Análisis de Datos Ómicos - PEC1

Vanesa Paramá Pérez

2024-10-28

Introducción

La interrogante clave que motivó este estudio fue identificar fosfopéptidos que permitan diferenciar entre los dos grupos de tumores, MSS y PD. Explorar estas diferencias moleculares es crucial para entender los mecanismos subyacentes que distinguen a estos subtipos de cáncer y puede proporcionar información valiosa para el desarrollo de tratamientos específicos. Los datos se obtuvieron de un archivo Excel descargado de Github, que contiene la abundancia normalizada de señales MS para aproximadamente 1400 fosfopéptidos.

Las muestras fueron seleccionadas de modelos PDX de dos subtipos diferentes, con dos réplicas técnicas para cada muestra, usando técnicas de enriquecimiento de fosfopéptidos. Los investigadores utilizaron un diseño experimental que incluyó muestras de diferentes individuos dentro de cada grupo (MSS y PD) para asegurar una representatividad adecuada y reducir el sesgo experimental.

Posibles influencias del diseño: Tanto el diseño experimental como la selección de individuos y la técnica de enriquecimiento de fosfopéptidos, podrían haber influido en la variabilidad y la reproducibilidad de los resultados. La inclusión de réplicas técnicas es una forma de minimizar el impacto de la variabilidad técnica.

El uso de la técnica de enriquecimiento de fosfopéptidos seguido del análisis LC-MS ha sido fundamental para obtener datos específicos y relevantes de los fosfopéptidos. Este enfoque ha permitido una exploración detallada de las modificaciones posraduccionales que diferencian a los subtipos de tumores.

Los posibles desafíos relacionados con la calidad de los datos crudos incluyen el manejo de valores NA y la variabilidad técnica entre las réplicas.

Descarga de los Datos

El dataset utilizado en este análisis se descargó del archivo TIO2+PTYR-human-MSS+MSIvsPD.XLSX, que contiene dos hojas: originalData y targets. La hoja originalData contiene los datos cuantitativos, mientras que la hoja targets contiene los metadatos de las muestras.

Descripción de los Datos

Los datos se dividen en dos partes principales: los datos cuantitativos y los metadatos.

originalData

- **SequenceModifications**: Modificaciones de las secuencias de los fosfopéptidos.
- **Accession**: Identificación de acceso del fosfopéptido.
- **Description**: Descripción del fosfopéptido.
- **Score**: Puntuación de calidad del fosfopéptido.
- **Muestras (M1_1_MSS, M1_2_MSS, etc.)**: Abundancia normalizada de señales MS para cada fosfopéptido en cada muestra.
- **CLASS**: Clase del fosfopéptido.
- **PHOSPHO**: Estado de fosforilación del fosfopéptido.

targets

- **Sample**: Identificador de muestra.
- **Individual**: Identificador del individuo al que pertenece la muestra.
- **Phenotype**: Fenotipo del individuo (MSS o PD).

Carga y Preparación de los Datos

Para cargar y preparar los datos, primero se instalaron y cargaron las bibliotecas necesarias:

```
library(readxl)
library(dplyr)

##
## Adjuntando el paquete: 'dplyr'

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

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

library(SummarizedExperiment)

## Cargando paquete requerido: MatrixGenerics

## Cargando paquete requerido: matrixStats

##
## Adjuntando el paquete: 'matrixStats'

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

##
## Adjuntando el paquete: '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

## Cargando paquete requerido: GenomicRanges

## Cargando paquete requerido: stats4

## Cargando paquete requerido: BiocGenerics

##
## Adjuntando el paquete: '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

## Cargando paquete requerido: S4Vectors

##
## Adjuntando el paquete: '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

## Cargando paquete requerido: IRanges

##
## Adjuntando el paquete: 'IRanges'

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

## The following object is masked from 'package:grDevices':
##
##     windows

## Cargando paquete requerido: GenomeInfoDb

## Cargando paquete requerido: Biobase

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

##
## Adjuntando el paquete: 'Biobase'

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

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

Luego, se especificó la ruta al archivo y se cargaron los datos de las dos hojas:

