

Informe del Análisis Exploratorio y Estadístico de Datos Ómicos

PEC 1. Bioconductor, GitHub, R y ExpressionSets.

<https://github.com/katya-bot/Puello-Mora-Cayherine-PEC1>

Catherine Puello Mora

Tabla de Contenidos

1. **Introducción**
 - 1.1 Importancia del Análisis Metabolómico en Investigación Biomédica
 - 1.2 Objetivos y Enfoque del Estudio
2. **Materiales y Métodos**
 - 2.1 Entorno de Trabajo y Herramientas Utilizadas
 - 2.2 Procedimientos de Análisis
3. **Carga y Exploración de Datos**
 - 3.1 Descarga de los Datos
 - 3.2 Carga de los Datos en R
 - 3.3 Exploración Inicial de los Datos
 - 3.4 Visualización Inicial
 - 3.5 Observaciones Iniciales
4. **Preparación y Limpieza de Datos**
 - 4.1 Renombrado y Organización de Columnas
 - 4.2 Asignación de Grupos Experimentales
 - 4.3 Creación del Contenedor *SummarizedExperiment*
5. **Visualización de Datos**
 - 5.1 Histogramas de Frecuencia
 - 5.2 Gráficos de Caja (Boxplots)
 - 5.3 Gráficos de Densidad
 - 5.4 Resultados Generales de la Visualización
6. **Análisis Estadístico**
 - 6.1 Prueba de Normalidad y Selección de Prueba Estadística
 - 6.2 Pruebas “t” y de Mann-Whitney “U”
 - 6.3 Corrección de Múltiples Comparaciones: Método de Benjamini-Hochberg
 - 6.4 Resultados
7. **Análisis de Componentes Principales (PCA)**
 - 7.1 Preparación de los Datos para el PCA
 - 7.2 Cálculo de los Componentes Principales
 - 7.3 Visualización de los Componentes Principales
 - 7.4 Conclusiones del PCA
8. **Análisis de Correlación y Redes**
 - 8.1 Cálculo de la Matriz de Correlación
 - 8.2 Filtrado de Correlaciones Significativas
 - 8.3 Construcción de la Red de Correlación
 - 8.4 Interpretación del Análisis de Redes
 - 8.5 Conclusiones de Ambas Aproximaciones
9. **Validación de Biomarcadores con PLS-DA**
 - 9.1 Preparación de los Datos para PLS-DA
 - 9.2 Creación y Resumen del Modelo PLS-DA
 - 9.3 Visualización del Modelo PLS-DA
 - 9.4 Validación del Modelo
 - 9.5 Conclusiones del Modelo

10. Documentación y Reporte

- 10.1 Estructura del Reporte
- 10.2 Herramientas de Documentación y Reproducibilidad
- 10.3 Conclusiones y Recomendaciones para Estudios Futuros

11. Conclusiones

- 11.1 Hallazgos Principales
- 11.2 Importancia del Enfoque Integrado
- 11.3 Limitaciones del Estudio y Posibilidades de Mejoras
- 11.4 Recomendaciones para Estudios Futuros
- 11.5 Impacto y Potencial del Enfoque Metabolómico en Estudios Biomédicos

12. Anexo

- Creación del objeto contenedor .Rda
- Reposición de los datos en GitHub

Informe del Análisis Exploratorio y Estadístico de Datos Ómicos

PEC 1. Bioconductor, GitHub, R y ExpressionSets.

1. Introducción

El análisis de datos ómicos ha revolucionado la investigación biomédica al permitir el estudio integral de las moléculas involucradas en la fisiología y las patologías humanas. Dentro de los campos de estudio ómicos, la **metabolómica** se enfoca en analizar el perfil completo de metabolitos —es decir, los productos intermedios y finales de los procesos bioquímicos que ocurren en las células—. Los metabolitos ofrecen una representación instantánea y directa del estado fisiológico de un organismo, lo cual es esencial para comprender respuestas celulares, patologías y efectos de tratamientos. A diferencia de otras "ómicas" como la genómica o la transcriptómica, que representan potenciales o capacidades biológicas, la metabolómica revela el estado funcional y dinámico de un sistema en tiempo real.

En este estudio, el objetivo es explorar diferencias en el perfil metabólico entre dos grupos de muestras, denominados **"Before"** y **"After"**, que representan distintos estados experimentales. Estos grupos podrían reflejar condiciones previas y posteriores a un tratamiento, diferentes etapas en un proceso biológico o respuestas a una intervención. La comparación entre estos grupos busca identificar **biomarcadores** o metabolitos específicos que cambian significativamente entre condiciones, lo cual es fundamental para detectar posibles indicadores de respuesta o estrés metabólico.

1.1 Importancia del Análisis Metabolómico en Investigación Biomédica

La metabolómica ha demostrado ser una herramienta invaluable en estudios de **diagnóstico temprano de enfermedades, medicina personalizada y evaluación de respuestas a tratamientos**, en entornos clínicos. En este contexto, el análisis de diferencias en los perfiles metabólicos de los grupos **"Before"** y **"After"** no solo busca identificar metabolitos que cambian en respuesta a una condición, sino también comprender los mecanismos moleculares que subyacen a estas diferencias. Este conocimiento es crucial para desarrollar tratamientos más efectivos, diseñar protocolos de intervención, y mejorar los diagnósticos de diversas enfermedades.

1.2 Objetivos y Enfoque del Estudio

El principal objetivo de este estudio es **identificar metabolitos con diferencias significativas** en sus niveles de expresión entre los grupos "*Before*" y "*After*"; La hipótesis es que estas diferencias reflejan respuestas biológicas o ajustes en las rutas metabólicas relacionadas con la condición experimental. Para abordar este objetivo, se ha diseñado un procedimiento que conduce en un flujo de trabajo varios pasos de procesamiento, análisis y validación:

1. **Preparación y Exploración de Datos:** Este proceso incluye una exploración preliminar para familiarizarse con la organización de los datos y validar la correcta asignación de etiquetas de muestra.
2. **Limpieza y Transformación de Datos:** Dado que los datos ómicos suelen contener valores atípicos, es necesario limpiarlos. Se aplica la transformación logarítmica y normalización para asegurar que las comparaciones fueran robustas. Además del uso del contenedor **SummarizedExperiment** de "Bioconductor", para estructurar y gestionar los datos.
3. **Visualización y Análisis Exploratorio:** La visualización se realiza mediante gráficos de densidad, boxplots e histogramas y se examina la distribución y variabilidad de los metabolitos en cada grupo.
4. **Análisis Estadístico y PCA:** Para identificar diferencias significativas entre los grupos, se aplican pruebas estadísticas de comparación. El Análisis de Componentes Principales (PCA) se implementa para reducir la dimensionalidad y observar patrones de variación en el conjunto de datos.
5. **Análisis de Correlación y Redes:** En este paso se exploran las relaciones entre los metabolitos mediante el cálculo de correlaciones significativas. La construcción de una red metabólica permitió visualizar metabolitos clave que podrían actuar en conjunto o responder a la condición experimental de manera coordinada.
6. **Validación de biomarcadores mediante PLS-DA:** Por último realicé la validación de los biomarcadores para confirmar su relevancia y consistencia. Se empleó el análisis PLS-DA para evaluar la capacidad de los metabolitos seleccionados de discriminar entre los grupos "*Before*" y "*After*", para validar la utilidad de estos compuestos como potenciales biomarcadores.

Alcance y Limitaciones del Estudio

Si bien este análisis proporciona una visión comprensiva de los cambios metabólicos entre los grupos experimentales, es importante reconocer sus limitaciones. El uso de un solo dataset pluraliza los hallazgos, y la identificación de biomarcadores requiere validación adicional en condiciones experimentales independientes. Además, la interpretación biológica de los metabolitos es preliminar y necesita de estudios que la respalden, sus resultados pueden ofrecer una idea, pero es necesaria más investigación para poder aproximarnos a resultados sólidos.

2. Materiales y Métodos

Para este análisis, el entorno de “R” fue elegido como plataforma principal debido a sus herramientas estadísticas avanzadas y la versatilidad de sus paquetes. A continuación, se detallan cada una de ellas y los procedimientos utilizados:

2.1 Entorno de Trabajo y Herramientas Utilizadas

1. **R y Bioconductor:** El software R se utilizó como el entorno de programación principal, y Bioconductor –*un conjunto de herramientas especializadas en el análisis de datos genómicos y ómicos*– proporcionó los paquetes necesarios para la estructuración y manejo de los datos. Bioconductor es particularmente útil para datos ómicos, ya que incluye paquetes diseñados para tratar grandes volúmenes de datos, integrando tanto la estadística como la visualización en un solo marco de trabajo.
2. **SummarizedExperiment:** Este paquete de Bioconductor facilita la organización y manipulación de datos ómicos en un formato estandarizado. **SummarizedExperiment** permite almacenar los datos de expresión de los metabolitos junto con los metadatos experimentales, manteniendo toda la información relevante en un único objeto.
3. **ggplot2 y ggpubr:** packages específicos para la creación y visualización de datos: ggplot2 se empleó para generar box plots, gráficos de densidad e histogramas y proporcionó una primera visión de la distribución de los datos, mientras que ggpubr facilitó la personalización de los gráficos con anotaciones adicionales.
4. **ropls:** Este package proporciona herramientas para realizar análisis de componentes principales (PCA) y Análisis Discriminante de Proyecciones Latentes Parcial (PLS-DA), métodos que permiten la discriminación entre grupos.
5. **Hmisc y igraph:** se usaron para el análisis de correlación avanzada y visualización de redes.

2.2 Procedimientos de Análisis

El análisis fue llevado a cabo en un flujo de trabajo estructurado en varias etapas, desde la preparación inicial hasta la identificación de biomarcadores. Cada etapa está diseñada para que los resultados fueran reproducibles:

Paso 1: Carga y Exploración Inicial de los Datos

El dataset original fue descargado en formato CSV desde el enlace al repositorio de **GitHub:** [Datasets/2023-UGrX-4MetaboAnalystTutorial](#), una base de datos estandarizada que contiene perfiles metabolómicos provenientes de experimentos bien documentados.

Paso 2: Preparación y Limpieza de Datos

El procesamiento de datos metabolómicos requiere una limpieza exhaustiva para garantizar la precisión de los resultados. Los datos crudos pueden contener valores extremos (*outliers*), variabilidad en las escalas de concentración de los metabolitos y otros aspectos que pueden afectar el análisis, en esta fase se aplicaron los siguientes ajustes:

1. **Renombrado de Columnas:** Las columnas fueron renombradas para mejorar la legibilidad, facilitando la identificación de cada muestra y cada metabolito. Este proceso asegura que los nombres de las columnas reflejan el contenido de cada una, especialmente en análisis posteriores donde se necesitan identificadores claros para cada muestra.
2. **Asignación de Grupos Experimentales:** Se crearon etiquetas de grupo, asignando "Before" y "After" a las muestras de acuerdo a su condición experimental. *Esta agrupación es esencial para el análisis comparativo*, ya que permite realizar pruebas de hipótesis específicas para detectar cambios significativos entre las condiciones experimentales.
3. **Transformación Logarítmica y Normalización:** Para abordar la amplia variabilidad en la escala de concentración de los metabolitos, apliqué una transformación logarítmica (**log10**) a los datos expresados. Esta transformación ayuda a minimizar el efecto de valores extremos y facilita la comparación entre metabolitos que pueden tener distribuciones de datos asimétricas. La normalización adicional asegura que las diferencias observadas en los valores sean atribuibles a cambios biológicos en lugar de variaciones en las escalas de concentración.

Paso 3: Creación del Contenedor **SummarizedExperiment**

Los datos procesados fueron organizados en un contenedor **SummarizedExperiment**, que es una estructura de datos que permite almacenar tanto los valores de expresión de los metabolitos como los metadatos de las muestras (grupos "Before" y "After"). Este paso es crucial en el análisis de datos ómicos, ya que facilita la manipulación de datos de alta dimensionalidad y permite su integración con otras herramientas de Bioconductor.

