Informe

Primera prueba de evaluación continua. Análisis de datos ómicos

Paula María Ruiz Real

Índice

Resumen

El objeto SummarizedExperiment en R nos permite manejar datos junto a información relevante sobre estos (cómo se organizan, anotaciones de interés) de manera eficiente. En este trabajo se crea un objeto de estas características a partir del conjunto de datos 2018-Phosphoproteomics. Este contiene información sobre abundancia medida mediante análisis LC-MS de fosfoproteínas en 12 muestras diferentes. Pertenecen a dos grupos tumorales: MSS y PD, existiendo dos réplicas por muestra. Tras la exploración y preprocesamiento de los datos se aplica sobre él un análisis de abundancia diferencial de fosfopéptidos limma que permite identificar 96 fosfolípidos de interés para la diferenciación de los dos grupos tumorales.

Objetivos

1. Crear un objeto de clase SummarizedExperiment que contenga los datos y los metadatos del conjunto de datos seleccionado.

(a) Explorar los datos para comprender su estructura.

(b) Definir los datos de abundancia, la información sobre las columnas y la información sobre las filas.

(c) Crear el objeto SummarizedExperiment con los datos anteriores.

2. Llevar a cabo el objetivo indicado en los metadatos del conjunto de datos seleccionado: detectar fosfopéptidos que permitan la diferenciación de los dos grupos tumorales.

(a) Realizar un análisis exploratorio de los datos.

(b) Buscar los fosfopéptidos de interés mediante análisis estadístico.

(c) Mostrar visualmente el resultado del análisis.

Métodos

Para la resolución de actividad se ha utilizado el programa RStudio en el cual se ha utilizado el lenguaje R para llevar a cabo todas las tareas. Se ha utilizado Git para el control de versiones y los archivos generados han sido publicados en el repositorio de GitHub "Ruiz-Real-PaulaMaria-PEC1".

En primer lugar se cargaron los datos desde un archivo Excel. Se realizó una exploración básica de este así como la lectura del archivo adjunto "description" para poder comprender el estudio y la estructura de los datos y así poder crear el objeto SummarizedExperiment.

Se definieron las diferentes partes que lo componen:

- Matriz de abundancia: los valores de abundancia de fosfopéptidos por muestra.
- Metadatos de las muestras (colData): se creó un data frame con el nombre de la muestra, el grupo al que pertenece (MSS o PD) y la información sobre réplicas.
- Metadatos de los fosfopéptidos (rowData): se incluyeron columnas con información relevante sobre cada fosfopéptido como la secuencia, el código de acceso, la descripción, el score, la clase y la fosforilación. El archivo "description" indicaba que podían omitirse pero considero interesante contar con la mayor cantidad de información por si en un momento determinado pudiera resultar útil.

El objeto SummarizedExperiment se creó a partir de la matriz y los data frame definidos. También se añadieron metadatos a este con el nombre de los datos, año y el área al que pertenece.

Las principales diferencias entre SummarizedExperiment (SE) y ExpressionSet (ES) es que SE es más flexible con los tipos de datos que maneja mientras que ES es sobre todo para expresión génica (principalmente para microarrays). SE presenta tres niveles de metadatos y ES dos **(respuesta a la pregunta propuesta en la tarea 2)**.

Se exploraron los datos para verificar su estructura y detectar la presencia de valores faltantes. Se realizaron análisis descriptivos básicos, como el resumen estadístico de las abundancias o la comprobación de su dimensión. Para evaluar la distribución de los datos se generaron gráficos de tipo boxplot y QQ-plot. El boxplot se utilizó para observar la distribución de las abundancias de fosfopéptidos entre las diferentes muestras y el QQ-plot para evaluar la normalidad de los datos.

