PEC 1

RETO 1: LAS ÓMICAS

ANDREU RUIZ GASPARÍN

Tabla de Contenidos

Abstract

En este PEC 1 de la asignatura de Análisis de datos ómicos, se ha llevado a cabo el análisis de datos metabolómicos mediante herramientas bioinformáticas en R, usando la clase SummarizedExperiment para estructurar y analizar un conjunto de datos descargado de Metabolomics Workbench, Concretamente el 2024-fobitools-UseCase-1, recuperado del estudio ST000291 de metabolomicsworkbench.org.

El estudio involucró a 18 mujeres sanas, jóvenes y con un índice de masa corporal (IMC) normal. Se les pidió que evitaran consumir alimentos ricos en proantocianidinas, como arándanos y manzanas, durante los primeros días del estudio. Posteriormente, fueron asignadas al azar a dos grupos que consumieron zumo de arándano o de manzana, con el objetivo de investigar los cambios metabólicos asociados a estos concentrados de proantocianidinas. Las muestras de orina y sangre se recolectaron antes y después de la intervención para su análisis.

Se realizaron análisis exploratorios que revelaron patrones de variabilidad y diferencias preliminares en metabolitos específicos entre los grupos. Los resultados sugieren posibles biomarcadores asociados al consumo de arándanos, y el estudio refuerza el uso de *SummarizedExperiment* como herramienta clave en la gestión de datos ómicos. El código y los datos se han almacenado en un repositorio de GitHub para promover la transparencia y la reproducibilidad.

Objetivos del Estudio

El objetivo de este trabajo es llevar a cabo una versión simplificada de un análisis de datos ómicos, enfocado en metabolómica, con el propósito de explorar herramientas bioinformáticas clave en R como la clase *SummarizedExperiment*. Este ejercicio incluye la selección de un conjunto de datos de metabolómica de un repositorio público, seguido de su preparación y análisis exploratorio utilizando técnicas estadísticas y visuales. Este flujo de trabajo tiene como objetivo familiarizarse con la estructura y gestión de datos ómicos a través de herramientas de bioinformática, desarrollando habilidades en la organización de datos, la implementación de métodos de preprocesado, y la exploración de patrones moleculares. Además, se busca integrar los resultados obtenidos en un repositorio de GitHub, fomentando las prácticas de transparencia y reproducibilidad en la ciencia de datos ómicos.

Materiales y métodos

Inicialmente see seleccionó un conjunto de datos de metabolómica disponible en el repositorio GitHub, en concreto el 2024-fobitools-UseCase-1, recuperado del estudio ST000291 de metabolomicsworkbench.org, que contiene datos de perfiles metabólicos obtenidos mediante espectrometría de masas (MS) de muestras de suero/tejidos/otros en condiciones experimentales específicas. Estos datos incluyen información sobre metabolitos y metadatos asociados, así como el identificador de las muestras y los tratamientos experimentales.

Se han consultado los detalles del estudio en el paper, sacado de:
https://pubmed.ncbi.nlm.nih.gov/32133462/

Posteriormente fueron procesados en R, utilizando la clase *SummarizedExperiment* para organizar el contenido en un contenedor adecuado. Se generaron tres elementos principales:

Datos de características (assay) que contienen las concentraciones de metabolitos en formato de matriz, con filas correspondientes a metabolitos y columnas a muestras.

Metadatos de columnas (colData), con información sobre las muestras, como identificadores y grupos de tratamiento.

Metadatos de filas (rowData), que incluye información de los metabolitos, como el nombre, identificadores en bases de datos (ej., PubChem y KEGG) y propiedades bioquímicas relevantes.

Una vez preparado el objeto SummarizedExperiment, se realizó una exploración inicial del dataset para evaluar la estructura general y verificar la consistencia. Esta exploración incluyó análisis estadísticos descriptivos de los metabolitos y los grupos experimentales, visualización de datos mediante gráficos de distribución y boxplots para observar patrones y posibles variaciones.

