PAC 1 Análisis de datos ómicos

Ester Pol Ferrer

2024-11-01

Contents

Abstract

Este informe presenta un análisis de un conjunto de datos de fosfopéptidos obtenidos del repositorio metaboData. Los datos se organizan en un contenedor SummarizedExperiment que incluye tanto las abundancias como los metadatos. Se realiza una exploración exhaustiva mediante visualización, análisis de componentes principales (PCA) y pruebas estadísticas para identificar diferencias significativas entre grupos de muestras tumorales (MSS y PD).

Los resultados sugieren diferencias en los perfiles de fosforilación entre los grupos, lo cual podría indicar distintas características biológicas. Finalmente, se crea un repositorio en GitHub que incluye el código y los datos en formato .Rda para facilitar la reproducibilidad del estudio.

Objetivos del estudio

El objetivo principal es identificar diferencias en la fosforilación de proteínas entre los dos tipos de tumores (MSS y PD). La fosforilación es un proceso mediante el cual se añade un grupo fosfato a una proteína, lo cual puede cambiar su función. Estas modificaciones son clave en muchos procesos celulares, y diferencias en fosforilación pueden estar relacionadas con cómo crecen los tumores, cómo responden a ciertos tratamientos o su agresividad. Específicamente, se busca:

- Buscar patrones de fosforilación específicos que sean distintos entre los grupos MSS y PD. Esto puede proporcionar pistas sobre diferencias en las vías biológicas que están activas en cada tipo de tumor.
- Visualizar estas diferencias usando gráficos como el boxplot y el PCA, para observar si los fosfopéptidos (las proteínas modificadas con fosfato) permiten distinguir entre los tumores MSS y PD.

Si se encuentran fosfopéptidos con niveles de abundancia significativamente distintos entre MSS y PD, esto podría ayudar a identificar potenciales biomarcadores o incluso nuevos objetivos para tratamiento. Los biomarcadores son características que pueden indicar la presencia de una enfermedad o su tipo, y son muy útiles en medicina para hacer diagnósticos o planificar tratamientos.

Materiales y métodos

Origen y tipo de datos

El dataset utilizado en este estudio se obtiene de la carpeta “2018-Phosphoproteomics” en metaboData, y contiene abundancias normalizadas de señales de fosfopéptidos. Este conjunto de datos se utiliza para diferenciar entre subtipos tumorales (MSS y PD) en modelos de PDX humanos.

Herramientas informáticas y bioinformáticas utilizadas:

El análisis se realiza en R, empleando paquetes específicos de Bioconductor como readxl para la carga de datos, SummarizedExperiment para organizar el contenedor de datos, y ggplot2 para la visualización.

Métodos utilizados:

Preparación de los datos

En primer lugar, se cargan las abundancias de fosfopéptidos y se organizan en un objeto SummarizedExperiment que permite integrar datos y metadatos en un único contenedor. Posteriormente, se lleva a cabo un análisis exploratorio utilizando gráficos de caja para observar las distribuciones de abundancias y un análisis de componentes principales (PCA) para evaluar posibles agrupamientos.

Análisis

1. El primer paso de nuestro análisis va a ser cargar el dataset y preprocesarlo:

Instalamos y cargamos el paquete "readxl":

```
#install.packages("readxl")
library(readxl, quietly = TRUE)
```

Cargamos el archivo en R:

```
# Cargamos el dataset
dataset_path <- "/Users/esterpolferrer/Desktop/master/analisi d dades omiques/PAC1/TIO2+PTYR-human-MSS+
data <- read_excel(dataset_path)
```

Examinamos el dataset (filas y columnas):

```
# Examinamos las primeras filas y nombres de columna
head(data)
```

```
## # A tibble: 6 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.
## 6 STGPGASLGTGYDR[12] Pho~ O15551    Claudin-3 ~  63.7 644513.   261938.  187023.
## # 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>
```

```
colnames(data)
```

```
##  [1] "SequenceModifications" "Accession"             "Description"
##  [4] "Score"                 "M1_1_MSS"              "M1_2_MSS"
##  [7] "M5_1_MSS"              "M5_2_MSS"              "T49_1_MSS"
## [10] "T49_2_MSS"             "M42_1_PD"              "M42_2_PD"
## [13] "M43_1_PD"              "M43_2_PD"              "M64_1_PD"
## [16] "M64_2_PD"              "CLASS"                 "PHOSPHO"
```

Estructura del Dataset: Columna SequenceModifications: Contiene los identificadores de los fosfopéptidos, que nos permitirá saber a qué fosfopéptido pertenece cada fila. Columnas M1_1_MSS, M5_2_MSS, etc.: Estas son las columnas de muestras que contienen las abundancias de los fosfopéptidos. Son las que vamos a analizar para ver diferencias entre los dos grupos (MSS y PD). Otras Columnas (Accession, Description, Score, CLASS, PHOSPHO): No son necesarias para el análisis de abundancias, así que las omitiremos en los pasos siguientes.