```
file_path <- "C:/Users/Vane/Desktop/Archivos R/Datos Ómicos/TIO2+PTYR-human-MSS+MSIvsPD.XLSX"
data_matrix <- read_excel(file_path, sheet = "originalData")
metadata <- read_excel(file_path, sheet = "targets")

## New names:
## • `Sample` -> `Sample...1`
## • `Sample` -> `Sample...2`
```

```r
metadata <- as.data.frame(metadata)
```

Procesamiento y Limpieza de los Datos

Para procesar y limpiar los datos, se extrajeron los datos cuantitativos y se gestionaron los valores NA:

```r
quantitative_data <- as.matrix(data_matrix[, 7:ncol(data_matrix)])
quantitative_data <- apply(quantitative_data, 2, function(x)
as.numeric(as.character(x)))

## Warning in FUN(newX[, i], ...): NAs introducidos por coerción
## Warning in FUN(newX[, i], ...): NAs introducidos por coerción

if (any(is.na(quantitative_data))) {
  print("Existen valores NA en los datos. Se reemplazarán con 0.")
  quantitative_data[is.na(quantitative_data)] <- 0
}

## [1] "Existen valores NA en los datos. Se reemplazarán con 0."
```

Se asignaron nombres de filas y columnas:

```r
rownames(quantitative_data) <- data_matrix$Accession
colnames(quantitative_data) <- metadata$Sample
```

Creación del Contenedor SummarizedExperiment

```r
se <- SummarizedExperiment(
  assays = list(counts = quantitative_data),
  colData = metadata,
  rowData = data_matrix[, 1:6]  # Información de las filas
)
```

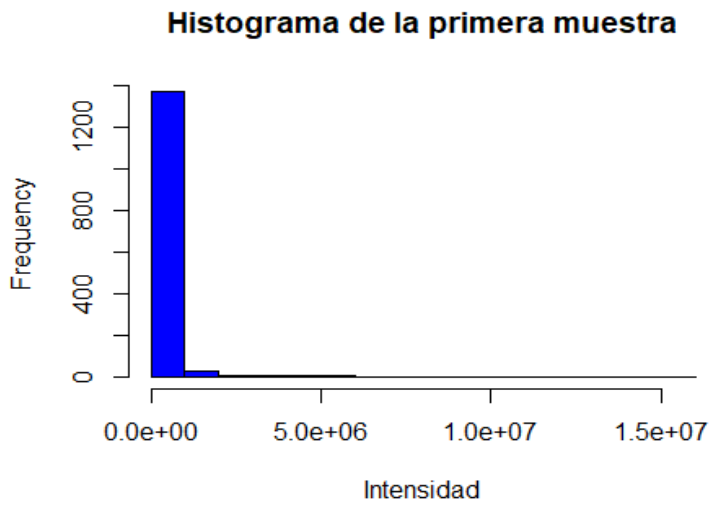
Exploración de los Datos

Se realizaron varias visualizaciones para explorar los datos.

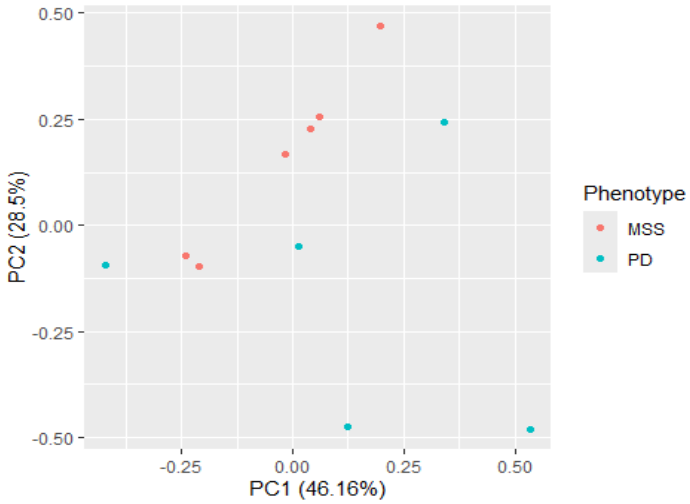
```r
print(summary(se))

## [1] "SummarizedExperiment object of length 1438 with 6 metadata
columns"

summary(assay(se))