Todos los resultados, incluyendo el objeto *SummarizedExperiment*, el código R utilizado para la carga y exploración de los datos, y archivos de texto que describen el conjunto de datos y sus características, fueron organizados y almacenados en un repositorio de GitHub.

Resultados

Se descargaron los archivos *features*, *metadata* y *metabolite names* y del repositorio y se creó un objeto *SummarizedExperiment* que integra los datos de metabolitos y metadatos correspondientes. Este objeto contiene un total de 1541 metabolitos (filas) y 45 muestras (columnas), facilitando el análisis de datos ómicos y la posterior exploración de las características del metaboloma.

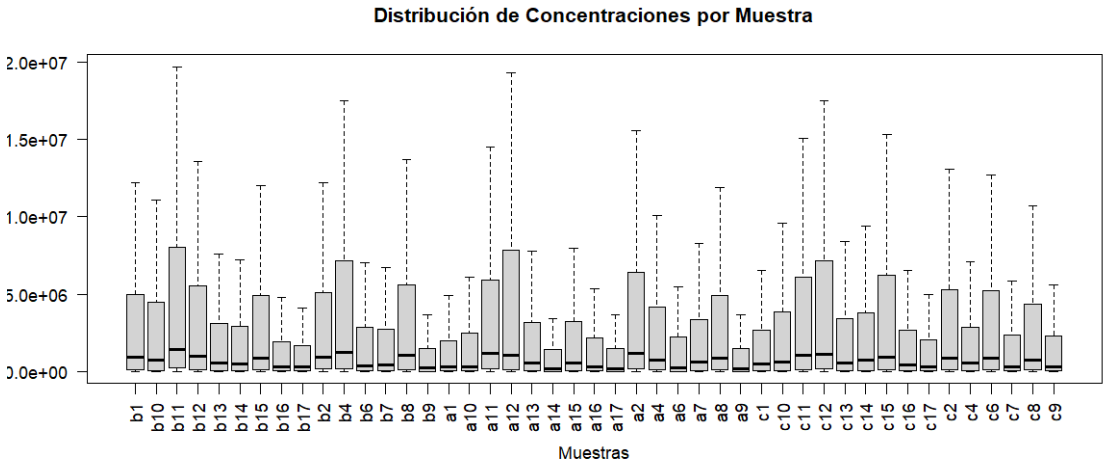
```
class: SummarizedExperiment
dim: 1541 45
metadata(0):
assays(1): counts
rownames(1541): 443489 107754 ... 53297445 11954209
rowData names(0):
colnames(45): b1 b10 ... c8 c9
colData names(2): ID Treatment
```

```
> # Verificación del número de variables (metabolitos) y muestras (sujetos)
> dim(se)
[1] 1541   45
```

