PEC 1 ANALISIS DE DATOS OMICOS

Victor Tablado Alonso

01/11/2024

Contenido

ABSTRACT

Durante este estudio se han estudiado los fosfopéptidos que se expresan en dos subtipos de cáncer, con el fin de encontrar alguno de estos fosfopéptidos que nos permita diferenciar ambos subtipos de cáncer. Los fosfopéptidos se obtuvieron mediante cromatografía liquida acoplada a un espectrómetro de masas (LC-MS), y el tratamiento estadístico se llevó a cabo con R.

Durante el análisis estadístico, se realizaron diagramas de caja e histogramas en fases tempranas del análisis para poder ver la distribución de los datos. Se transformaron los datos a base logarítmica para facilitar la representación de los mismos. Los análisis realizados fueron: correlación entre las muestras con representación en matriz de distancias, análisis de componentes principales (PCA), t-test para filtrar los péptidos característicos de cada uno de los tumores, y un análisis diferencial basado en un modelo lineal con el mismo objetivo que el t-test anterior. Finalmente se compararon los fosfopéptidos obtenidos del t-test y del análisis diferencial para obtener los fosfopéptidos característicos.

Los resultados obtenidos son bastante ambiguos, obtenemos 130 péptidos comunes entre las dos pruebas realizadas, pero cada una de las pruebas nos dan resultados algo contradictorios sobre el tipo de cáncer en el que cada uno de estos péptidos tienen una representación mayor. Encontramos varios fosfopéptidos que son característicos de el subtipo de cáncer PD, y tan solo uno en el que ambos análisis concuerdan que es característico para el subtipo de cáncer MSS.

OBJETIVOS DEL ESTUDIO

El objetivo del estudio es identificar fosfopéptidos diferenciales entre dos subtipos específicos de tumores. Las muestras de estos tumores se analizan mediante espectrometría de masas acoplado a una cromatografía liquida (LC-MS). Partimos de 1400 fosfopéptidos detectados en ambas muestras tumorales.

En el estudio se emplean análisis estadísticos y herramientas bioinformáticas para identificar aquellos péptidos cuya abundancia varía de manera significativa entre ambos subtipos de tumor. Estos fosfopéptidos significativos o diferenciales podrían servir como potenciales biomarcadores, permitiendo una diferenciación más sencilla y clara entre ambos tipos de tumores y proporcionando una nueva perspectiva sobre los mecanismos moleculares de cada tipo de cáncer.

MATERIALES Y MÉTODOS

Partimos de un archivo Excel (.xlsx) en el que se encuentra la matriz de datos con la que vamos a trabajar, que tiene la estructura que se ve en la **figura 1.** Lo primero que se ha hecho es descargar el archivo del repositorio de github en el que se encuentra y cargarlo en RStudio, que es el programa que se va a utilizar para su análisis.

SequenceModifications	Accession	Description	Score	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	CLASS	PHOSPHO
LYPELSQYMGLSLNEEEIR[2] Phospho|[9] Oxida	O00560	Syntenin-1 OS=H	48,07	24,29438	44475,96	0	6269,141	1135,817	21933,9	0	0	772,9056	2136,746	1820,724	1727,91	H	Y
VDKVIQAQTAFSANPANPAILSEASAPIPHDGNLY	O00560	Syntenin-1 OS=H	67,05	0	43138,9	2102,056	50355,05	248,9275	3239,16	1315,904	0	0	0	0	892,3565	H	Y
VIQAQTAFSANPANPAILSEASAPIPHDGNLYPR[3	O00560	Syntenin-1 OS=H	77,71	3412,603	172143	77323,02	307637,4	98442,28	192982,4	24851,34	16547,95	5565,282	0	3264,563	5901,958	H	Y
HADAEMTGYVVTR[6] Oxidation|[9] Phospho	O15264	Mitogen-activate	44,87	220431,2	145656,9	104287,8	75887,37	773377,5	481165,5	1027196	1163747	4080239	4885818	3093787	2759105	H	Y
HADAEMTGYVVTR[9] Phospho	O15264	Mitogen-activate	67,42	18254,78	8529,755	35955,9	44102,32	57145,17	34638,01	21231,26	49499,7	666107	379313,6	255792,1	579765	H	Y
STGPGASLGTGYDR[12] Phospho	O15551	Claudin-3 OS=Hc	63,69	644513,3	261938	187023,5	124867,7	4487444	2572575	535809,2	434645,9	91361,88	65997,91	243250,4	206632,6	H	Y

