

Análisis de datos ómicos (M0-157)

Primera prueba de evaluación continua

Jesús Castro Puntero

28 de marzo, 2025

Contents

1. Abstract	2
2. Objetivos	2
3. Métodos	2
4. Resultados	3
4.1 Carga del objeto <code>SummarizedExperiment</code>	3
4.2 Análisis de la estructura de datos	4
4.3 Análisis exploratorio inicial de los datos	6
5. Discusión	9
6. Conclusiones	10
7. Anexos	10
7.1 Repositorio GitHub	10
7.2 Referencias	10

1. Abstract

En este trabajo se ha llevado a cabo un análisis de datos metabolómicos utilizando principalmente el paquete POMA [1], adoptando el workflow propio de éste, para la exploración y procesamiento de los datos. Se ha trabajado con el resultado de los análisis llevados a cabo en el estudio ST000291[2], los cuales han sido integrados y trabajados a través de un objeto **SummarizedExperiment**. Se ha estudiado el origen de estos datos y se han procesado a partir de los datos curados proporcionados, preparándolos para un análisis más profundo y aplicando sPSL-DA como técnica de reducción de la dimensionalidad de los datos.

Se han identificado diferencias en el metabolismo tras la ingesta de distintos zumos, lo que podría afectar a la biosíntesis de procianidinas, compuestos con potenciales beneficios para la salud metabólica. Dado su posible papel en la sensibilidad a la insulina, sería necesario un estudio más profundo para evaluar su impacto.

2. Objetivos

En esta PEC ejecutaremos un análisis exploratorio de los datos del estudio *LC-MS Based Approaches to Investigate Metabolomic Differences in the Urine of Young Women after Drinking Cranberry Juice or Apple Juice* (ID del estudio: ST000291). Este estudio tiene como objetivo principal investigar los cambios metabólicos generales inducidos por la ingesta de concentrados de procianidinas provenientes de zumo de arándanos y zumo de manzana. Para ello, se utilizó un enfoque metabolómico basado en cromatografía líquida y espectrometría de masas (LC-MS), que permite una visión global de los metabolitos presentes en las muestras biológicas [3]. Este trabajo de la Universidad de Florida es un caso de estudio comparativo o de expresión diferencial ya que su objetivo es buscar las diferencias entre los perfiles metabólicos provenientes de los dos tipos de zumo.

Las procianidinas son un grupo importante de moléculas bioactivas conocidas por sus beneficios para la salud humana, mostrando un gran potencial en el tratamiento de enfermedades metabólicas crónicas como el cáncer, la diabetes y las enfermedades cardiovasculares, ya que previenen el daño celular relacionado con el estrés oxidativo [4]. Estudios como los de Y. Liu et al. [5] y N. Gonzalez-Abuin et al. [6] apoyan esta idea, sugiriendo que las procianidinas tienen efectos beneficiosos en la mejora a la resistencia a la insulina y la regulación de la glucosa, en parte a través de la modulación de la microbiota intestinal y la inhibición de vías inflamatorias.

Conocer el comportamiento y la absorción de estas moléculas puede ser de gran importancia dado su efecto en la salud humana y su amplia disponibilidad como residuo de productos agro-industriales [4]. El objetivo de este documento será realizar una exploración inicial de los datos, enfocada dentro del proceso de análisis de datos ómicos, construyendo un objeto **SummarizedExperiment**, poner en práctica el workflow típico del paquete POMA que nos permita generar representaciones gráficas, procesar los datos y realizar un análisis superficial de lo obtenido con el fin de generar una primera hipótesis de trabajo.

