Análisis de datos ómicos (M0-157) Primera prueba de evaluación continua

Informe

Gabriel Peña Peña

2025-04-02

Contents

Resumen

Este trabajo realizado en esta PEC presenta un análisis exploratorio de datos de metabolómica asociados a la caquexia, una condición clínica caracterizada por la pérdida de masa muscular. Se utilizó el conjunto de datos 2024-Cachexia, que incluye concentraciones de metabolitos en pacientes con y sin pérdida muscular. Los datos fueron organizados mediante la clase SummarizedExperiment, lo que permitió una gestión estructurada de las matrices de expresión y metadatos. Se aplicaron técnicas estadísticas básicas para explorar la variabilidad y la correlación entre metabolitos, así como pruebas de ANOVA para detectar diferencias significativas entre grupos. Entre los hallazgos más relevantes se identificaron metabolitos como valina, leucina y pirroglutamato, cuyas concentraciones difieren significativamente entre condiciones clínicas. A pesar de las limitaciones

metodológicas, los resultados ofrecen una primera aproximación útil para futuras investigaciones sobre este dataset.

Objetivos

El objetivo de este trabajo fue llevar a cabo un análisis exploratorio de datos de metabolómica utilizando herramientas del entorno Bioconductor, con el fin de familiarizarse con el manejo de datos ómicos y aplicar métodos estadísticos básicos en un contexto biológico real.

Para ello, se trabaja con el conjunto de datos "2024-Cachexia", que contiene concentraciones de metabolitos en pacientes con y sin pérdida muscular. A partir de estos datos, se busca identificar patrones de variabilidad, correlaciones y diferencias entre grupos que puedan estar asociadas a la condición clínica.

Los objetivos específicos del análisis son:

- Cargar y organizar el conjunto de datos en un objeto de clase SummarizedExperiment, separando correctamente los datos cuantitativos y los metadatos asociados.
- Realizar un análisis descriptivo y exploratorio de las concentraciones de metabolitos en las distintas muestras.
- Evaluar la variabilidad de los metabolitos entre pacientes mediante el cálculo del coeficiente de variación.
- Explorar la correlación entre metabolitos para identificar posibles agrupaciones o comportamientos conjuntos.
- Aplicar un análisis ANOVA para detectar diferencias significativas en la concentración de metabolitos entre los grupos con y sin pérdida muscular.
- Representar visualmente los resultados obtenidos mediante gráficos como histogramas, boxplots, heatmaps y barplots.

Hipótesis

La condición de caquexia está asociada a alteraciones en el perfil metabólico de los pacientes, reflejadas en diferencias significativas en la concentración de ciertos metabolitos en comparación con individuos sin pérdida muscular.

Métodos

Origen y naturaleza de los datos

El conjunto de datos utilizado en este análisis proviene del repositorio de GitHub del proyecto nutrimetabolomics/metaboData, específicamente del dataset denominado 2024-Cachexia. Este conjunto incluye datos de metabolómica obtenidos de 77 pacientes, de los cuales 47 presentan caquexia, la caul es una condición clínica caracterizada por pérdida severa de masa muscular y 30 corresponden a controles. El archivo contiene concentraciones relativas de 63 metabolitos, así como información clínica adicional, entre la que destaca la variable MuscleLoss, utilizada como referencia para clasificar a los pacientes.

Metodología empleada

El análisis comenzó con la carga de los datos desde un archivo CSV. Se separaron los metadatos relevantes (como el identificador del paciente y su clasificación clínica) de la matriz de datos que contiene las concentraciones de metabolitos. A continuación, la información fue organizada utilizando la clase SummarizedExperiment del paquete Bioconductor, una estructura diseñada específicamente para almacenar datos ómicos junto con sus metadatos de forma eficiente y estructurada.

Esta clase permitió mantener, en un solo objeto, tanto la matriz de concentraciones de metabolitos como la información clínica de los pacientes, facilitando así el acceso a los datos y su análisis posterior.

Herramientas estadísticas y bioinformáticas

Para llevar a cabo el análisis se utilizaron principalmente herramientas del entorno R, incluyendo los siguientes paquetes:

- SummarizedExperiment: para estructurar y gestionar los datos.
- Funciones base de R (summary, apply, cor, aov, etc.): para realizar cálculos estadísticos y análisis exploratorios.
- Visualización: se emplearon funciones de R base (boxplot, hist, heatmap, barplot) para representar los resultados de forma gráfica y facilitar su interpretación.