2. El segundo paso va a ser seleccionar las columnas relevantes y preparar el dataset

De la descripción sabemos que hay dos grupos (MSS y PD) y seis muestras con dos réplicas técnicas cada una. Creamos un dataframe simplificado con las abundancias de fosfopéptidos y las etiquetas de los grupos (nos quedaremos solo con SequenceModifications y las columnas de muestras).

Filtramos la columna con las abundancias y creamos una estructura de datos adecuada:

```
# Seleccionamos la columna relevante y ajusta el dataframe
#install.packages("dplyr")
library(dplyr, quietly = TRUE)
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
data_abundances <- data %>% select(SequenceModifications, starts_with("M"), starts_with("T"))
head(data_abundances)
```

```
## # A tibble: 6 x 13
##   SequenceModifications    M1_1_MSS M1_2_MSS M5_1_MSS M5_2_MSS M42_1_PD M42_2_PD
##   <chr>                       <dbl>    <dbl>    <dbl>    <dbl>    <dbl>    <dbl>
## 1 LYPELSQYMGLSLNEEEIR[2] ~     24.3   44476.       0     6269.       0        0
## 2 VDKVIQAQTAFSANPANPAILSE~      0     43139.    2102.   50355.    1316.       0
## 3 VIQAQTAFSANPANPAILSEASA~   3413.   172143.   77323.  307637.   24851.   16548.
## 4 HADAEMTGYVVTR[6] Oxidat~ 220431.   145657.  104288.   75887. 1027196. 1163747.
## 5 HADAEMTGYVVTR[9] Phospho  18255.     8530.   35956.   44102.   21231.   49500.
## 6 STGPGASLGTGYDR[12] Phos~ 644513.   261938.  187023.  124868.  535809.  434646.
## # i 6 more variables: M43_1_PD <dbl>, M43_2_PD <dbl>, M64_1_PD <dbl>,
## #   M64_2_PD <dbl>, T49_1_MSS <dbl>, T49_2_MSS <dbl>
```

```
colnames(data_abundances)
```

```
##  [1] "SequenceModifications" "M1_1_MSS"              "M1_2_MSS"
##  [4] "M5_1_MSS"              "M5_2_MSS"              "M42_1_PD"
##  [7] "M42_2_PD"              "M43_1_PD"              "M43_2_PD"
## [10] "M64_1_PD"              "M64_2_PD"              "T49_1_MSS"
## [13] "T49_2_MSS"
```

data_abundances tendrá la columna SequenceModifications junto con las columnas de abundancias de cada muestra.

Añadimos metadatos de grupos. Definimos los grupos MSS y PD para cada muestra (3 de cada grupo, con dos réplicas). Creamos un dataframe de metadatos con esta información:

```r
library(S4Vectors, quietly = TRUE)
```

```
##
## Attaching package: 'BiocGenerics'

## The following objects are masked from 'package:dplyr':
##
##     combine, intersect, setdiff, union

## The following objects are masked from 'package:stats':
##
##     IQR, mad, sd, var, xtabs

## The following objects are masked from 'package:base':
##
##     anyDuplicated, aperm, append, as.data.frame, basename, cbind,
##     colnames, dirname, do.call, duplicated, eval, evalq, Filter, Find,
##     get, grep, grepl, intersect, is.unsorted, lapply, Map, mapply,
##     match, mget, order, paste, pmax, pmax.int, pmin, pmin.int,
##     Position, rank, rbind, Reduce, rownames, sapply, setdiff, table,
##     tapply, union, unique, unsplit, which.max, which.min

##
## Attaching package: 'S4Vectors'

## The following objects are masked from 'package:dplyr':
##
##     first, rename

## The following object is masked from 'package:utils':
##
##     findMatches

## The following objects are masked from 'package:base':
##
##     expand.grid, I, unname
```

```r
# Creamos los metadatos de grupo
sample_names <- c("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")
groups <- c("MSS", "MSS", "MSS", "MSS", "MSS", "MSS", "PD", "PD", "PD", "PD", "PD", "PD")
metadata <- DataFrame(Sample = sample_names, Group = groups)
```