3. Métodos

Para la obtención de las muestras biológicas se reclutaron 18 mujeres universitarias sanas (21-29 años, IMC 18.5-25) y se aplicó un protocolo para la obtención de las muestras biológicas. Durante los primeros seis días, las participantes evitaron alimentos ricos en procianidinas. El séptimo día, tras un ayuno nocturno, se recogieron muestras basales de orina. Las participantes fueron asignadas aleatoriamente en dos grupos (n=9) para consumir zumo de arándano o de manzana durante tres días (días 7-9). El décimo día, tras un nuevo ayuno nocturno, se recogieron muestras de orina. Tras un período de lavado de dos semanas, las participantes cambiaron al otro tipo de zumo y repitieron el protocolo. Se excluyeron tres participantes del análisis: una por falta de asistencia y dos por no proporcionar muestras de orina. Cabe destacar que, aunque en el estudio se menciona la obtención de muestras biológicas de plasma incluidas dentro de el protocolo, no se publicaron los datos relativos al análisis de las mismas.

Las muestras biológicas fueron analizadas mediante un enfoque metabolómico basado en cromatografía líquida acoplada a espectrometría de masas (LC-MS) para el análisis de los metabolitos. Para la cromatografía líquida se utilizó el sistema Thermo Scientific-Dionex Ultimate 3000 con una columna ACE Excel 2 C18-PFP (100 x 2.1mm, 2µm). Para la espectrometría de masas se utilizó el instrumento Thermo Q Exactive Orbitrap con los modos de ionización positivos y negativos. En el protocolo de preparación de muestras (Sampleprep ID: SP000319) se detalla que los datos crudos obtenidos del análisis se convierten a archivos .MZxml y procesados con MZmine y Metaboanalyst. De esta manera, el estudio proporciona dos análisis complementarios: el análisis AN000464 en modo positivo y el análisis AN000465 en modo negativo.

Previo al trabajo de este documento, los nombres de los metabolitos han sido identificados a través de la herramienta de web scrapping `rvest` para obtener los identificadores genéricos de PubChem, KEGG y sus nombres. Los datos, la información de los metabolitos y la información de las covariables se han almacenado en el archivo ST000291.xlsx. Para la realización de esta PEC, se usará el archivo ST000291_curated.xlsx que identifica los datos con el metabolito correspondiente.

Para el posterior análisis, se usarán principalmente el paquete POMA [1] que nos permitirá construir objetos de `SummarizedExperiment` y aplicar métodos predefinidos para la elaboración de análisis y gráficos. También recalamos la importancia del paquete `metabolomicsWorkbenchR` [7] que nos permite recuperar, entre otros, la información de los metadatos del análisis original. Al final del documento se muestra en detalle el conjunto de paquetes usados para ejecución del código.

4. Resultados

4.1 Carga del objeto `SummarizedExperiment`

Como se ha comentado en el apartado anterior, para llevar a cabo nuestra exploración de los datos, utilizaremos principalmente el paquete POMA, que facilitará en gran medida la generación de gráficos y estadísticas de los mismos.

En primer lugar, será necesario cargar los datos del archivo facilitado ‘ST000291curated.xlsx’ y generar un objeto `SummarizedExperiment` para poder aplicar las funciones destinadas a ese objeto.

Una vez cargados los datos y almacenados en las distintas variables, podemos utilizar la función de `PomaCreateObject()` para generar el objeto `SummarizedExperiment` que, al igual que los objetos `ExpressionSet`, son un tipo de estructura de datos utilizadas en R para almacenar y manipular datos ómicos y metadatos relacionados. Ambas pertenecen a diferentes paquetes de Bioconductor, pero mientras `ExpressionSet` solo permite una matriz de datos en cada objeto, `SummarizedExperiment` puede tener múltiples matrices, las cuales pueden ser accedidas a través de `assays()`. Además, el uso de las funciones y objetos relacionadas con el paquete `SummarizedExperiment`, permite el uso de muchos otros paquetes más modernos desarrollados con métodos enfocados con esta estructura de datos. Como veremos a continuación, los métodos relacionados con el acceso a la información mostrados en la Figura 1, difieren de los utilizados para los `ExpressionSet`.

```
# Generamos el objeto SE utilizando la función incluida en  
# la librería POMA  
  
# Nótese que tenemos que transponer features para encajar  
# con el formato esperado de la función  
se <- PomaCreateObject(metadata = target, features = t(features),  
  group)
```

