Informe PEC1: Análisis del dataset de metabolómica 2023-CIMCBTutorial

Sara Suárez Hernández

2025-04-02

Contents

Todo el material necesario para reproducir esta PEC se encuentra en el siguiente repositorio: https://github.com/sara-suarez93/Suarez-Hernandez-Sara-PEC1

1. Selección del dataset

He elegido el dataset *2023-CIMCBTutorial* por interés en comparar muestras de cáncer vs sanas y por disponer de un tutorial [1] para Python (lenguaje que me gustaría aplicar para datos ómicos aunque para la PEC haya elegido R).

En este estudio, analizaron con espectroscopia de resonancia nuclear magnética 1H (1H-NMR) muestras de orina de 43 pacientes con cáncer de estómago (GC), 40 con tumores benignos (BN), y 40 participantes sanos (HE) emparejados (supongo que por edad y sexo) [2].

2. Objecto summarizedExperiment

Para crear el objeto summarizedExperiment, primero descargué los datos directamente del repositorio de Github (https://github.com/nutrimetabolomics/metaboData/blob/main/Datasets/2023-CIMCBTutorial/GastricCancer_NMR.xlsx). Es una documento de Excel compuesto de dos hojas:

```
## [1] "Data" "Peak"
```

- Data: incluye 140 filas correspondientes a cada una de las muestras analizadas y 153 columnas.
  - Las primeras 5 columnas incluyen información de las muestras.
  - Las siguientes 149 columnas se corresponden a los 149 metabolitos estudiados.

[1] Tutorial repositorio
[2] Metabolomics Workbench

- Peak: incluye 149 filas correspondientes a los metabolitos estudiados y 5 columnas con infromación sobre los metabolitos.

Para construir el objeto *SummarizedExperiment* he definido 1 matriz (data.metabolomics) con la expresión de cada metabolito (columnas) para cada una de las muestras (filas) y 2 dataFrames (metadatos_muestras y metadatos_metabolitos) que incluyen información de las muestras (en columnas) y de los metabolitos (en filas) respectivamente.

Este es el objeto:

```
## class: SummarizedExperiment
## dim: 149 140
## metadata(0):
## assays(1): metabolomics
## rownames(149): M1 M2 ... M148 M149
## rowData names(3): nombre_metabolito Perc_missing QC_RSD
## colnames(140): sample_1 sample_2 ... sample_139 sample_140
## colData names(2): SampleType Class
```

Diferencias principales con ExpressionSet

La diferencia principal es que mientras el objeto *ExpressionSet* es específico para un *output* de un experimento concreto (i.e., resultados crudos de RNAseq) para un conjunto de muestras, en el que cada fila se corresponde con un *transcript*; el objeto *SummarizedExperiment* permite englobar varios *outputs* del mismo experimento (i.e., datos de expresión crudos y normalizados) o resumir información para cada uno de los *transcripts* con el objeto RangedSummarizedExperiment (i.e., rango de exones de cada *transcript*). Es decir, *SummarizedExperiment* permite almacenar más información sobre las observaciones de un experimento [3].

La manera de construir y trabajar con ambos objetos es similar, aunque se diferencian en los métodos o funciones necesarias para acceder a los datos y en los nombres de los *slots*. La siguiente tabla resume algunas de las diferencias en los *slots* de cada objeto:

Datos en slot	ExpressionSet	SummarizedExperiment
Datos (observaciones) de las muestras	assayData (matriz)	assays (lista de matrices)
Información muestras	phenoData	colData
Información datos (observaciones)	featureData	rowData

