, la mitad. Considero que se debería establecer un criterio de

reproducibilidad. Por ejemplo, en la actividad 1.3 de la asignatura vimos que para cada gen promediaban las réplicas técnicas A y B por separado y si mostraban una variación de más del 20% las eliminaban del estudio (Cui et al. 2006). Un criterio que vaya en esta línea debería establecerse en este estudio. Además, aunque en una proporción pequeña (2.2547077×10^{-5} %), hemos visto que hay valores 0, pero ningún NA, lo que probablemente indique que si un péptido solo se identifica en un subgrupo de muestras al resto se les da un valor 0. Probablemente la mayoría de 0 desapareceran durante el preprocesado, pero habría que revisarlo.

Distinguir entre la variación biológica y la debida a artefactos técnicos es importante durante el preprocesado de los datos ya que en un análisis posterior puede afectar a un PCA o un ANOVA, que miden la varianza. De hecho, una vez fijado un criterio para excluir péptidos con alta variación artefactual sería interesante llevar a cabo estos análisis junto con un volcano plot para ver que péptidos están diferencialmente fosforilados. También sería interesante constatar si para las proteínas con varios sitios de de fosforilación todos ellos se encuentran fosforilados o solo algunos (implicaciones en regulación, señalización…) y finalmente, usar herramientas como Gene Ontology para identificar que pathways pueden estar diferencialmente activados en estos PDX, lo que puede ayudar a entender mejor los distintos tumores y a la larga acercarnos más a la medicina personalizada.

Por último, sería también interesante revisar los scores que se han obtenido para identificación de las proteínas a partir de los fosfopéptidos, especialmente para aquellas para las que se encuentren diferencias significativas.

Conclusiones

- Los objetos Summarized Experiment almacenan la información de los estudios ómicos de una forma clara y flexible, permitiendo tener por separado los datos, la información de las muestras y la información de las variables(péptidos, genes…).
- Es importante controlar la reproducibilidad técnica en los estudios ómicos durante el preprocesado de los datos. En este caso en concreto destacaremos:
  - Es necesario filtrar los 0 exactos si son un artefacto técnico (no se ha identificado el péptido).
  - Es importante establecer un criterio de reproducibilidad técnica y eliminar los fosfopéptidos con una alta variabilidad entre las réplicas (probablemente en algunos casos relacionado con el punto anterior).

References

Código disponible en https://github.com/espelm/Lopez-Merino-Esperanza-PEC1

Datos disponibles en https://github.com/nutrimetabolomics/metaboData/tree/main/Datasets/2018-Phosphoproteomics

Cui, Dapeng, Kimberly J Dougherty, David W Machacek, Michael Sawchuk, Shawn Hochman, and Deborah J Baro. 2006. "Divergence Between Motoneurons: Gene Expression Profiling Provides a Molecular Characterization of Functionally Discrete Somatic and Autonomic Motoneurons." *Physiological Genomics* 24 (3): 276–89.

Liu, Yihan, Wantao Wu, Changjing Cai, Hao Zhang, Hong Shen, and Ying Han. 2023. "Patient-Derived Xenograft Models in Cancer Therapy: Technologies and Applications." *Signal Transduction and Targeted Therapy* 8 (1): 160.