Figura 1: Estructura del database

Una vez cargado, se descargan todos los paquetes necesarios que se van a utilizar durante el análisis (Bioconductor, BiocManager, ggplot2, SummarizedExperiment y readxl) y se abren sus respectivas librerías para poder acceder a todos los códigos e información que almacenan.

Una vez hecho esto, se crea un contenedor del tipo Summarized Experiment, el que se nos pide en el enunciado de la actividad, que contiene los datos y metadatos del archivo (información del dataset, filas y columnas).

Se realizan algunos análisis básicos del dataset, como su estructura o sus dimensiones. sHay que tener en cuenta que según la descripción del dataset, los datos ya están normalizados, por lo que en ese sentido no se hace ningún tratamiento estadístico. Se representan los datos mediante un diagrama de caja y un histograma para ver mas detalladamente la distribución de la muestra.

Debido a que con esta muestra es difícil interpretar los diagramas realizados, se transforman los datos para trabajar con ellos en base logarítmica.

Una vez se tienen los datos adaptados, se comienza el análisis estadístico del dataset. Primero se realiza un análisis de correlación entre las muestras, se calcula la matriz de correlaciones y se representa mediante un mapa de distancias con la función heatmap. Con este análisis se ve si algunas las abundancias de cada uno de los fosfopéptidos están relacionadas entre sí, o son independientes.

A continuación, se lleva a cabo un análisis de componentes principales con el fin de identificar si las variables están agrupadas de alguna forma, o hay alguna variable atípica. Como los dos componentes principales no tienen valores especialmente altos, ya que es un dataset complejo, vamos a utilizar también el tercer componente principal del data en un gráfico 3D. Se hace una representación bidimensional de los dos primeros componentes principales y una representación tridimensional de los tres primeros componentes principales.

Se lleva a cabo una prueba t test con la cual queremos filtrar que fosfopéptidos se diferencian entre ambos tumores. Se realiza el test, se calculan los valores significativos obtenidos y guardamos los fosfopéptidos que nos da el estudio. Se representan los datos en un volcano plot.

Por último, se crea un modelo lineal para realiza un análisis diferencial para modelar la relación entre cada tipo de cáncer y los niveles de expresión de los fosfopéptidos. Se representa mediante un volcano plot.

Para finalizar, vamos a comparar los fosfopéptidos obtenidos en ambos análisis, el t-test y el análisis diferencial, para obtener los fosfopéptidos que podemos decir que son diferenciales en cada uno de los cánceres.

CONEXIÓN A GITHUB

Subimos el archivo con el código a github mediante comandos en la consola de R a pesar de haberlo subido una vez terminado para, en caso de modificar el código tener mayor comodidad y facilidad. Primero se crea el repositorio de git mediante el comando use_git(), y se conecta posteriormente con github mediante el comando use_github().

Hay que tener en cuenta que antes de utilizar estos comandos, hay que tener una cuenta creada en github, instalar git y conectar nuestra cuenta de github con el proyecto con el que estamos trabajando mediante un token. También es importante tener instalado la librería usethis que es la librería que contiene los comandos de git y github mencionados anteriormente.

La URL del directorio de github con el archivo de código y todos los documentos es el siguiente:

[Directorio de github]

Figura 2: Archivos de Github

En la figura 2, podemos ver todos los archivos cargados finalmente en Github.

