Informe PEC1

Paula Galiana Haro

2025-03-31

Contents

2. Resumen

En este trabajo se ha realizado un análisis exploratorio de un conjunto de datos metabolómicos compuesto por 77 muestras de orina de pacientes control y caquéxicos. Para ello se ha empleado principalmente el paquete POMA de Bioconductor, trabajando sobre un objeto de clase `SummarizedExperiment`. Tras el preprocesamiento de los datos (normalización y eliminación de outliers),se aplicó un Análisis de Componentes Principales (PCA) que mostró cierta separación entre grupos. También se realizó un análisis univariante (test de Mann-Whitney),identificando múltiples metabolitos con diferencias significativas entre ambos grupos, representados en un Volcano Plot. Finalmente, se utilizó una técnica supervisada, sPLS-DA, para maximizar la separación de los grupos, donde N.N.dimetilgicina, Valina y Quinolato fueron algunos de los metabolitos con mayor capacidad discriminativa entre grupos. Los resultados sugieren que el perfil metabolómico en orina podría emplearse para distinguir a pacientes caquéxicos.

3. Objetivos

El objetivo principal de esta actividad es realizar un proceso de análisis de datos ómicos sobre un conjunto de datos de metabolómica. Dentro de este, se plantean los siguientes objetivos:

- Seleccionar y preparar un dataset de metabolómica, creando un objeto de clase `SummarizedExperiment` para el análisis de los datos.

- Realizar un preprocesamiento de los datos, incluyendo normalización y búsqueda de outliers, para garantizar la calidad del dataset previamente a su análisis.
- Exploración global del conjunto de datos mediante técnicas multivariantes como el Análisis de Componentes Principales (PCA), buscando posibles agrupaciones entre las muestras y que metabolitos contribuyen más a explicar la variabilidad del dataset.

- Análisis univariante de los datos, utilizando tests estadísticos para identificar los metabolitos con diferencias significativas entre grupos de pacientes control y caquéxicos.
- Aplicar técnicas supervisadas (sPLS-DA) para maximizar la separación entre grupos y detectar los metabolitos que más facilitan la discriminación de los grupos de pacientes.

4. Métodos

- El dataset utilizado es el `human_cachexia.csv` proviene del repositorio GitHub aportado en la actividad. Recoge información metabolómica de 77 muestras de orina, donde 47 pertenecen a pacientes con caquexia y 30 a pacientes control. Cada muestra contiene valores de concentración de `dim(object)[1]` metabolitos.

- Se utilizó el paquete POMA Bioconductor para todo el análisis, ya que ofrece un conjunto de herramientas que facilitan el procesamiento y análisis estadístico de datos ómicos, utilizando la clase `SummarizedExperiment`.

- Se empleó la función `PomaCreateObject()` para construir un objeto de clase `SummarizedExperiment`, que contiene la matriz de los datos (valores de los metabolitos) como los metadatos asociados a las muestras (ID del paciente y grupo al que pertenecen). Esta clase, en comparación con `ExpressionSet`, ofrece una estructura más flexible y moderna, ya que utiliza objetos S4, que tienen una mejor compatibilidad con grandes conjuntos de datos. Además, permite trabajar con más de una matriz de datos simultáneamente, siendo una herramienta muy utilizada en análisis ómicos. Una vez creado el objeto, se guarda con el nombre `cachexia_summarized_experiment.Rda`. Vemos que tiene 63 filas, que corresponden a los metabolitos y 77 columnas, que corresponden a los pacientes.

- Para el **preprocesamiento** de los datos, primero se aplicó una normalización tipo pareto, utilizando la función `PomaNorm(method = "log_pareto")`. Esta realiza una transformación logarítmica para reducir la asimetría de las distribuciones y un escalado tipo pareto. Este escalado es ampliamente utilizado en estudios metabolómicos y usa la raíz cuadrada de la desviación estándar como factor de escalado, lo que permite homogeneizar la varianza. En segundo lugar, se realizó una detección de outliers utilizando la función PomaOutliers(), para detectar muestras atípicas y poder eliminarlas del conjunto de datos a analizar.