Los análisis realizados incluyeron:

- Análisis descriptivo: Se obtuvieron estadísticas resumen de las concentraciones de metabolitos y se visualizaron sus distribuciones con histogramas y boxplots.
- Análisis de correlación: Se calculó una matriz de correlación para detectar relaciones entre metabolitos, con el objetivo de identificar patrones conjuntos o agrupamientos.
- Análisis de variabilidad: Se utilizó el coeficiente de variación (CV) para evaluar la variabilidad relativa de cada metabolito entre pacientes, identificando aquellos con mayor dispersión.
- Análisis de diferencias entre grupos: Se aplicó un modelo ANOVA para cada metabolito, con el fin de detectar si existían diferencias estadísticamente significativas en sus concentraciones entre pacientes con y sin pérdida muscular.

Procedimiento general de análisis

1. Carga y preprocesamiento de datos: lectura del archivo CSV, separación de metadatos y concentración de metabolitos.
2. Estructuración de datos: construcción del objeto SummarizedExperiment con los datos organizados para análisis.
3. Exploración inicial: visualización general de los datos, revisión de distribuciones y comprobación de la variable de clasificación (MuscleLoss).
4. Análisis estadístico: cálculo del CV y aplicación de ANOVA para evaluar la relación entre los metabolitos y la pérdida muscular.
5. Visualización e interpretación: uso de gráficos (boxplots, heatmaps, barplots) para mostrar los resultados y facilitar su análisis biológico.

Resultados

Análisis exploratorio de los datos

Una vez creado el objeto SummarizedExperiment, se procedió a la inspección inicial de su contenido. Este objeto incluye 63 metabolitos (filas) cuantificados en 77 pacientes (columnas), junto con una variable categórica llamada MuscleLoss, que clasifica a los individuos en dos grupos: “cachexic” (con pérdida muscular) y “control” (sin pérdida muscular). La estructura interna del objeto se verificó correctamente, asegurando la coherencia entre los datos numéricos y los metadatos asociados.

También se verificó el contenido de los metadatos, confirmando que cada muestra cuenta con una etiqueta asociada a su condición clínica. Del total de pacientes, 47 (61.04%) fueron clasificados como cachexic y 30 (38.96%) como control, lo que proporciona un tamaño de muestra razonable para realizar comparaciones entre grupos (figura en anexo).

La matriz de concentraciones de metabolitos mostró una organización adecuada, con valores que varían ampliamente entre compuestos. Por ejemplo, el metabolito X1.6.Anhydro.beta.D.glucose presentó concentraciones desde 4.71 hasta 685.40, con una media de 105.63, mientras que X3.Aminoisobutyrate alcanzó un valor máximo de 1480.30, indicando que algunos metabolitos tienen rangos de variación muy amplios. Este comportamiento fue confirmado por el resumen estadístico de los datos, que permitió identificar metabolitos con mayor dispersión y potencial presencia de valores atípicos (figura en anexo).

Distribución de la concentración de metabolitos

Como primer paso exploratorio, se analizó la distribución de concentraciones del metabolito X1.6.Anhydro.beta.D.glucose, que corresponde a la primera fila del conjunto de datos. Aunque no fue seleccionado por su significancia estadística ni por su variabilidad, se utilizó como ejemplo inicial para visualizar cómo se comporta un metabolito individual en la muestra, permitiendo interpretar con claridad los tipos de gráficos empleados.

El histograma mostró que la mayoría de las muestras (alrededor del 50%) se encuentran en el rango de 0 a 100, seguido por una disminución de frecuencia en los rangos superiores. Solo unas pocas muestras presentan concentraciones por encima de los 300, lo que sugiere que este metabolito tiende a concentraciones bajas en la mayoría de los pacientes, con algunos casos atípicos.