RESULTADOS

Una vez cargados los datos en R, lo primero que se hizo fue preparar los datos y crear el SummaryExperiment que tiene la estructura que vemos en la **figura 3**.

	M1_1_MSS	M1_2_MSS	M5_1_MSS	M5_2_MSS	T49_1_MSS	T49_2_MSS	M42_1_PD	M42_2_PD
[1,]	2.429438e+01	4.447596e+04	0.000000e+00	6.269141e+03	1.135817e+03	2.193390e+04	0.000000e+00	0.000000e+00
[2,]	0.000000e+00	4.313890e+04	2.102056e+03	5.035505e+04	2.489275e+02	3.239160e+03	1.315904e+03	0.000000e+00
[3,]	3.412603e+03	1.721430e+05	7.732302e+04	3.076374e+05	9.844228e+04	1.929824e+05	2.485134e+04	1.654795e+04
[4,]	2.204312e+05	1.456569e+05	1.042878e+05	7.588737e+04	7.733775e+05	4.811655e+05	1.027196e+06	1.163747e+06
[5,]	1.825478e+04	8.529755e+03	3.595590e+04	4.410232e+04	5.714517e+04	3.463801e+04	2.123126e+04	4.949970e+04

Figura 3: Estructura del SummarizedExperiment

Se representa la distribución de los datos en un boxplot y en un histograma de las 5 primeras muestras, los cuales podemos ver en las **figuras 4 y 5.**

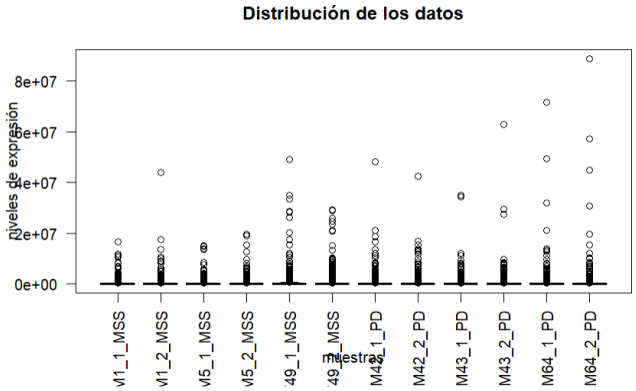

Figura 4: Distribución de los datos representada mediante un boxplot

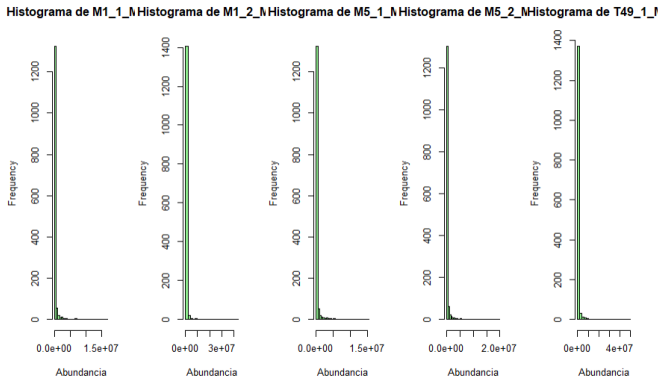

Figura 5: Distribución de los datos representada mediante un histograma de las 5 primeras muestras del dataset

Después de la transformación de los datos a base logarítmica, repetimos estos mismos gráficos para ver la nueva distribución (**Figuras 6 y 7)**

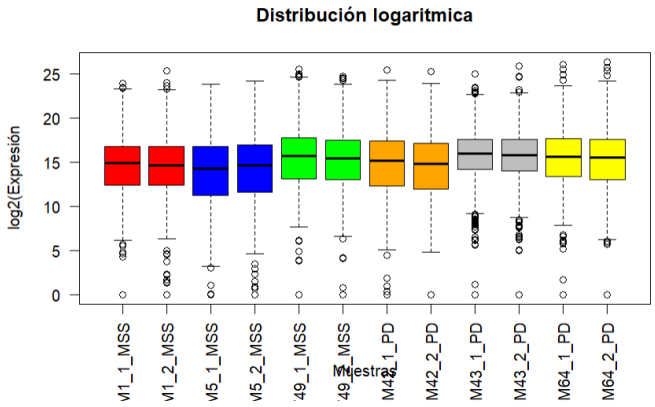

Figura 6: Distribución de los datos en base logarítmica representada mediante un boxplot