Se realizó un análisis de abundancia diferencial utilizando el paquete limma y el enfoque voom para normalizar las abundancias de los fosfopéptidos. Los pasos que se siguieron fueron:

- Construcción de un objeto DGEList y normalización de los datos.
- Definición de un modelo que incluyó los factores de grupo (MSS vs. PD) y las réplicas.
- Utilización de la función voom para ajustar las varianzas.
- Ajuste del modelo utilizando lmFit, realización del contraste entre los grupos y ajuste bayesiano.

Por último, se identificaron los fosfopéptidos con un p-valor ajustado menor a 0.05.

Para visualizar los resultados, se generaron dos tipos de gráficos:

Volcano plot: se creó para mostrar los fosfopéptidos considerados como significativos. Muestra a cada uno en base al cambio en su abundancia y su p-valor ajustado.

Heatmap: se generó un heatmap para visualizar la abundancia de los fosfopéptidos significativos entre las muestras, etiquetándolas con el nombre del grupo al que pertenecen.

Finalmente, el objeto SummarizedExperiment se guardó en un archivo .rda y los datos de abundancia fueron exportados a un archivo CSV.

Todo el código que realiza los pasos descritos se encuentra en el archivo "Código.R".

Se creó un archivo markdown "Metadatos" que incluye una descripción redactada sobre el archivo "Objeto_SummarizedExperiment" explicando sus datos y metadatos.

Resultados

Incluyo el objeto SummarizedExperiment creado.

```
## class: SummarizedExperiment
## dim: 1438 12
## metadata(3): dataset_name year research_field
## assays(1): abundance
## rownames(1438): LYPELSQYMGLSLNEEEIR[2] Phospho|[9] Oxidation
##   VDKVIQAQTAFSANPANPAILSEASAPIPHDGNLYPR[35] Phospho ...
##   YQDEVFGGFVTEPQEESEEEVEEPEER[17] Phospho YSPSQNSPIHHIPSRR[1]
##   Phospho|[7] Phospho
## rowData names(6): SequenceModifications Accession ... Class
##   Phosphorilation
## colnames(12): M1_1_MSS M1_2_MSS ... M64_1_PD M64_2_PD
## colData names(3): Sample Group Replicate
```

Gráfico de la abundancia de los fosfopéptidos por muestra para visualizar de forma general la distribución.