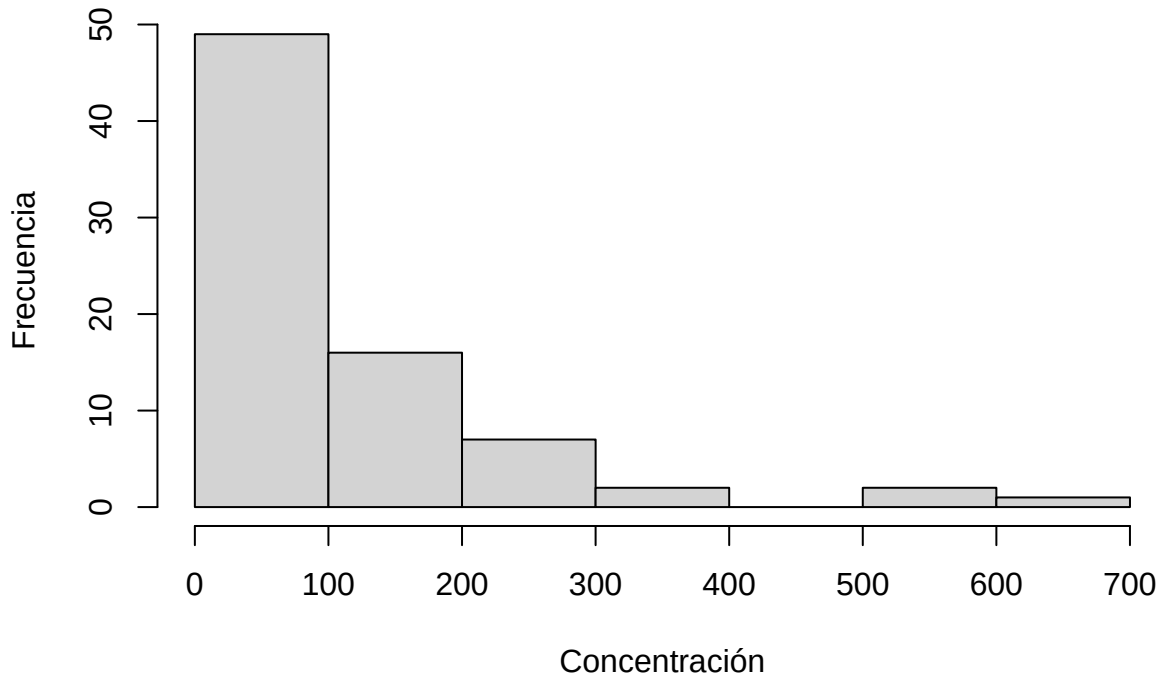

Figure 1: Distribución de las concentraciones del primer metabolito en el conjunto de datos.

Al comparar este metabolito por condición (MuscleLoss), se observaron diferencias en la dispersión de los valores. En el grupo control, la caja del boxplot fue estrecha, con la mayoría de los valores concentrados entre 50 y 80, y unos pocos valores atípicos que alcanzan hasta los 200, y un caso cercano a 520. En contraste, en el grupo caquéxico, la caja fue más amplia (aproximadamente entre 70 y 180), y se observaron valores extremos por encima de 600, incluso cercanos a 700. Estos resultados sugieren una mayor variabilidad en la concentración de este metabolito entre los pacientes con caquexia.