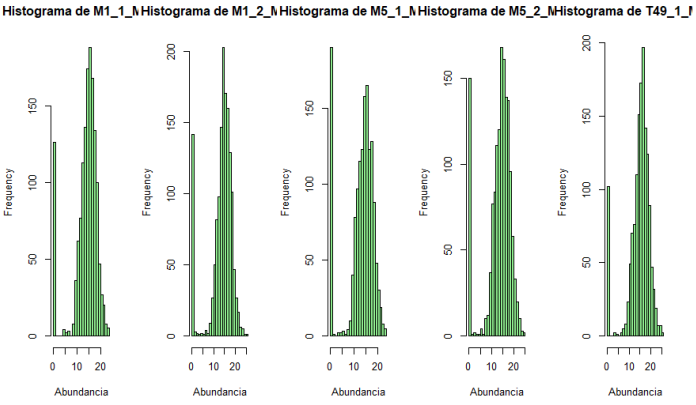

Figura 7: Distribución de los datos en base logarítmica representada mediante un histograma de las primeras 5 muestras

Como vemos en las figuras 5 y 6, con la transformación a base logarítmica la distribución se puede interpretar de forma mucho más fácil, ya que se asemeja más a una distribución normal.

A continuación, se lleva a cabo el análisis de correlación entre las muestras, que se representa mediante una matriz de distancias (Figura 8). En este punto se pueden comenzar a sacar algunas conclusiones sobre las variables con las que estamos trabajando.

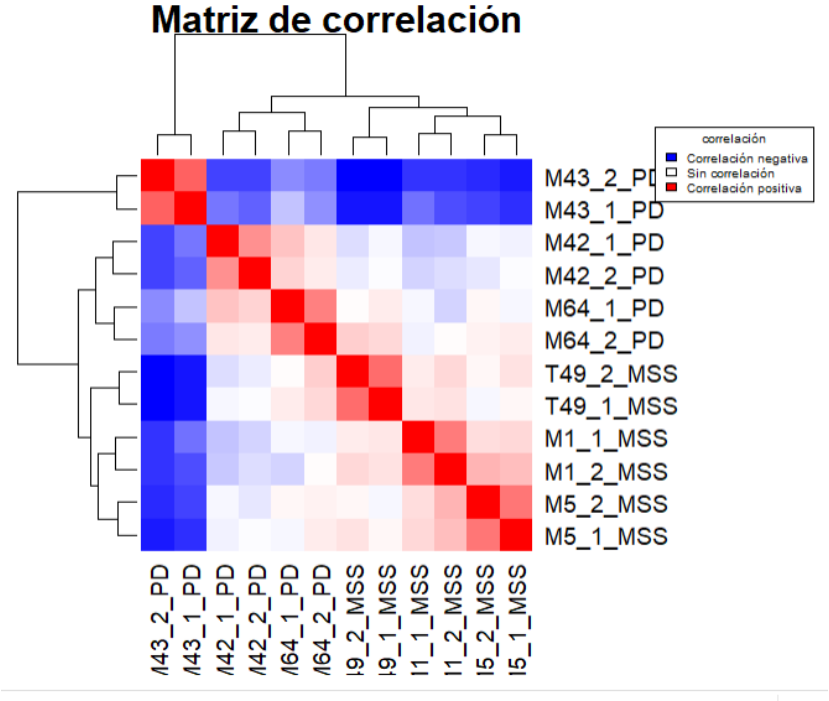

Figura 8: Matriz de distancias de las muestras del dataset

Con este gráfico podemos decir que las muestras no están muy relacionadas entre si, ya que solo vemos cierta similitud entre las replicas técnicas de cada una de las muestras. Cabe destacar las muestras M43_1_PD y su replica M43_2_PD, que parecen tener una correlación negativa con el resto de muestras.