##        V1                  V2                  V3                  V4
##  Min.   :       0   Min.   :       0   Min.   :       0   Min.   :
0
##  1st Qu.:    2573   1st Qu.:    3273   1st Qu.:    9306   1st Qu.:
8611
```

```
##  Median :   20801   Median :   26241   Median :   55641   Median :
46110
##  Mean   :  232967   Mean   :  261067   Mean   :  542449   Mean   :
462616
##  3rd Qu.:  113958   3rd Qu.:  130132   3rd Qu.:  223103   3rd Qu.:
189141
##  Max.   :15135170   Max.   :19631820   Max.   :49218870   Max.
:29240210
##        V5                  V6                  V7                  V8
##  Min.   :       0   Min.   :       0   Min.   :       0   Min.   :
0
##  1st Qu.:    5341   1st Qu.:    4216   1st Qu.:   19641   1st Qu.:
17299
##  Median :   36854   Median :   30533   Median :   67945   Median :
59607
##  Mean   :  388424   Mean   :  333587   Mean   :  349020   Mean   :
358822
##  3rd Qu.:  180252   3rd Qu.:  152088   3rd Qu.:  205471   3rd Qu.:
201924
##  Max.   :48177680   Max.   :42558110   Max.   :35049400   Max.
:63082980
##        V9                 V10                 V11         V12
##  Min.   :       0   Min.   :       0   Min.   :0   Min.   :0
##  1st Qu.:   11038   1st Qu.:    8660   1st Qu.:0   1st Qu.:0
##  Median :   52249   Median :   47330   Median :0   Median :0
##  Mean   :  470655   Mean   :  484712   Mean   :0   Mean   :0
##  3rd Qu.:  209896   3rd Qu.:  206036   3rd Qu.:0   3rd Qu.:0
##  Max.   :71750330   Max.   :88912730   Max.   :0   Max.   :0
```

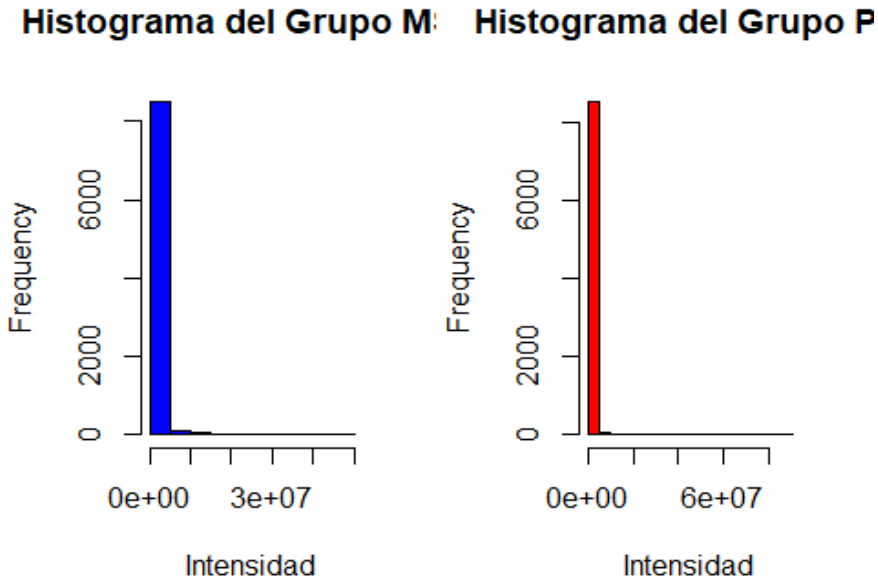
```r
hist(assay(se)[, 1], main = "Histograma de la primera muestra", xlab =
"Intensidad", col = "blue")
```



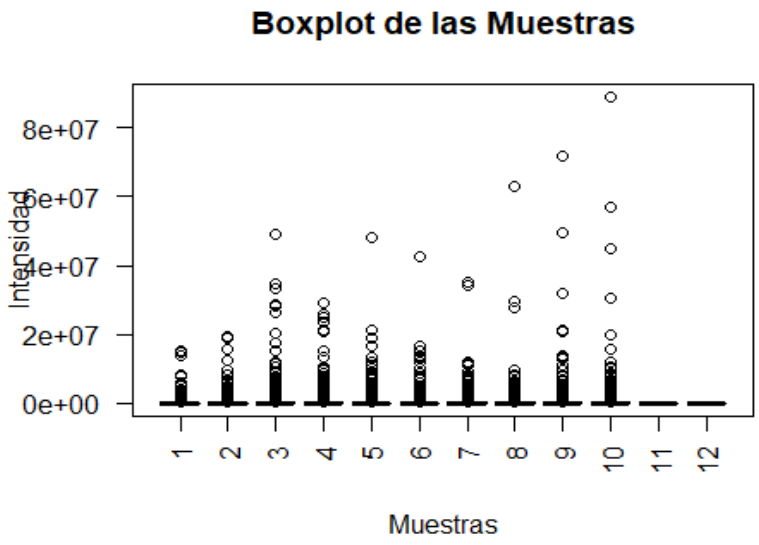
```r
library(ggplot2)
library(ggfortify)
autoplot(prcomp(t(assay(se))), data = as.data.frame(colData(se)), colour
= 'Phenotype')
```