La exploración inicial del objeto *SummarizedExperiment* reveló un resumen estadístico de las concentraciones de metabolitos. Se realizó un análisis descriptivo que mostró la distribución de los metabolitos en las muestras, permitiendo identificar variaciones en los niveles de metabolitos entre los diferentes grupos de tratamiento.

```
> # Resumen estadístico de los valores de metabolitos
> summary(assay(se, "counts"))
       b1                  b10                 b11                 b12                 b13                 b14                 b15          
 Min.   :0.000e+00   Min.   :0.000e+00   Min.   :0.000e+00   Min.   :0.000e+00   Min.   :0.000e+00   Min.   :0.000e+00   Min.   :0.000e+00  
 1st Qu.:1.235e+05   1st Qu.:8.145e+04   1st Qu.:2.240e+05   1st Qu.:1.390e+05   1st Qu.:5.810e+04   1st Qu.:4.435e+04   1st Qu.:1.070e+05  
 Median :9.100e+05   Median :7.200e+05   Median :1.450e+06   Median :1.010e+06   Median :5.380e+05   Median :4.890e+05   Median :8.560e+05  
 Mean   :3.245e+07   Mean   :2.800e+07   Mean   :4.107e+07   Mean   :3.606e+07   Mean   :2.452e+07   Mean   :2.227e+07   Mean   :3.086e+07  
 3rd Qu.:4.980e+06   3rd Qu.:4.500e+06   3rd Qu.:8.050e+06   3rd Qu.:5.545e+06   3rd Qu.:3.095e+06   3rd Qu.:2.925e+06   3rd Qu.:4.920e+06  
 Max.   :1.920e+10   Max.   :1.550e+10   Max.   :2.070e+10   Max.   :2.300e+10   Max.   :1.130e+10   Max.   :1.240e+10   Max.   :1.650e+10  
 NA's   :182         NA's   :182         NA's   :182         NA's   :182         NA's   :182         NA's   :182         NA's   :182        
      b16                 b17                  b2                  b4                  b6                  b7                  b8          
 Min.   :0.000e+00   Min.   :0.000e+00   Min.   :0.000e+00   Min.   :0.000e+00   Min.   :0.000e+00   Min.   :0.000e+00   Min.   :0.000e+00  
 1st Qu.:2.560e+04   1st Qu.:2.635e+04   1st Qu.:1.565e+05   1st Qu.:1.620e+05   1st Qu.:3.615e+04   1st Qu.:4.225e+04   1st Qu.:1.435e+05  
 Median :2.940e+05   Median :2.910e+05   Median :9.160e+05   Median :1.250e+06   Median :3.740e+05   Median :4.540e+05   Median :1.070e+06  
 Mean   :1.551e+07   Mean   :1.270e+07   Mean   :3.124e+07   Mean   :4.695e+07   Mean   :2.116e+07   Mean   :1.688e+07   Mean   :2.841e+07  
 3rd Qu.:1.940e+06   3rd Qu.:1.675e+06   3rd Qu.:5.070e+06   3rd Qu.:7.130e+06   3rd Qu.:2.830e+06   3rd Qu.:2.710e+06   3rd Qu.:5.615e+06  
 Max.   :7.320e+09   Max.   :6.810e+09   Max.   :1.700e+10   Max.   :2.610e+10   Max.   :1.270e+10   Max.   :7.870e+09   Max.   :1.430e+10  