3. El tercer paso va a ser crear el objeto SummarizedExperiment

Con los datos de abundancia y los metadatos (metadata) listos, ahora podemos crear el objeto SummarizedExperiment. Este paso combina la información de abundancia y los metadatos en un solo objeto; prepararemos un objeto estructurado que contenga: - Las abundancias de fosfopéptidos en formato de matriz. - Los metadatos de las muestras, especificando a qué grupo pertenece cada muestra.

Convertimos las abundancias en matriz: Extraemos únicamente las columnas de abundancias (omitiendo SequenceModifications) para convertirlas en una matriz, ya que el objeto SummarizedExperiment requiere que las abundancias estén en formato de matriz.

```r
# Creamos una matriz solo con las columnas de abundancias (sin SequenceModifications)
abundances_matrix <- as.matrix(data_abundances %>% select(-SequenceModifications))
```

Creamos el objeto SummarizedExperiment: Usamos la matriz de abundancias y el DataFrame de metadatos (metadata).

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

library(SummarizedExperiment, quietly = TRUE)
```

```
##
## Attaching package: 'matrixStats'

## The following object is masked from 'package:dplyr':
##
##     count

##
## Attaching package: 'MatrixGenerics'

## The following objects are masked from 'package:matrixStats':
##
##     colAlls, colAnyNAs, colAnys, colAvgsPerRowSet, colCollapse,
##     colCounts, colCummaxs, colCummins, colCumprods, colCumsums,
##     colDiffs, colIQRDiffs, colIQRs, colLogSumExps, colMadDiffs,
##     colMads, colMaxs, colMeans2, colMedians, colMins, colOrderStats,
##     colProds, colQuantiles, colRanges, colRanks, colSdDiffs, colSds,
##     colSums2, colTabulates, colVarDiffs, colVars, colWeightedMads,
##     colWeightedMeans, colWeightedMedians, colWeightedSds,
##     colWeightedVars, rowAlls, rowAnyNAs, rowAnys, rowAvgsPerColSet,
##     rowCollapse, rowCounts, rowCummaxs, rowCummins, rowCumprods,
##     rowCumsums, rowDiffs, rowIQRDiffs, rowIQRs, rowLogSumExps,
##     rowMadDiffs, rowMads, rowMaxs, rowMeans2, rowMedians, rowMins,
##     rowOrderStats, rowProds, rowQuantiles, rowRanges, rowRanks,
##     rowSdDiffs, rowSds, rowSums2, rowTabulates, rowVarDiffs, rowVars,
##     rowWeightedMads, rowWeightedMeans, rowWeightedMedians,
##     rowWeightedSds, rowWeightedVars

##
## Attaching package: 'IRanges'

## The following objects are masked from 'package:dplyr':
##
##     collapse, desc, slice

## Welcome to Bioconductor
##
##     Vignettes contain introductory material; view with
##     'browseVignettes()'. To cite Bioconductor, see
##     'citation("Biobase")', and for packages 'citation("pkgname")'.
```

```
##
## Attaching package: 'Biobase'

## The following object is masked from 'package:MatrixGenerics':
##
##     rowMedians

## The following objects are masked from 'package:matrixStats':
##
##     anyMissing, rowMedians
```

```
# Creamos el objeto SummarizedExperiment
SE <- SummarizedExperiment(assays = list(counts = abundances_matrix), colData = metadata)