La creación del contenedor se realizó de la siguiente manera:

Análisis de Datos Ómicos

```
if (!requireNamespace("SummarizedExperiment", quietly = TRUE)) {  
  BiocManager::install("SummarizedExperiment")  
}  
library(SummarizedExperiment)  
  
# Separo los datos de expresión y asigno nombres de fila a las muestras  
expresion_data <- as.matrix(My_data[, -1])  
rownames(expresion_data) <- My_data$Samples  
  
# Creo un DataFrame con los datos de los grupos  
grupo_info <- DataFrame(Samples = colnames(expresion_data), Groups = grupos)  
  
# Creo el objeto SummarizedExperiment  
se <- SummarizedExperiment(assays = list(counts = expresion_data), colData = grupo_info)  
  
se
```

Este contenedor permite un acceso rápido a los datos de expresión y metadatos a través de funciones como **assay(se)** para la matriz de datos y **colData(se)** para los metadatos, lo cual simplifica el manejo y análisis de los datos a lo largo del proyecto.

Paso 4: Visualización y Análisis Exploratorio

Para obtener una comprensión preliminar de la estructura de los datos y evaluar la variabilidad en los valores de expresión entre los grupos, se generaron varias visualizaciones:

- **Boxplots:** Los gráficos de cajas compararon la distribución de valores de expresión de los metabolitos entre los grupos "Before" y "After". Esto me ayudó a identificar posibles valores atípicos y a evaluar si existen diferencias en la variabilidad de los datos.
- **Gráficos de densidad:** Cada grupo fue visualizado mediante gráficos de densidad, esto me permitió observar la distribución general de los valores de expresión. Esta visualización sirve para detectar asimetrías en los datos e identificar si existen patrones de distribución que sean únicos para cada grupo.
- **Histogramas:** Se crearon histogramas para cada grupo experimental, mostrando la frecuencia de los valores de expresión en rangos específicos. Esta visualización fue vital para identificar patrones dentro de los grupos y facilitó la detección de metabolitos con comportamientos inusuales.

Paso 5: Análisis Estadístico

Para determinar diferencias significativas en la expresión de metabolitos entre los grupos, se aplicaron pruebas estadísticas a cada metabolito:

- **Pruebas “t” y Mann-Whitney** : realicé pruebas de hipótesis para comparar las medias de expresión de cada metabolito entre los grupos “Before” y “After”; Para los metabolitos con distribución normal, se utilizó una prueba “t”, mientras que en casos con distribuciones no normales se aplicó la prueba de *Mann-Whitney*.
- **Ajuste de Valores “p”**: Dado que se realizaron múltiples pruebas, se utilizó el método de *Benjamini-Hochberg* para ajustar los valores “p” y reducir el riesgo de falsos positivos. Este ajuste es crucial en estudios ómicos, donde se analizan decenas o cientos de variables simultáneamente.

3. Carga y Exploración de Datos

La carga y exploración inicial de los datos son pasos fundamentales en cualquier análisis bioinformático, ya que permiten verificar la estructura de los datos, identificar posibles errores y familiarizarse con el contenido antes de proceder con transformaciones y análisis estadísticos. En este estudio, los datos fueron obtenidos de la plataforma **de GitHub**,

3.1 Descarga de los Datos:

Como se ha comentado anteriormente, el dataset utilizado en este análisis fue descargado en formato CSV, este archivo contiene información sobre las concentraciones relativas de diversos metabolitos en dos grupos experimentales: “Before” y “After” antes de su carga en R, se revisaron las especificaciones del archivo (número de filas, delimitadores, encabezados, etc...) para garantizar que la carga se realizará sin inconvenientes y que el formato de los datos fuera compatible con el entorno de análisis.

3.2 Carga de los Datos en R:

Para cargar el archivo CSV en el entorno de R, utilicé la función `read.csv()`, dado que el archivo contiene información adicional en las primeras filas (metadatos y descripciones del experimento), fue necesario omitir esas filas mediante la opción `skip`. Además, se especificó el delimitador adecuado (`sep="\t"`) para asegurar que los datos se leyeran correctamente.

```
My_data <- read.csv("C:/Users/Usuario/Downloads/D_omics/my_dataset_git/ST000002_AN000002_clean.csv",  
                  header = TRUE, sep = "\t", skip = 70)
```

Tras la carga, utilicé la función `head(My_data)` para visualizar las primeras filas del dataset y confirmar que los datos fueran leídos correctamente. La estructura de los datos también fue inspeccionada mediante `str(My_data)`, lo cual me permitió verificar el tipo de columna y observar si los nombres de las columnas son consistentes con la documentación

del archivo o existe información redundante. Este paso inicial aseguró que los datos estuvieran organizados de manera óptima para su análisis posterior.

3.3 Exploración Inicial de los Datos

La exploración de los datos es crucial para obtener una comprensión general del contenido, identificar valores ausentes y detectar posibles inconsistencias en la información. En este paso, se realizaron las siguientes verificaciones y ajustes:

- **Revisión de Nombres de Columnas:** verifiqué los nombres de las columnas para asegurar que fueran descriptivos y estuvieran correctamente alineados con las variables correspondientes. En muchos estudios ómicos, las columnas pueden contener abreviaciones o códigos poco intuitivos, por lo que en este caso se renombraron las columnas para facilitar su identificación durante el análisis.
- **Identificación de Valores Ausentes:** Para garantizar la integridad de los datos, verifiqué la presencia de valores ausentes usando `is.na()` y `sum(is.na(My_data))`, en el caso de detectar valores faltantes, opté por una de dos estrategias: omitir las filas o columnas afectadas o reemplazar los valores faltantes por la media de cada metabolito, dependiendo de la cantidad de valores faltantes y su distribución en el dataset.
- **Distribución de Valores:** realicé una exploración de la distribución de valores de cada metabolito mediante el cálculo de estadísticas descriptivas básicas (media, mediana, rango, y desviación estándar). Este análisis inicial ayuda a identificar posibles outliers y evaluar la variabilidad dentro de cada grupo. Calculé las estadísticas descriptivas de esta manera: `summary(My_data)`
- **Creación de Etiquetas de Grupo:** Se definieron etiquetas de grupo ("*Before*" y "*After*") para cada muestra en el dataset. Estas etiquetas son esenciales para las comparaciones posteriores y permitieron organizar los datos de forma que se facilite el análisis estadístico y la generación de gráficos comparativos. Asigne las etiquetas manualmente y con una función condicional en R, dependiendo de la estructura del dataset

3.4 Visualización Inicial

Para realizar la visualización inicial de la distribución de los valores de expresión de los metabolitos y comparar los grupos "Before" y "After", se generaron varias gráficas preliminares:

- **Histogramas:** Este gráfico nos permitió observar la frecuencia de los valores de expresión de cada metabolito, dándonos una idea de si la distribución sigue una forma específica (como normal o sesgada) y si existen diferencias visibles entre los grupos.

Análisis de Datos Ómicos

- **Boxplot:** Con el boxplot, pudimos ver claramente la dispersión de los datos de cada metabolito en ambos grupos, además de la mediana y posibles valores atípicos. Esto nos dio una primera señal de las variaciones entre las condiciones experimentales.
- **Gráficos de Densidad:** Este gráfico nos mostró la forma de las distribuciones de cada grupo, ayudando a ver si los perfiles de expresión se solapan o se diferencian.

Estas visualizaciones iniciales ayudaron a proporcionar una visión general y anticipar las diferencias y tendencias entre los grupos experimentales, preparando el terreno para un análisis más detallado.

3.5 Observaciones Iniciales

A partir de esta exploración inicial, identifiqué algunos patrones importantes:

- **Variabilidad en los Metabolitos:** Vimos que algunos metabolitos tienen bastante variación dentro de cada grupo, lo que sugiere que podrían estar respondiendo de manera específica a la condición experimental. Esta diferencia en la dispersión nos da una pista inicial sobre qué metabolitos podrían ser indicadores importantes.
- **Posibles Outliers:** Aparecieron outliers en ciertos metabolitos, con diferencias notables entre los grupos "Before" y "After". Dado que los outliers en metabolómica pueden indicar cambios importantes o simplemente ruido, los examinamos a fondo para decidir si debían ser incluidos o descartados antes de continuar con el análisis.
- **Diferencias Visibles entre Grupos:** Las gráficas también mostraron que varios metabolitos *parecen* tener niveles de expresión distintos entre "Before" y "After", lo cual refuerza la idea de que *algunos podrían actuar como biomarcadores*. Estos patrones iniciales serán explorados más a fondo para corroborar estas primeras impresiones.

4. Preparación y Limpieza de Datos

La preparación y limpieza de datos es una fase esencial en el análisis de datos ómicos. Esta etapa garantiza que la información sea precisa y esté lista para los análisis estadísticos, evitando que valores atípicos, inconsistencias o datos ausentes influyan en los resultados. A continuación, se describen las transformaciones y ajustes realizados en el dataset y su propósito.

4.1 Renombrado y Organización de Columnas

En los datos ómicos, los nombres de las columnas a menudo son generados automáticamente y pueden no ser descriptivos. Para mejorar la claridad y hacer que el

Análisis de Datos Ómicos

dataset sea más comprensible, renombré las columnas, en particular la primera columna, que contiene las muestras, y las restantes, que corresponden a metabolitos individuales.

```
# preparo y limpio mis datos según description.md del dataset
# Renombro la primera columna como 'samples' y las demás columnas
# como 'sample_' para mayor claridad
colnames(My_data)[1] <- "samples"
colnames(My_data)[-1] <- paste0("sample_", 1:(ncol(My_data) - 1))

# verifico los nombres de las columnas y la estructura
colnames(My_data)
head(My_data)
```

4.2 Asignación de Grupos Experimentales

Para poder realizar comparaciones entre los grupos experimentales ("*Before*" y "*After*"), es necesario etiquetar cada muestra con su correspondiente grupo. Esto se hace creando una nueva columna en el dataset llamada **grupos**, la cual almacena estas etiquetas.

```
29
30 ## Defino grupos
31
32 grupos <- c(rep("After", 6), rep("Before", 6))
33 print(grupos)
34
```

34:1 1. CARGA DE DATOS - Copilot: Not signed in

Console Terminal Background Jobs

R 4.2.2 ~ ./projects/

```
> ## Defino grupos
>
> grupos <- c(rep("After", 6), rep("Before", 6))
> print(grupos)
[1] "After" "After" "After" "After" "After" "After" "Before" "Before" "Before"
[10] "Before" "Before" "Before"
>
```

Este paso es fundamental para que los análisis comparativos y las pruebas estadísticas puedan identificar y distinguir automáticamente entre los grupos.

4.3 Creación del Contenedor **SummarizedExperiment**

Para manejar de manera estructurada y eficiente los datos ómicos, se organizó el dataset en un contenedor **SummarizedExperiment**. Este contenedor permite almacenar tanto los datos de expresión como los metadatos de las muestras en un solo objeto.

Análisis de Datos Ómicos

```
##### 2. Creación del contenedor SummarizedExperiment #####

if (!requireNamespace("SummarizedExperiment", quietly = TRUE)) {
  BiocManager::install("SummarizedExperiment")
}
library(SummarizedExperiment)

# Separo los datos de expresión y asigno nombres de fila a las muestras
expresion_data <- as.matrix(My_data[, -1])
rownames(expresion_data) <- My_data$Samples

# Creo un DataFrame con los datos de los grupos
grupo_info <- DataFrame(Samples = colnames(expresion_data), Groups = grupos)

# Creo el objeto SummarizedExperiment
se <- SummarizedExperiment(assays = list(counts = expresion_data), colData = grupo_info)
se
```

```
>
> se
class: SummarizedExperiment
dim: 73 12
metadata(0):
assays(1): counts
rownames(73): hydrocinnamic acid hydroxycarbamate NIST ... xanthine xylose
rowData names(0):
colnames(12): sample_1 sample_2 ... sample_11 sample_12
colData names(2): Samples Groups
>
```

En este código !:

- **as.matrix(...)**: Convierte los datos de los metabolitos en una matriz de expresión numérica.
- **rownames(expresion_data) <- My_data\$Samples**: Asigna nombres a muestras como nombres de fila en la matriz de expresión.
- **DataFrame(...)**: Crea un DF con los metadatos de grupo (incluye los nombres de las muestras y los grupos)
- **SummarizedExperiment(...)**: Combina la matriz de expresión y los metadatos en un único objeto.