3. Análisis exploratorio de los datos

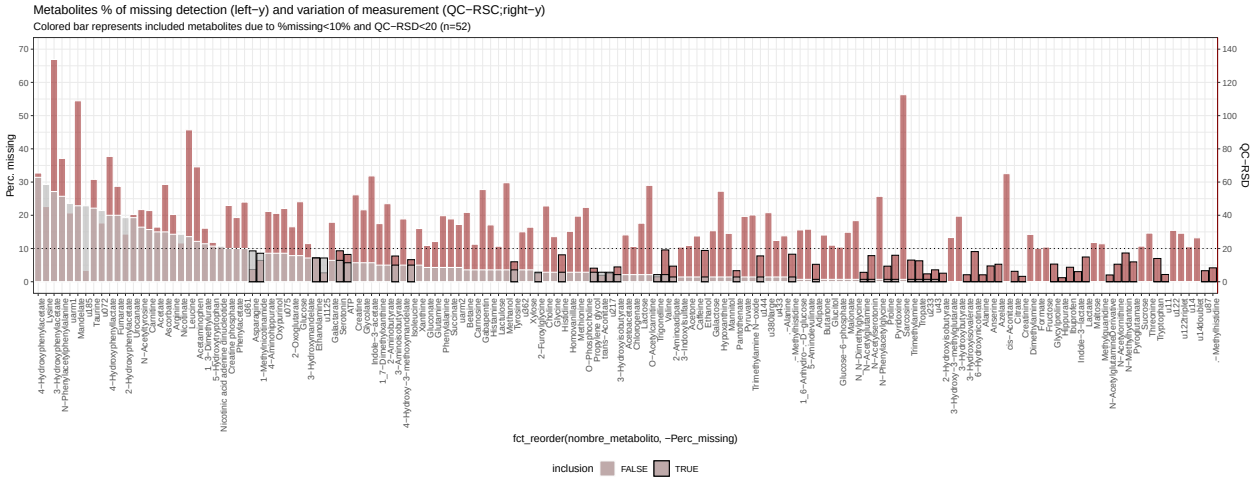
En resumen, estos son los pasos seguidos para hacer un análisis exploratorio de los datos:

1. Revisar datos faltantes (en cuántas muestra falta información de ciertos metabolitos)
2. Explorar la distribución de los datos de los metabolitos (en las muestras)
3. Imputar datos faltantes y normalizar los datos
4. Seleccionar el dataset para análisis (excluyendo metabolitos con alto porcentaje de datos faltantes en las muestras y las muestras que son outliers)
5. Análisis