# Guardamos el objeto para análisis posterior
save(SE, file = "dataset_metabolomics.Rda")
```

Resultados

En este paso vamos a explorar si existen diferencias en la abundancia de fosfopéptidos (proteínas fosforiladas) entre los grupos MSS y PD.

Esto se hace mediante: - Boxplot: Nos permitirá ver si las distribuciones de abundancia son diferentes entre MSS y PD. - PCA: Nos ayudará a ver si las muestras de MSS y PD se agrupan de forma diferente, lo que podría indicar que tienen perfiles de fosforilación distintos. - Prueba t para comparar grupos MSS y PD: Para cada fosfopéptido, podemos realizar una prueba t para comparar las abundancias entre los grupos MSS y PD, que nos ayudará a identificar si existen fosfopéptidos con diferencias significativas entre los grupos.

1. Boxplot de Abundancias de Fosfopéptidos

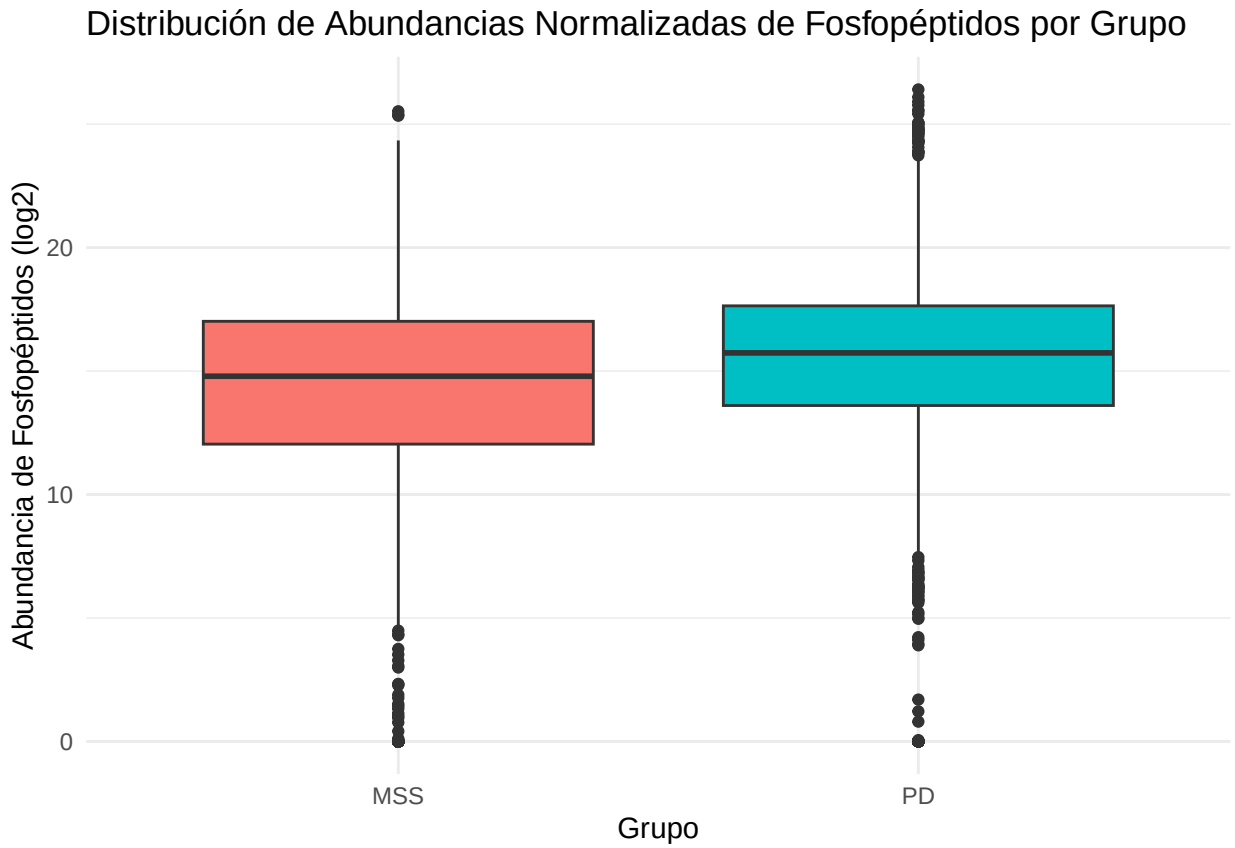
Este gráfico muestra la distribución de abundancias de fosfopéptidos en cada grupo (MSS y PD). Nos permitirá ver si hay diferencias en la cantidad de fosforilación entre los grupos.