A continuación, para obtener los datos acerca del estudio y los análisis, accederemos a los datos originales disponibles en www.metabolomicsworkbench.org a través del paquete `metabolomicsWorkbenchR`. Como el estudio ST000291 contiene dos análisis, accederemos a cualquiera de ellos para llegar a los metadatos relativos

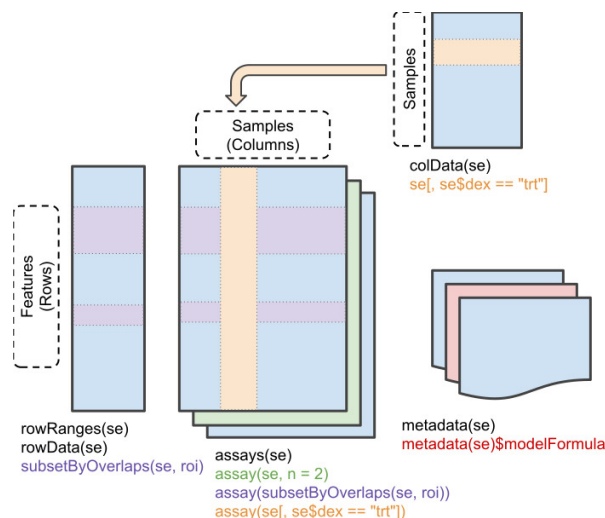


Figure 1: Estructura de datos de un **SummarizedExperiment** Morgan M. 2023 <https://bioconductor.org/packages/release/bioc/vignettes/SummarizedExperiment/inst/doc/SummarizedExperiment.html>

al experimento. Finalmente, lo añadiremos a nuestro **SummarizedExperiment**, que contiene los datos de ambos análisis integrados.

```
# Utilizando la librería metabolomicsWorkbenchR previamente
# cargada Generaremos el objeto SE og_se, proveniente del
# análisis original para recuperar los metadatos
og_se = do_query(context = "study", input_item = "study_id",
  input_value = "ST000291", output_item = "SummarizedExperiment")

# Seleccionamos uno de los dos análisis que contiene el
# estudio
og_se1 <- og_se$AN000464

# Configuramos los metadatos de nuestro se, que integra
# ambos análisis del estudio ST000291
metadata(se)$data_source <- metadata(og_se1)$data_source
metadata(se)$study_id <- metadata(og_se1)$study_id
metadata(se)$analysis_id <- "AN000464 and AN000465"
metadata(se)$analysis_summary <- "ESI Negative mode and ESI Negative mode"
metadata(se)$units <- metadata(og_se1)$units
metadata(se)$description <- metadata(og_se1)$description
```

4.2 Análisis de la estructura de datos

En primer lugar, nos fijamos en las dimensiones de la matriz de datos cargada para verificar que el número de muestras es correcto.

```
# Dimensiones del SummarizedExperiment
dim(se)
```

```
## [1] 1541 45
```

Observamos que hemos obtenido 45 muestras ubicadas en las columnas y un total de 1541 metabolitos identificados en las filas. Las 45 muestras se corresponden a las muestras obtenidas a las 15 participantes incluidas en el experimento en cada una de las tres extracciones del protocolo.

A continuación, analizaremos la cantidad de datos faltantes que podría dificultarnos análisis posteriores.

```
# La matriz de datos la encontramos como el primer objeto
# en la lista de assays de nuestro SummarizedExperiment
sum(is.na(assays(se)[[1]])) # Suma de valores NA
```

```
## [1] 8190
```

Observamos que, para poder seguir haciendo análisis exploratorios sobre los datos, tendremos que lidiar con los valores faltantes, cosa que será abordada más adelante.