[3] Summarized Experiment

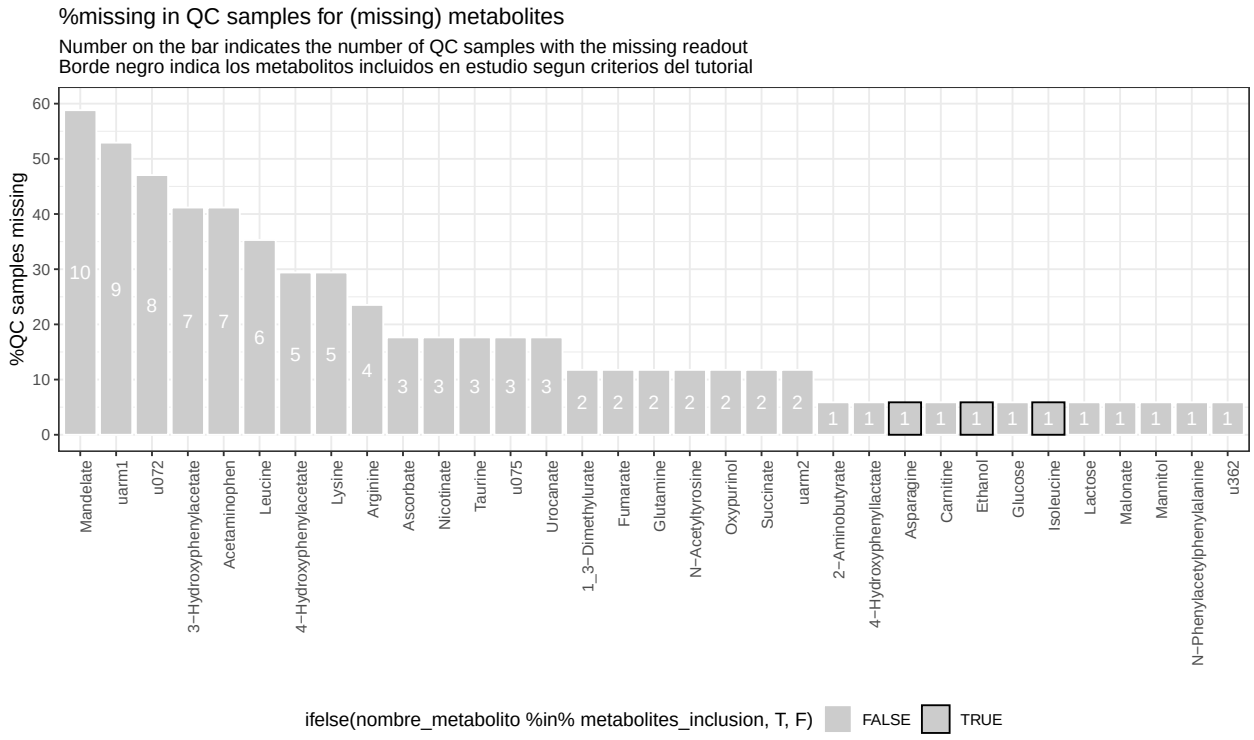
3.1 Datos faltantes

Como se menciona en el tutorial [4], los metadatos incluyen dos variables (Perc_missing y QC_RSD) que utilizan como criterios de inclusión. En concreto, seleccionan los metabolitos que se han detectado en más del 90% de las muestras (Perc_missing<10%) y cuyas medidas tienen poca variación (QC_RSD<20). Así, nos aseguramos que los metabolitos que analicemos se encuentran comunmente en este tipo de muestras con unos valores consistentes, de tal manera que las conclusiones no se deben a metabolitos que se encuentran en pocas muestras y o con medidas de poca calidad.



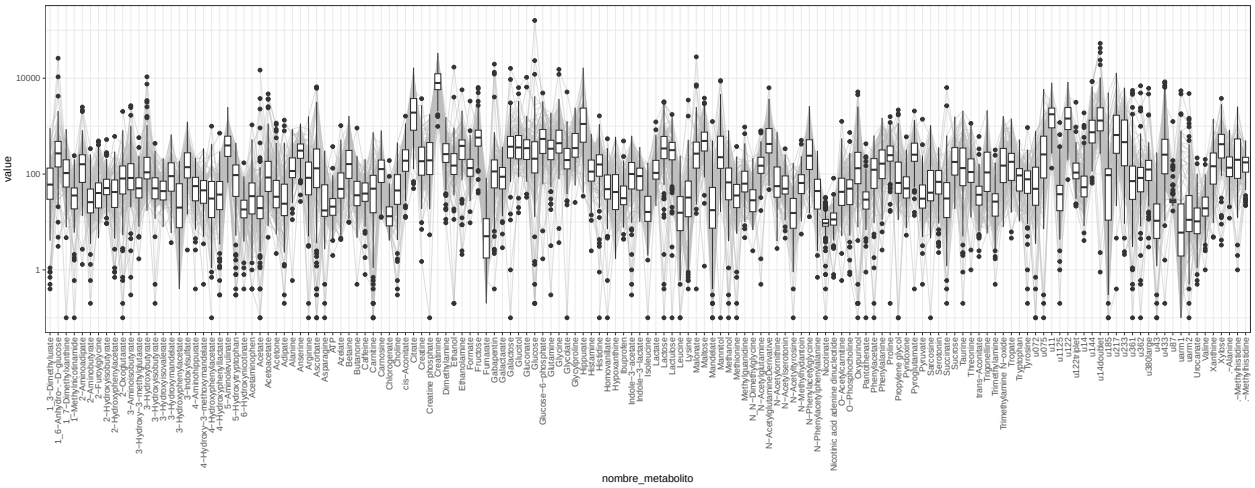
Personalmente, al no haber trabajado nunca con datos ómicos, mi primera idea fue explorar los metabolitos con datos faltantes en las muestras de control (QC). Desconozco que tipo de muestras son las QC, pero entiendo que proporcionan cierta seguridad de detectar los metabolitos analizados para p. ej., descartar que no se detecten en las muestras experimentales por algún fallo técnico. Los metabolitos en la siguiente tabla son los que no se detectaron en el 100% de las muestras control, y por tanto, puede ser que no se detecten en las muestras experimentales por causas diferentes a la hipótesis de estudio. Sin embargo, vemos que 3 de estos metabolitos fueron incluídos en el análisis del tutorial. Entiendo que al solo faltar en una de las muestras control, estar presente en un 90% o más de todas las muestras analizadas y tener medidas consistentes (QC_RSD<20) son argumentos suficientes para incluirlos.