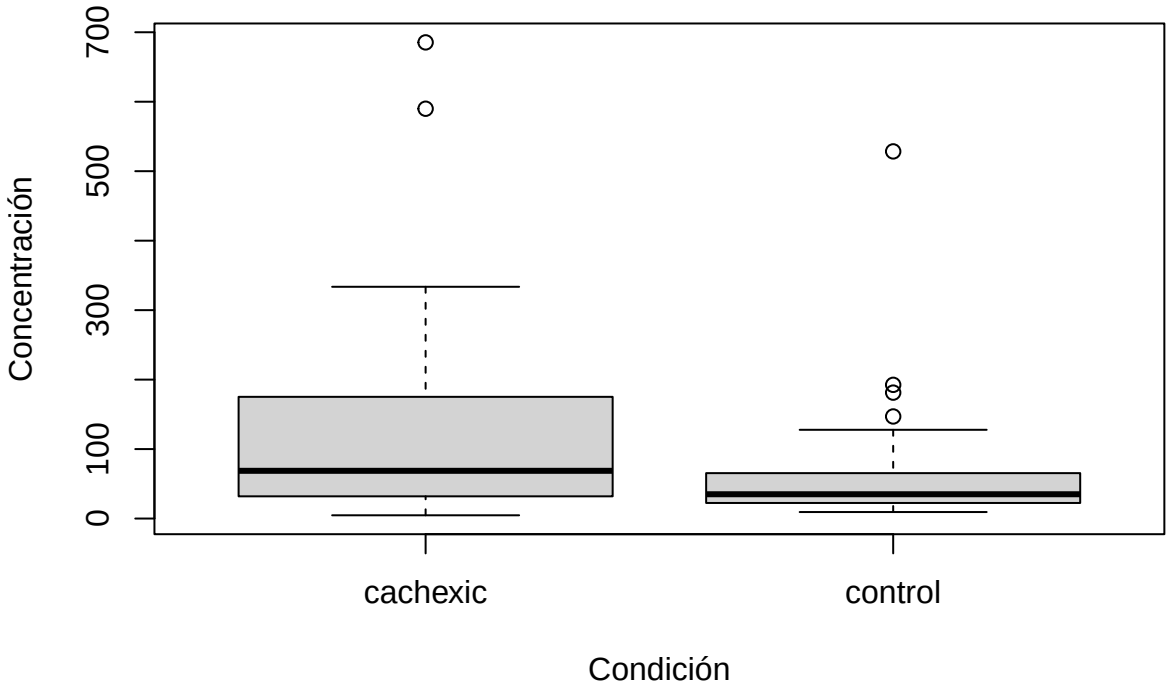

Figure 2: Comparación de la concentración del primer metabolito entre pacientes con y sin caquexia.

En cuanto a la distribución global de metabolitos, se observaron perfiles diferentes, algunos pacientes presentaron valores relativamente bajos y peuqeños, mientras que otros mostraron concentraciones más altas, con bigotes que se extendían por encima de 1000 o 1500. También se observaron casos con distribuciones muy reducidas. Esta representación proporciona una visión general de la variabilidad entre pacientes, aunque no permite analizar metabolitos específicos, sino más bien la dispersión interna por muestra (figura en anexo).

Análisis de correlación entre metabolitos

A partir del cálculo de la matriz de correlación y su representación gráfica en un mapa de calor, se observó que ciertos metabolitos mostraron correlaciones positivas intensas, como por ejemplo la hipoxantina con valina, asparagina, X3-hidroxibutirato e histidina. Por otro lado, metabolitos como xilosa, pantotenato, sacarosa o carnitina mostraron correlaciones débiles o nulas respecto a esos mismos compuestos.

La interpretación del heatmap se realizó en base a su escala de colores: los tonos más cálidos (rojos/naranjas) indican correlaciones positivas, los tonos fríos (azules) corresponden a correlaciones negativas, y los colores claros o neutros (amarillo pálido o blanco) reflejan ausencia o baja correlación. En este caso, no se observaron correlaciones negativas marcadas (figura en anexo).