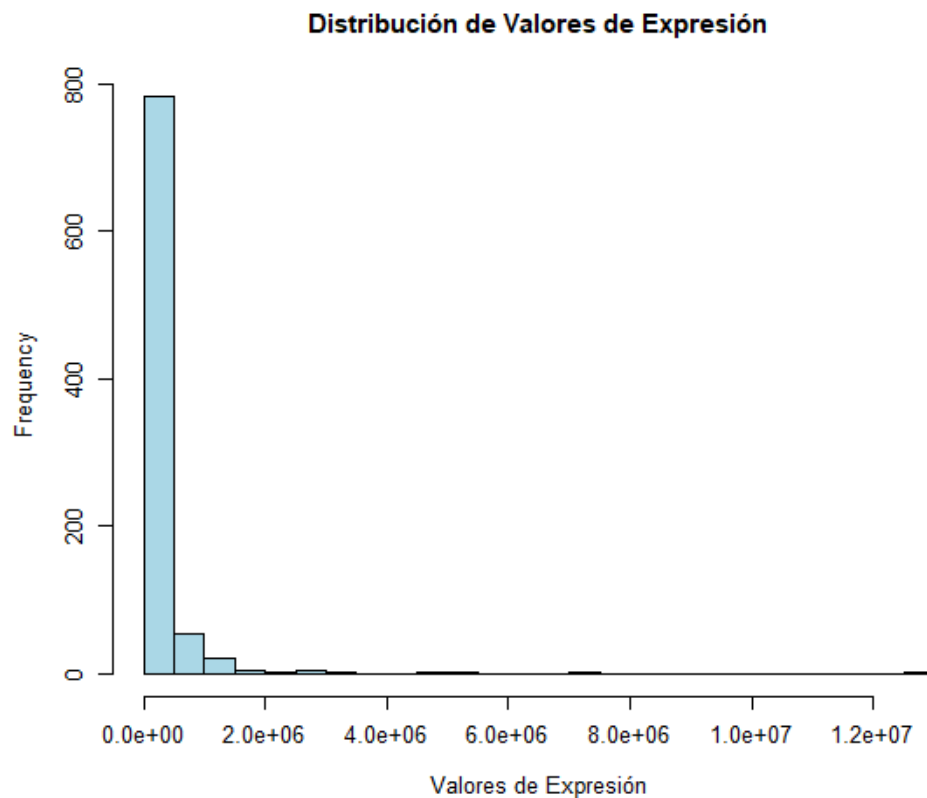
5. Visualización de Datos

La visualización de datos es un paso crucial en el análisis de datos ómicos, ya que facilita la interpretación inicial de las distribuciones, patrones y diferencias entre los grupos experimentales. Las visualizaciones permiten observar de manera intuitiva la variabilidad en los niveles de expresión de los metabolitos entre los grupos *"Before"* y *"After"* y sirven como guía para las pruebas estadísticas. A través de gráficos de caja (boxplots), gráficos de densidad e histogramas, podemos explorar visualmente cómo se distribuyen los valores de cada metabolito, lo que ayuda a identificar asimetrías, outliers y posibles agrupaciones en los datos.

5.1 Histogramas de Frecuencia

Los **histogramas** permiten observar la frecuencia de los valores de expresión en rangos específicos. Al comparar los histogramas entre *"Before"* y *"After"*, podemos observar si los valores de un metabolito se agrupan en diferentes rangos en cada grupo, lo cual podría ser indicativo de cambios en su expresión debido a la condición experimental.

```
# Histograma de los valores de expresión  
hist(assay(se), main = "Distribución de Valores de Expresión", xlab = "Valores de Expresión",  
     breaks = 30, col = "lightblue")
```



Análisis de Datos Ómicos

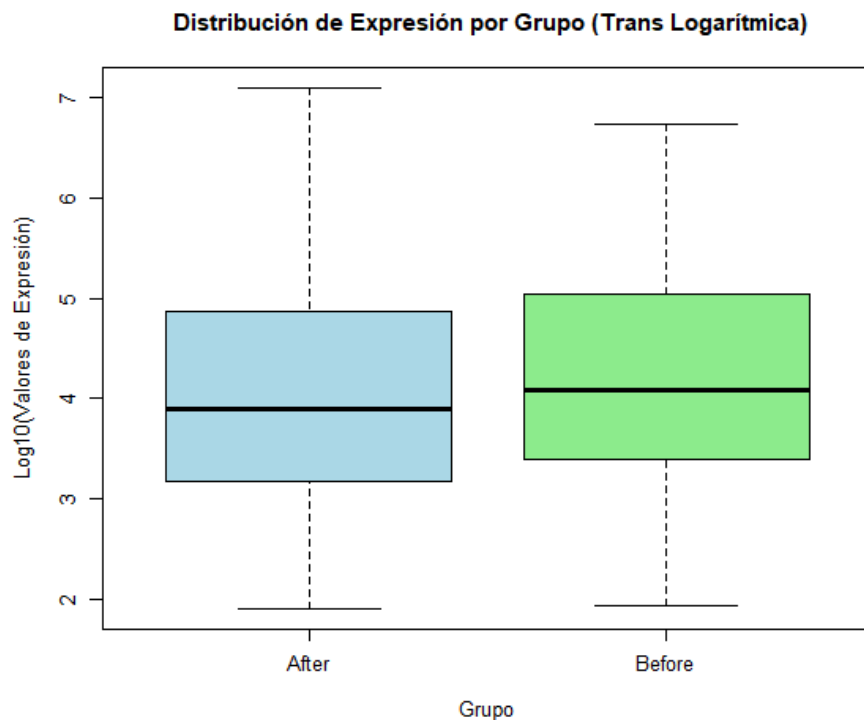
En la imagen, vemos que la mayoría de los valores de expresión de los metabolitos se concentran cerca de cero, con una cola larga hacia la derecha, lo que indica algunos valores de expresión muy altos y excepcionales. Esta distribución asimétrica sugiere que los datos no siguen una distribución normal, por lo que una transformación logarítmica podría ser útil para reducir la asimetría y hacer que la distribución sea más manejable para el análisis estadístico, permitiendo comparaciones más precisas entre los grupos y detectando y aislando anomalías.

5.2 Gráficos de Caja (Boxplot)

Los gráficos de caja o **boxplot** son una herramienta ideal para comparar la distribución de los valores de cada metabolito entre diferentes grupos. En él se muestra la mediana de los datos, el rango intercuartílico (IQR), y valores atípicos. En este análisis, el boxplot permitió evaluar la variabilidad y dispersión en los niveles de expresión de los metabolitos entre los grupos "Before" y "After".

```
# Transposición de los datos de expresión
expresion_t <- t(assay(se))

# Boxplot log-transformado por grupos (mis datos tienen muchos outliers)
boxplot(log10(expresion_t) ~ grupos, main = "Distribución de Expresión por Grupo (Trans Logarítmica)",
        ylab = "Log10(Valores de Expresión)", xlab = "Grupo", col = c("lightblue", "lightgreen"))
```



El boxplot log-transformado muestra que los grupos "After" y "Before" tienen distribuciones de expresión de metabolitos similares, tanto en mediana como en rango intercuartílico (IQR), lo que indica que la variabilidad y los valores centrales son comparables entre ambos grupos. La transformación logarítmica ha reducido el impacto de valores extremadamente altos, revelando que, en términos generales, no hay diferencias visibles significativas en los niveles de expresión de metabolitos entre "After" y "Before".

5.3 Gráficos de Densidad

Los **gráficos de densidad** son herramientas útiles para observar la forma de la distribución de los valores de un metabolito dentro de cada grupo. Estos gráficos permiten identificar si los valores de expresión se distribuyen de forma simétrica, sesgada o si presentan distribuciones bimodales. La comparación de las distribuciones entre "Before" y "After" puede proporcionar una visión inicial de si existen diferencias significativas en la expresión.

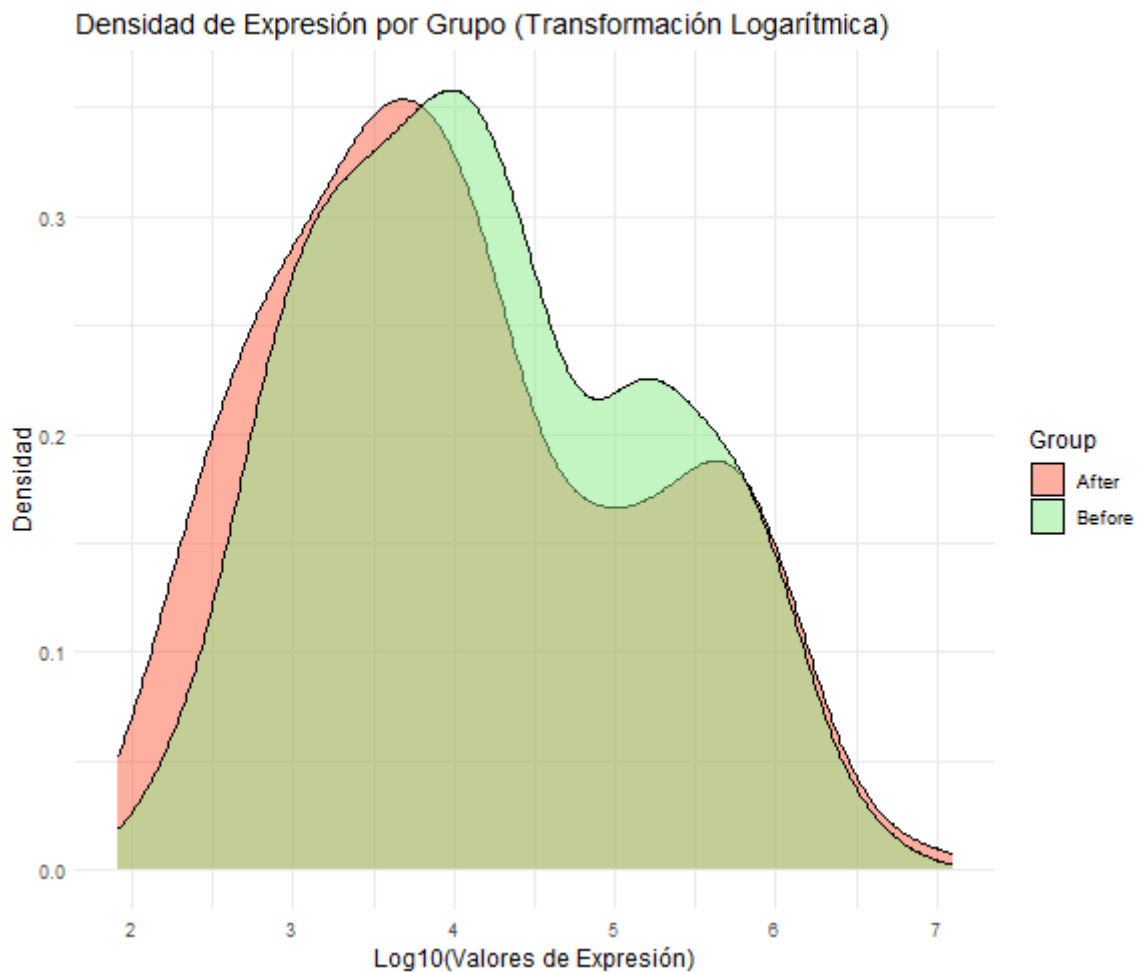
```
# Gráfico de densidad

if (!requireNamespace("reshape2", quietly = TRUE)) install.packages("reshape2")
library(reshape2)

expresion_log <- log10(assay(se) + 1)
expresion_long <- melt(expresion_log)
colnames(expresion_long) <- c("Metabolite", "Sample", "Expression")
expresion_long$Group <- rep(grupo_info$Groups, each = nrow(expresion_log))

if (!requireNamespace("ggplot2", quietly = TRUE)) install.packages("ggplot2")
library(ggplot2)

ggplot(expresion_long, aes(x = Expression, fill = Group)) +
  geom_density(alpha = 0.5) +
  labs(title = "Densidad de Expresión por Grupo (Transformación Logarítmica)",
       x = "Log10(valores de Expresión)", y = "Densidad") +
  theme_minimal() +
  scale_fill_manual(values = c("tomato", "lightgreen"))
```