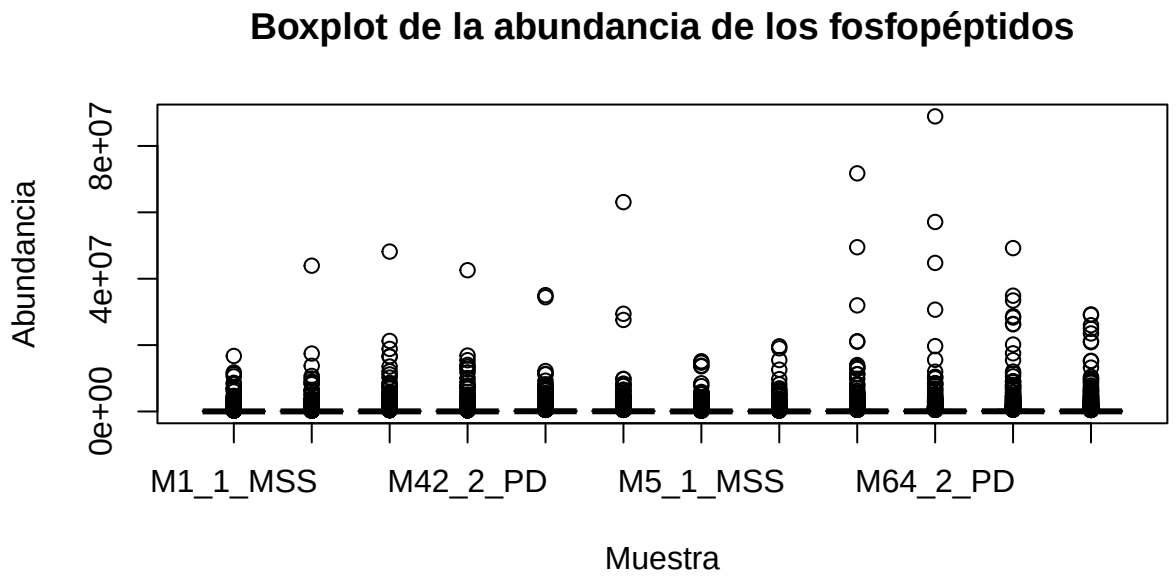

Figura 1: Distribución de los valores de abundancia de fosfopéptidos por muestra.

Q-Q Plot para la comprobación de la normalidad de los datos. La línea naranja representa la referencia para la normalidad. No muestran una distribución normal.

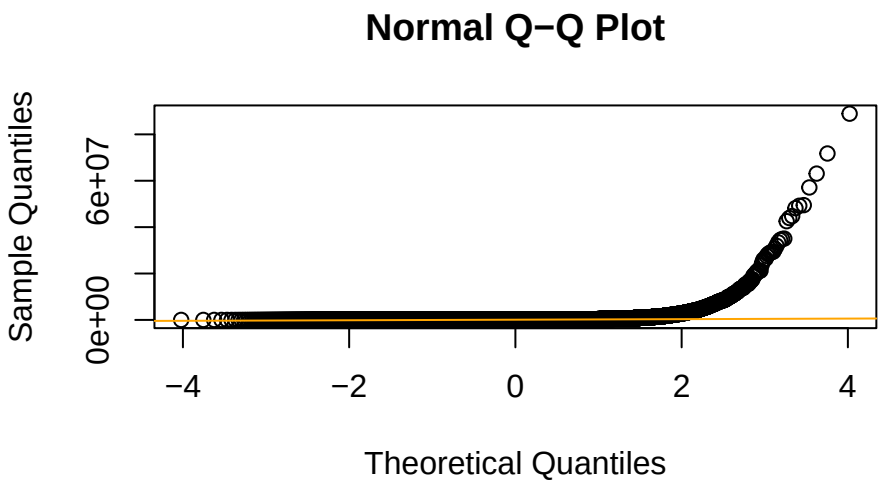

Figura 2: Gráfico Q-Q de la abundancia de fosfopéptidos.