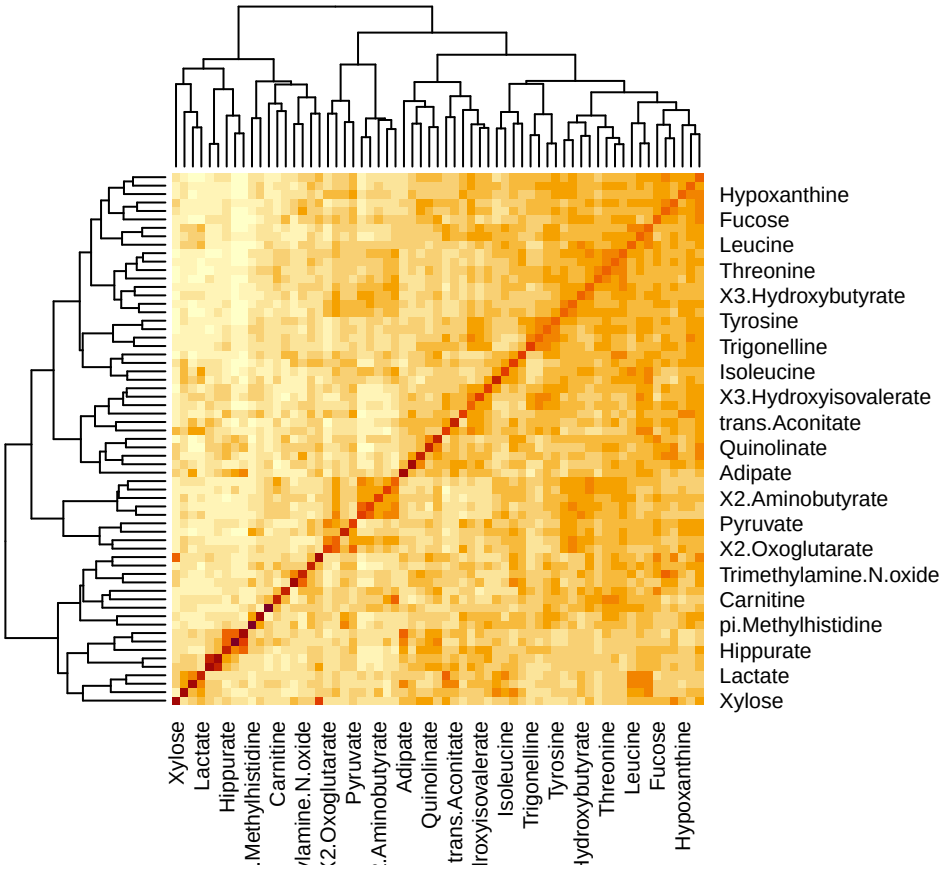

Figure 3: Mapa de calor de la correlación entre metabolitos. Los colores indican el grado y signo de la correlación: tonos rojos intensos representan correlaciones positivas fuertes, tonos azules intensos indican correlaciones negativas fuertes, y los colores pálidos o intermedios (cerca del blanco) corresponden a correlaciones débiles o nulas

Análisis de variabilidad de metabolitos

A partir del calculo del coeficiente de variación para cada metabolito, se seleccionaron los 20 metabolitos más variables y se representaron en un mapa de calor para visualizar sus patrones de concentración.

En este, se incluyeron metabolitos como succinate, tartrato, xilosa, lactato, oxoglutarato y O-acetilcarnitina, entre otros. En general, la mayoría de las celdas del gráfico presentaron un color amarillo pálido, lo que indica valores cercanos al promedio tras la estandarización. Sin embargo, se observaron puntos aislados, en uno o dos pacientes por metabolito, donde se indicaban valores mucho más altos o bajos que la media. Este patrón sugiere que, si bien estos metabolitos presentan alta variabilidad global, esta puede estar influida por casos atípicos concretos más que por una dispersión homogénea en todo el grupo.