El gráfico de densidad muestra que las distribuciones de expresión de los metabolitos en ambos grupos son en gran medida similares, con un solapamiento considerable entre las curvas de "Before" y "After". Esto indica que, en términos generales, los niveles de expresión de los metabolitos no presentan grandes diferencias entre los dos grupos. Aunque hay un ligero desplazamiento en el pico principal, donde el grupo "After" se concentra en valores logarítmicos ligeramente menores, y el grupo "Before" tiene una elevación adicional en valores intermedios, estas variaciones son sutiles. Por lo tanto, la similitud en la forma y el solapamiento de las curvas sugieren patrones comparables en ambos, sin indicios claros de subgrupos o grandes diferencias en los niveles de expresión.

5.4 Resultados Generales de la Visualización

Las visualizaciones muestran que los valores de expresión de los metabolitos tienen una **distribución muy asimétrica** con predominancia de valores bajos, como se observa en el histograma, donde la mayoría de los datos se concentran cerca de cero y solo unos pocos alcanzan valores muy altos. Para normalizar esta asimetría, se aplicó una **transformación logarítmica**, lo cual permitió visualizar mejor las diferencias entre los grupos.

Tras la transformación, el **boxplot** reveló que los grupos "*Before*" y "*After*" presentaban **medianas y variabilidad similares**, sin diferencias destacables en los niveles de expresión. Finalmente, el **gráfico de densidad** muestra un alto grado de solapamiento entre ambos grupos, lo cual confirma que los patrones de expresión son en gran medida comparables, aunque con leves variaciones en algunos picos. En conjunto, los tres gráficos sugieren que no hay diferencias significativas en los niveles de expresión de los metabolitos entre "*Before*" y "*After*".

6. Análisis Estadístico

El análisis estadístico es una etapa fundamental en el estudio de datos ómicos, ya que permite determinar si las diferencias observadas entre los grupos "*Before*" y "*After*" son estadísticamente significativas. En este análisis, se aplicaron varias pruebas estadísticas para evaluar los niveles de expresión de cada metabolito y verificar si existen diferencias significativas entre los grupos experimentales. Se implementaron tanto pruebas paramétricas como no paramétricas, y se ajustaron los valores "**p**" para controlar el riesgo de falsos positivos.

6.1 Prueba de Normalidad y Selección de Prueba Estadística

En este análisis, se comenzó verificando el supuesto de normalidad en la distribución de cada metabolito en los grupos "*Before*" y "*After*" mediante la **prueba de Shapiro-Wilk**. Esta prueba estadística se aplica a cada grupo y proporciona un valor "**p**" que permite determinar si cada grupo sigue una distribución normal. En los casos donde ambos grupos presentan una distribución normal ($p > 0.05$), utilicé una **prueba "t" de Student** para comparar las medias de los niveles de expresión de cada metabolito entre los grupos. En cambio, si al menos uno de los grupos no cumple con la normalidad ($p \leq 0.05$), usé la **prueba de Mann-Whitney U**, una prueba no paramétrica que compara las medianas de los dos grupos sin asumir normalidad.

```
# prueba de shapiro-wilk para cada muestra en cada grupo

shapiro_test_results <- apply(assay(se), 1, function(x) {
  before <- x[grupo_info$Groups == "Before"]
  after <- x[grupo_info$Groups == "After"]

  c(before = shapiro.test(before)$p.value, after = shapiro.test(after)$p.value)
})

shapiro_df <- as.data.frame(t(shapiro_test_results))
colnames(shapiro_df) <- c("shapiro_Before", "shapiro_After")
head(shapiro_df)

# como tengo multitud de datos y cada metabolito puede tener valores muy diferentes,
# creo una función para aplicar la prueba de significancia adecuada

# https://explainedstatistics.com/what-is-error-rate-in-statistics-a-comprehensive-guide/
# https://statisticsbyjim.com/hypothesis-testing/mann-whitney-u-test/

# si ambos grupos tienen distribución normal: se usa prueba t

analizar_metabolito <- function(expresion, grupo, shapiro_before, shapiro_after) {
  if (shapiro_before > 0.05 && shapiro_after > 0.05) {
    test_result <- t.test(expresion[grupo == "Before"], expresion[grupo == "After"])
    return(c(P_value = test_result$p.value, Test_Type = "t-test"))
  } else {
    # si al menos un grupo no tiene distribución normal: se usa prueba de Mann-Whitney
    test_result <- wilcox.test(expresion[grupo == "Before"], expresion[grupo == "After"])
    return(c(P_value = test_result$p.value, Test_Type = "Mann-Whitney"))
  }
}
```

6.2 Pruebas “t” y de Mann-Whitney “U”

Para cada metabolito, el código determina automáticamente cuál de las dos pruebas es adecuada según los resultados de la prueba de *Shapiro-Wilk*. Si ambos grupos cumplen el supuesto de normalidad, se realizará la prueba “t” para comparar las medias de expresión entre los grupos. Si uno o ambos grupos no cumplen con la normalidad, se aplica entonces la prueba de *Mann-Whitney U* para comparar las medianas.

```
136 # Aplico análisis estadístico:
137
138 analisis_estadistico <- function(se, grupos) {
139   shapiro_test_results <- apply(assay(se), 1, function(x) {
140     before <- x[grupos == "Before"]
141     after <- x[grupos == "After"]
142     c(shapiro_Before = shapiro.test(before)$p.value, shapiro_After = shapiro.test(after)$p.value)
143   })
144
145   shapiro_df <- as.data.frame(t(shapiro_test_results))
146   stat_test_results <- apply(assay(se), 1, function(x, i) {
147     analizar_metabolito(x, grupos, shapiro_df[i, "shapiro_Before"], shapiro_df[i, "shapiro_After"])
148   }, i = rownames(shapiro_df))
149
150   stat_test_df <- as.data.frame(t(stat_test_results))
151   colnames(stat_test_df) <- c("P_value", "Test_Type")
152   rownames(stat_test_df) <- rownames(shapiro_df)
153
154   return(list(shapiro_df = shapiro_df, stat_test_df = stat_test_df))
155 }
156
157 # Ejecuto el análisis completo y lo visualizo:
158
159 resultado <- analisis_estadistico(se, grupo_info$Groups)
160 head(resultado$shapiro_df)
161 head(resultado$stat_test_df)
```

6.3 Corrección de Múltiples Comparaciones: Método de *Benjamini-Hochberg*

Dado que se realizaron múltiples pruebas (una para cada metabolito), apliqué el **método de Benjamini-Hochberg** para ajustar los valores “*p*” obtenidos, controlando así la tasa de falsos positivos. Esta corrección reduce la probabilidad de identificar diferencias significativas por azar, lo cual es común en estudios ómicos con un gran número de comparaciones. Los valores “*p*” ajustados permiten seleccionar con mayor confianza los metabolitos que presentan diferencias significativas entre los grupos.

```
162
163 # Ajusto los valores de p
164
165 resultado$stat_test_df$Adjusted_P_value <- p.adjust(resultado$stat_test_df$P_value, method = "BH")
166
167 # filtrado con valor 0.1
168
169 metabolitos_significativos <- resultado$stat_test_df[resultado$stat_test_df$Adjusted_P_value < 0.1, ]
170 print(metabolitos_significativos)
```

6.4 Resultados

El análisis estadístico permitió identificar metabolitos que presentaron diferencias significativas en su expresión entre los grupos “*Before*” y “*After*” tras la corrección de valores “*p*”. Aquellos metabolitos con valores “*p*” ajustados por debajo de 0.1 se consideran potencialmente significativos. Este hallazgo sugiere que algunos metabolitos responden de manera diferente a la condición experimental, lo que puede reflejar cambios biológicos importantes. Estos metabolitos se analizarán en mayor profundidad en los pasos posteriores para evaluar su relevancia como posibles biomarcadores.

7. Análisis de Componentes Principales (PCA)

El Análisis de Componentes Principales (PCA) se aplica aquí para observar patrones en los **metabolitos que mostraron diferencias significativas** entre los grupos “*Before*” y “*After*” tras el análisis estadístico. Esto permite visualizar si estos metabolitos se agrupan de manera que refleje la diferencia entre las condiciones experimentales.

7.1 Preparación de los Datos para el PCA

Antes de realizar el PCA, se filtran los metabolitos significativos en función de los valores ajustados obtenidos en el análisis estadístico. Luego, se transponen los datos de expresión para que las muestras estén organizadas como filas, (paso necesario para el análisis de PCA en R).

```
##### PCA #####  
  
# Filtro los datos significativos:  
significant_metabolites <- rownames(metabolitos_significativos)  
expresion_significativa <- assay(se)[significant_metabolites, ]
```

7.2 Cálculo de los Componentes Principales

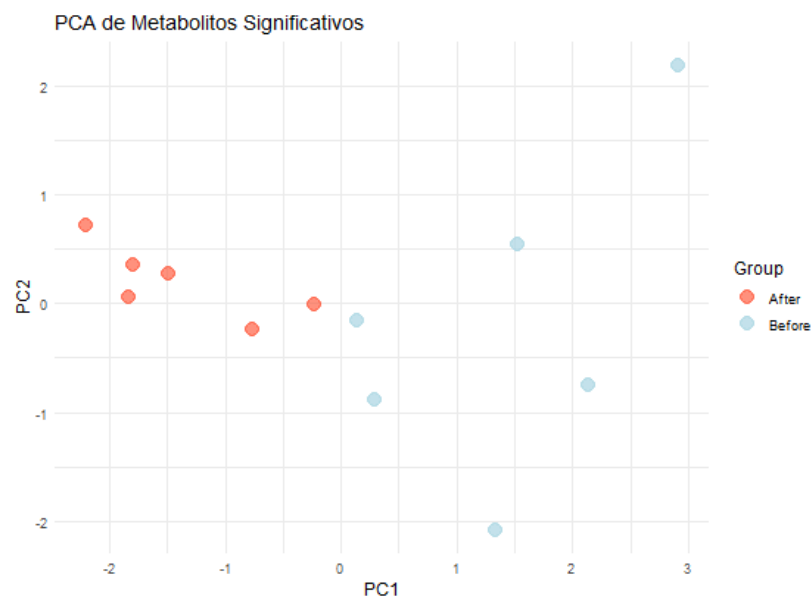
Para el PCA, trabajamos solo con los metabolitos que mostraron diferencias significativas en los análisis estadísticos. Primero, transponemos estos datos para que las muestras se organicen como filas, un formato necesario para el análisis de PCA en R. Luego, usamos la función **prcomp()** con la opción de escalado (**scale = TRUE**), lo que asegura que todos los metabolitos contribuyan de manera equilibrada al análisis.

```
pca_result <- prcomp(t(expresion_significativa), scale = TRUE)
```

7.3 Visualización de los Componentes Principales

Para visualizar los resultados, creamos un gráfico de dispersión que representa las muestras en el espacio de los dos primeros componentes principales (PC1 y PC2). Este gráfico permite ver si los grupos "Before" y "After" tienden a separarse de forma natural en función de los metabolitos significativos.

```
# gráfico:  
pca_df <- data.frame(PC1 = pca_result$x[, 1], PC2 = pca_result$x[, 2], Group = grupo_info$Groups)  
ggplot(pca_df, aes(x = PC1, y = PC2, color = Group)) +  
  geom_point(size = 4, alpha = 0.7) +  
  labs(title = "PCA de Metabolitos Significativos") +  
  scale_color_manual(values = c("tomato", "lightblue")) +  
  theme_minimal()
```



7.4 Conclusiones del PCA

En el gráfico, vemos cómo se distribuyen las muestras en un espacio reducido, donde los dos primeros componentes principales (PC1 y PC2) capturan la mayor parte de la variabilidad en los metabolitos significativos. Los componentes principales están contruidos a partir de las cargas de cada metabolito, que representan cuánto contribuye cada uno a la variabilidad capturada en PC1 y PC2. Aunque algunos metabolitos pueden estar impulsando estas direcciones, la ausencia de una separación clara entre los puntos de los grupos sugiere que las diferencias metabólicas entre "Before" y "After" no son suficientemente fuertes como para crear dos grupos bien definidos.