Tal y como se ha mencionado más arriba, las funciones para acceder a la información contenida en el `SummarizedExperiment` son diferentes a las que usaremos para los objetos `ExpressionSet`. Concretamente, utilizaremos la función `colData()` para acceder a la información contenida en las muestras, que forman las columnas de la matriz de datos. Aquí podremos ver la covariable `Treatment` asociada a cada una de las muestras biológicas recogidas: “Baseline” para la orina recogida el séptimo día como muestra de nivel basal, “Apple” para las muestras de orina tras el período de ingesta de zumo de manzana y “Cranberry” para las muestras tras el período de ingesta de zumo de arándano.

```
# Mostramos las muestras biológicas en cada nivel de
# tratamiento
summary(colData(se)$Treatment)
```

```
##      Apple  Baseline Cranberry
##         15         15         15
```

Por otro lado, con `rowData()` podremos acceder a la información relativa a las filas, es decir, a los metabolitos. Aquí, gracias al trabajo previo de webscrapping mencionado en el apartado de métodos, tenemos identificado el nombre y el ID de PubChem y de KEGG de cada uno de los 1541 metabolitos.

```
# Mostramos la información cargada de los metabolitos
head(rowData(se))
```

```
## DataFrame with 6 rows and 3 columns
##           names      PubChemID      KEGG
##           <character> <character> <character>
## V1 10-Deacetyl-2-debenz..    443489    C11899
## V2 1,1-Diethyl-2-hydrox..    107754    C13773
## V3 1,2-Dihydroxynaphtha..    9543071   C16196
## V4 12-trans-Hydroxy juv..    11011465   C16508
## V5 14-Dihydroxycornestin     5281160   C08483
## V6 1-(5-Phosphoribosyl)..    440341   C04437
```

Con la función `metadata()` podemos acceder a la información relativa al estudio previamente cargada, como el origen de los datos o las unidades:

```
metadata(se)$data_source
```

```
## [1] "Metabolomics Workbench"
```

```
metadata(se)$units
```

```
## [1] "Peak area"
```

4.3 Análisis exploratorio inicial de los datos

El análisis de estudios ómicos suele estar caracterizado por una alta dimensionalidad en los datos, lo que puede dificultar la comprensión a través de resúmenes estadísticos donde se muestren todos los componentes de cada muestra. Para proceder, en primer lugar deberemos eliminar la presencia de errores faltantes ya que puede comprometer la calidad de las métricas estadísticas o inutilizar el uso de muchas funciones.

4.3.1 Imputación de valores faltantes

Para llevar a cabo la imputación de valores en lugares de la matriz de datos con valores faltantes, usaremos a cabo la función `PomaImpute` de la librería `POMA`. A través de los parámetros de opciones del método, podemos elegir que no trate a los 0 como valores faltantes, ya que simplemente indican el valor obtenido para ese metabolito en la muestra biológica y que sí que elimine aquellas filas con un exceso de valores faltantes, siendo en este caso el límite un 20% de los datos. Dado que nuestros datos están agrupados, cada muestra biológica perteneciendo a un tipo de tratamiento, y disponemos de muchos metabolitos, podemos parametrizar que elimine la fila entera para todos los grupos. Finalmente, el método especificado para la imputación de valores será el algoritmo `k-Nearest Neighbours`, que estimará los valores faltantes en función de los valores vecinos más cercanos.

```
# Curamos los datos de nuestro SE mediante la función PomaImpute
imp_se <- se %>%PomaImpute(zeros_as_na = FALSE, # Los 0 no son NA
  remove_na = TRUE, # Eliminamos features con más de un/
  cutoff = 20, # /20% de valores perdidos (cutoff)
  group_by = TRUE,
  method = "knn") # Imputamos a los valores perdidos según
```

```
## 182 features removed.
```

Con la previa imputación hemos eliminado un total de 182 metabolitos, reduciendo las dimensiones de nuestra matriz de datos.