- Se aplicó un **Análisis de Componentes Principales (PCA)**, una técnica multivariante no supervisada que reduce la dimensionalidad de nuestro conjunto de datos preservando gran parte de la varianza. Esta técnica nos ayuda a realizar una primera exploración de los datos y poder detectar y observar si existe agrupamiento entre las muestras. Se utilizó la función `PomaPCA()`, especificando únicamente que centre las variables (`center = TRUE`), pues ya se realizó el escalado anteriormente. Si ordenamos los loadings de la PC1 de mayor a menor, podemos seleccionar los 5 metabolitos que más contribuyen a explicar la varianza global del conjunto de datos.

- Se ejecutó un **análisis univariante** para identificar los metabolitos con diferencias significativas entre ambos grupos de pacientes. Para ello se utilizó la función `PomaUnivariate()`, especificando el test no paramétrico de **Mann-Whitney**. Pese a haber normalizado los datos no hemos estudiado si siguen una distribución normal o no, por lo que me pareció el test más adecuado.

- Los resultados se visualizaron mediante un **Volcano plot**, utilizando la función `PomaVolcano()`. Esta gráfica nos permite observar la magnitud del cambio y la significancia estadística, visualizando los

metabolitos más signficativos y si encuentran a concentraciones más altas o más bajas en el grupo de caquéxicos.

- Finalmente se utilizó una técnica supervisada, **sPLS-DA(Sparte Partial Least Squares Discriminant Analysis)**, para intentar maximizar la separación entre grupos utilizando solo los metabolitos con un p valor menor a 0,05. Los resultados se visualizaron mediante el gráfico de factores y se seleccionaron los 5 metabolitos que mejor separan a los pacientes en los grupos control y caquéxicos.

5. Resultados

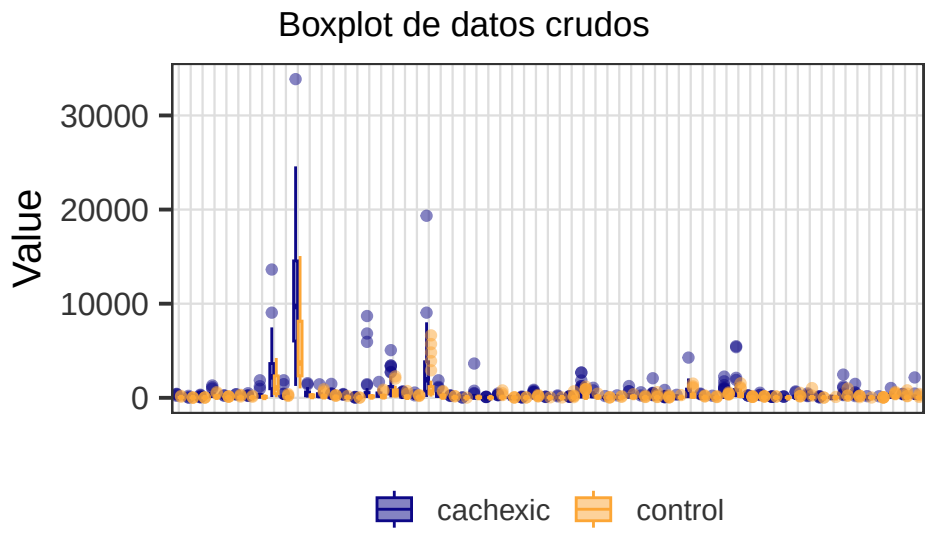
No se comentará nada a cerca del código, únicamente las salidas de los gráficos. En el documento `código-PEC1.R`, se puede consultar el código empleado. Aquí se muestra el objeto `SummarizedExperiment`, el cual contiene 63 filas correspondientes a los metabolitos y 77 columnas, que corresponden a las muestras de orina de cada paciente.

```
object
```

```
## class: SummarizedExperiment
## dim: 63 77
## metadata(0):
## assays(1): ''
## rownames(63): X1.6.Anhydro.beta.D.glucose X1.Methylnicotinamide ...
##   pi.Methylhistidine tau.Methylhistidine
## rowData names(0):
## colnames(77): PIF_178 PIF_087 ... NETL_003_V1 NETL_003_V2
## colData names(1): Muscle.loss
```