[4] Tutorial repositorio



3.2 Distribución metabolitos en muestras

A continuación, grafiqué los datos de todos los metabolitos en todas las muestas con boxplots para tener una idea de las escalas de medida y de las distribución de los datos. Añadí líneas para tener una idea rápida de si las medidas *outlier* (puntos) provienen de las mismas muestras o no.



La distribución de los datos naturales está sesgada a la derecha, es decir, que hay pocas muestras con valores extremadamente altos que no permiten observar la distribución de la mayoría de muestras con valores similares y más bajos. (No muestro el gráfico). Para corregir este sesgo y poder observar todos los datos, he aplicado una transformación log10 al eje-y. Vemos que para distintos metabolitos hay varios órdenes de magnitud de diferencia, por lo que los datos necesitan ser normalizados antes de hacer un análisis multivariante. También vemos que los outliers parecen venir de las mismas muestras, pero los analizaremos más adelante.

3.3 Imputación y normalización de los datos

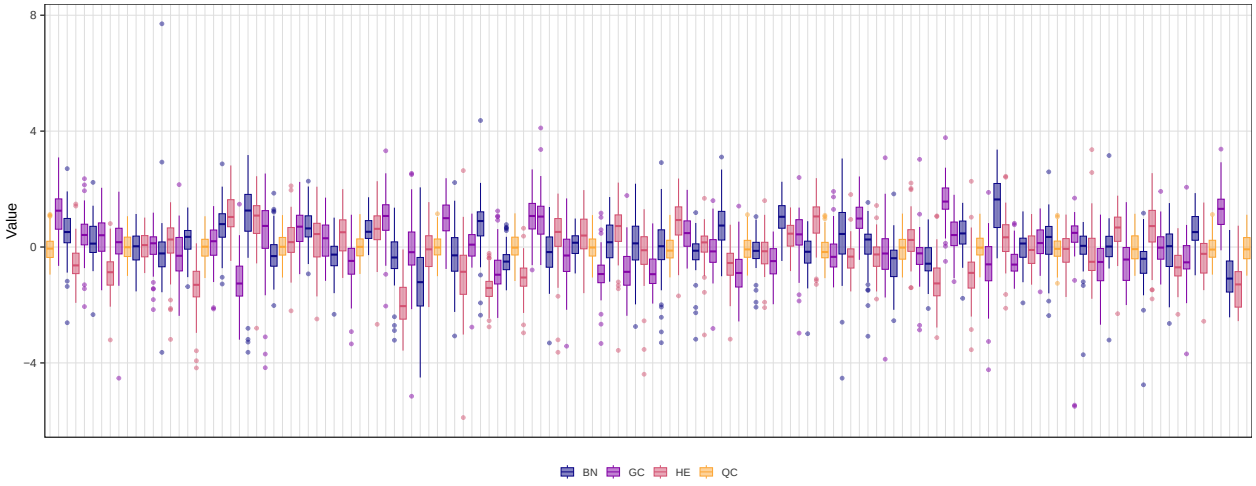
Nota: a partir de este apartado, utilicé el paquete POMA de Bioconductor para analizar directamente los datos contenidos en SummarizedExperiment (ver script, [5] *).*

Imputé los datos siguiendo el procedimiento de *k-nearest neighbours (K-nn)*, estableciendo el número de vecinos en k=3 como en el Tutorial. Este método asigna la media de los 3 valores más cercanos (en las tres muestras más similares).

A continuación, normalicé los datos con la transformación log. También exploré las log-pareto (ver script). Hice uso de la característica del objeto SummarizedExperiment para alojar los datos transformados de distintas maneras dentro del mismo objeto. Este es el objeto incluyendo los datos naturales (raw_data), inputados, y con las dos transformaciones exploradas. Tiene una dimensión de 52 metabolitos (que pasaron los criterios de inclusión de datos faltantes) en las 140 muestras.

```
## class: SummarizedExperiment
## dim: 52 140
## metadata(0):
## assays(4): metabolomics imputed_data normalized_log
##   normalized_logpareto
## rownames(52): M4 M5 ... M148 M149
## rowData names(3): nombre_metabolito Perc_missing QC_RSD
## colnames(140): sample_1 sample_2 ... sample_139 sample_140
## colData names(2): SampleType Class
```

Y así resultan los datos completos y normalizados comparando muestras de control (QC) y experimentales (HE: Healthy, BN: Benign tumor; GC: Gastric cancer):



Vemos que los QC tienen valores y distribuciones similares para todos los metabolitos (eje-x), mientras que varían en las muestras. Como *outliers*, destaca un metabolito con valores positivos extremos en una muestra BN y otro con valores negativos extremos en una muestra HE (muy bajos o con valores <1 en la escala original que resultan en log<0). Los outliers los identificamos en el siguiente apartado.