En el análisis de componentes principales (Figura 9) no nos revela una gran cantidad de información. De este nuevo gráfico podemos decir que las componentes principales no tienen un peso muy elevado, debido a que es un dataset muy complejo. No parece haber una separación clara de las muestras MSS y PD, por lo que parece que ambos subtipos de cáncer tienen perfiles de expresión bastante similares.

Las únicas muestras que parecen estar más alejadas del resto son las muestras T49_1_MSS y M_64_1_PD, pero esto no nos da ninguna información muy relevante sobre los datos.

Por último, se realizan dos análisis diferentes, una prueba t-test y un análisis diferencial basado en un modelo linear y se comparan los datos.

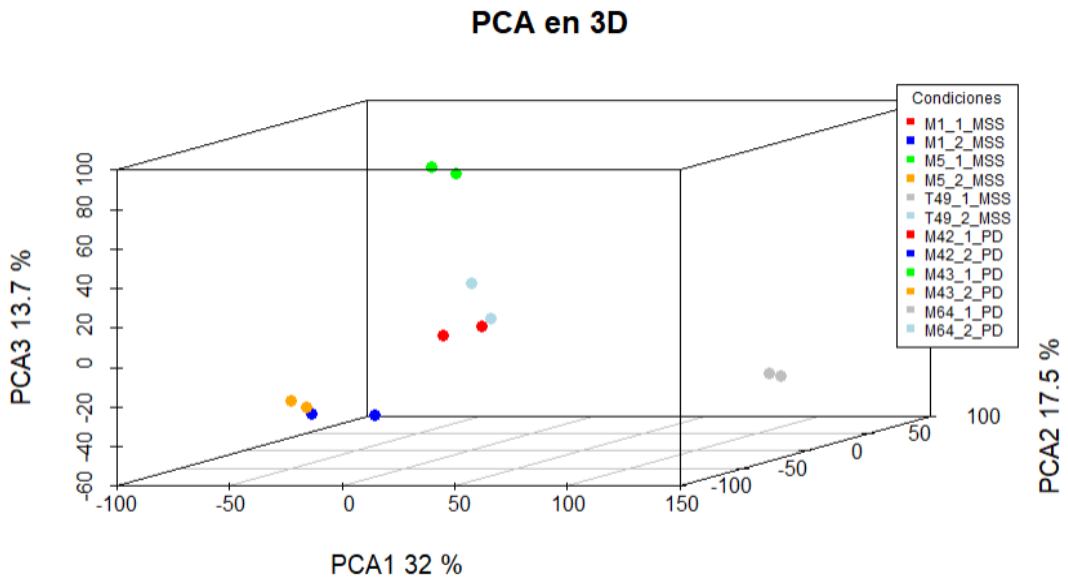

Figura 9: Analisis de componentes principales (PCA)

En el caso del t-test (Figura 10), encontramos 350 péptidos con un p-value menor a 0.05, lo que significa que son fosfopéptidos con diferencias significativas sobre el resto, por lo que podemos suponer que pueden ser característicos de uno de los dos subtipos de tumor