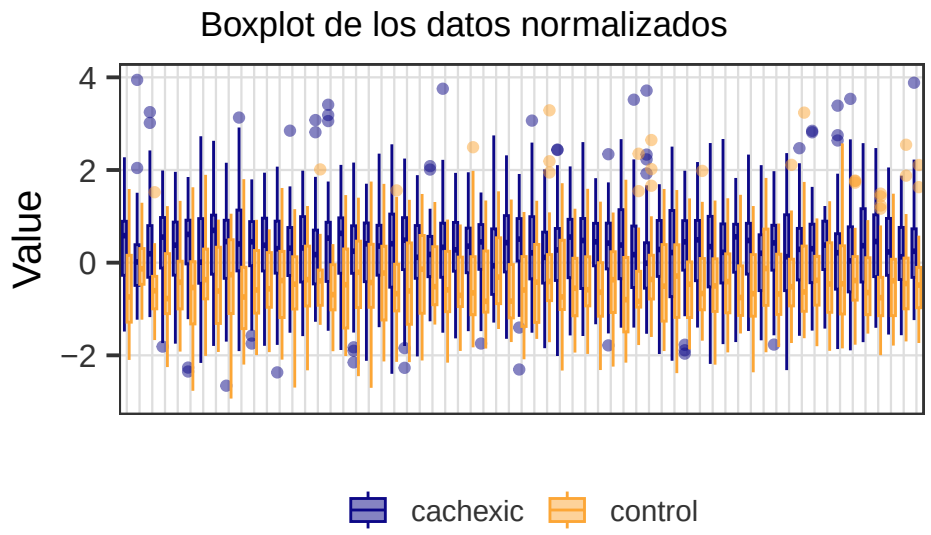
En primer lugar, se exploraron los datos crudos mediante un boxplot, observando una alta variabilidad entre metabolitos y presencia de posibles outliers. La asimetría de algunas distribuciones nos indica que algunos datos estan en escalas diferentes.

```
grafica_boxplot
```



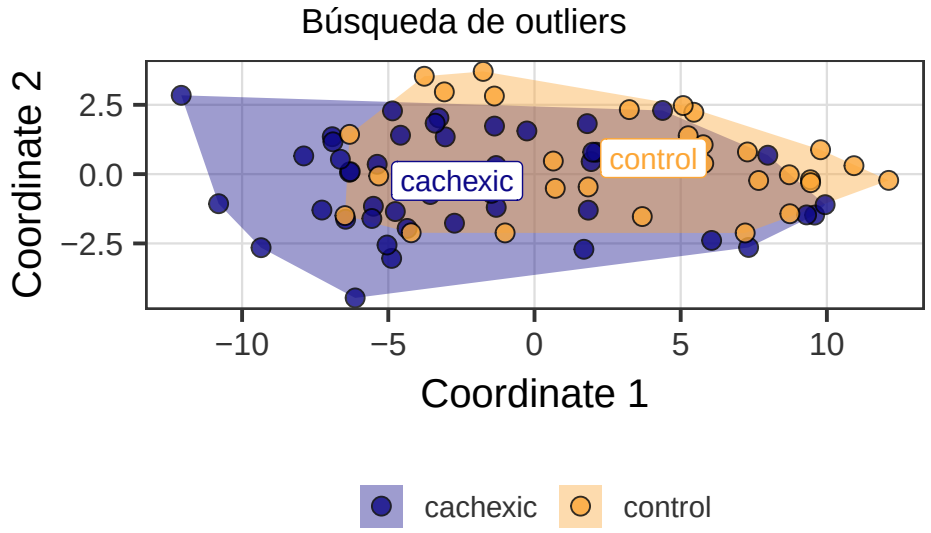
Para igualar la escala de todos los metabolitos y que puedan ser comparables entre sí, se aplicó una transformación logarítmica y un escalado tipo pareto. En el boxplot con los datos normalizados, se puede observar una distribución de los metabolitos mucho más uniforme, pero vemos posibles outliers que podrían modificar la tendencia general del conjunto de datos. También se observa que todos los metabolitos presentan valores más elevados en los pacientes caquéxicos que en los control.

```
boxplot_norm
```