4.3.2 Normalización de los datos

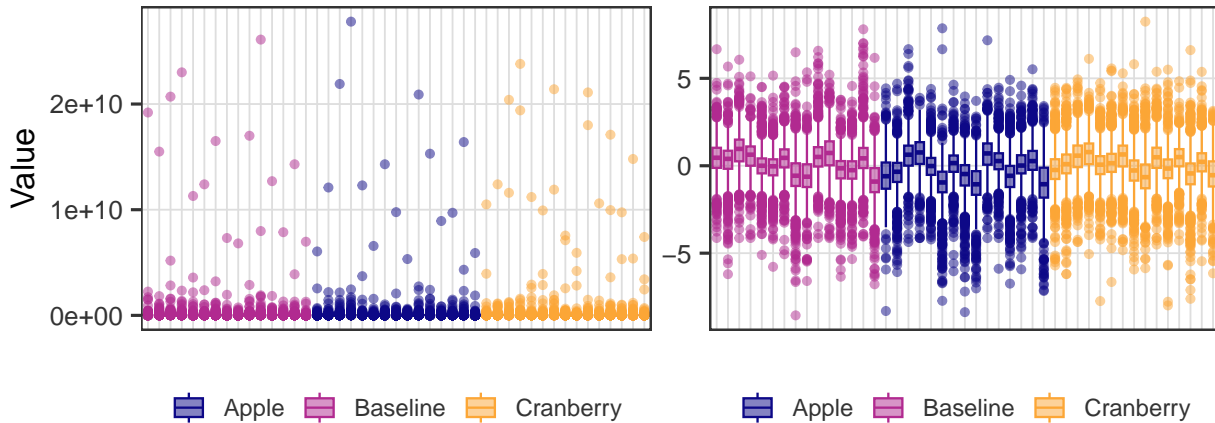
A continuación y siguiendo el workflow típico de un análisis utilizando las funciones del paquete `POMA`, normalizaremos los datos y observaremos como afecta esto a los mismos. Observamos que podemos conseguir una normalización de los datos con una transformación del logaritmo natural de Pareto. Finalmente, compararemos gráficamente con un diagrama de caja las distribuciones de los tres grupos antes y después de la normalización.

```
# Generamos la variable con los datos normalizados
norm_se <- imp_se %>%
  PomaNorm(method = "log_pareto")

# Y comparamos gráficamente como son los datos antes y
# después de la normalización
p1 <- PomaBoxplots(imp_se, x = "samples", theme_params = list(axistext = "y"))
```

```
p2 <- PomaBoxplots(norm_se, x = "samples", theme_params = list(legend_title = FALSE,
  axistext = "y", axistitle = "X"))

# Imprimimos ambos gráficos juntos usando el paquete
# patchwork
p1 + p2
```



De los gráficos anteriores observamos, antes de la normalización, que para la mayoría de metabolitos los niveles detectados son 0 o muy bajos, y sólo algunos metabolitos obtienen valores de área elevados como resultado del análisis. Estos metabolitos podrían estar relacionados con nuestro objeto de interés. Por otro lado, tras la normalización, observamos que aunque las distribuciones son relativamente similares y aproximadamente normales a simple vista, parece haber bastante varianza intragrupal.

4.3.3 Detección de outliers

La identificación y eliminación de outliers es un punto importante en el análisis estadístico ya que puede afectar a los resultados. No obstante, también pueden aportar información clave si estos provienen de información genuina y no de un error de medición o de procedimiento. Para la identificación y eliminación de outliers, encontramos la función `PomaOutliers` que nos permitirá escoger el tipo de método de medición de distancia del outlier y el tipo de medida central. También establecemos el valor de corte por defecto.

```
outlier_results <- imp_se %>%
  PomaOutliers(method = "euclidean", type = "median", coef = 2,
    labels = FALSE)

outlier_results$outliers

## # A tibble: 1 x 4
##   sample groups distance_to_centroid limit_distance
##   <chr>   <fct>         <dbl>         <dbl>
## 1 a12     Apple           15917686960.    14376816373.
```