 NA's   :182         NA's   :182         NA's   :182         NA's   :182         NA's   :182         NA's   :182         NA's   :182        
      b9                  a1                 a10                 a11                 a12                 a13                 a14          
 Min.   :0.000e+00   Min.   :0.000e+00   Min.   :0.000e+00   Min.   :0.000e+00   Min.   :0.000e+00   Min.   :0.000e+00   Min.   :0.000e+00  
 1st Qu.:1.520e+04   1st Qu.:3.375e+04   1st Qu.:3.025e+04   1st Qu.:1.480e+05   1st Qu.:1.430e+05   1st Qu.:5.950e+04   1st Qu.:9.860e+03  
 Median :2.090e+05   Median :3.000e+05   Median :2.860e+05   Median :1.180e+06   Median :1.030e+06   Median :5.350e+05   Median :1.640e+05  
 Mean   :1.288e+07   Mean   :1.371e+07   Mean   :2.108e+07   Mean   :3.527e+07   Mean   :4.372e+07   Mean   :2.336e+07   Mean   :1.336e+07  
 3rd Qu.:1.475e+06   3rd Qu.:2.005e+06   3rd Qu.:2.475e+06   3rd Qu.:5.895e+06   3rd Qu.:7.840e+06   3rd Qu.:3.155e+06   3rd Qu.:1.400e+06  
 Max.   :6.970e+09   Max.   :6.050e+09   Max.   :1.210e+10   Max.   :2.190e+10   Max.   :2.780e+10   Max.   :1.230e+10   Max.   :6.570e+09  
 NA's   :182         NA's   :182         NA's   :182         NA's   :182         NA's   :182         NA's   :182         NA's   :182        
     a15                 a16                 a17                  a2                  a4                  a6                  a7          
 Min.   :0.000e+00   Min.   :0.000e+00   Min.   :0.000e+00   Min.   :0.000e+00   Min.   :0.000e+00   Min.   :0.000e+00   Min.   :0.000e+00  
 1st Qu.:7.320e+04   1st Qu.:2.300e+04   1st Qu.:1.085e+04   1st Qu.:1.670e+05   1st Qu.:9.705e+04   1st Qu.:1.870e+04   1st Qu.:7.105e+04  
 Median :5.380e+05   Median :2.790e+05   Median :1.520e+05   Median :1.150e+06   Median :7.080e+05   Median :2.420e+05   Median :6.410e+05  
 Mean   :2.305e+07   Mean   :1.758e+07   Mean   :1.145e+07   Mean   :3.704e+07   Mean   :2.767e+07   Mean   :1.619e+07   Mean   :2.039e+07  
 3rd Qu.:3.230e+06   3rd Qu.:2.150e+06   3rd Qu.:1.475e+06   3rd Qu.:6.380e+06   3rd Qu.:4.140e+06   3rd Qu.:2.215e+06   3rd Qu.:3.355e+06  
 Max.   :1.430e+10   Max.   :9.770e+09   Max.   :5.330e+09   Max.   :2.090e+10   Max.   :1.530e+10   Max.   :8.930e+09   Max.   :9.710e+09  
 NA's   :182         NA's   :182         NA's   :182         NA's   :182         NA's   :182         NA's   :182         NA's   :182        
      a8                  a9                  c1                 c10                 c11                 c12                 c13          