Se realizó un estudio de estos posibles outliers donde se identificaron 3 muestras con valores extremos, todas de pacientes caquéxicos. Estas fueron eliminadas del conjunto de datos para evitar desviaciones en análisis posteriores. En la representación bidimensinal observamos una gran dispersión de los puntos de cada grupo, lo que nos puede indicar una elevada heterogeneidad de valores, probablemente asociado a la naturaleza de la muestra.

```
grafica_outliers
```



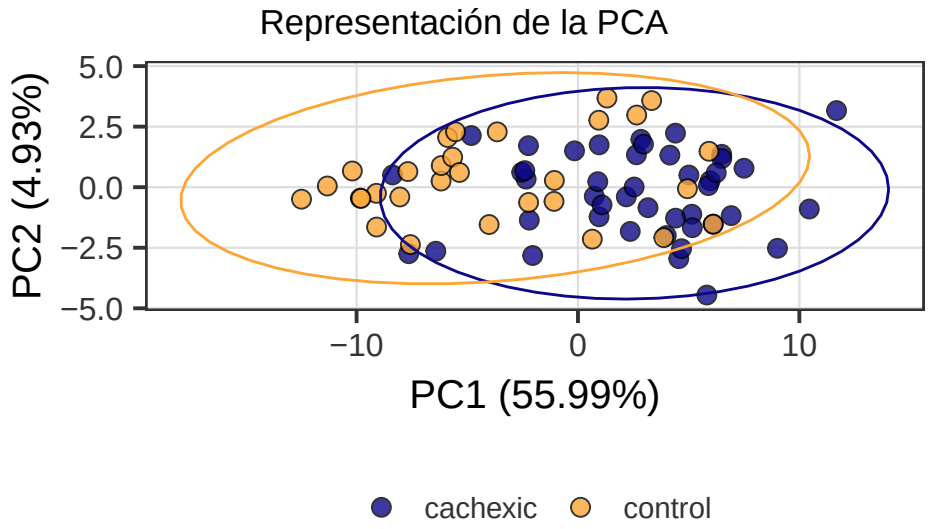
```
#Estos son los 3 outliers
outliers
```

```
## # A tibble: 3 x 4
##   sample       groups   distance_to_centroid limit_distance
##   <chr>        <fct>                   <dbl>          <dbl>
## 1 PIF_119      cachexic                 13.0           12.7
## 2 PIF_099      cachexic                 12.8           12.7
## 3 NETCR_003_V1 cachexic                 13.2           12.7
```

Aunque en el gráfico de outliers ya se intuía cierta separación entre grupos, se aplicó un **Análisis de Componentes Principales (PCA)** para explorar la estructura global del dataset y observar si las muestras se agrupan dependiendo de si son de pacientes control o caquéxicos.

La primera componente principal (PC1) explica el 56% de la variabilidad global del conjunto de datos, lo que indica que existe una fuente dominante de variación. Además, en el plot existe cierta separación entre los grupos, aunque hay un elevado solapamiento entre individuos. Esto puede deberse a que existen otros factores que explican la variabilidad de global del dataset o que hay mucho ruido debido al tipo de muestra con el que estamo trabajando. Si ordenamos los datos de las cargas de mayor a menor, observamos que cis.Aconitate, Glutamine, Alanine, Succinate, Histidine son los metabolitos que más contribuyen a explicar la varianza global de los datos.

```
grafica_pca
```