```r
# Datos del grupo MSS
mss_data <- assay(se)[, colData(se)$Phenotype == "MSS"]
pd_data <- assay(se)[, colData(se)$Phenotype == "PD"]

par(mfrow = c(1, 2))  # Dividir la ventana gráfica en 2 columnas
hist(mss_data, main = "Histograma del Grupo MSS", xlab = "Intensidad",
col = "blue")
hist(pd_data, main = "Histograma del Grupo PD", xlab = "Intensidad", col
= "red")
```



```r
boxplot(assay(se), main = "Boxplot de las Muestras", xlab = "Muestras",
ylab = "Intensidad", col = "lightblue", las = 2)
```



Análisis de Varianza (ANOVA) Identificación de fosfopéptidos con diferencias significativas entre grupos.

```r
# Realizamos un ANOVA para cada fosfopéptido
anova_results <- apply(assay(se), 1, function(x) {
  fit <- aov(x ~ colData(se)$Phenotype)
  summary(fit)[[1]][["Pr(>F)"]][1]
})

head(anova_results)
```

```
##     O00560     O00560     O00560     O15264     O15264     O15551
## 0.31077853 0.27739478 0.02813961 0.05515099 0.03967544 0.10541859
```

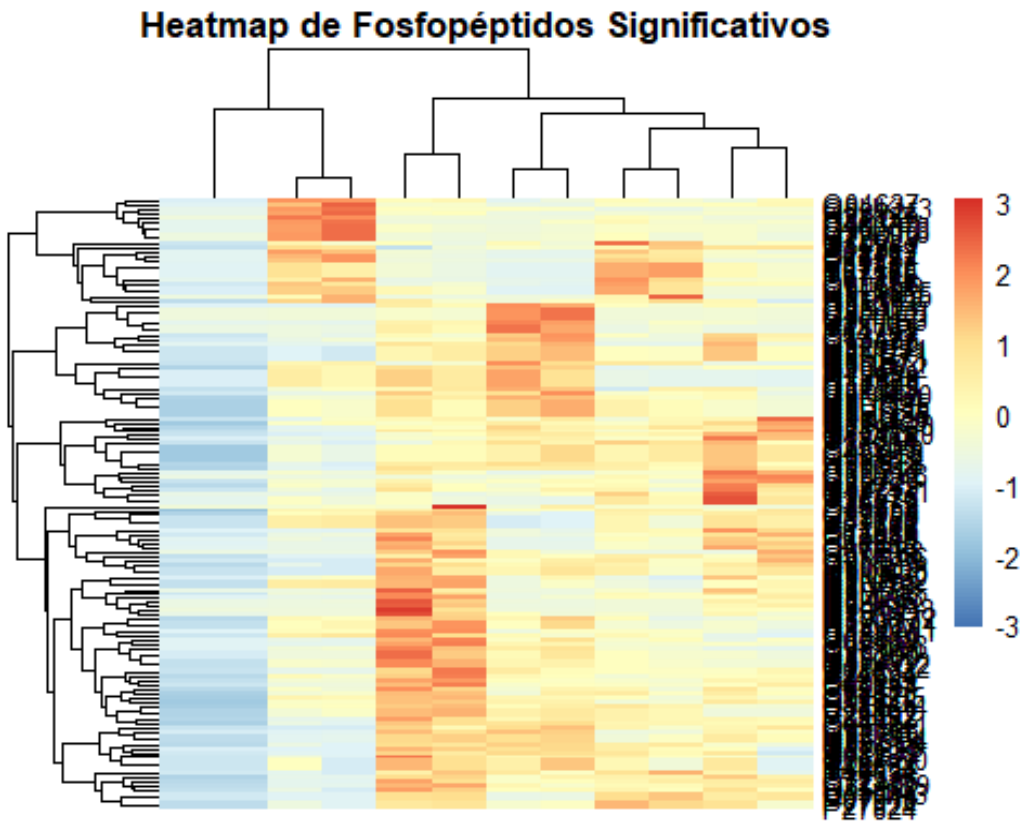
```r
# Seleccionamos los fosfopéptidos con p-valor significativo y válidos
signif_peptides <- rownames(assay(se))[!is.na(anova_results) &
anova_results < 0.05]