Observamos que, para los datos sin normalizar, la función detecta que la muestra a12 perteneciente al grupo del tratamiento tras la ingesta de zumo de manzana es un outlier. Podemos acceder al nuevo `SummarizedExperiment` con la matriz de datos sin esta muestra. Observaremos que tras la operación de imputación y la eliminación de outliers hemos reducido las dimensiones de nuestra matriz de datos, habiendo eliminado 182 metabolitos y 1 muestra.

```
clean_se <- outlier_results$data
dim(clean_se)
```

```
## [1] 1359 44
```

4.3.4 Análisis sparse Partial Least Squares Discriminant Analysis

Como se ha comentado al comienzo del apartado 4.3, una manera de poder analizar datos con miles de variables, como es el caso de estos datos metabolómicos, es a través de la reducción de la dimensionalidad. A través de la técnica *sparse Partial Least Squares Discriminant Analysis* (sPLS-DA) podemos ejecutar la elección de variables, con costes computacionales competitivos y resultados gráficos interpretables [8].

En el paquete POMA tenemos funciones que nos ayudan a realizar este análisis.

```
poma_splsda <- clean_se %>%
  PomaPLS(method = "splsda", y = "Treatment", ncomp = 5, labels = FALSE,
    ellipse = FALSE)

# Función pull() para pasar a vector Fuente:
# https://stackoverflow.com/questions/21618423/extract-a-dplyr-tbl-column-as-a-vector

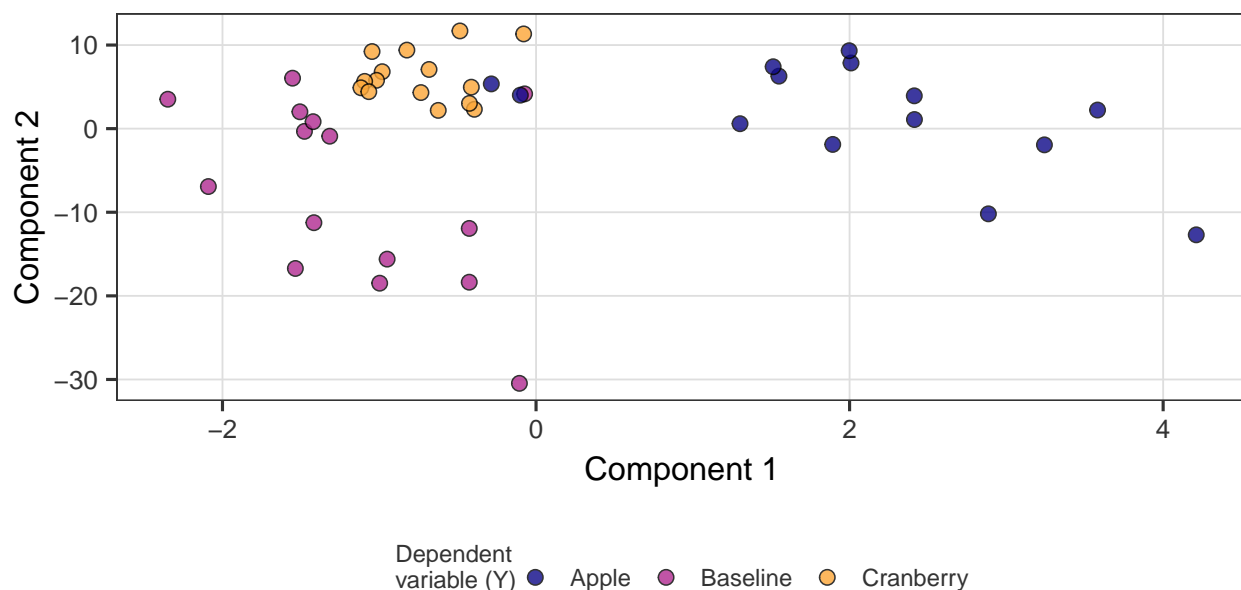
vec_features <- pull(poma_splsda$selected_features[, 1])
vec_info <- rowData(se)[vec_features, 1]
vec_info -> poma_splsda$selected_features[, 3]
names(poma_splsda$selected_features)[3] <- "Metabolito"

kable(poma_splsda$selected_features, caption = "Features seleccionadas")
```