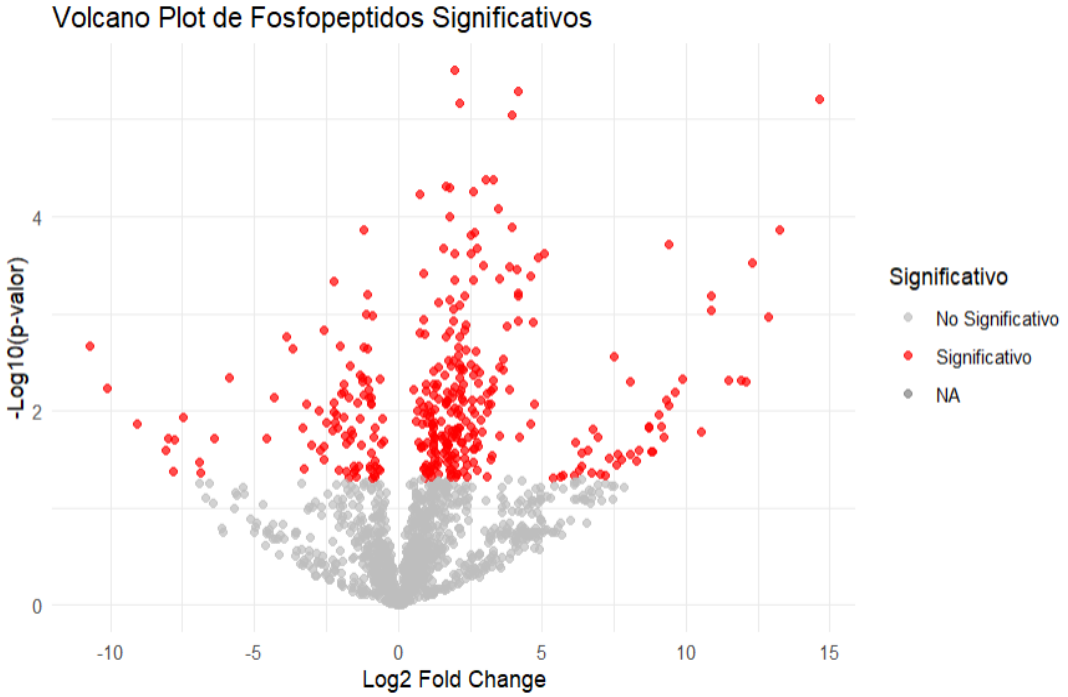

Figura 10: Representación gráfica mediante volcano plot de la prueba t-test realizada sobre el dataset

En el análisis diferencial, creamos un modelo linear y lo ajustamos con el Summarized Experiment en base logarítmica. De forma similar al t-test nos quedamos con los fosfopéptidos que tienen un p-value ajustado menor de 0.05, en este caso vemos que

hay 130 fosfopéptidos significativos. Se representan los datos obtenidos en un volcano plot (Figura 11). Vemos que con este análisis encontramos una cantidad notablemente menor de fosfopéptidos significativos.