```
library(ggplot2, quietly = TRUE)

# Extraemos la matriz de abundancias del objeto SummarizedExperiment
abundances_matrix_boxplot <- assay(SE, "counts")

# Normalizamos logarítmicamente para reducir el rango de valores
abundances_matrix_boxplot <- log2(abundances_matrix_boxplot + 1)

# Convertimos la matriz en un dataframe y añadir los grupos
plot_data <- as.data.frame(t(abundances_matrix_boxplot))  # Transponemos para tener una columna por fos
plot_data$Group <- colData(SE)$Group  # Añadir los grupos como columna

library(tidyr, quietly = TRUE)
```

```
##
## Attaching package: 'tidyr'
```

```
## The following object is masked from 'package:S4Vectors':
##
##     expand
```

```r
# Convertimos abundances_matrix en formato largo
plot_data_long <- plot_data %>%
    pivot_longer(cols = starts_with("V"), names_to = "Fosfopeptido", values_to = "Abundancia")

# Creamos el boxplot con múltiples fosfopéptidos
ggplot(plot_data_long, aes(x = Group, y = Abundancia, fill = Group)) +
    geom_boxplot() +
    theme_minimal() +
    labs(title = "Distribución de Abundancias Normalizadas de Fosfopéptidos por Grupo",
         x = "Grupo",
         y = "Abundancia de Fosfopéptidos (log2)") +
    theme(legend.position = "none")
```



- Rango y Mediana: Podemos ver que ambos grupos tienen medianas similares en términos de abundancia. La mediana es la línea dentro de cada caja, que representa el valor medio de abundancia en cada grupo. - Variabilidad: Aunque la mediana es similar, el grupo "PD" parece tener una mayor variabilidad (o dispersión) en las abundancias de fosfopéptidos, como se puede ver en el rango intercuartílico (el tamaño de la caja) y en las líneas que se extienden hacia los valores atípicos (outliers). - Outliers: Ambos grupos presentan valores atípicos, pero el grupo "PD" parece tener una mayor cantidad de estos valores extremos, lo que indica que algunos fosfopéptidos en este grupo tienen abundancias significativamente diferentes de la mediana.

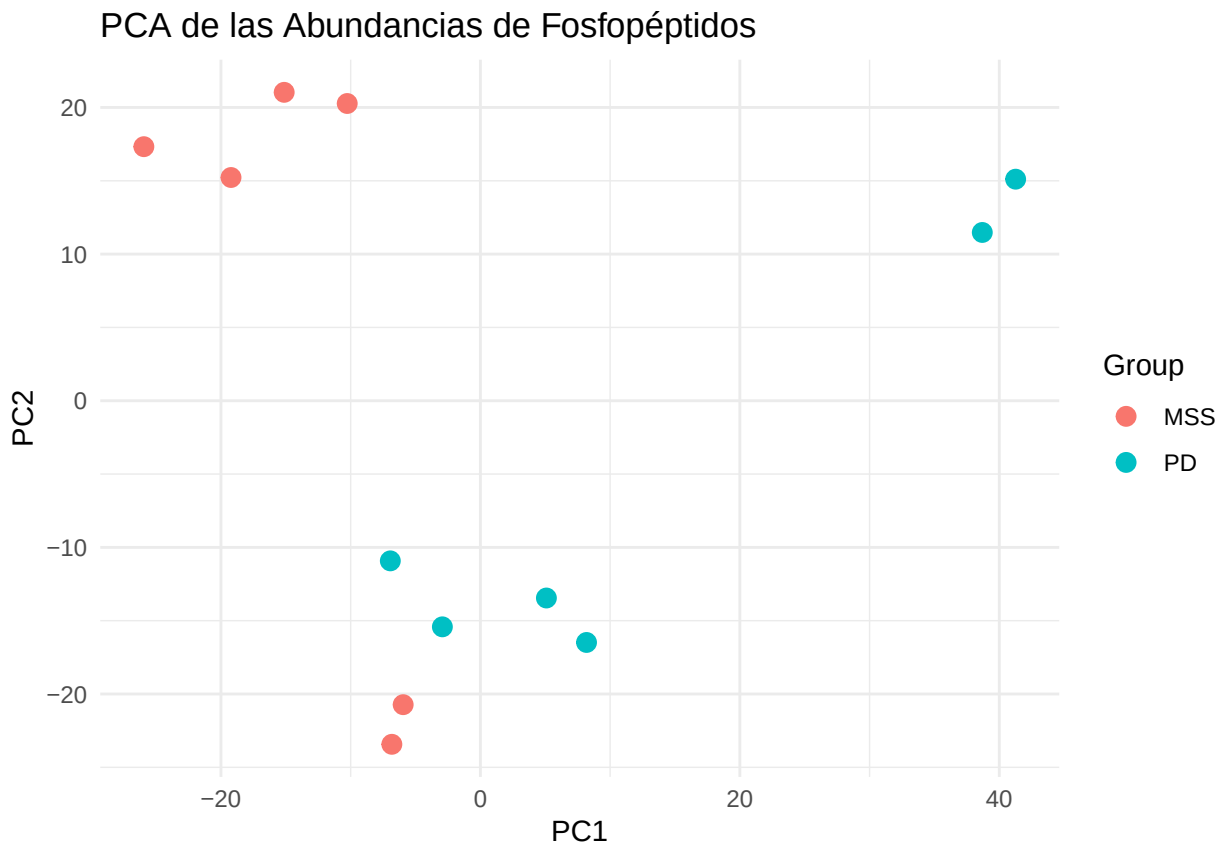
Aunque la mediana de abundancia es similar entre los dos grupos, la mayor dispersión y la presencia de más

valores atípicos en el grupo “PD” sugieren que hay variaciones en la abundancia de ciertos fosfopéptidos que podrían ser relevantes para la comparación entre ambos grupos.

2. Análisis de Componentes Principales (PCA)

El PCA nos ayuda a reducir la dimensionalidad de los datos, lo que facilita observar agrupamientos entre las muestras. Cada punto representa una muestra, y la separación entre MSS y PD sugiere diferencias en los perfiles de fosfopéptidos.