3.4 Criterios de exclusión de metabolitos y muestras

Los criterios de exclusión de metabolitos son los comentados en el apartado 3.1 (ausentes en más de un 10% de las muestras y con una variación de las medidas inferior a 20).

- Al aplicar estos filtros, los 149 metabolitos iniciales se reducen a 52.

[5] POMA Workflow

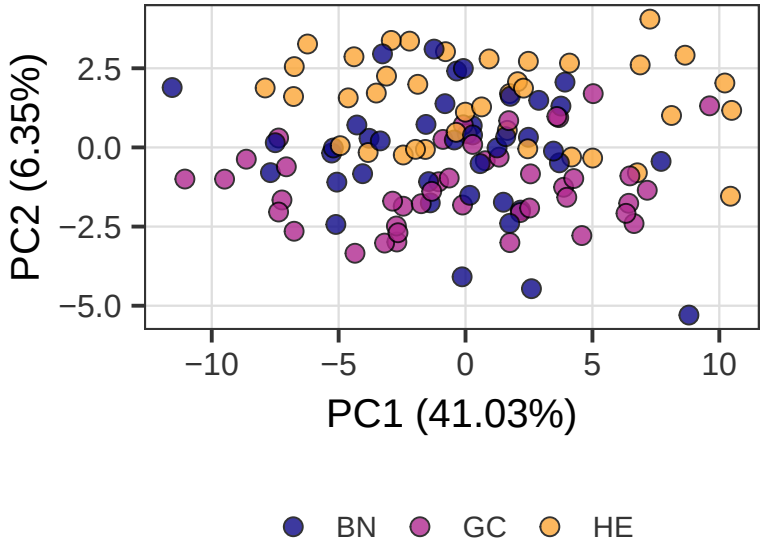
Para las muestras, utilicé la función PomaOutliers del paquete POMA para identificar los *outliers* multivariantes, es decir, que los valores para todas sus muestras se alejan significativamente de la distribución multivariante de los datos.

- Descartamos la muestra HE *sample_42* por *outlier*. Podríamos considerar dejarla en el dataset y tener en cuenta que puede estar afectando a las conclusiones (por ejemplo, interpretar que es un metabolito menos común en muestras HE comparado con GC cuando la observación está causada por esta muestra y no por el grupo entero). Para este análisis preliminar, la he excluído.

El gráfico de coordenadas nos muestra como la varianza de los datos deriva principalmente de las muestras experimentales, mientras que las muestras de control son muy similares entre ellas (indicador de buenos controles). Una vez tenemos el dataset listo para analizar (con datos completos y normalizados, y con metabolitos y muestras que tienen una distribución y varianza similar y comparable), podemos empezar a analizar los datos.