 Min.   :0.000e+00   Min.   :0.000e+00   Min.   :0.000e+00   Min.   :0.000e+00   Min.   :0.000e+00   Min.   :0.000e+00   Min.   :0.000e+00  
 1st Qu.:1.005e+05   1st Qu.:1.135e+04   1st Qu.:4.840e+04   1st Qu.:6.490e+04   1st Qu.:1.305e+05   1st Qu.:1.625e+05   1st Qu.:6.530e+04  
 Median :8.590e+05   Median :1.740e+05   Median :4.920e+05   Median :5.980e+05   Median :1.050e+06   Median :1.120e+06   Median :5.770e+05  
 Mean   :3.144e+07   Mean   :1.317e+07   Mean   :2.021e+07   Mean   :2.269e+07   Mean   :4.042e+07   Mean   :5.147e+07   Mean   :2.468e+07  
 3rd Qu.:4.900e+06   3rd Qu.:1.470e+06   3rd Qu.:2.680e+06   3rd Qu.:3.880e+06   3rd Qu.:6.120e+06   3rd Qu.:7.135e+06   3rd Qu.:3.415e+06  
 Max.   :1.640e+10   Max.   :5.900e+09   Max.   :1.050e+10   Max.   :1.240e+10   Max.   :2.040e+10   Max.   :2.380e+10   Max.   :1.120e+10  
 NA's   :182         NA's   :182         NA's   :182         NA's   :182         NA's   :182         NA's   :182         NA's   :182        
     c14                 c15                 c16                 c17                  c2                  c4                  c6          
 Min.   :0.000e+00   Min.   :0.000e+00   Min.   :0.000e+00   Min.   :0.000e+00   Min.   :0.000e+00   Min.   :0.000e+00   Min.   :0.000e+00  
 1st Qu.:8.165e+04   1st Qu.:1.150e+05   1st Qu.:4.185e+04   1st Qu.:2.320e+04   1st Qu.:1.095e+05   1st Qu.:4.110e+04   1st Qu.:1.180e+05  
 Median :7.260e+05   Median :9.290e+05   Median :4.350e+05   Median :2.700e+05   Median :8.590e+05   Median :5.200e+05   Median :8.790e+05  
 Mean   :2.447e+07   Mean   :4.430e+07   Mean   :2.168e+07   Mean   :1.682e+07   Mean   :4.522e+07   Mean   :1.774e+07   Mean   :3.881e+07  
 3rd Qu.:3.810e+06   3rd Qu.:6.195e+06   3rd Qu.:2.655e+06   3rd Qu.:2.055e+06   3rd Qu.:5.310e+06   3rd Qu.:2.865e+06   3rd Qu.:5.245e+06  
 Max.   :9.930e+09   Max.   :2.140e+10   Max.   :7.550e+09   Max.   :5.920e+09   Max.   :2.110e+10   Max.   :1.060e+10   Max.   :1.710e+10  
 NA's   :182         NA's   :182         NA's   :182         NA's   :182         NA's   :182         NA's   :182         NA's   :182        
      c7                  c8                  c9          
 Min.   :0.00e+00   Min.   :0.000e+00   Min.   :0.000e+00  
 1st Qu.:2.31e+04   1st Qu.:8.745e+04   1st Qu.:2.650e+04  
 Median :3.06e+05   Median :7.260e+05   Median :3.020e+05  
 Mean   :2.12e+07   Mean   :2.937e+07   Mean   :1.814e+07  
 3rd Qu.:2.35e+06   3rd Qu.:4.345e+06   3rd Qu.:2.265e+06  
 Max.   :9.75e+09   Max.   :1.480e+10   Max.   :7.420e+09  
 NA's   :182        NA's   :182         NA's   :182        
```