Listado de los péptidos identificados como significativos para identificar el tipo tumoral.

```
##  [1] "LSLEGDHSTPPSAYGSVK[14] Phospho"
##  [2] "SASPDDDLGSSNWEAADLGNEER[3] Phospho"
##  [3] "HADAEMTGYVVTR[6] Oxidation|[9] Phospho"
##  [4] "GVGYETILK[4] Phospho"
##  [5] "HTDDEMTGYVATR[6] Oxidation|[9] Phospho"
##  [6] "PYQYPALTPEQK[4] Phospho"
##  [7] "FTDKDQQPSGSEGEDDDAEAALKK[9] Phospho"
##  [8] "GIPLATGDTSPEPELLPGAPLPPPK[9] Phospho"
##  [9] "YGPADVEDTTGSGATDSKDDDDIDLFGSDDEEESEEAKR[28] Phospho"
## [10] "NEEENIYSVPHDSTQGK[7] Phospho"
## [11] "LPSGSGAASPTGSAVDIR[3] Phospho"
## [12] "RLEISPDSSPER[5] Phospho"
## [13] "NLDNGGFYISPR[8] Phospho"
## [14] "TTVYVVEDQR[4] Phospho"
## [15] "KPATPAEDDEDDDIDLFGSDNEEEDKEAAQLR[19] Phospho"
## [16] "ALELPLAASSIPRPRTPSPESHR[16] Phospho|[18] Phospho"
## [17] "AYTNFDAER[2] Phospho"
## [18] "EVAATEEDVTRLPSPTSPFSSLSQDQAATSK[16] Phospho"
## [19] "LDQPVSAPPSPR[10] Phospho"
## [20] "KASPEPPDSAEGALK[3] Phospho"
## [21] "HKAPGSADYGFAPAAGR[9] Phospho"
## [22] "AHSPASTLPNSPGSTFER[3] Phospho"
## [23] "RLEISPDSSPER[5] Phospho|[9] Phospho"
## [24] "SSPEQSYQGDMYPTR[7] Phospho|[11] Oxidation"
## [25] "IHDLEDDLEMSSDASDASGEEGGRVPK[10] Oxidation|[18] Phospho"
## [26] "GLLAQGLRPESPPPAGPLLNGAPAGESPQPK[11] Phospho"
## [27] "LTAGVPDTPTR[8] Phospho"
## [28] "RQVSASELHTSGILGPETLR[4] Phospho"
## [29] "KDPEDTGAEKSPTTSADLK[14] Phospho"
## [30] "KHSPSPPPPTPTESR[3] Phospho|[5] Phospho"
## [31] "AEDMYSAQSHQAATPPK[4] Oxidation|[5] Phospho"
```

```
## [32] "FAGDKGYLTK[7] Phospho"
## [33] "HTDDEMTGYVATR[9] Phospho"
## [34] "GPPQSPVFEGVYNNSR[5] Phospho"
## [35] "LSVPTSDEEDEVPAPKPR[6] Phospho"
## [36] "LSLEGDHSTPPSAYGSVK[16] Phospho"
## [37] "GSGTASDDEFENLR[6] Phospho"
## [38] "AEVPGATGGDSPHLQPAEPPGEPR[11] Phospho"
## [39] "ASPSPQPSSQPLQIHR[4] Phospho"
## [40] "LKTEKEPDATPPSPR[13] Phospho"
## [41] "ENPPVEDSSDEDDKR[9] Phospho"
## [42] "KGAGDGSDEEVDGKADGAEAKPAE[7] Phospho"
## [43] "LAEETEGPSDGESAAEKR[9] Phospho"
## [44] "NVYYELNDVR[3] Phospho"
## [45] "SSPEQSYQGDMYPTR[12] Phospho"
## [46] "IGEGTYGVVYK[10] Phospho"
## [47] "AHLTVGQAAAGGSGNLLTER[13] Phospho"