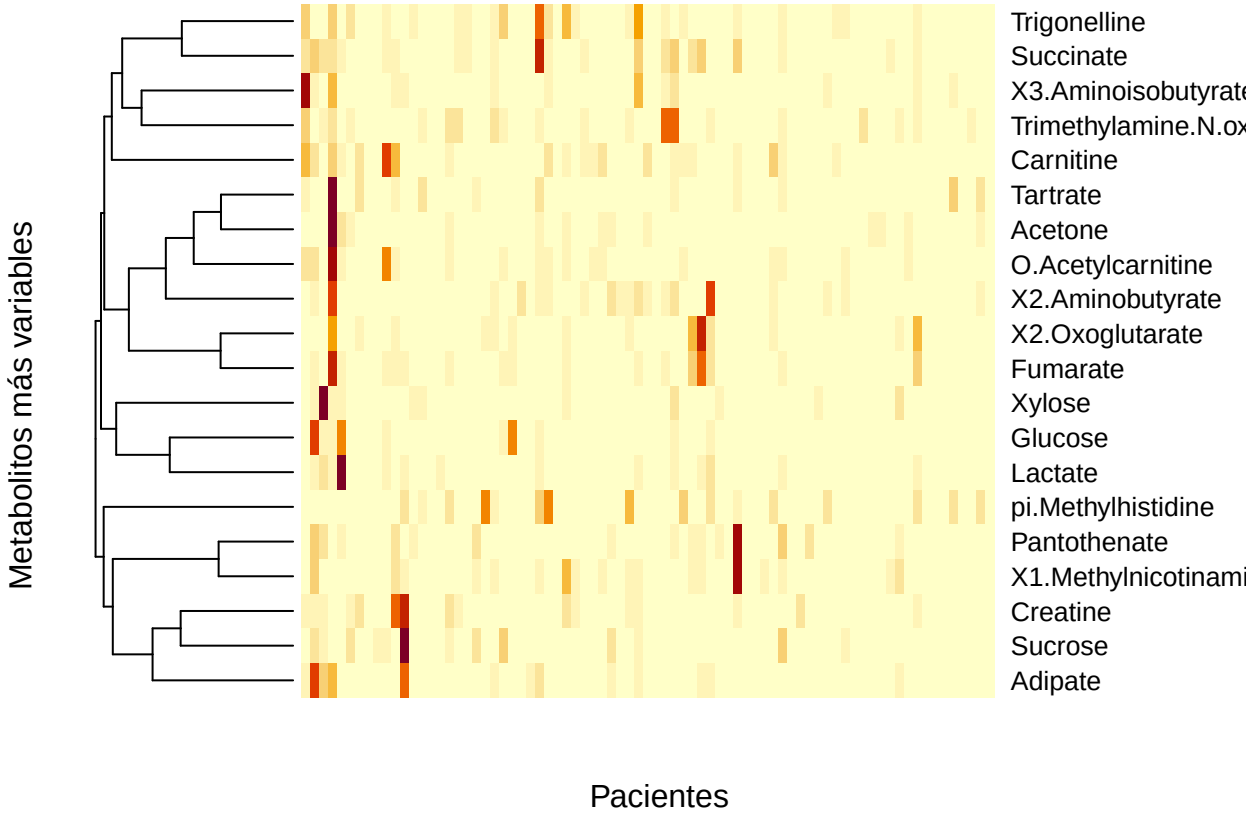

Figure 4: Mapa de calor de los 20 metabolitos con mayor coeficiente de variación (CV), considerando aquellos con CV 1.39.

Análisis ANOVA para comparar grupos

Con el objetivo de evaluar la hipótesis planteada (que la condición de caquexia se asocia a cambios en los perfiles de concentración de metabolitos) se aplicó un análisis ANOVA para cada uno de los metabolitos del conjunto de datos.

Los p-valores resultantes fueron ordenados para identificar aquellos metabolitos con diferencias estadísticamente significativas entre los grupos de pacientes con y sin pérdida muscular. A continuación se muestran los diez metabolitos con menor p-valor:

```
##            Metabolito      p_valor
## 1          Quinolinate 0.0001185108
## 2              Valine 0.0001394238
```

```
## 3  N.N.Dimethylglycine 0.0001554346
## 4               Leucine 0.0002695046
## 5         Dimethylamine 0.0004460069
## 6         Pyroglutamate 0.0004845363
## 7            Creatinine 0.0005129808
## 8             Glutamine 0.0010607678
## 9               Alanine 0.0011788609
## 10  X3.Hydroxybutyrate 0.0011853671
```

Estos resultados sugieren que varios aminoácidos (como valina, leucina, alanina, glutamina) y compuestos relacionados con el metabolismo energético y del nitrógeno presentan diferencias significativas entre ambos grupos. Si bien este análisis no permite establecer relaciones causales, sí indica una posible asociación entre la condición clínica de caquexia y alteraciones en el metabolismo de ciertos compuestos.