Se generó un *boxplot* que ilustra la distribución de las concentraciones de metabolitos por muestra. Este análisis gráfico permitió observar diferencias en la variabilidad de metabolitos entre las distintas muestras, sugiriendo la existencia de patrones que podrían relacionarse con los tratamientos administrados.

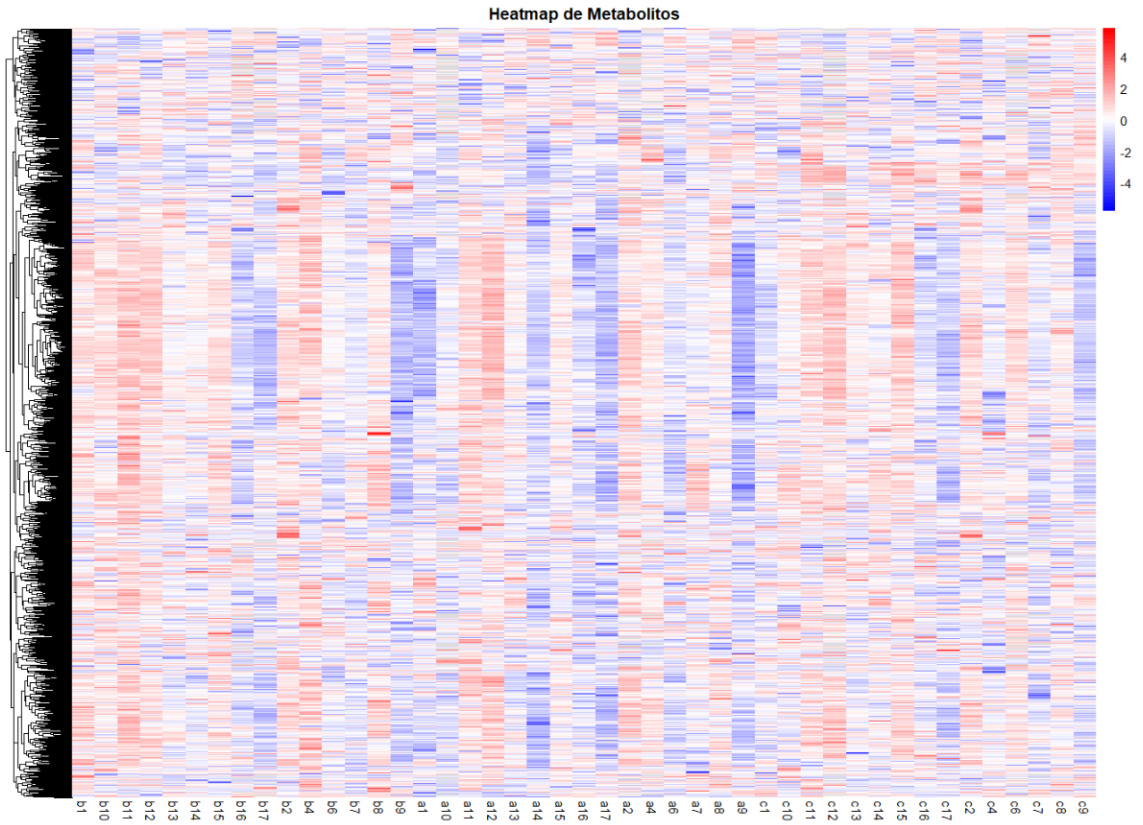


1. Box Plot de la distrubución de concentraciones

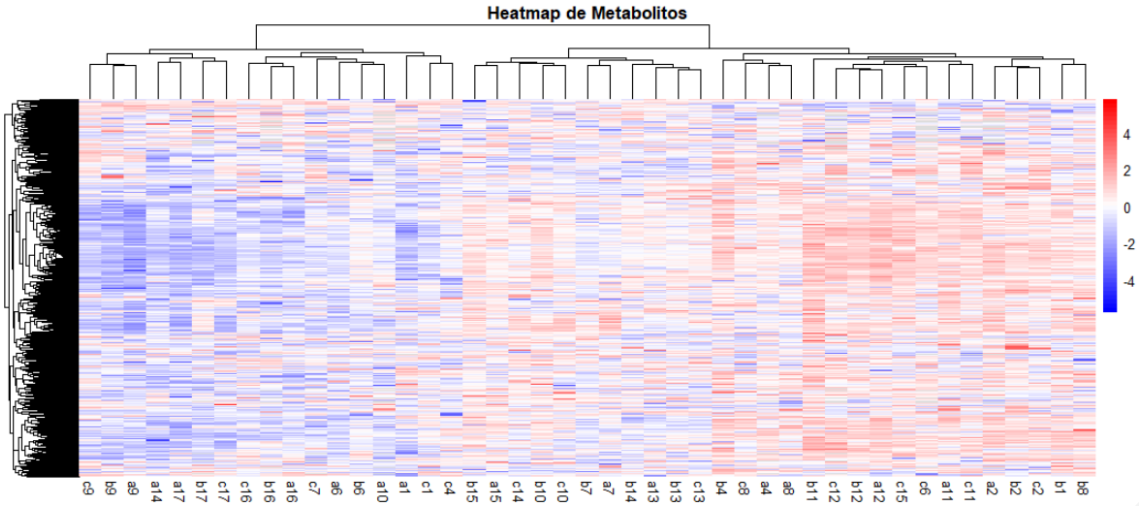
Las diferencias que se observan en la mediana y la dispersión entre muestras podrían indicar efectos del tratamiento o variabilidad biológica entre grupos. se sugiere que los tratamientos pueden tener un efecto medible en los niveles de algunos metabolitos.

Con tal de complementar la información y revelar patrones de agrupamiento que reflejen diferencias en los perfiles de expresión de los metabolitos entre las muestras, se construyó un *heatmap* utilizando la función *pheatmap*, donde se representaron los metabolitos, previamente log-transformados para estabilizar la varianza.