La falta de una separación evidente indicaría que los cambios metabólicos entre "Before" y "After" son demasiado sutiles, o que existe una variabilidad significativa dentro de cada grupo, lo que enmascara las diferencias experimentales. Esta dispersión puede reflejar tanto la variabilidad natural de las muestras como la influencia de factores individuales que afectan el perfil metabólico. Por lo tanto, aunque los metabolitos significativos capturan parte de la variabilidad entre las muestras, esta no es lo suficientemente consistente para diferenciar de manera clara los dos grupos en el análisis.

8. Análisis de Correlación y Redes

El análisis de correlación entre metabolitos nos ayuda a entender cómo estos se relacionan entre sí, lo que puede darnos pistas sobre rutas metabólicas compartidas o respuestas coordinadas.

Primero, visualizamos estas relaciones en un **mapa de calor de la matriz de correlación** y luego construimos **una red de interacciones** para centrarnos en las conexiones más fuertes y relevantes.

8.1 Cálculo de la Matriz de Correlación

Para analizar cómo se relacionan los metabolitos entre sí, calculé una **matriz de correlación** que mide la intensidad y dirección de la relación entre cada par de metabolitos significativos. Utilicé el método de *Spearman*, que es adecuado para capturar relaciones en datos que no necesariamente siguen una distribución normal y es menos sensible a valores extremos.

La matriz generada muestra valores que van de -1 a 1:

- Los valores cercanos a 1 indican una correlación positiva fuerte (los metabolitos aumentan o disminuyen juntos)
- Los valores cercanos a -1 indican una correlación negativa fuerte (los metabolitos tienen un comportamiento opuesto)
- Los valores cercanos a 0 sugieren una relación débil o inexistente.

Análisis de Datos Ómicos

Esta matriz de correlación nos da una visión inicial sobre cómo pueden agruparse o diferenciarse los metabolitos en función de sus patrones de expresión, y sienta las bases para el filtrado y la visualización de conexiones relevantes en la red.

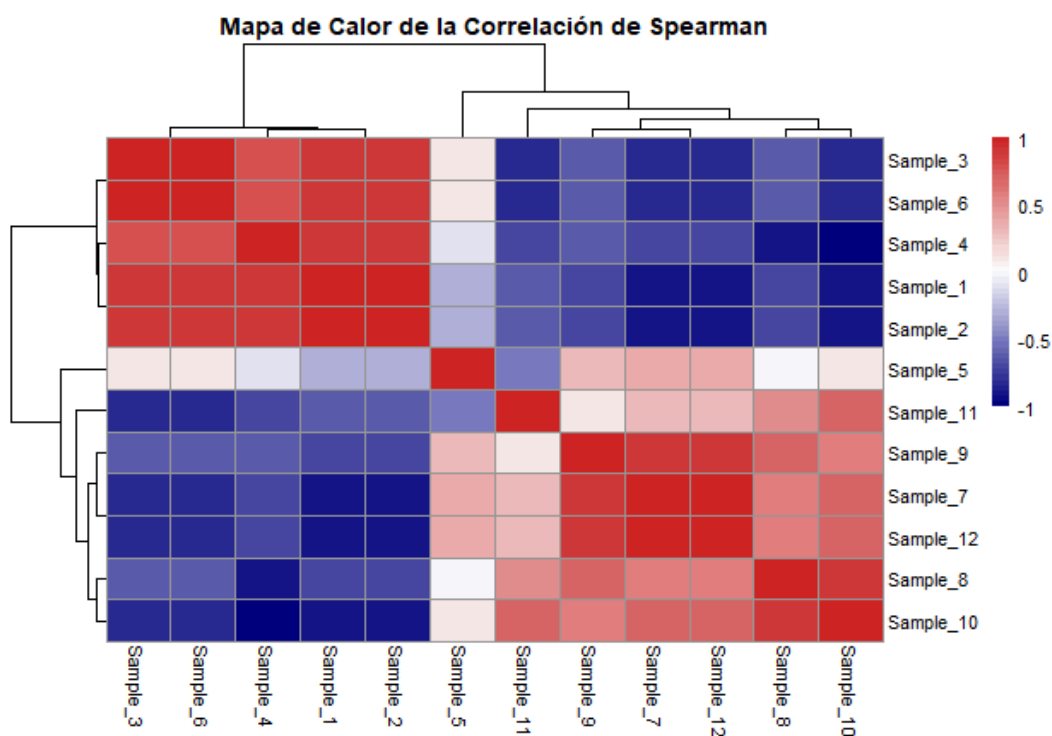
```
191
192- ##### Análisis de Correlación y Redes Metabólicas #####
193
194 # matriz de correlación y mapa de calor
195
196 correlation_matrix <- cor(t(expresion_significativa), method = "spearman")
197 head(correlation_matrix)
198
199 library(pheatmap)
200 pheatmap(correlation_matrix, color = colorRampPalette(c("navy", "white", "firebrick3"))(50),
201          main = "Mapa de calor de la correlación de Spearman")
202
203
```

204:1 Análisis de Correlación y Redes Metabólicas : Copilot: Not signed in.

Console Terminal Background Jobs

R 4.2.2 ~/projects/

```
> # matriz de correlación y mapa de calor
>
> correlation_matrix <- cor(t(expresion_significativa), method = "spearman")
> head(correlation_matrix)
      Sample_1 Sample_2 Sample_3 Sample_4 Sample_5 Sample_6 Sample_7 Sample_8 Sample_9 Sample_10
Sample_1      1.0      1.0      0.9      0.9     -0.3      0.9     -0.9     -0.7     -0.7     -0.9
Sample_2      1.0      1.0      0.9      0.9     -0.3      0.9     -0.9     -0.7     -0.7     -0.9
Sample_3      0.9      0.9      1.0      0.8      0.1      1.0     -0.8     -0.6     -0.6     -0.8
Sample_4      0.9      0.9      0.8      1.0     -0.1      0.8     -0.7     -0.9     -0.6     -1.0
Sample_5     -0.3     -0.3      0.1     -0.1      1.0      0.1      0.4      0.0      0.3      0.1
Sample_6      0.9      0.9      1.0      0.8      0.1      1.0     -0.8     -0.6     -0.6     -0.8
      Sample_11 Sample_12
Sample_1     -0.6     -0.9
Sample_2     -0.6     -0.9
Sample_3     -0.8     -0.8
Sample_4     -0.7     -0.7
Sample_5     -0.5      0.4
Sample_6     -0.8     -0.8
> library(pheatmap)
> pheatmap(correlation_matrix, color = colorRampPalette(c("navy", "white", "firebrick3"))(50),
+          main = "Mapa de calor de la correlación de Spearman")
>
```



La visualización en un **mapa de calor** facilita la identificación de patrones generales de correlación entre los metabolitos. Los colores nos muestran cómo se relacionan las muestras entre sí. Los tonos rojos indican correlaciones positivas, mientras que los tonos azules representan correlaciones negativas, sugiriendo que una muestra aumenta mientras la otra disminuye. Los tonos más claros en ambos lados reflejan correlaciones más débiles. Así, podemos ver agrupaciones de muestras con patrones similares, como **Sample_3**, **Sample_6**, **Sample_4**, y **Sample_2**, lo cual sugiere que podrían estar involucradas en rutas metabólicas similares. En cambio, **Sample_5** y **Sample_10** parecen estar menos alineadas con el resto, mostrando patrones de correlación distintos que podrían indicar respuestas diferentes.

También destacan **Sample_9** y **Sample_12**, que tienen correlaciones positivas con varias otras muestras, lo que podría indicar un rol central en la red metabólica, como si fueran nodos. **Sample_5**, por otro lado, muestra varias correlaciones negativas, lo que sugiere que responde de manera opuesta al resto.

8.2 Filtrado de Correlaciones Significativas

Para hacer el análisis más claro, apliqué un filtro que conserva solo las correlaciones fuertes y estadísticamente significativas. Nos quedamos con correlaciones con valores absolutos mayores de 0.7 (positivas o negativas) para eliminar relaciones débiles que puedan añadir ruido a la red.

Este filtrado permite que la red resultante represente únicamente las conexiones más relevantes entre los metabolitos.


```
# Filtrado de correlaciones
cor_matrix <- correlation_matrix
cor_matrix[abs(cor_matrix) < 0.7] <- 0 # así solo conserva correlaciones fuertes
```

8.3 Construcción de la Red de Correlación

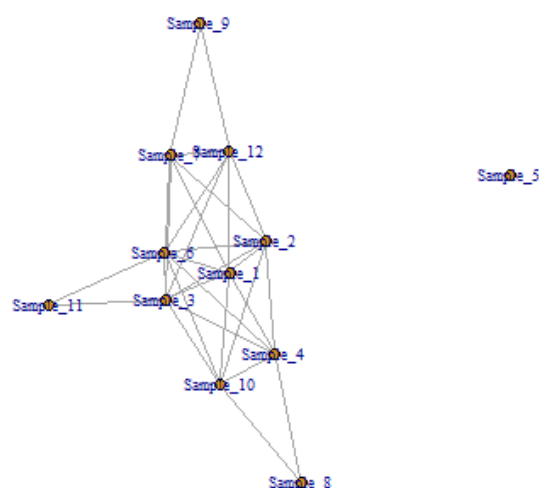
Con la matriz de correlación filtrada, construí una **red de correlación** usando el paquete **igraph**. En esta red, cada nodo representa un metabolito, y las conexiones entre ellos indican correlaciones fuertes y significativas. Los nodos con muchas conexiones pueden ser de especial interés, ya que podrían actuar como “centros” en la red metabólica y reflejar un rol importante en la respuesta experimental. El gráfico de la red muestra cómo los metabolitos se interconectan, destacando aquellos que tienen un papel clave debido a su alta conectividad.

```
# igraph y construcción de la red
if (!requireNamespace("igraph", quietly = TRUE)) install.packages("igraph")
library(igraph)

cor_graph <- graph_from_adjacency_matrix(cor_matrix, mode = "undirected", weighted = TRUE, diag = FALSE)
cor_graph <- simplify(cor_graph)

# visualización de la red
plot(cor_graph, vertex.label = v(cor_graph)$name, vertex.size = 5, vertex.label.cex = 0.7,
     main = "Red Metabólica Basada en Correlación Significativa")
```

Red Metabólica Basada en Correlación Significativa



8.4 Interpretación del análisis de redes

La red de correlación revela cómo se relacionan los metabolitos entre sí, destacando patrones clave de interacción. Las conexiones fuertes sugieren que ciertos metabolitos actúan en conjunto dentro de rutas metabólicas o responden de manera coordinada a la condición experimental.