```
# Filtramos filas con baja variación o constantes
abundances_matrix_pca <- abundances_matrix[apply(abundances_matrix, 1, function(x) {
    sd(x) > 1e-5  # Filtra filas con desviación estándar significativa (ajusta si es necesario)
}), ]

# Normalizamos logarítmicamente para reducir el rango de valores
abundances_matrix_pca <- log2(abundances_matrix_pca + 1)

# Realizamos el PCA en la matriz filtrada
pca <- prcomp(t(abundances_matrix_pca), scale = TRUE)
pca_data <- data.frame(pca$x, Group = colData(SE)$Group)

# Graficamos el PCA
library(ggplot2)
ggplot(pca_data, aes(x = PC1, y = PC2, color = Group)) +
    geom_point(size = 3) +
    theme_minimal() +
    labs(title = "PCA de las Abundancias de Fosfopéptidos",
         x = "PC1",
         y = "PC2")
```



- Agrupación: Las muestras de cada grupo (MSS y PD) parecen estar parcialmente separadas en el espacio de los componentes principales, con algunas muestras de "MSS" agrupadas hacia un extremo y algunas de "PD" en el otro. Sin embargo, hay solapamientos entre los dos grupos. - PC1 y PC2: PC1 es la componente principal que captura la mayor variabilidad en los datos, seguida de PC2. La separación a lo largo de PC1 o PC2 indica que hay diferencias en las abundancias de fosfopéptidos que varían entre ambos grupos, pero no es una separación completa, por lo que es posible que otros factores o fosfopéptidos específicos influyan en la variabilidad observada.

Aunque los grupos no se separan completamente, la distribución sugiere que existen ciertas diferencias en la abundancia de fosfopéptidos que están asociadas con cada grupo. Esto apoya la idea de que las abundancias de algunos fosfopéptidos son diferenciales entre "MSS" y "PD".

3. Test t para comparar grupos MSS y PD

Para cada fosfopéptido, podemos realizar una prueba t para comparar las abundancias entre los grupos MSS y PD; esto nos ayudará a identificar si existen fosfopéptidos con diferencias significativas entre los grupos.

```
# Extraemos y normalizamos la matriz de abundancias (si no está ya normalizada)
abundances_matrix <- assay(SE, "counts")
abundances_matrix <- log2(abundances_matrix + 1)

library(tidyr, quietly = TRUE)
library(dplyr, quietly = TRUE)

# Creamos dataframe para realizar comparaciones
long_data <- as.data.frame(t(abundances_matrix))  # Transponer para tener una columna por fosfopéptido
long_data$Group <- colData(SE)$Group  # Añadir la información de grupo
```

```r
long_data <- pivot_longer(long_data, cols = starts_with("V"), names_to = "Fosfopeptido", values_to = "Ab