Este análisis es particularmente útil en el contexto del ejercicio ya que proporciona una visión general de los patrones de expresión que pueden estar asociados a los tratamientos. Al identificar grupos de metabolitos con perfiles específicos, se puede comenzar a especular sobre la relación entre los tratamientos y ciertos metabolitos. Esto es clave en estudios de metabolómica, donde el objetivo es generalmente identificar biomarcadores o vías metabólicas afectadas por un tratamiento o condición experimental.



2. heatmap de metabolitos sin Clustering



3. heatmap de metabolitos con Clustering

El clustering de las columnas (muestras) agrupa muestras con perfiles de metabolitos similares, y esto podría correlacionarse con los tratamientos recibidos. Las zonas de color en el heatmap indican niveles altos o bajos de metabolitos específicos en ciertos grupos de muestras, sugiriendo que algunos metabolitos podrían estar especialmente regulados o suprimidos en funciónm del tratamiento. Concretamente, las muestras b4, c8, a4, a8, b11, c12, b12, a12, c15, c6, a11, c11, a2, b2, c2, b1, y b8 presentan unos niveles más elevados de metabolitos.

Discusión y Conclusiones

En este estudio, se ha llevado a cabo una exploración detallada de un conjunto de datos metabolómicos utilizando herramientas de análisis ómico, con el fin de identificar patrones en la expresión de metabolitos entre diferentes muestras y grupos de tratamiento. A través de la construcción de un contenedor *SummarizedExperiment*, ha sido posible organizar y estructurar los datos de manera eficiente, permitiendo el uso de herramientas avanzadas de análisis en R. La implementación de un análisis descriptivo, seguido de visualizaciones como el boxplot y el heatmap, ha proporcionado una visión general de la distribución y variabilidad de los metabolitos en las muestras así como de las relaciones entre ellas.

El boxplot ha revelado variaciones en las concentraciones de metabolitos entre muestras individuales, lo cual indica que pueden existir diferencias significativas en los perfiles metabólicos que podrían estar asociadas con los tratamientos aplicados, siendo relevante ya que indica heterogeneidad en las respuestas metabólicas. Podría estar relacionada con factores específicos de cada grupo o condición experimental.

El heatmap, para el cual se ha aplicado una transformación logarítmica para estabilizar la varianza, ha mostrado patrones de agrupamiento entre muestras que comparten perfiles metabólicos similares. Este análisis ha permitido observar que ciertas muestras, agrupadas según su tratamiento, presentan patrones de expresión de metabolitos característicos, lo cual indica una relación potencial entre los tratamientos y los cambios en los niveles de los metabolitos. Las muestras con tratamientos similares han tendido a agruparse en el *heatmap*, lo cual refuerza la hipótesis de que el tratamiento influye en el perfil metabolómico.

El análisis de *clustering* realizado en el *heatmap* ha facilitado la identificación de grupos de muestras con perfiles similares. Los patrones de agrupamiento sugieren que, los tratamientos aplicados, afectan la expresión de los metabolitos, que pueden reflejar adaptaciones metabólicas específicas a los tratamientos. Esta información nos es útil para identificar biomarcadores potenciales o metabolitos específicos que podrían estar relacionados con los efectos de los tratamientos, siendo relevante para estudios posteriores de validación.

Como conclusión, el uso de técnicas de visualización y análisis exploratorio de datos en metabolómica permite obtener información muy interesante sobre la relación entre el tratamiento y el perfil metabolómico de las muestras. Este tipo de análisis facilita la comprensión de la variabilidad biológica y ayuda a identificar patrones que podrían ser útiles para estudios de intervención o para la detección de biomarcadores. evaluar la significancia de las diferencias observadas y explorar modelos predictivos que puedan identificar metabolitos clave asociados con cada tratamiento.