Las **correlaciones positivas** indican que algunos metabolitos aumentan o disminuyen de manera conjunta. Por otra parte, las **correlaciones negativas** pueden reflejar respuestas opuestas o funciones divergentes.

Al observar los nodos individuales, noté que algunos, como **Sample_5** y **Sample_11**, están más aislados en la red, con menos conexiones, lo que sugiere que estos metabolitos podrían responder de manera menos coordinada o estar en rutas más independientes, en contraste con nodos como **Sample_3** y **Sample_9** aparecen como **nodos centrales** debido a su alta conectividad, lo que los convierte en posibles reguladores clave en la red metabólica. Estos metabolitos centrales son especialmente interesantes como candidatos a biomarcadores o reguladores de procesos específicos en esta condición experimental. En conjunto, la red de correlación nos ofrece una representación clara de las interacciones y resalta tanto los metabolitos clave como los agrupamientos de interés, estableciendo una base sólida para explorar en detalle las rutas y procesos metabólicos afectados en este contexto.

8.5 Conclusiones de ambas aproximaciones

Las aproximaciones de análisis, tanto mediante el mapa de calor como la red de correlación, me han permitido comprender mejor las relaciones entre los metabolitos y resaltar aquellos con potencial para ser claves en la respuesta metabólica. El **mapa de calor** nos brinda una visión global de las correlaciones, permitiéndonos identificar patrones de co-expresión y agrupamientos de muestras que podrían estar participando en procesos metabólicos comunes o respondiendo de manera similar.

Por otro lado, la **red de correlación** complementa esta visión al mostrarnos las conexiones más significativas entre los metabolitos. En la red, algunos nodos como **Sample_3** y **Sample_9** aparecen como **nodos centrales** con muchas conexiones, sugiriendo su rol como posibles reguladores metabólicos clave (y confirmando los resultados del mapa de calor). Otros, como **Sample_5** y **Sample_11**, están más aislados, lo que podría indicar una respuesta más específica o independiente de la red general (**Sample_5** se muestra desconectada de la red, indicando más aún su lejanía respecto a las otras muestras). En conjunto, ambas aproximaciones ofrecen una representación clara de las interacciones y patrones, resaltando tanto los metabolitos clave como los agrupamientos de interés.

9. Validación de Biomarcadores con PLS-DA

El Análisis de Discriminante de Proyecciones Latentes Parcial (PLS-DA) es una técnica supervisada que nos permite identificar metabolitos que contribuyen a la diferenciación entre grupos. En este caso, se ha aplicado PLS-DA para verificar si los metabolitos seleccionados permiten distinguir con precisión entre los grupos *"Before"* y *"After"* y determinar cuáles pueden ser considerados como biomarcadores relevantes.

9.1 Preparación de los Datos para PLS-DA

Antes de aplicar el PLS-DA, seleccioné los metabolitos que mostraron diferencias significativas en sus niveles de expresión entre los grupos. Dado que el PLS-DA maximiza la separación entre los grupos utilizando estas variables predictoras, es importante que los datos estén bien normalizados. Aquí, utilicé el paquete **ropls** en R para implementar el modelo y generar visualizaciones que faciliten la interpretación de los resultados.

```
##### validación de Biomarcadores con PLS-DA #####  
  
# me aseguro que los paquetes se han cargado  
  
if (!requireNamespace("ropls", quietly = TRUE)) BiocManager::install("ropls")  
library(ropls)  
  
# hago el escalado:  
  
expresion_significativa <- scale(t(expresion_significativa))
```

9.2 Creación y Resumen del Modelo PLS-DA

Con los datos preparados, construí el modelo PLS-DA para ver cómo se comportan los grupos *"Before"* y *"After"* en función de los metabolitos seleccionados. Usé el paquete **ropls** para crear el modelo y obtener un resumen que nos muestra el nivel de ajuste del modelo y su capacidad para predecir la clasificación de nuevas muestras.

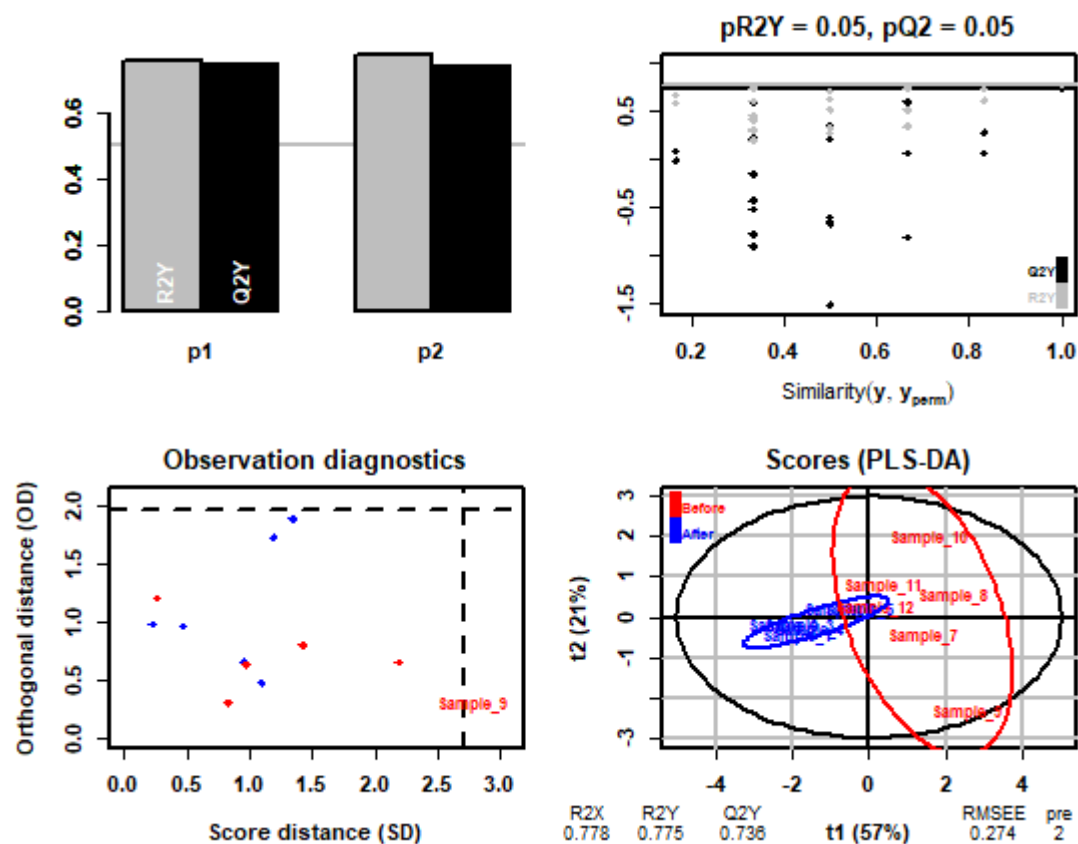
```
# creo el modelo:  
  
plsda_model <- oplr(expresion_significativa, grupo_info$Groups, predI = 2)  
summary(plsda_model)
```

9.3 Visualización del Modelo PLS-DA

Las visualizaciones del modelo PLS-DA muestran cómo los metabolitos seleccionados permiten diferenciar claramente entre los grupos "Before" y "After".

```
#visualizo los resultados:

#resumen del modelo
plot(plsda_model, typevc = "Resumen del Modelo")
#visualización de las cargas
plot(plsda_model, typevc = "Cargas")
```



En el resumen del Modelo, las métricas **R2Y** (0.778) y **Q2Y** (0.736) indican que el modelo captura bien la variabilidad entre los grupos y tiene una capacidad predictiva adecuada. Los valores "**p**" de 0.05 para R2Y y Q2Y refuerzan esta robustez estadística. En el gráfico de **Diagnóstico de Observaciones (Observation Diagnostics)**, **Sample_9** se destaca como un valor atípico, lo que sugiere un perfil metabólico distinto. Finalmente, el gráfico de **Scores (Scores PLS-DA)** muestra una separación clara entre "Before" y "After", con cada grupo en áreas distintas y elipses que reflejan su variabilidad interna, que respalda la capacidad de los metabolitos para diferenciar las condiciones experimentales.

9.4 Validación del Modelo

La validación del modelo confirma que el PLS-DA es sólido, con valores altos de **R²_Y** y **Q²_Y**, y valores “**p**” de 0.05 que avalan su significancia estadística. Esto sugiere que el modelo no solo ajusta bien los datos, sino que también puede predecir correctamente nuevas muestras.

En conjunto, estos resultados destacan la capacidad de los metabolitos seleccionados para distinguir entre las distintas condiciones experimentales. Metabolitos clave como *levanbiosa*, *piruvato* y *ácido láctico* aparecen como responsables de esta diferenciación, y **Sample_9** es un valor atípico a considerar en futuras investigaciones.

9.5 Conclusiones del Modelo

El análisis PLS-DA ha sido clave para identificar metabolitos que diferencian de forma efectiva los grupos “*Before*” y “*After*”.

Los altos valores de **R²_Y** y **Q²_Y** indican que el modelo no solo explica bien las diferencias entre los grupos, sino que también tiene una capacidad predictiva sólida, lo que refuerza la fiabilidad de estos resultados. Así mismo, metabolitos como *levanbiosa*, *piruvato* y *ácido láctico* se destacan como factores importantes en la separación entre grupos, convirtiéndose en biomarcadores prometedores para estudios futuros. La aparición de **Sample_9** como un valor atípico sugiere que podría haber respuestas individuales o particularidades únicas que merecen un análisis más profundo.

En conjunto, estos resultados proporcionan una base sólida para investigar más a fondo los cambios metabólicos en contextos experimentales similares y explorar el potencial de estos biomarcadores en aplicaciones clínicas o experimentales.

10. Documentación y Reporte

La documentación de este análisis de datos ómicos es clave para que los métodos y hallazgos sean claros y fáciles de seguir. He escrito un reporte detallado que cubre desde la carga de datos hasta la validación de biomarcadores con PLS-DA, que servirá como protocolo de referencia tanto para futuros estudios como para quienes busquen replicar o aprender de este análisis.

10.1 Estructura del Reporte

El reporte sigue una estructura sencilla que cubre cada paso del análisis:

Análisis de Datos Ómicos

- **Introducción:** Define los objetivos y la relevancia de este estudio.
- **Materiales y Métodos:** Explica cómo se procesaron los datos y las herramientas empleadas.
- **Carga y Exploración de Datos:** Presenta la carga de datos y la exploración inicial con gráficos para entender las distribuciones de los metabolitos.
- **Preparación y Limpieza de Datos:** Describe el tratamiento de valores ausentes, transformación y normalización de los datos.
- **Visualización de Datos:** Muestra gráficos como histogramas y boxplots con interpretaciones.
- **Análisis Estadístico:** Explica las pruebas realizadas, como “*t*” de *Student* y ajustes de *p*-valor.
- **Análisis PCA y Redes de Correlación:** Documenta la reducción de dimensionalidad y el análisis de redes de correlación.
- **Validación de Biomarcadores con PLS-DA:** Detalla el análisis PLS-DA y sus resultados, confirmando la eficacia de ciertos metabolitos para distinguir entre “*Before*” y “*After*”.
- **Conclusiones y Recomendaciones:** Resume los hallazgos clave y sugiere pasos para estudios futuros.

10.2 Herramientas de Documentación y Reproducibilidad

Para facilitar la claridad y la posibilidad de repetir el análisis, se usaron las siguientes herramientas:

- **R Markdown:** Integra el código con el reporte en un solo documento reproducible, generando archivos fáciles de compartir y revisar.
- **GitHub:** Almacena el código y el reporte, permitiendo control de versiones y colaboración.
- **Comentarios en el Código:** Cada bloque de código está comentado, lo que facilita su comprensión para otros usuarios.