Table 1: Features seleccionadas

feature	value	Metabolito
V308	-0.7333580	12-Hydroxydodecanoic acid_1
V1236	0.4107719	(R)-Mevalonate_1
V766	0.2975272	Digitalose
V639	-0.2913641	Butanoic acid_1
V1011	-0.2447678	Methyl farnesoate_1
V762	-0.1757723	Diethylene glycol_1
V457	0.1354120	4-Hydroxybutanoic acid_1
V813	0.0802793	epsilon-Caprolactam
V173	-0.0652055	Hypoglycin B
V464	-0.0140313	4-Isopropylbenzaldehyde

```
poma_splsda$factors_plot
```

Del análisis anterior observamos cómo, tras la selección de variables, las muestras biológicas posteriores a la ingesta de zumo de manzana según el protocolo establecido, presenta unas puntuaciones diferenciadamente mayores que las obtenidas con zumo de arándano y las muestras basales. Observando los metabolitos seleccionados para el cálculo del componente 1, vemos que esta información se traduce principalmente, en menor presencia de ácido 12-hidroxidodecanoico y mayor presencia de (R)-Mevalonato. El resto de metabolitos y su contribución en el componente 1 puede ser observado en la tabla 1.

5. Discusión

Los resultados anteriores sugieren un metabolismo diferenciado entre la ingesta de zumo de manzana respecto a las muestras con niveles basales o provenientes del tratamiento con zumo de arándanos. Debería realizarse un estudio de los metabolitos mostrados en la tabla 1 para analizar su relación con la biosíntesis de procianidinas en rutas metabólicas del organismo así como de su relación con la diabetes y la resistencia o sensibilidad a la insulina. Una mayor presencia de aquellos metabolitos beneficiosos por participar en rutas de biosíntesis de procianidina o de mejora de la sensibilidad a la insulina y una menor presencia de aquellos metabolitos que generen resistencia a la insulina podría sugerir un impacto positivo del consumo de zumo de manzana en el contexto de la prevención y manejo de la diabetes.

También hemos podido observar en los resultados que mientras que los resultados en el análisis por sPLS-DA de la ingesta de zumo de arándano tienen un componente 1 parecido a los de las muestras basales, presentan un componente 2 mucho más elevado. También parece haber menor variabilidad en los resultados. Aunque el estudio del componente 2 y los metabolitos relacionados quedan fuera del alcance de este estudio, sería necesaria la comprobación de los mismos y su influencia en la síntesis de procianidinas y su relación con la sensibilidad o resistencia a la insulina para sacar ideas más concluyentes.

Hay que tener presente que, en este estudio, se han contemplado un total de 15 individuos y la generación de 45 muestras biológicas, realizando un estudio cruzado, lo cual permite una comparación más precisa entre los individuos, reduciendo variabilidad. No obstante, el tamaño muestral puede no ser lo suficientemente grande para sacar conclusiones reproducibles. Tampoco se ha especificado la fuente de los zumos y su contenido, lo cual puede afectar profundamente los resultados. Tampoco se han incluido los análisis de plasma mencionados en el resumen del estudio[2].