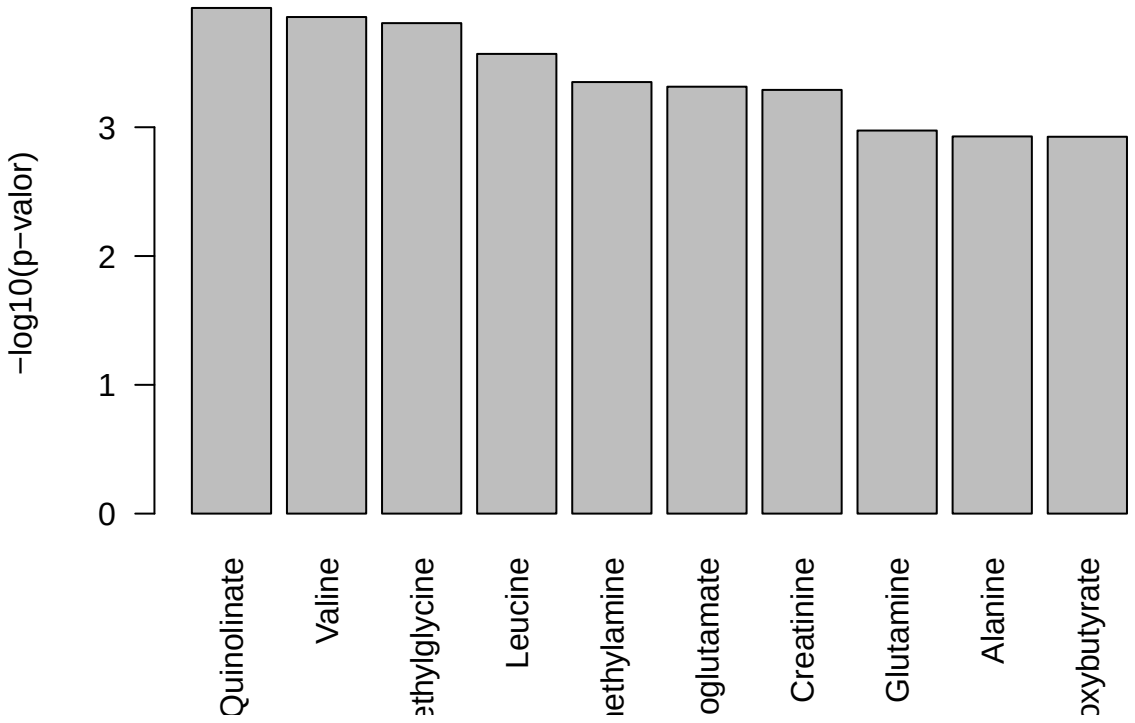

Figure 5: Metabolitos con mayor significancia estadística según ANOVA. Se muestra el valor -log10(p) de los diez compuestos con menor p-valor

Discusión

El trabajo realizado en esta PEC permitió llevar a cabo una exploración inicial de un conjunto de datos de metabolómica en el contexto de la caquexia, una condición caracterizada por la pérdida de masa muscular y peso corporal, frecuente en enfermedades crónicas avanzadas como el cáncer. Aunque no se profundizó en los mecanismos biológicos específicos de esta enfermedad, el análisis realizado permite una primera aproximación a las posibles alteraciones metabólicas asociadas a esta condición clínica, que fue objeto de estudio.

Uno de los hallazgos más destacados fue la identificación de diferencias significativas en la concentración

de varios metabolitos entre los grupos caquéxico y control, como por ejemplo la valina, la leucina y el pirroglutamato. Muchos de estos compuestos son aminoácidos o moléculas relacionadas con el metabolismo energético y del nitrógeno, por lo que podrían desempeñar un papel relevante en la fisiopatología de la caquexia, o bien reflejar adaptaciones metabólicas del organismo frente al desgaste muscular. Aunque los resultados son preliminares, pueden dar un primer aproximamiento sobre la enfermedad.

Desde el punto de vista metodológico, la utilización del objeto SummarizedExperiment permitió organizar los datos y metadatos de forma estructurada, facilitando significativamente tanto los análisis exploratorios como los estadísticos. Además, las visualizaciones mediante mapas de calor, boxplots y análisis de correlación ayudaron a interpretar patrones de variación y relaciones entre metabolitos de forma clara y comprensible.

Este análisis presenta, sin embargo, algunas limitaciones importantes. No se consideraron posibles variables de confusión (como edad, sexo u otras condiciones clínicas), que podrían haber influido en los niveles de ciertos metabolitos. Asimismo, la interpretación biológica fue limitada: ya que no se incorporaron análisis funcionales o de enriquecimiento debido a la complejidad metodológica que implican. Además, los resultados del ANOVA no fueron corregidos por pruebas múltiples, lo que podría incrementar el riesgo de falsos positivos.

Desde una perspectiva técnica, algunas figuras presentaron dificultades de visualización debido a la densidad de datos o restricciones de formato en el informe. Por este motivo, se optó por trasladarlas a un anexo, conservando así su valor informativo sin comprometer la legibilidad del documento principal.

En conjunto, el trabajo realizado permitió cumplir con los objetivos planteados: explorar un conjunto de datos reales de metabolómica, aplicar herramientas bioinformáticas del entorno Bioconductor y extraer conclusiones preliminares sobre la asociación entre perfiles metabólicos y una condición clínica concreta como la caquexia. Quedan abiertas líneas para análisis más avanzados que permitan esclarecer posibles mecanismos con mayor solidez y profundidad.

Conclusiones

El análisis exploratorio realizado sobre el conjunto de datos de metabolómica del estudio 2024-Cachexia permitió identificar diferencias significativas en la concentración de varios metabolitos entre los grupos caquéxico y control. Entre los compuestos más destacados se encuentran la valina, la leucina y el pirroglutamato, relacionados con el metabolismo energético y del nitrógeno, lo que sugiere una posible implicación en la fisiopatología de la caquexia. Asimismo, se observaron patrones de variabilidad y correlaciones relevantes entre metabolitos, lo que refuerza la utilidad de este tipo de análisis como herramienta inicial de exploración biológica.

Desde el punto de vista metodológico, el uso de la clase SummarizedExperiment facilitó una adecuada organización de los datos y metadatos, permitiendo desarrollar visualizaciones e inferencias estadísticas de manera estructurada, facil y reproducible.

Aunque si bien se cumplieron los objetivos propuestos, se reconocen ciertas limitaciones, como la falta de ajuste por variables clínicas adicionales y la ausencia de corrección por comparaciones múltiples, lo que deberia ser considerado al interpretar los resultados.