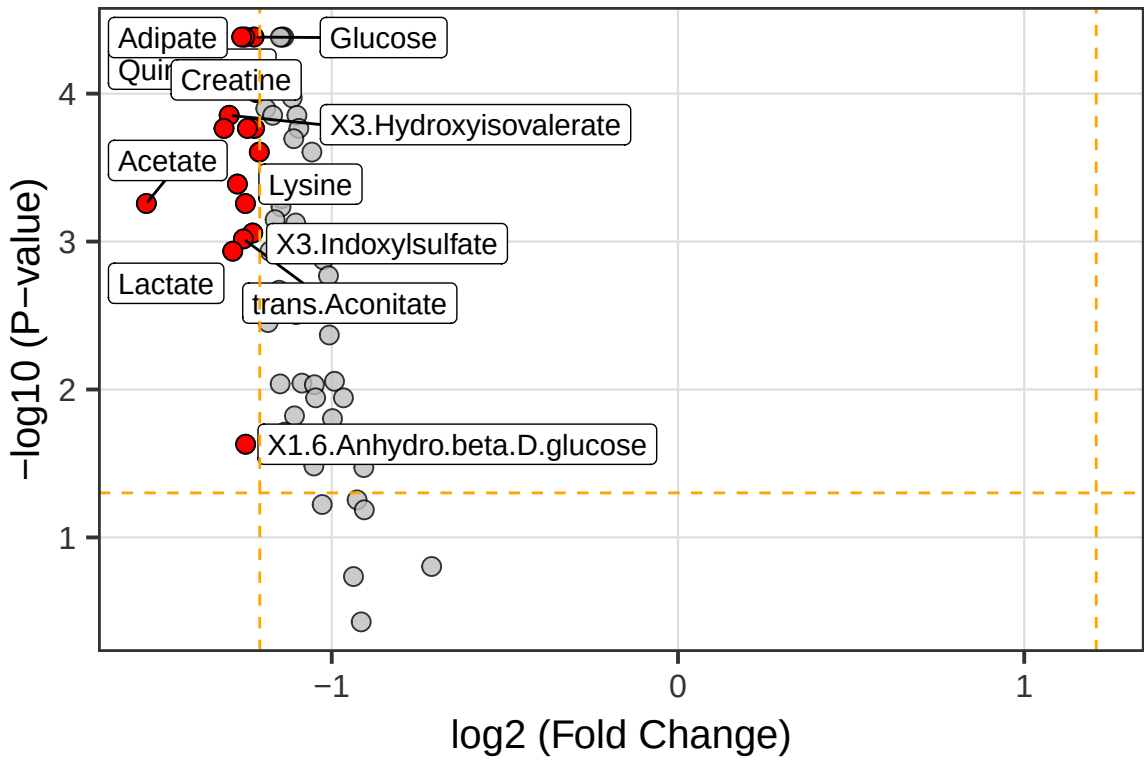
Dado que la separación entre grupos observada en la PCA no era del todo clara, se consideró aplicar un método supervisado para maximizar la separación entre grupos. Primero, se aplicó un análisis univariante mediante el **test de Mann-Whitney** para evaluar cada metabolito individualmente y comprobar si existen diferencias significativas entre grupos. Se seleccionaron los metabolitos con un p-valor ajustado menor a 0.05, obteniendo un grupo de 57 metabolitos con diferencias significativas entre pacientes control y caquéxicos. Observamos que los metabolitos Glucose, Quinolinate, Valine, N.N.Dimethylglycine, Adipate son los que presentan las diferencias más significativas entre grupos.

Se representaron los resultandos mediante un **Volcano plot**. Vemos que todos los metabolitos tienen un log2 fold change negativo, lo que confirma lo que habíamos observado en los boxplots anteriores: los metabolitos se encuentran a concentraciones mayores en caquéxicos que en pacientes control. El cálculo

del fold_change se realiza dividiendo el segundo grupo (control) / el de referencia (caquéxicos), por eso observamos valores negativos. Los metabolitos por encima del eje Y nos indican qu tienen un p-valor $<$ 0,05 y que son significativos. Cuanto más altos, menor p-valor presentan y por tanto, mayor signficancia estadística.