Finalmente, y siguiendo la estructura de la actividad 1.3 de este curso, se da respuesta a las siguientes cuestiones sobre el estudio:

1. Pregunta Biológica

El estudio ST000291 se centra en investigar cómo el consumo de proantocianidinas, presente en zumos de arándano y manzana, afecta el perfil metabólico en humanos. La pregunta biológica principal va dirigida a entender las alteraciones metabólicas inducidas por estos zumos y su impacto potencial en la salud. El estudio nos permite explorar si el consumo de proantocianidinas sería beneficioso en la prevención de enfermedades metabólicas e identificar biomarcadores de respuesta a este tipo de intervención dietética.

2. Diseño Experimental

El estudio utilizó un diseño controlado aleatorizado, con un grupo de mujeres sanas que consumieron zumo de arándano o manzana en diferentes momentos en el tiempo. Este diseño de tipo cruzado, en el cual cada participante actúa como su propio control, favorece en minimizar la variabilidad interindividual y aumentar la sensibilidad del análisis. Aun así, al limitarse a una población específica (mujeres sanas), los resultados pueden tener restricciones en su generalización a otras poblaciones. Además, la falta de un grupo de control sin intervención puede limitar la interpretación de los efectos observados, ya que no se pueden descartar otros factores externos que afecten el metabolismo.

3. Obtención de Datos Crudos

La obtención de datos crudos mediante metabolómica es crítica para reflejar los cambios en el metabolismo de forma precisa. En este estudio, el uso de técnicas avanzadas, como la microdisección de captura láser, ayuda a seleccionar las áreas anatómicas relevantes, aunque puede introducir sesgos si no se seleccionan las regiones de manera óptima. La calidad de las muestras es un aspecto fundamental, ya que la variabilidad en el procesamiento y manejo de las muestras puede afectar los niveles de los metabolitos y, por lo tanto, los resultados finales.

4. Control de Calidad y Preprocesado

Para asegurar la fiabilidad de los resultados, el control de calidad y preprocesamiento de los datos es esencial. Entre los posibles desafíos de calidad están la variabilidad en la recolección de muestras y posibles interferencias en la extracción de metabolitos. En el preprocesado, se realizaron pasos como la normalización para reducir variaciones técnicas y el filtrado de metabolitos con baja abundancia, importante para mejorar la robustez del análisis estadístico y disminuir el ruido en los datos.

5. Análisis Estadístico y Metabolitos Diferencialmente Expresados

Para evaluar la significancia de los cambios en el metabolismo, es probable que el análisis estadístico incluyera métodos como el ANOVA o pruebas t, ajustadas por múltiples pruebas para corregir la tasa de falsos positivos. Esto permitió identificar metabolitos diferencialmente expresados entre los grupos. El criterio para selección de metabolitos clave pudo incluir un p-valor ajustado y un umbral mínimo de cambio en la abundancia, garantizando que solo se seleccionen metabolitos con cambios biológicamente relevantes.

6. Análisis de Significación Biológica

El análisis de la significación biológica de los metabolitos identificados se llevó a cabo utilizando herramientas como MetaboAnalyst, las cuales ayudan a ubicar los metabolitos en el contexto de vías metabólicas y procesos biológicos específicos. Este tipo de análisis permitió identificar rutas metabólicas que podrían estar influenciadas por el consumo de proantocianidinas, facilitando una comprensión más profunda de su impacto en la fisiología humana.

7. Respuesta a la Pregunta Biológica

Los resultados del estudio sugieren que el consumo de proantocianidinas tiene un efecto mensurable en ciertos metabolitos que pueden estar relacionados con la regulación del metabolismo. Estos hallazgos apoyan la hipótesis de que los zumos ricos en proantocianidinas (arándano y manzana) pueden influir en la salud metabólica, y podrían ser utilizados en un futuro para formular recomendaciones dietéticas personalizadas y estrategias preventivas contra enfermedades metabólicas; ya que los metabolitos identificados podrían actuar como biomarcadores de respuesta, lo cual tiene implicaciones positivas tanto en la investigación biomédica como en la salud pública.

El enlace al repositorio Github es el siguiente:

https://github.com/aruizg20/Entregable-PEC-1-micas/new/main