signif_peptides
```

```
##   [1] "O00560" "O15264" "P07355" "P07355" "P48960" "P57739" "Q09666"
"Q12929"
##   [9] "Q13113" "Q13443" "Q15758" "Q5VW32" "Q6UXY8" "Q8IVI9" "Q12929"
"P29323"
##  [17] "P54753" "P54753" "Q9UQB8" "P07948" "Q5T5P2" "P07355" "P07355"
"Q8TE68"
##  [25] "O00264" "O14976" "O15231" "O43491" "O60832" "O60841" "O75152"
"O75976"
##  [33] "P05388" "P07910" "P07910" "P08727" "P11388" "P13861" "P18615"
```

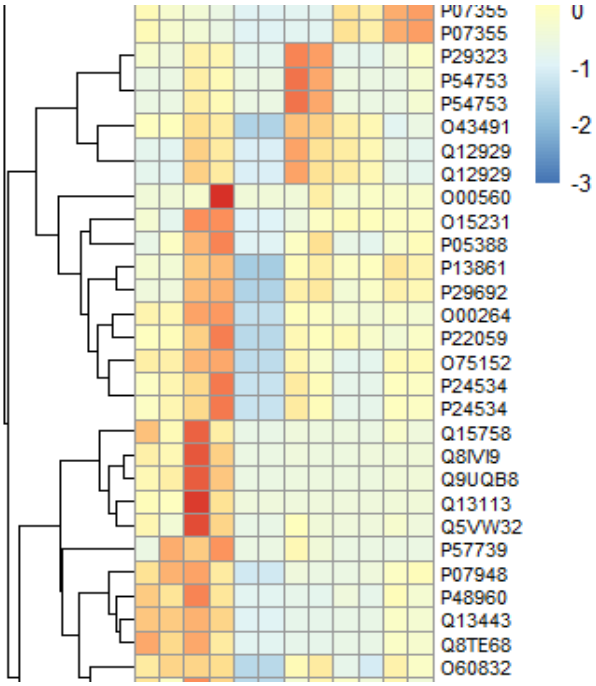
```
"P19338"
##  [41] "P22059" "P24534" "P24534" "P25205" "P27824" "P27824" "P29692"
"P35269"
##  [49] "P35579" "P35579" "P48634" "P48634" "P51858" "P51858" "P51858"
"P52948"
##  [57] "P54727" "Q04637" "Q08945" "Q12906" "Q12929" "Q13200" "Q13200"
"Q13427"
##  [65] "Q13428" "Q13610" "Q14676" "Q15149" "Q15424" "Q5T5U3" "Q5TAQ9"
"Q6PKG0"
##  [73] "Q6QNY0" "Q6ZRV2" "Q7Z406" "Q8IU81" "Q8IVT2" "Q8N8A6" "Q8NE71"
"Q8NE71"
##  [81] "Q8NE71" "Q8NE71" "Q8WWI1" "Q8WWI1" "Q92733" "Q92766" "Q92922"
"Q96RS0"
##  [89] "Q96SB4" "Q96SB4" "Q96ST2" "Q96ST2" "Q96ST2" "Q96ST2" "Q99733"
"Q9BRD0"
##  [97] "Q9BY44" "Q9C0C2" "Q9C0C9" "Q9H1E3" "Q9H1E3" "Q9H3N1" "Q9H4A3"
"Q9H6F5"
## [105] "Q9NR30" "Q9NR30" "Q9NRL2" "Q9NX94" "Q9NXG2" "Q9NXG2" "Q9NZT2"
"Q9P0P8"
## [113] "Q9P2G1" "Q9UKV3" "Q9ULT8" "Q9UQ35" "Q9UQ35" "Q9UQ35" "Q9UQ35"
"Q9Y3T9"
## [121] "P29692" "Q9H1E3" "Q8TCJ2" "P08238" "P48634" "Q7Z6Z7" "Q14103"
"Q9NYF8"
## [129] "P05387" "Q12929" "P17096" "Q7Z5L9" "Q15773" "O76080" "P55327"
"Q9NYF8"
## [137] "P46937" "Q9UKE5" "P21127" "Q7Z5L9" "Q9C040" "Q92625" "Q8IVT2"
"O95394"
## [145] "Q9UKE5" "Q9NZ63"
```

Los fosfopéptidos con un p-valor menor a 0.05 se consideraron significativamente diferentes entre los grupos.