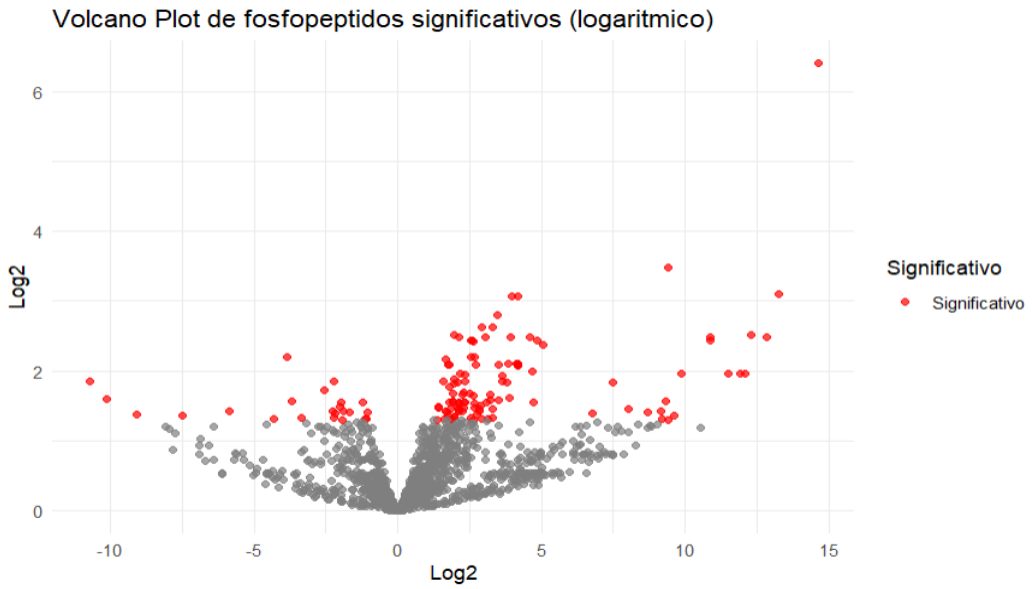

Figura 11: Representación gráfica mediante volcano plot del análisis diferencial del dataset

Debido a esta disparidad en los resultados de ambas pruebas, vamos a comprobar cuales de los fosfopéptidos son comunes en los resultados de las pruebas. Generamos dos dataframes para cada uno de los resultados que contiene el fosfopéptido y el cáncer en el que tienen una mayor representación. El subtipo de cáncer en el que cada fosfopéptido tiene un mayor peso se obtiene a partir de el valor logFC que se obtiene en cada una de las anteriores pruebas.

Fosfopeptido <chr>	Tipo_Cancer_fosfopeptidos <chr>	Tipo_Cancer_volcano <chr>
ACPDSLGSPAPSHAYHGGVL[2] Carbamidomethyl|[15] Phospho	PD	PD
AEDMYSAQSHQAATPPK[4] Oxidation|[5] Phospho	MSS	PD
AEDMYSAQSHQAATPPK[5] Phospho	PD	PD
AEDMYSAQSHQAATPPKDGK[5] Phospho	PD	PD
APGSADYGFAPAAGR[7] Phospho	PD	PD
AYTNFDAER[2] Phospho	PD	PD
DASRPHVVKVYSEDGACR[11] Phospho|[17] Carbamidomethyl	PD	MSS
FAGDKGYLTK[7] Phospho	PD	PD
GSHFFPGNNVIYEK[12] Phospho	MSS	PD
HADAEMTGYVVTR[6] Oxidation|[9] Phospho	PD	PD

1-10 of 32 rows Previous 1 2 3 4 Next

Figura 12: Fosfopeptidos característicos comunes a los dos análisis estadísticos realizados y el tipo de cáncer al que pertenecen según ambas pruebas

En los resultados que vemos en la figura 12, y si recorremos el resto del dataframe obtenido, vemos que hay bastante disparidad entre el tipo de cáncer al que pertenece o al que representa cada uno de los fosfopéptidos, esto puede deberse a la forma que tiene cada una de estas pruebas de tratar los datos y los ajustes que se hacen en el modelo lineal.

DISCUSIÓN, LIMITACIONES Y CONCLUSIÓN DEL ESTUDIO

Una vez puestos sobre la mesa todos los resultados del estudio, podemos concluir que tan solo hay un fosfopéptidos característico del subtipo de tumor MSS, mientras que hay una mayor cantidad para el subtipo PD. Sin embargo, serían necesarios más análisis para concretar los fosfopéptidos característicos de cada uno de los tumores, y es que al estar tratando con datos de abundancia y la mayoría de los fosfopéptidos están presentes en ambos tipos de tumores, el análisis es complicado.

Depende del tipo de tratamiento estadístico que utilicemos nos pueden variar desde el número de fosfopéptidos significativos, hasta el tipo de cáncer del que es característico cada fosfopéptido. Este es uno de los grandes retos de la búsqueda de biomarcadores, la gran variabilidad de las muestras dependiendo del paciente o la evolución del cáncer y la oscilación de los resultados dependiendo del tratamiento estadístico. A día de hoy no hay una “receta” perfecta para encontrar biomarcadores para enfermedades tan complejas como el cáncer, alzheimer o la esclerosis lateral, y es necesario una ardua investigación y trabajo estadístico para encontrar biomarcadores o chekpoints que nos indiquen como se desarrolla la enfermedad.