Informe PEC1

Erick Marcos Castillo Vargas

Índice

1. Resumen

El estudio partió de un dataset descargado del Metabolomics Workbench y analizado a través de un objeto `SummarizedExperiment`, lo que permitió integrar de forma eficiente datos experimentales, metadatos, matrices de datos del metaboloma y la información de las muestras. Se compararon las diferencias entre `SummarizedExperiment` y `ExpressionSet`, destacando la versatilidad del primero al manejar múltiples matrices de datos y metainformación. Posteriormente, se realizó un análisis exploratorio que incluyó la inspección de metadatos, imputación de valores faltantes, resumen numérico univariante y normalización de datos por *log-scaling*. Finalmente, se aplicó un análisis de componentes principales (PCA) que, sin evidenciar una separación clara entre las condiciones experimentales, sugiere que los cambios metabolómicos inducidos por el consumo de jugos no son evidentes, de modo que se requieren análisis estadísticos más avanzados para poder detectarlos, si es que los hay.

2. Objetivos

Para el presente trabajo se utilizó un dataset proveniente del Metabolomics Workbench, el cual ha sido analizado mediante un objeto de clase `SummarizedExperiment` en R. Específicamente, se busca:

1. Realizar un análisis exploratorio que permita obtener una visión global del dataset.
2. Aprender a trabajar con objetos `SummarizedExperiment` en R, facilitando la integración y manipulación de datos experimentales.
3. Determinar las principales diferencias entre la clase `SummarizedExperiment` y la clase `ExpressionSet`.

3. Métodos

Se inició cargando los paquetes necesarios, entre ellos *metabolomicsWorkbenchR* para la descarga de datos, *SummarizedExperiment* para gestionar la estructura del dataset, y *POMA* junto a *ggplot2* y *ggtext* para realizar imputación, normalización y visualización. El dataset fue descargado directamente desde el *Metabolomics Workbench*, utilizando el paquete *metabolomicsWorkbenchR*. La información se importó en un objeto de clase `SummarizedExperiment`, lo que facilitó la integración de datos experimentales, metadatos, datos de metabolitos y características de las muestras en una única estructura. La descarga se realizó a través de la función *do_query*, especificando el identificador del estudio (ST000291) y solicitando el objeto `SummarizedExperiment`.

Se procedió a inspeccionar el contenido general del objeto, extrayendo los metadatos experimentales de cada análisis y organizándolos en una tabla comparativa. Además, se examinaron las matrices de datos (usando la función `assay`) y se extrajeron tanto la información de los metabolitos (a través de `rowData`) como la de las muestras (utilizando `colData`). Se determinó la distribución de muestras según la condición experimental y se evaluó la cantidad de valores faltantes en cada muestra mediante el cálculo de sumas por columnas.

Ante la presencia de valores faltantes, se aplicó una imputación utilizando la función específica de POMA para generar versiones completas de los datos. Posteriormente, se calculó un resumen numérico univariante para cada muestra (media, mediana, desviación estándar y coeficiente de variación) para evaluar la variabilidad inherente en los datos.

Posteriormente, observando el alto grado de variabilidad y como procedimiento estándar en el análisis de este tipo de datos, se procedió con una normalización basada en logaritmo (*log-scaling*) que permite estabilizar la varianza y reducir la dispersión. Se generaron gráficos de boxplots y curvas de densidad tanto antes como después de la normalización, lo que facilitó la comparación visual del efecto de esta transformación en la distribución de los datos.

Asimismo, se evaluó la presencia de *outliers* utilizando herramientas específicas de POMA. Se identificaron los puntos atípicos en los datos sin normalizar y se observó que, tras la normalización, dichos *outliers* dejaban de considerarse problemáticos.

Finalmente, se realizó un análisis de componentes principales (PCA) sobre los datos normalizados. Se calculó el porcentaje de varianza explicada por cada componente y se construyeron gráficos de dispersión (PC1 versus PC2) en los que se visualizaban los patrones y agrupaciones de las muestras según la condición experimental. Este enfoque permitió comprobar la estructura subyacente de los datos y la posible separación entre condiciones.

4. Resultados

Primeramente, conviene comentar las principales diferencias entre la clase `SummarizedExperiment` y la clase `ExpressionSet`. Por un lado, `ExpressionSet` está diseñado para almacenar un único conjunto de datos (usualmente una matriz de expresión), mientras que `SummarizedExperiment` permite integrar múltiples matrices de datos (ensayos) en un solo objeto. Además, en `ExpressionSet` los metadatos se gestionan a través de construcciones como `phenoData` y `featureData` (definidos en el paquete Biobase), mientras que en `SummarizedExperiment` se utilizan las ranuras `colData` y `rowData`. Asimismo, es importante destacar que `SummarizedExperiment` tiene una subclase, denominada `RangedSummarizedExperiment`, donde la información de las filas son rangos genómicos (algo inexistente en la clase `ExpressionSet`), haciendo a este tipo de clase ideal para experimentos basados en secuenciación, como los RNA-seq.