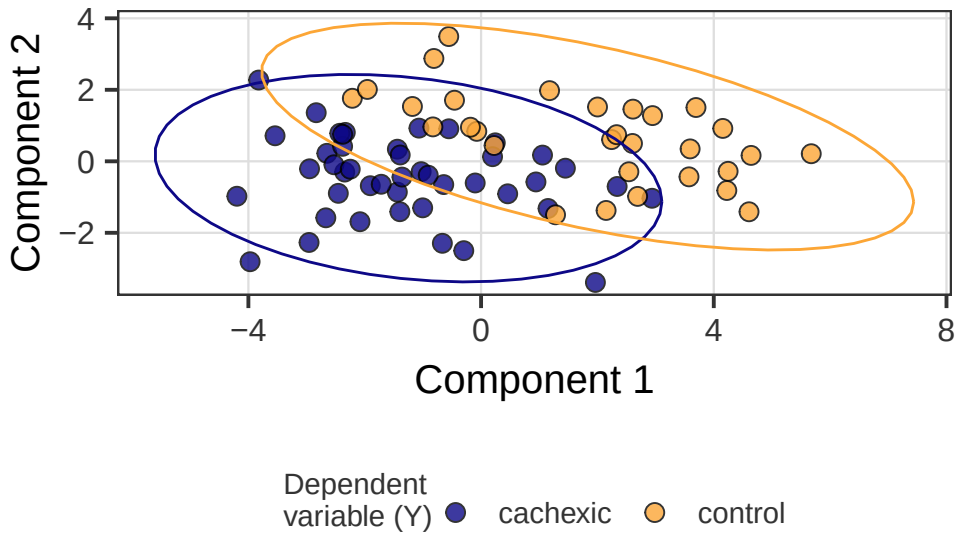
```
grafica_volcano
```

```
## Warning: ggrepel: 5 unlabeled data points (too many overlaps). Consider
## increasing max.overlaps
```



Como ya hemos evaluado cada metabolito por separado, ahora completaremos el análisis aplicando la técnica **sPLS-DA (Partial Least Squares Discriminant Analysis)**, que permite maximizar la separación entre los grupos conocidos. Se construyó el modelo con los metabolitos significativos identificados en el análisis univariante. En el gráfico de factores se aprecia una separación más definida entre los grupos en comparación con la observada en la PCA. Aunque todavía existe solapamiento entre grupos, este es mucho menor,por lo que hemos conseguido maximizar la separación de los grupos. Los metabolitos N.N.Dimethylglycine, Quinolinate, Leucine, Valine, Glutamine son los que mayor capacidad para discriminar entre ambos grupos presentan.

```
grafica_pls_factors
```



6. Discusión

La caquexia es un síndrome metabólico complejo asociado con una enfermedad subyacente, como puede ser el cáncer,y que está caracterizado por pérdida de masa muscular con o sin pérdida de masa grasa. Existe una escasez de datos sobre alteraciones globales del metaboloma en pacientes caquéxicos, por lo que la caracterización de perfiles metabolómicos puede resultar muy útil para comprender las vías metabólicas desreguladas en este síndrome. También podría servir para identificar biomarcadores diferenciales, que ayudarían a realizar un diagnóstico precoz y una terapia más dirigida y efectiva.

A lo largo del análisis se identificaron diferencias en los perfiles metabolómicos de los pacientes control y caquéxicos. A pesar de que en la PCA se observó solapamiento entre ambos grupos, sí se observaron dos agrupaciones claras, además de que entre la PC1 y PC2 se explica hasta un 60% de la varianza total, siendo cis.Aconitate, Glutamine, Alanine, Succinate, Histidine los que más contribuyen.

El análisis univariante mostró que 57 metabolitos presentan diferencias signficativas entre grupos, lo que nos confirma que existen perfiles metabolómicos diferentes entre pacientes caquéxicos y control. Tanto en los boxplots exploratorios como en el Volcano plot, se puede observar que los metabolitos presentan una concentración más elevada en pacientes caquéxicos que en pacientes control, siendo Glucose, Quinolinate, Valine, N.N.Dimethylglycine, Adipate los que presentan diferencias más significativas.