## [48] "STIGVMVTASHNPEEDNGVK[6] Oxidation|[10] Phospho"
## [49] "KLSVPTSDEEDEVPAPKPR[7] Phospho"
## [50] "TPPSTTVGSHSPPETPVLTR[11] Phospho"
## [51] "TPSFLKK[3] Phospho"
## [52] "APAPAPPGTVTQVDVR[9] Phospho"
## [53] "SAESPTSPVTSETGSTFK[4] Phospho"
## [54] "LDSSPSVSSTLAAK[4] Phospho"
## [55] "KPVTVSPTTPTSPTEGEAS[12] Phospho"
## [56] "ESLKEEDESDDDNM[9] Phospho"
## [57] "NVPQEESLEDSDVDADFK[11] Phospho"
## [58] "EGMNPSYDEYADSDEDQHDAYLER[3] Oxidation|[13] Phospho"
## [59] "RPDPDSDEDEDYERER[6] Phospho"
## [60] "LDIDSPPITAR[5] Phospho"
## [61] "SSKASLGSLEGEAEAEASSPK[5] Phospho|[8] Phospho"
## [62] "TTVYVVEDQRR[4] Phospho"
## [63] "FKAEAPLPSPK[9] Phospho"
## [64] "LTFDSSFSPNTGKK[8] Phospho"
## [65] "NHSDSSTSESEVSSVSPLK[16] Phospho"
## [66] "MGAPESGLAEYLFDKHTLGDSDNES[1] Oxidation|[25] Phospho"
## [67] "IEDSEPHIPLIDDTDAEDDAPTKR[4] Phospho"
## [68] "SFLDSGYR[7] Phospho"
## [69] "VDEEDSDEESHHDEMSEQEEELEDDPTVVK[6] Phospho|[15] Oxidation|[16] Phospho"
## [70] "TASLPGYGR[3] Phospho"
## [71] "EYVSNDAAQSDDEEKLQSQPTDTDGGR[10] Phospho"
## [72] "SPGHMVILDQTK[1] Phospho|[5] Oxidation"
## [73] "ASPVPAPSSGLHAAVR[2] Phospho"
## [74] "EFITGDVEPTDAESEWHSENEEEEKLAGDMK[18] Phospho|[30] Oxidation"
## [75] "AEDMYSAQSHQAATPPK[5] Phospho"
## [76] "NSHTDNVSYEHSFNK[9] Phospho"
## [77] "KPATPAEDDEDDDIDLFGSDNEEEDKEAAQLR[4] Phospho|[19] Phospho"
## [78] "WAHDKFSGEEGEIEDDESGTENREEK[18] Phospho"
## [79] "QVEPPAKKPATPAEDDEDDDIDLFGSDNEEEDKEAAQLR[26] Phospho"
## [80] "AFGYYGPLR[4] Phospho"
## [81] "IHSPIIR[3] Phospho"
## [82] "QVSASELHTSGILGPETLR[3] Phospho"
## [83] "IGSDPLAYEPK[3] Phospho"
## [84] "GLAGPPASPGK[8] Phospho"
## [85] "SVIEPLPVTPTR[9] Phospho"
```

```
## [86] "VKGEYDMTVPK[5] Phospho|[7] Oxidation"
## [87] "QKSEEPSVSIPFLQTALLR[9] Phospho"
## [88] "HQSFGAAVLSR[3] Phospho"
## [89] "FLEDDPSDPTYTSSLGGK[11] Phospho"
## [90] "KGPGEGVLTLR[9] Phospho"
## [91] "SSKASLGSLEGEAEAEASSPK[1] Phospho|[2] Phospho"
## [92] "VKAQTPPGPSLSGSK[5] Phospho"
## [93] "SFDYNYR[1] Phospho"
## [94] "TTSISPALAR[5] Phospho"
## [95] "SNYYDAYQAQPLATR[3] Phospho"
## [96] "TTSPPLSIPTTHLIHQPAGSR[3] Phospho"
```