A futuro, sería recomendable ampliar este análisis incorporando variables clínicas relevantes, aplicar metodologías multivariantes más robustas y profundizar en el análisis funcional de rutas metabólicas alteradas, con el fin de comprender mejor los mecanismos implicados en la caquexia y su posible valor diagnóstico o terapéutico.

Referencias

Referencias El código utilizado para llevar a cabo el análisis se encuentra disponible en el siguiente repositorio de GitHub, con control de versiones mediante Git y comentarios explicativos en el archivo .Rmd:

https://github.com/Gaban1998/Pena-Pena-Gabriel-PEC1

Anexo: Figuras complementarias

Se incluyen algunas figuras cuya visualización dentro del cuerpo principal del informe presentaba limitaciones de formato o espacio.

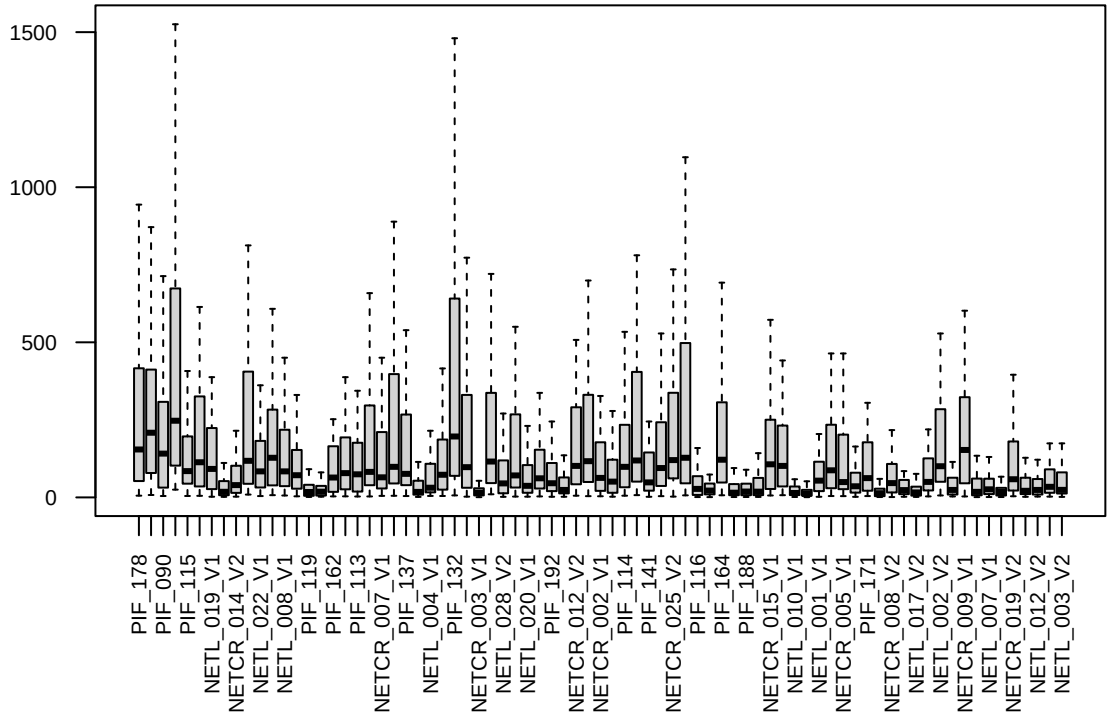

Figure 6: Distribución global de concentraciones de metabolitos por paciente. Se aprecian perfiles metabólicos heterogéneos, con algunas muestras que presentan valores notablemente más altos que otras.

Table 1: Distribución porcentual de la variable MuscleLoss.

Condición	Porcentaje
cachexic	61.04
control	38.96

Table 2: Resumen estadístico de concentraciones para los primeros tres metabolitos.

X1.6.Anhydro.beta.D.glucose	X1.Methylnicotinamide	X2.Aminobutyrate
Min. : 4.71	Min. : 6.42	Min. : 1.28
1st Qu.: 28.79	1st Qu.: 15.80	1st Qu.: 5.26
Median : 45.60	Median : 36.60	Median : 10.49
Mean :105.63	Mean : 71.57	Mean : 18.16
3rd Qu.:141.17	3rd Qu.: 73.70	3rd Qu.: 19.49
Max. :685.40	Max. :1032.77	Max. :172.43