Mediante el análisis supervisado se consiguió maximizar la discriminación entre grupos, donde se observó que N.N.Dimethylglycine, Quinolinate, Leucine, Valine, Glutamine son los metabolitos con mayor poder discriminativo.

Antes de analizar los resultados es importante remarcar que trabajamos con muestras de orina, donde se recogen los productos finales de las diferentes vías metabólicas y por tanto, alteraciones en su concentración nos puede dar información relevante acerca del los cambios de estas rutas metabólicas. La concentración de estos metabolitos también puede variar entre individuos del mismo grupo dependiendo de la dieta, estado de hidratación o ingesta de medicamentos, cosa que explica la enorme variabilidad que hemos observado en los datos y la importancia de normalizarlos y quitar los valores atípicos.

Los resultados obtenidos concuerdan con los mecanismos fisiopatológicos de los pacientes caquéxicos descritos en la literatura. La presencia de elevadas concentraciones de aminoácidos en orina (como Valina, Alanina o Glutamina) podría explicarse por el estado hipercatabólico en el que se encuentran estos pacientes, donde hay pérdida de proteínas musculares entre otras patologías. De forma similar, altas concentraciones de adipato en orina pueden reflejar la activación de la lipólisis del tejido adiposo.

Elevación de quinolinato en orina, un metabolito intermedario en el catabolismo del triptófano asociado a respuestas inflamatorias y neurotoxicidad, nos podría indicar alteraciones de la respuesta inflamatoria y hormonal. Por último, N.N-Dimetilglicina está relacionada con el metabolismo de la colina, podría actuar como un marcador de estrés oxidativo, pues son alteraciones frecuentes en cáncer.

Es importante considerar algunas limitaciones del estudio. El tamaño muestral es relativamente bajo 77 aunque suficiente para un análisis exploratorio de los datos, por lo que se deberían validar los resultados obtenidos con grupos muestrales mayores y más controladas. Se deberían tener en cuenta otros factores que podrían tener un impacto en el perfil metabolómico como la edad, sexo, si los pacientes presentan cáncer y de qué tipo o algunas variables relacionadas con hábitos (como si es fumador o no). La caquexia es un síndrome multifactorial, por lo que es de vital importancia un análisis completo de todas las variables biológicas posibles.

Con los resultados obtenidos y la información de otros factores ya comentados, se podrían construir modelos predictivos que permitan clasificar a los pacientes caquéxicos en función de su perfil metabolómico. Estos modelos podrían entrenarse y testarse con muestras de nuevos pacientes, para intentar mejorar la identificación de este grupo de pacientes.

7. Conclusiones

A través del análisis exploratorio del conjunto de datos de 77 muestras de orina de pacientes control y caquéxicos, se ha podido identificar perfiles metabolómicos diferentes entre grupos. Mediante el análisis univariante,se detectaron 57 metabolitos con diferencias signficativas en pacientes control y caquéxicos y se observó esta separación entre grupos mediante PCA. Utilizando técnicas supervisadas como sPLS-DA, se maximizó la discriminación entre pacientes.

Los metabolitos N.N.dimetilglicina, Quinolinato y Valina no solo presentan diferencias significativas entre grupos, sino que también tiene una gran capacidad discriminativa, por lo que podrían ser buenos biomarcadores de caquexia en muestras de orina. Los resultados observados en la exploración de los datos están alineados con las alteraciones metabólicas observadas en pacientes con caquexia, como el estado hipercatabólico, aumento de la lipólisis y la respuesta inflamatoria y hormonal alterada.

Aunque los resultados son prometedores, se debería recoger información de otros factores clínicos que podrían afectar al perfil metabolómico, además de aumentar la cohorte para mejorar la robustez estadística del análisis.

8. Referencias

Este es el enlace al repositorio GitHub, donde encontrareis el código utilizado para abordar el análisis, así como el dataset y el summarizedExperiment: https://github.com/paulagaliana8/Galiana-Haro-Paula-PEC1.git