En lo que respecta al dataset seleccionado, mencionar que este proviene del proyecto titulado *LC-MS Based Approaches to Investigate Metabolomic Differences in the Urine and Plasma of Young Women after Drinking Cranberry Juice or Apple Juice*, que pretendía investigar los cambios metabólicos globales causados por el consumo de jugo de arándano o jugo de manzana, utilizando un enfoque metabolómico global basado en LC-MS (cromatografía líquida-espectrometría de masas). El motivo de haberlo seleccionado fue realmente simple, me pareció curioso que hubiera un estudio que analizara cambios en el metaboloma urinario inducidos por el consumo de diferentes tipos de zumo. Posteriormente, al documentarme sobre el tema, entendí la relevancia del trabajo, pues esas diferencias observadas en el metaboloma pueden servir como biomarcadores del consumo de jugo de arándanos y explicar sus propiedades beneficiosas para la salud humana (como la prevención de ciertas infecciones).

Pasando al análisis exploratorio, tras la descarga del dataset como objeto `SummarizedExperiment`, se observó que contenía 2 análisis, cada uno con un modo de ionización diferente (positivo o negativo): AN000464 (positivo) y AN000465 (negativo). Al realizar ambos modos de ionización, se maximiza la cobertura de los metabolitos detectados en la muestra. Algunos compuestos son más sensibles en modo positivo, mientras que otros se detectan mejor en modo negativo. Utilizar ambos modos garantiza una detección más completa del perfil metabolómico de las muestras. Los metadatos de cada análisis (Cuadro 1) se obtuvieron usando la función `metadata` sobre el objeto `SummarizedExperiment`. De la misma manera, con las funciones `assay`, `rowData` y `colData` se puede acceder e inspeccionar la matriz de datos, los metadatos de los metabolitos y los metadatos de las muestras, respectivamente. Con esta información se puede determinar que:

- El número de muestras totales es 45 (las mismas para ambos análisis), divididas en 3 condiciones: orina basal, orina después de tomar jugo de manzana y orina después de tomar jugo de arándanos.
- El número total de metabolitos detectados para AN000464 es 1786 y para AN000465 es 747.

Cuadro 1: Metadatos experimentales de cada análisis

Campo	AN000464	AN000465
data_source	Metabolomics Workbench	Metabolomics Workbench
study_id	ST000291	ST000291
analysis_id	AN000464	AN000465
analysis_summary	ESI Positive mode	ESI Negative mode
units	Peak area	Peak area

name	ST000291:AN000464	ST000291:AN000465
description	LC-MS Based Approaches to Investigate Metabolomic Differences in the Urine of Young Women after Drinking Cranberry Juice or Apple Juice	LC-MS Based Approaches to Investigate Metabolomic Differences in the Urine of Young Women after Drinking Cranberry Juice or Apple Juice
subject_type	NA	NA

Después del examen superficial de la estructura del dataset, es imprescindible realizar un análisis descriptivo del mismo. Para ello, primero se comprobó el número de muestras por condición experimental, un total de 9 (Cuadro 2), seguido del número de valores faltantes por muestra en la matriz de datos de cada análisis (Cuadro 3, solo se muestran las 8 primeras muestras). Acto seguido, usando la función `PomaImpute` se realizó su imputación usando el método k vecinos más cercanos y un umbral de 20, donde los metabolitos que tenían valores faltantes en un número de muestras superior a este eran eliminados. La imputación es necesaria para la exploración porque otros métodos que se emplearán más adelante exigen que no haya valores faltantes en la matriz de datos.

Cuadro 2: Número de muestras por condición experimental

Condición	Nº muestras
Baseline urine	15
Urine after drinking apple juice	15
Urine after drinking cranberry juice	15

Cuadro 3: Número de valores faltantes por muestra

Muestra	AN000464	AN000465
a1	74	17
a10	81	23
a11	40	10
a12	50	13
a13	58	22
a14	123	44
a15	51	11
a16	90	31

Continuando con el análisis descriptivo, se llevó a cabo un resumen numérico univariante de cada muestra para cada análisis experimental, donde se calculó su media, mediana, desviación estándar y coeficiente de variación (Cuadros 4 y 5, solo se muestran las 8 primeras muestras). Se puede observar como la variabilidad es notablemente elevada.