3.5 Análisis preliminar de los datos

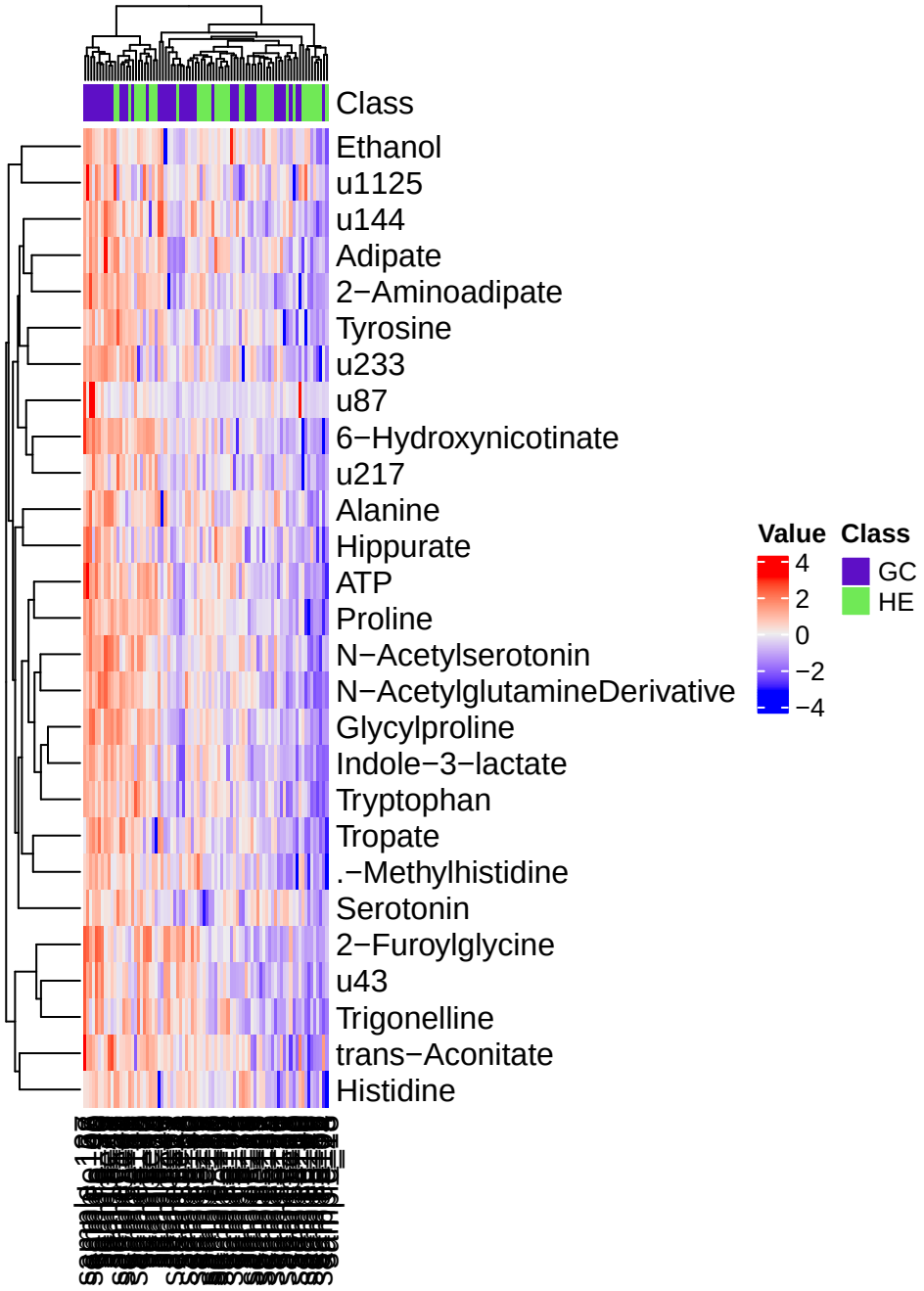
Como análisis preliminar, en primer lugar he realizado un análisis de componentes principales para tener una primera idea de si hay diferencia entre HE, GC y BN. Al ver que las muestras HE se agrupan separadas de las GC, añadí un análisis de expresión diferencial para identificar qué metabolitos distinguen las muestras sanas de las cancerígenas.



PCA

En primer lugar, vemos que la mayoría de la varianza de los datos está explicada por la primera componente (41%), mientras que la segunda componente apenas explica el 6%. Es decir, que las dos primeras componentes no llegan a explicar ni la mitad de la varianza de los datos. Esto indica que las diferencias metabólicas entre muestras HE, BN o GC no son claras, ya que hay mucha variación. Vemos que según la primera componente, hay diferencias entre las muestras que se sitúan a la derecha y a la izquierda, aunque los metabolitos que determinan esas posiciones no son diferentes para las distintas condiciones. Sin embargo, vemos que la mayoría de muestras cancerígenas se localizan en la parte negativa del eje-y, mientras que la mayoría de las sanas se sitúan en la parte positiva. Esto indica, que a pesar de ser una proporción pequeña de la varianza, hay ciertos metabolitos que permiten diferencias muestras sanas de cancerígenas.