También se incluyeron las versiones de R y los paquetes utilizados, junto con enlaces a los datos originales en Metabolomics Workbench, para que cualquiera pueda replicar o comparar el análisis en un entorno similar.

10.3 Conclusiones y Recomendaciones para Estudios Futuros

El reporte concluye con un resumen de los hallazgos y algunas recomendaciones para futuros estudios:

- **Validación en Nuevas Cohortes:** Recomendamos validar estos biomarcadores en otros conjuntos de datos para confirmar su utilidad.
- **Análisis Funcional de Metabolitos Clave:** Explorar en profundidad el papel de los metabolitos clave en rutas metabólicas.
- **Probar Otros Modelos Supervisados:** Experimentar con otros enfoques, como *Random Forest*, para comparar su precisión en la predicción de biomarcadores.

La creación de un protocolo nos asegura que el análisis de datos ómicos sea claro, reproducible y útil para otros investigadores. La estructura y las herramientas de documentación utilizadas por lo general de código abierto permiten que el análisis sea accesible y replicable, que los resultados puedan revisarse y aplicarse en estudios futuros, ofreciendo una base sólida para explorar cambios metabólicos y el uso de biomarcadores en investigaciones experimentales.

11. Conclusiones

El análisis metabolómico llevado a cabo en este estudio ha revelado diferencias importantes entre los grupos *"Before"* y *"After"*, identificando metabolitos con patrones de expresión específicos que pueden ser clave para entender la respuesta de los sistemas biológicos a la condición experimental.

Con la creación de un protocolo que integró varios métodos —desde pruebas estadísticas y PCA hasta PLS-DA y redes de correlación— logré construir un perfil completo de las interacciones y cambios metabólicos asociados. Este enfoque integral me ha permitido no solo identificar metabolitos de interés, sino también entender mejor la estructura y dinámica de las redes metabólicas implicadas.

Uno de los hallazgos más destacados fue la identificación de metabolitos que podrían servir como biomarcadores, ya que mostraron diferencias significativas en sus niveles de expresión entre los grupos. Metabolitos como la *levanbiosa* y el *piruvato* demostraron una capacidad alta para diferenciar los grupos, lo que los convierte en candidatos interesantes para futuros estudios sobre su potencial en el diagnóstico o monitoreo de condiciones experimentales similares, además, el análisis de redes de correlación ofreció una perspectiva visual de las interacciones metabólicas, donde se identificaron agrupaciones y nodos centrales que podrían representar rutas reguladoras coordinadas. Estas redes no solo revelan interacciones metabólicas, sino que también sugieren adaptaciones funcionales a la condición experimental.

El uso de métodos de validación, como el ajuste de valores *"p"* y la validación cruzada en el modelo PLS-DA, me permitió asegurar que los resultados fueran sólidos y minimizar el riesgo de falsos positivos. Esto añade una capa de robustez al análisis y refuerza la credibilidad de los hallazgos, haciendo que los biomarcadores propuestos tengan un respaldo estadístico firme y puedan servir como base para futuros estudios. Sin embargo,

Análisis de Datos Ómicos

es importante reconocer algunas limitaciones, como el tamaño de muestra limitado, que podría influir en la generalización de los resultados. Aumentar el tamaño de muestra en estudios futuros permitiría evaluar con mayor precisión la relevancia de estos biomarcadores y reducir la varianza en los análisis.

Pensando en futuros estudios, me gustaría validar estos biomarcadores en otras cohortes para confirmar su aplicabilidad en contextos distintos. También sería enriquecedor integrar estos datos metabolómicos con otras "ómicas", como la *transcriptómica* o la *proteómica*, para obtener una visión más global de las respuestas celulares y de cómo los cambios en los genes y proteínas se relacionan con los cambios metabólicos. Además, los metabolitos que actúan como nodos centrales en la red de correlación o presentan altas cargas en el PLS-DA que podría investigar a fondo para entender su papel en rutas metabólicas específicas, lo que posibilitará más apertura a nuevas terapias o intervenciones en el ámbito biomédico.

En conclusión, este estudio destaca el valor del análisis metabolómico como una herramienta poderosa para detectar cambios metabólicos y construir un panorama detallado de las respuestas celulares en condiciones experimentales. Con más validación, los biomarcadores identificados aquí podrían tener aplicaciones prácticas en el diagnóstico, monitoreo y tratamiento de condiciones asociadas al metabolismo, marcando un avance importante en la investigación biomédica. El enfoque integral en la creación de un protocolo y los métodos aplicados en capas no solo nos han permitido obtener resultados sólidos, sino que también sentaron las bases para futuras exploraciones en la identificación de perfiles metabólicos y biomarcadores con potencial terapéutico.

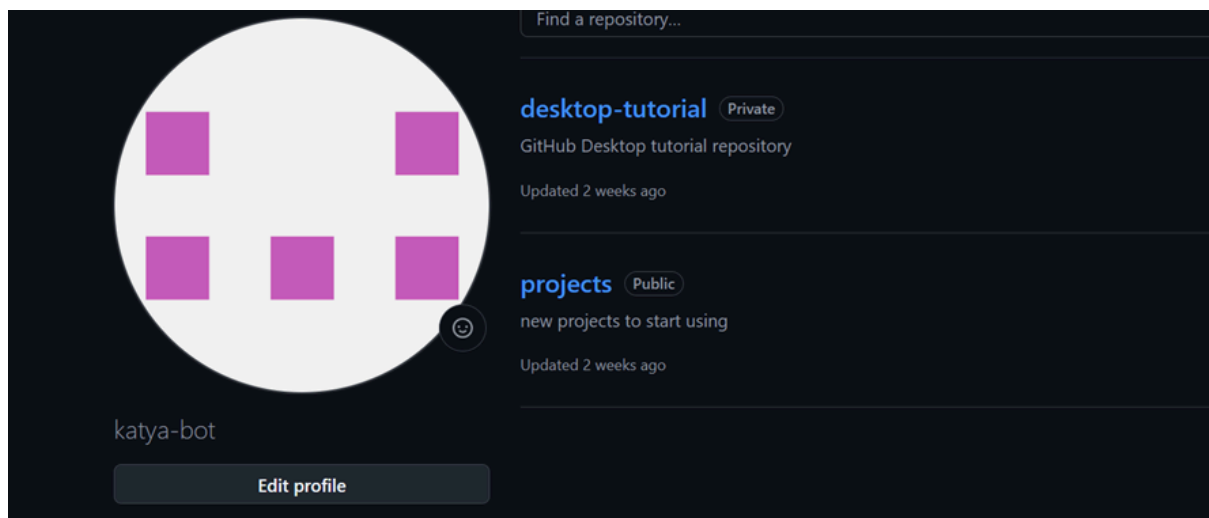
Anexo

❖ Creación del objeto contenedor .Rda

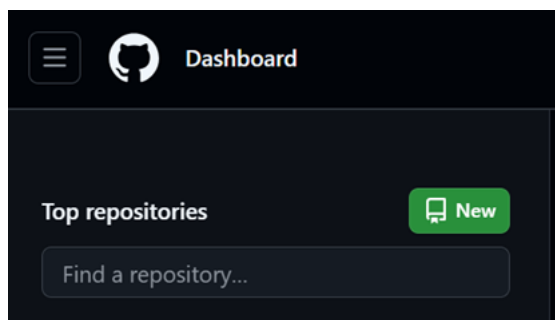
```
#####  
#####  
##### DESCARGA DEL CONTENEDOR #####  
  
# simplemente guardo el archivo con extensión .Rda y lo pongo en la ruta donde  
# tengo mi repositorio local vinculado a GitHub  
  
save(se, file = "c:/Users/Usuario/Desktop/Puello-Mora-Catherine-PEC1/  
SE_CatherinePM_PEC1.Rda")
```

❖ Reposición de los datos en GitHub

Para subir los datos a GitHub primero me creé una cuenta:



Después creé el repositorio, lo he llamado igual que el nombre de la entrega:




Create a new repository

A repository contains all project files, including the revision history. Already have a project repository elsewhere? [Import a repository.](#)

Required fields are marked with an asterisk (*).

Owner *

 katya-bot

Repository name *

Puello-Mora-Cayherine-PE

✔ Puello-Mora-Cayherine-PEC1 is available.

Great repository names are short and memorable. Need inspiration? How about [bookish-journey](#) ?

Description (optional)


OMICS PEC 1



Public

Anyone on the internet can see this repository. You choose who can commit.

Quick setup — if you've done this kind of thing before

 Set up in Desktop

or

HTTPS

SSH

<https://github.com/katya-bot/Puello-Mora-Cayherine-PEC1.git>

Get started by [creating a new file](#) or [uploading an existing file](#). We recommend every repository include a [README](#), [LICENSE](#), and [.gitignore](#).

...or create a new repository on the command line

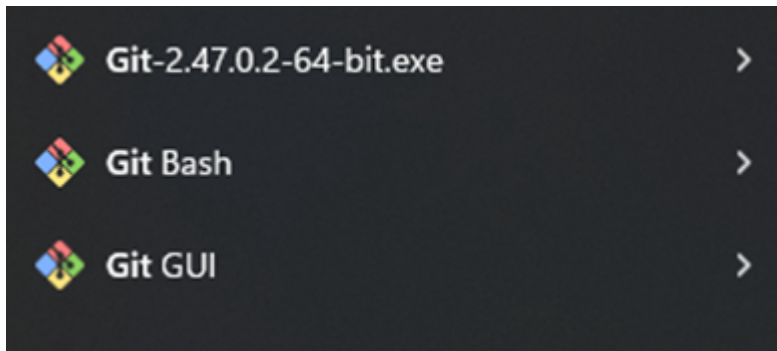
```
echo "# Puello-Mora-Cayherine-PEC1" >> README.md
git init
git add README.md
git commit -m "first commit"
git branch -M main
git remote add origin https://github.com/katya-bot/Puello-Mora-Cayherine-PEC1.git
git push -u origin main
```

...or push an existing repository from the command line

```
git remote add origin https://github.com/katya-bot/Puello-Mora-Cayherine-PEC1.git
git branch -M main
git push -u origin main
```

Una vez creado me dio los datos del repositorio con mis archivos locales, me conecté a mi Shell de Windows para a través de Git (ya lo tenía descargado) vincular ambas carpetas:

Análisis de Datos Ómicos



```
Windows PowerShell
PS C:\Users\Usuario\desktop> cd Puello-Mora-Catherine-PEC1
PS C:\Users\Usuario\desktop\Puello-Mora-Catherine-PEC1> git init
Initialized empty Git repository in C:/Users/Usuario/Desktop/Puello-Mora-Catherine-PEC1/.git/
PS C:\Users\Usuario\desktop\Puello-Mora-Catherine-PEC1> git init
Reinitialized existing Git repository in C:/Users/Usuario/Desktop/Puello-Mora-Catherine-PEC1/.git/
PS C:\Users\Usuario\desktop\Puello-Mora-Catherine-PEC1> git remote add origin https://github.com/k
atya-bot/Puello-Mora-Cayherine-PEC1.git
PS C:\Users\Usuario\desktop\Puello-Mora-Catherine-PEC1> git status
On branch master

No commits yet

nothing to commit (create/copy files and use "git add" to track)
PS C:\Users\Usuario\desktop\Puello-Mora-Catherine-PEC1> git add .
PS C:\Users\Usuario\desktop\Puello-Mora-Catherine-PEC1> git status
On branch master

No commits yet

Changes to be committed:
  (use "git rm --cached <file>..." to unstage)
    new file:   Catherine_Puello_OMICS_PEC_1.R
PS C:\Users\Usuario\desktop\Puello-Mora-Catherine-PEC1> git commit -m "agrego archivo PEC_1 R"
[master (root-commit) 955906a] agrego archivo PEC_1 R
 1 file changed, 215 insertions(+)
 create mode 100644 Catherine_Puello_OMICS_PEC_1.R
PS C:\Users\Usuario\desktop\Puello-Mora-Catherine-PEC1> git status
On branch master
nothing to commit, working tree clean
```