```
if (length(signif_peptides) > 0) {
  # Subset de fosfopéptidos significativos
  signif_data <- assay(se)[signif_peptides, ]

  # heatmap
  library(pheatmap)
  pheatmap(signif_data, cluster_rows = TRUE, cluster_cols = TRUE, scale =
"row",
           main = "Heatmap de Fosfopéptidos Significativos")
} else {
  print("No se encontraron fosfopéptidos significativos con un p-valor
menor a 0.05.")
}
```




```r
# Seleccionamos los 50 fosfopéptidos más significativos para que podamos visualizar mejor el heatmap
top_signif_peptides <- head(signif_peptides, 50)

top_signif_data <- assay(se)[top_signif_peptides, ]

pheatmap(top_signif_data, cluster_rows = TRUE, cluster_cols = TRUE, scale = "row",
         main = "Heatmap de los 50 Fosfopéptidos Más Significativos",
         fontsize_row = 8, fontsize_col = 8, cellwidth = 10, cellheight = 10)
```



El heatmap utiliza una escala de colores para representar la expresión de los fosfopéptidos, donde los colores más oscuros representan una mayor fosforilación y los colores más claros representan una menor fosforilación.Las filas y columnas del heatmap están agrupadas para mostrar patrones de similitud. Las muestras que están cercanas entre sí en el dendrograma tienen perfiles de fosforilación más similares, lo que puede sugerir una relación biológica entre ellas.

Conclusiones

El estudio ha permitido identificar fosfopéptidos que son diferencialmente fosforilados entre los grupos de tumores MSS y PD. Estos fosfopéptidos son prometedores como biomarcadores, facilitando la distinción entre los subtipos de tumores y contribuyendo al desarrollo de terapias dirigidas.

Los fosfopéptidos identificados pueden ser utilizados como biomarcadores en la clínica, mejorando el diagnóstico y la clasificación de los subtipos de tumores. Esto tendrá un impacto positivo en las decisiones clínicas y el manejo de los pacientes.

La creación del contenedor SummarizedExperiment ha facilitado significativamente el manejo y la visualización de los datos, demostrando ser una herramienta útil para futuros análisis de datos ómicos.

```
save(se, file = "PEC1_SummarizedExperiment.Rda")
write.csv(as.data.frame(assay(se)), file = "datos_cuantitativos.csv",
row.names = TRUE)
write.csv(as.data.frame(colData(se)), file = "metadatos.csv", row.names =
TRUE)
```

Repositorio en GitHub

El repositorio con todos los archivos está disponible en: https://github.com/vpp-uoc/DATOSOMICOS-PEC1