Volcano plot para la visualización de los datos seleccionados. Los puntos en morado indican fosfopéptidos con p-valor ajustado menor a 0.05, mientras que los puntos en naranja corresponden a aquellos con p-valor ajustado mayor. La línea horizontal marrón marca el umbral de significancia (p-valor ajustado = 0.05).

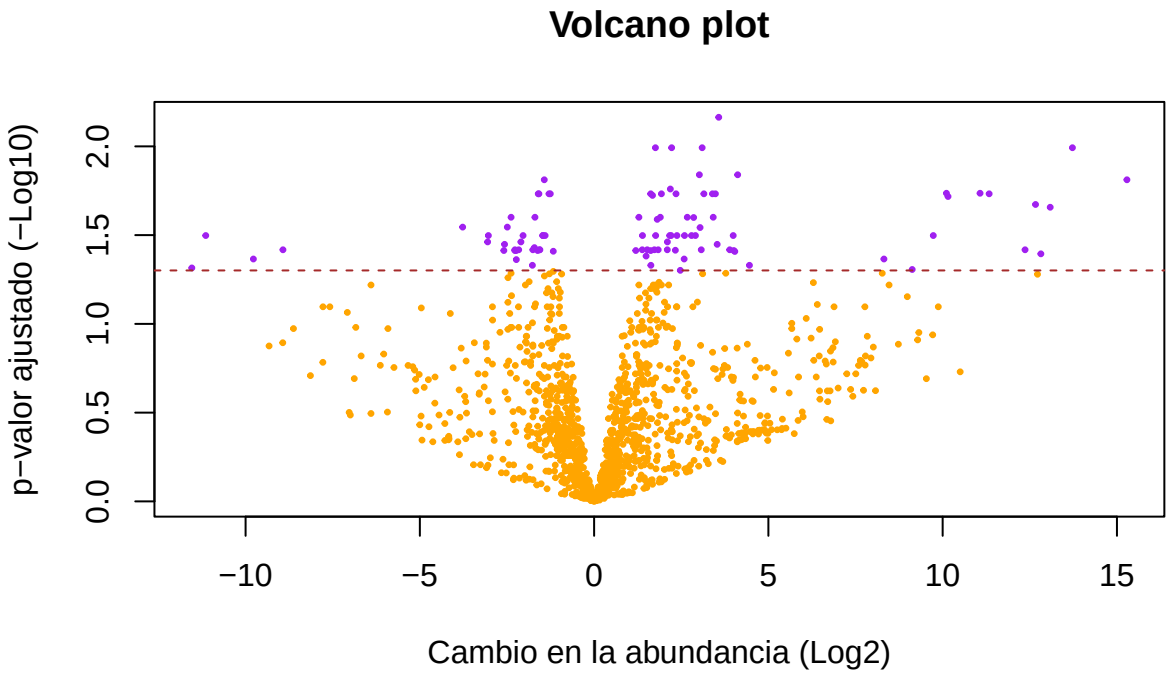

Figura 3: Volcano Plot mostrando la relación entre el cambio en la abundancia (Log2) y el p-valor ajustado (-Log10) para cada fosfopéptido.

Heatmap, gráfico que nos muestra mediante un código de color la abundancia de cada fosfopéptido considerado como significativo para cada muestra, la cual está etiquetada con el nombre del grupo al que pertenece. Los colores van de verde (baja abundancia) a rojo (alta abundancia), pasando por blanco que nos indica una abundancia media.