# Realizamos prueba t para cada fosfopéptido y obtener p-values
results <- long_data %>%
  group_by(Fosfopeptido) %>%
  summarise(p_value = t.test(Abundancia ~ Group)$p.value)

# Aplicamos corrección de p-values para múltiples pruebas (FDR)
results <- results %>%
  mutate(adj_p_value = p.adjust(p_value, method = "fdr"))

# Filtramos fosfopéptidos significativos
significant_results <- results %>% filter(adj_p_value < 0.05)

# Ordenamos los resultados por adj_p_value de menor a mayor
ordered_results <- significant_results %>% arrange(adj_p_value)

# Mostramos fosfopéptidos con diferencias significativas entre MSS y PD ordenados
print(ordered_results)
```

```
## # A tibble: 64 x 3
##    Fosfopeptido    p_value adj_p_value
##    <chr>             <dbl>       <dbl>
##  1 V811         0.00000298     0.00429
##  2 V1247        0.0000346      0.0124
##  3 V573         0.0000291      0.0124
##  4 V655         0.0000261      0.0124
##  5 V1076        0.0000893      0.0128
##  6 V1118        0.0000785      0.0128
##  7 V1134        0.0000645      0.0128
##  8 V1209        0.0000551      0.0128
##  9 V1394        0.0000977      0.0128
## 10 V153         0.0000677      0.0128
## # i 54 more rows
```

- P-Values: Cada p-value muestra la probabilidad de que las diferencias observadas en las abundancias entre MSS y PD sean debidas al azar. Los fosfopéptidos con valores p más bajos indican una mayor evidencia de diferencia significativa entre los grupos.
- Adjusted P-Values (adj_p_value): Estos valores ajustados controlan el error de tipo I en pruebas múltiples, de modo que fosfopéptidos con valores ajustados más bajos son los que muestran diferencias estadísticamente significativas entre los grupos. En tu tabla, vemos fosfopéptidos con un adj_p_value muy bajo, lo cual indica que la abundancia de estos fosfopéptidos difiere significativamente entre los grupos.

La lista de fosfopéptidos con valores p ajustados más bajos (como V811 o V1247) es especialmente importante porque estos son los fosfopéptidos cuya abundancia está más estrechamente asociada con la diferencia entre MSS y PD. Esto podría señalar que estos fosfopéptidos en particular tienen un papel relevante en las diferencias biológicas o de estado entre los dos grupos.

Discusión y limitaciones y conclusiones del estudio

Este estudio revela diferencias importantes en los perfiles de fosforilación entre los grupos MSS y PD. Aunque ambos grupos presentan medianas de abundancia similares, el grupo PD muestra una mayor variabilidad en las abundancias de fosfopéptidos, lo que indica una posible heterogeneidad en sus vías de señalización. El análisis de componentes principales sugiere una tendencia de agrupación parcial entre las muestras de MSS y PD, reflejando diferencias en los perfiles de fosforilación que, aunque no suficientemente pronunciadas para separar completamente los grupos, sí indican variaciones relevantes en su regulación molecular. Además, la prueba t identificó un conjunto de fosfopéptidos con diferencias estadísticamente significativas entre los grupos, destacando algunos fosfopéptidos específicos cuya abundancia podría estar particularmente asociada con las características biológicas distintivas de cada grupo. En conjunto, estos hallazgos sugieren que los perfiles de fosforilación diferenciales entre MSS y PD podrían tener implicaciones para la comprensión de los mecanismos moleculares subyacentes y podrían servir de base para la identificación de posibles biomarcadores o dianas terapéuticas.

Estos hallazgos abren la puerta a futuras investigaciones para:

- Explorar en profundidad los mecanismos reguladores que podrían explicar la mayor variabilidad en PD.
- Validar los fosfopéptidos diferenciales como posibles biomarcadores o dianas terapéuticas específicas para cada grupo.
- Ampliar el análisis incluyendo otros enfoques ómicos o muestras adicionales para confirmar la robustez de estos resultados.
- Este análisis contribuye al conocimiento sobre las diferencias moleculares entre MSS y PD, proporcionando una base sólida para futuras investigaciones en la caracterización de estos subtipos y su potencial relevancia clínica.

Repositorio

El código y los datos utilizados en este análisis están disponibles en el siguiente repositorio de GitHub: https://github.com/esterpol/Pol-Ferrer-Ester-PEC1.git