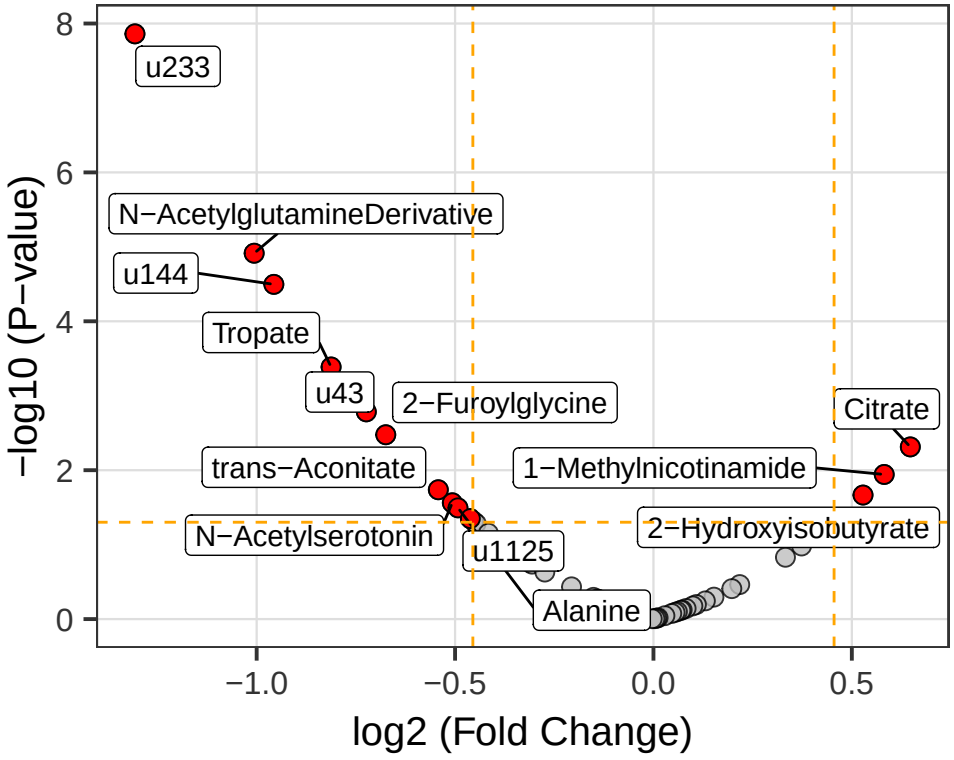
Por esto, identifiqué qué metabolitos se asocian con los valores negativos de la PC2 (n=27) y los visualicé en un histograma para identificar diferencias entre muestras HE y GC.