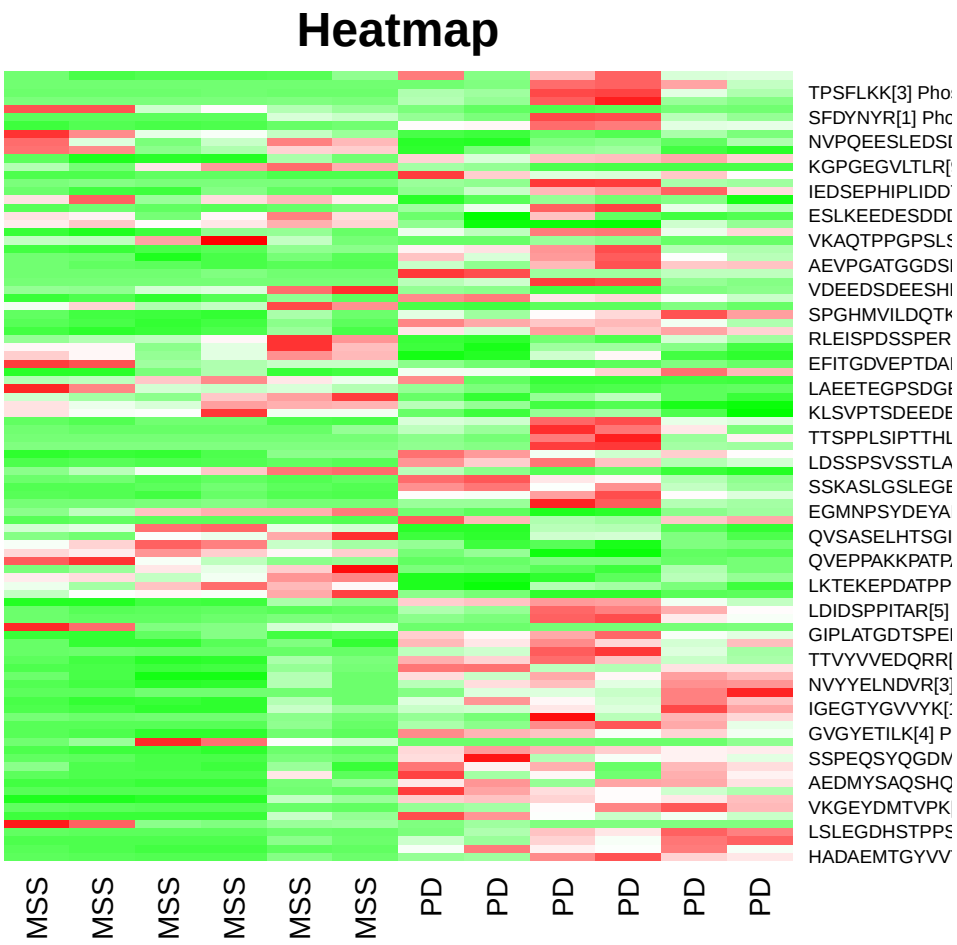

Figura 4: Heatmap que muestra la abundancia relativa de los fosfopéptidos significativos entre las muestras.

Discusión

En este estudio se ha analizado un conjunto de datos de fosfoproteómica con el objetivo de identificar fosfopéptidos que puedan diferenciar entre dos subtipos tumorales: MSS y PD. Se han identificado 96 fosfopéptidos con diferencias significativas en abundancia entre ambos grupos tumorales, lo que sugiere que ciertos fosfopéptidos pueden estar relacionados con características específicas de estos subtipos. Estos hallazgos podrían proporcionar una base para futuras investigaciones dirigidas a la identificación de posibles marcadores tumorales o a la comprensión de los mecanismos moleculares que dan lugar a las diferencias entre MSS y PD.

Para el análisis estadístico, se utilizó un enfoque basado en el modelo de diseño lineal limma y la normalización voom, lo que permitió ajustar las varianzas de los datos y detectar diferencias significativas con un p-valor ajustado de 0.05. La visualización mediante volcano plot reflejó la diferencia en la abundancia de los fosfopéptidos entre los grupos. El heatmap permitió observar patrones de abundancia de fosfopéptidos que podrían estar relacionados con diferencias biológicas entre estos subtipos tumorales.

Este estudio presenta varias limitaciones que deben tenerse en cuenta al interpretar los resultados. Primero, el tamaño de la muestra es limitado, lo que podría afectar la generalización de los resultados y reducir la potencia estadística del análisis. Segundo, solo se incluyeron dos réplicas por muestra, lo que puede influir en la robustez de los resultados. Tercero, el análisis realizado se basa en datos de espectrometría de masas, por lo que la identificación de fosfopéptidos requiere validación adicional mediante otras técnicas para confirmar su relevancia biológica en estos subtipos tumorales.

Conclusiones

El uso del objeto SummarizedExperiment ha facilitado el análisis y la organización de los datos de fosfoproteómica, permitiendo un manejo eficiente de la abundancia de fosfopéptidos en las diferentes muestras.

El análisis estadístico ha identificado diferencias significativas en la abundancia de fosfopéptidos entre los subtipos tumorales MSS y PD, lo que sugiere que algunos de estos fosfopéptidos podrían estar relacionados con características biológicas específicas de cada subtipo tumoral.

Aunque el estudio presenta limitaciones, como el tamaño reducido de la muestra y el número limitado de réplicas, los resultados obtenidos proporcionan un punto de partida para futuras investigaciones en la identificación de marcadores tumorales y la comprensión de los mecanismos moleculares de estos subtipos tumorales.

Referencias

Ruiz, P. M. (2025). Ruiz-Real-PaulaMaria-PEC1 (v.1.0) [Repositorio de GitHub]. GitHub. https://github.com/pruizreal/Ruiz-Real-PaulaMaria-PEC1