Subo los archivos vinculados:

Análisis de Datos Ómicos

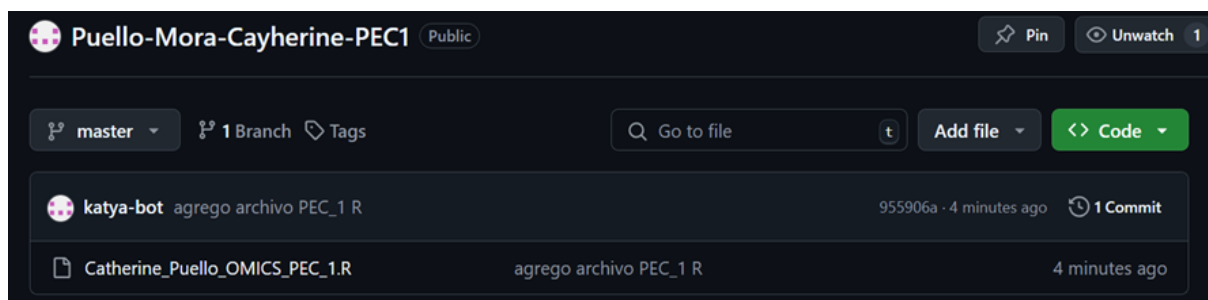
```
nothing to commit, working tree clean
PS C:\Users\Usuario\desktop\Puello-Mora-Catherine-PEC1> git push
fatal: The current branch master has no upstream branch.
To push the current branch and set the remote as upstream, use

    git push --set-upstream origin master

To have this happen automatically for branches without a tracking
upstream, see 'push.autoSetupRemote' in 'git help config'.

PS C:\Users\Usuario\desktop\Puello-Mora-Catherine-PEC1> git push --set-upstream origin master
Enumerating objects: 3, done.
Counting objects: 100% (3/3), done.
Delta compression using up to 16 threads
Compressing objects: 100% (3/3), done.
Writing objects: 100% (3/3), 3.03 KiB | 3.03 MiB/s, done.
Total 3 (delta 0), reused 0 (delta 0), pack-reused 0 (from 0)
To https://github.com/katya-bot/Puello-Mora-Catherine-PEC1.git
 * [new branch]      master -> master
branch 'master' set up to track 'origin/master'.
PS C:\Users\Usuario\desktop\Puello-Mora-Catherine-PEC1> |
```

Compruebo en GitHub si se ha subido mi cambio:

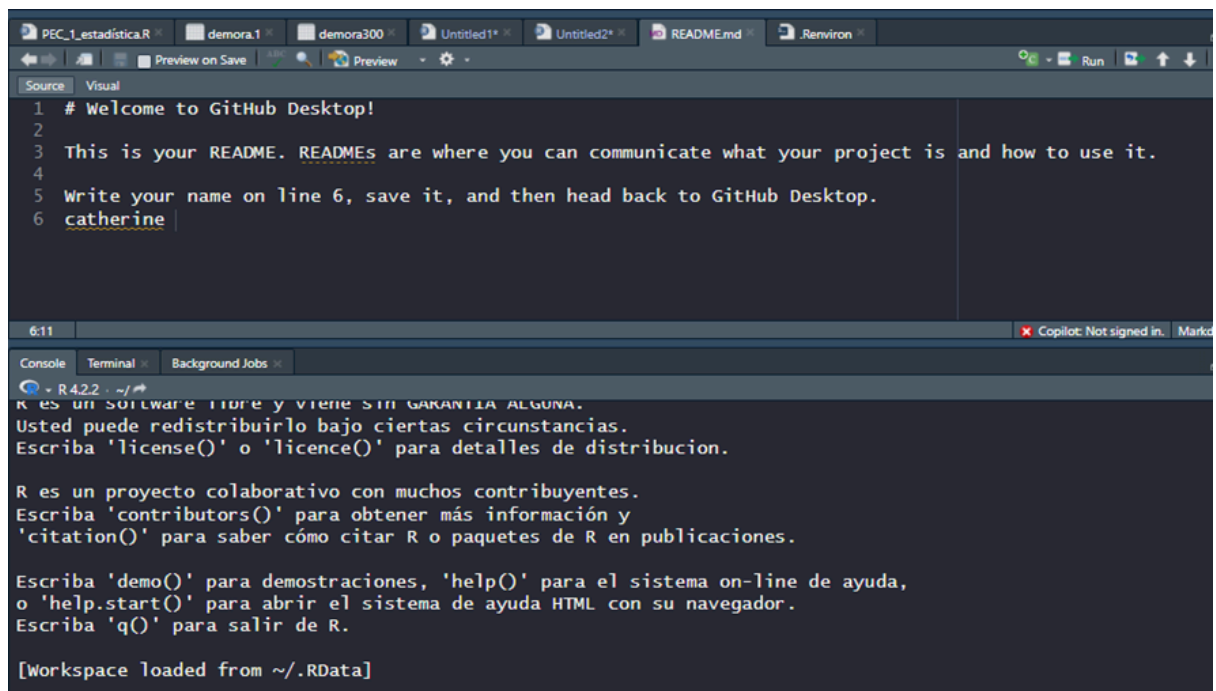


Otros datos:

Instalando el token para unir R con GitHub:

```
The downloaded source packages are in
      'C:\Users\Usuario\AppData\Local\Temp\RtmpojIMAz\downloaded_packages'
> usethis::edit_r_environ(token)
Error in match.arg(scope) : objeto 'token' no encontrado
> usethis::edit_r_environ()
❑ Modify C:/Users/Usuario/Documents/.Renvirom.
❑ Restart R for changes to take effect.
> |
```

Análisis de Datos Ómicos



The screenshot shows the GitHub Desktop interface. The top pane displays a README file with the following content:

```
1 # Welcome to GitHub Desktop!
2
3 This is your README. READMEs are where you can communicate what your project is and how to use it.
4
5 Write your name on line 6, save it, and then head back to GitHub Desktop.
6 catherine
```

The bottom pane shows the R console with the following output:

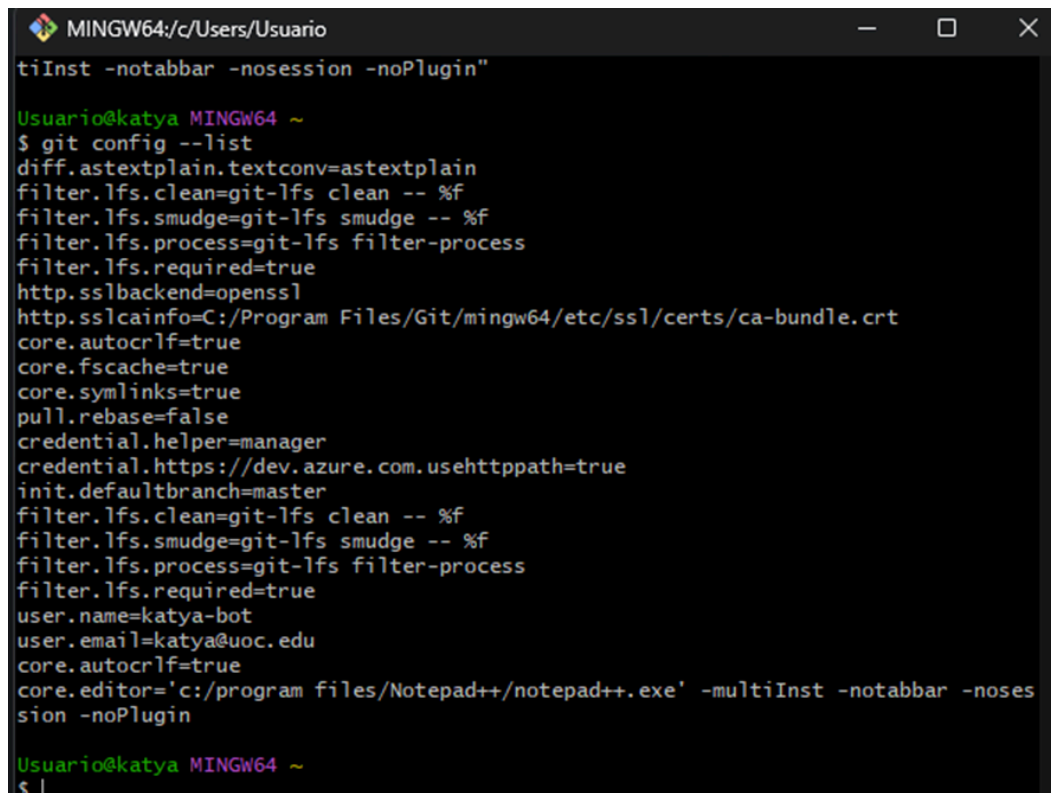
```
R 4.2.2 ~/>
R es un software libre y viene sin GARANTIA ALGUNA.
Usted puede redistribuirlo bajo ciertas circunstancias.
Escriba 'license()' o 'licence()' para detalles de distribucion.

R es un proyecto colaborativo con muchos contribuyentes.
Escriba 'contributors()' para obtener más información y
'citation()' para saber cómo citar R o paquetes de R en publicaciones.

Escriba 'demo()' para demostraciones, 'help()' para el sistema on-line de ayuda,
o 'help.start()' para abrir el sistema de ayuda HTML con su navegador.
Escriba 'q()' para salir de R.

[Workspace loaded from ~/.RData]
```

Configuración de git:



The screenshot shows a MINGW64 terminal window with the following commands and output:

```
tiInst -notabbar -nosession -noPlugin"

Usuario@katya MINGW64 ~
$ git config --list
diff.astextplain.textconv=astextplain
filter.lfs.clean=git-lfs clean -- %f
filter.lfs.smudge=git-lfs smudge -- %f
filter.lfs.process=git-lfs filter-process
filter.lfs.required=true
http.sslbackend=openssl
http.sslcainfo=C:/Program Files/Git/mingw64/etc/ssl/certs/ca-bundle.crt
core.autocrlf=true
core.fscache=true
core.symlinks=true
pull.rebase=false
credential.helper=manager
credential.https://dev.azure.com.usehttppath=true
init.defaultbranch=master
filter.lfs.clean=git-lfs clean -- %f
filter.lfs.smudge=git-lfs smudge -- %f
filter.lfs.process=git-lfs filter-process
filter.lfs.required=true
user.name=katya-bot
user.email=katya@uoc.edu
core.autocrlf=true
core.editor='c:/program files/Notepad++/notepad++.exe' -multiInst -notabbar -nosession -noPlugin

Usuario@katya MINGW64 ~
$
```

Análisis de Datos Ómicos

```
user.name=katya-bot
user.email=katya@uoc.edu
core.autocrlf=true
core.editor='c:/program files/Notepad++/notepad++.exe' -multiInst -notabbar -nosession -noPlugin

Usuario@katya MINGW64 ~
$ mkdir Omics

Usuario@katya MINGW64 ~
$ cd Omics

Usuario@katya MINGW64 ~/Omics
$ git init
Initialized empty Git repository in C:/Users/Usuario/Omics/.git/

Usuario@katya MINGW64 ~/Omics (master)
$ ls
$ ls -a
./ ../ .git/

Usuario@katya MINGW64 ~/Omics (master)
$ git status
On branch master

No commits yet

nothing to commit (create/copy files and use "git add" to track)

Usuario@katya MINGW64 ~/Omics (master)
$ cd

Usuario@katya MINGW64 ~
$ mkdir pec_1_omics

Usuario@katya MINGW64 ~
$ cd pec_1_omics

Usuario@katya MINGW64 ~/pec_1_omics
$ git init
Initialized empty Git repository in C:/Users/Usuario/pec_1_omics/.git/

Usuario@katya MINGW64 ~/pec_1_omics (master)
$
```