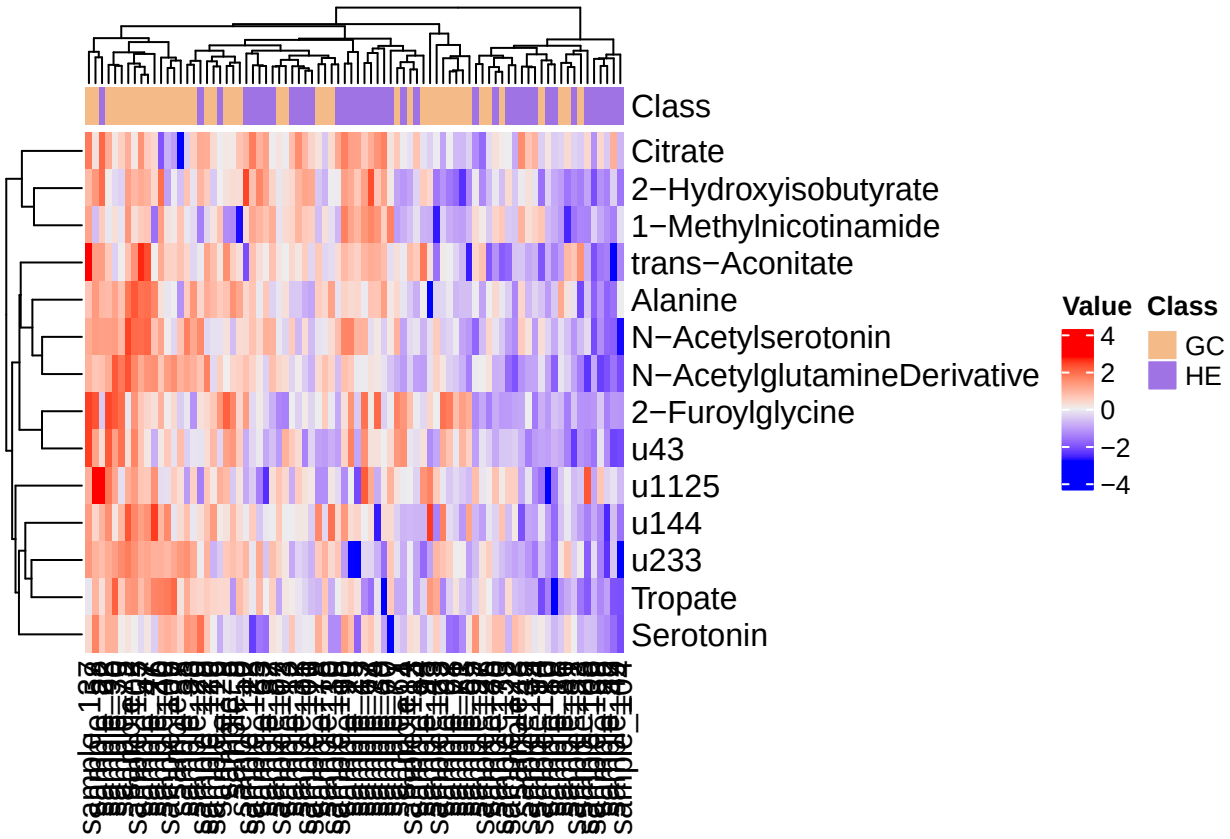
Según el dendograma de las columnas, vemos que hay dos *clusters* principales. El de la izquierda con una mayoría de GC (especialmente en el borde izquierdo) y el de la derecha con una mayoría HE (especialmente en el borde derecho). Sin embargo, vemos que los grupos no son claros y que hay muchas muestras HE y GC con valores similares de metabolitos.

Para intentar filtrar aún más los metabolitos que diferencian HE de GC, realicé un análisis de expresión diferencial comparado HE vs GC.

Análisis de expresión differencial El volcano-plot representa (coloreado y con el identificador del metabolito), aquellos que son significativamente diferentes entre HE y GC (n=14).



11/16 metabolitos (excepto “Citrate”, “1-Methylnicotinamide” y “2-Hydroxyisobutyrate”) fueron identificados en el análisis de la PCA.



En este caso, el cluster pequeño (a la derecha) está sobrerepresentado por muestras sanas, indicando que

para estos 11 metabolitos, las personas sanas son más similares (y se detectan en menor medida).

Este análisis no es suficiente para encontrar diferencias significativas entre ambos grupos, ya que el heatmap y el clustering aún muestran grupos heterogéneos a pesar de haber filtrado el análisis a los metabolitos significativamente diferentes entre ambos grupos.

En el tutorial, aplicaron modelos de aprendizaje automático *Partial Least Squares-Discriminant Analysis* (PLS-DA) y describen 20 metabolitos importantes para que el modelo identifique correctamente muestras del test dataset como HE o GC, indicando ser potenciales biomarcadores [6]. En el resumen del trabajo en Workbench, definen tres metabolitos (2-hydroxyisobutyrate, 3-indoxylsulfate, y alanine) como biomarcadores de muestras cancerígenas tras aplicar modelos LS-LASSO [7]. Excepto 3-indoxylsulfate, los otros dos biomarcadores fueron identificados en mi análisis y en el tutorial. Sin embargo, si miramos el heatmap, esos tres biomarcadores aún muestran cierto solapamiento entre muestras sanas y cancerígenas (valores similares del metabolito). Supongo, que ambas condiciones deben tener otras variables en común (p. ej. dieta) que explique la gran parte de la varianza en los datos reflejada en la PC1 que no distingue entre muestras sanas o enfermas. Si dispusiéramos de más información, podríamos intentar mejorar los modelos para explicar mejor la asociación entre los biomarcadores identificados y los fenotipos sano/cancerígeno, problemente ligado a otras características (p.ej., si una persona sana tiene una mala dieta, su *metabolome* se parecerá al de una persona con cáncer de estómago).

[6] Tutorial repositorio
[7] Metabolomics Workbench