6. Conclusiones

Los resultados obtenidos en este análisis resaltan la utilidad de herramientas como **Bioconductor** y, en particular, la clase **SummarizedExperiment**, que facilita el almacenamiento y la gestión de datos metabolómicos. Asimismo, el uso de **metabolomicsWorkbenchR** para la adquisición de datos y del paquete **POMA** para su procesamiento y exploración mejoran y facilitan la implementación del flujo de trabajo. Estas herramientas proporcionan un entorno eficiente para la manipulación y análisis de datos ómicos, permitiendo la aplicación de técnicas como imputación y eliminación de valores faltantes, normalización y herramientas enfocadas en la reducción de la dimensión de los datos.

Sin embargo, los análisis realizados en este estudio no pueden considerarse concluyentes. Para obtener una comprensión más profunda y veraz, sería necesario complementar estos resultados con otras técnicas de reducción de dimensión y profundizar en la interpretación biológica de los metabolitos identificados. Además, el impacto de estos compuestos en la salud, particularmente en relación con la diabetes y el metabolismo de las procianidinas, debe ser explorado en estudios adicionales. La validación de los efectos en la salud de estos alimentos podría derivar en desarrollos biomédicos o de fitomejoras más enfocados en la optimización de la absorción o producción de ciertos metabolitos.

7. Anexos

7.1 Repositorio GitHub

El informe, el código comentado y utilizado en este documento, el objeto **SummarizedExperiment** en formato binario y algunos archivos y explicaciones adicionales se pueden encontrar en el repositorio: <https://github.com/jcastropu/Castro-Puntero-Jesus-PEC1>

7.2 Referencias

- [1] “Castellano-escuder p, gonzález-domínguez r, carmona-pontaque f, andrés-lacueva c, sánchez-pla a (2021). ‘POMAShiny: A user-friendly web-based workflow for metabolomics and proteomics data analysis.’ *PLOS computational biology*, 17(7), 1-15. <https://doi.org/10.1371/journal.pcbi.1009148>.”
- [2] “This data is available at the NIH common fund’s national metabolomics data repository (NMDR) website, the metabolomics workbench, <https://www.metabolomicsworkbench.org>, where it has been assigned project ID PR000233. The data can be accessed directly via it’s project DOI: 10.21228/M8J590 this work is supported by NIH grant, U2C- DK119886.”
- [3] “NCI dictionary of cancer terms.” *Cancer.gov*, 2025. Accessed: Mar. 24, 2025. [Online]. Available: <https://www.cancer.gov/publications/dictionaries/cancer-terms/def/lc-ms>
- [4] L. J. Valencia-Hernandez, J. E. Wong-Paz, J. A. Ascacio-Valdés, M. L. Chávez-González, J. C. Contreras-Esquivel, and C. N. Aguilar, “Procyanidins: From agro-industrial waste to food as bioactive molecules,” *Foods*, vol. 10, p. 3152, Dec. 2021, doi: 10.3390/foods10123152.
- [5] Y. Liu *et al.*, “Procyanidins and its metabolites by gut microbiome improves insulin resistance in gestational diabetes mellitus mice model via regulating NF-kB and NLRP3 inflammasome pathway,” *Biomedicine & Pharmacotherapy*, vol. 151, p. 113078, Jul. 2022, doi: 10.1016/j.biopha.2022.113078.
- [6] N. Gonzalez-Abuin, M. Pinent, A. Casanova-Martí, L. Arola, M. Blay, and A. Ardevol, “Procyanidins and their healthy protective effects against type 2 diabetes,” *Current Medicinal Chemistry*, vol. 22, pp. 39–50, 2015, doi: 10.2174/0929867321666140916115519.
- [7] *Bioconductor*. 2025. Available: <https://www.bioconductor.org/packages/release/bioc/html/metabolomicsWorkbenchR.html>
- [8] K.-A. Lê Cao, S. Boitard, and P. Besse, “Sparse PLS discriminant analysis: Biologically relevant feature selection and graphical displays for multiclass problems,” *BMC Bioinformatics*, vol. 12, no. 1, Jun. 2011, doi: <https://doi.org/10.1186/1471-2105-12-253>.