Cuadro 4: Resumen numérico univariante del análisis experimental AN000464

Muestra	Media	Mediana	SD	VC
a1	10918716	266000	163323661	14.95814
a10	17163483	279000	304088980	17.71721
a11	29187119	952000	539384595	18.48023
a12	35328047	1010000	683717040	19.35338

a13	19053906	471000	317308315	16.65319
a14	10411846	140000	167233548	16.06185
a15	18265155	469000	350824206	19.20729
a16	14459352	279000	246835805	17.07101

Cuadro 5: Resumen numérico univariante del análisis experimental AN000465

Muestra	Media	Mediana	SD	VC
a1	12213695	645500	96514230	7.902132
a10	14930551	599000	109325723	7.322283
a11	29904245	2287651	206315710	6.899212
a12	28421365	1970000	156612725	5.510387
a13	17495020	952500	141299858	8.076576
a14	8132438	334500	61098836	7.512979
a15	18488647	933500	148567968	8.035632
a16	10250869	365000	69478968	6.777862

A este tipo de datos es conveniente aplicarles una normalización, como puede ser un *log-scaling*, que también ayuda a reducir la dispersión. Para ello se utilizó la función `PomaNorm` con el método mencionado. En los gráficos 1 y 2 se puede ver la estructura y distribución de los datos antes y después de la normalización (se muestra solo el análisis AN000464 a modo de ejemplo). Se puede ver como, efectivamente, la dispersión de los datos ha disminuido y ahora su distribución es bastantente normal.

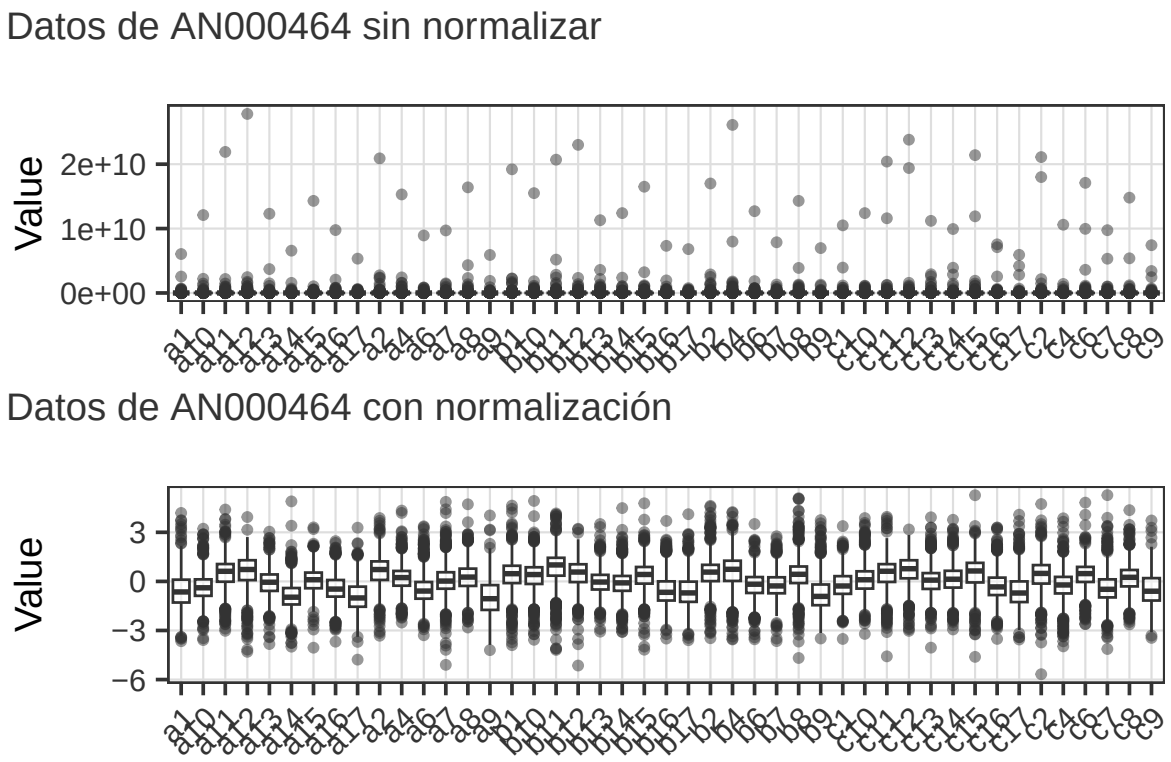

Figura 1: Boxplot de los datos de AN000464 antes y después de normalizar

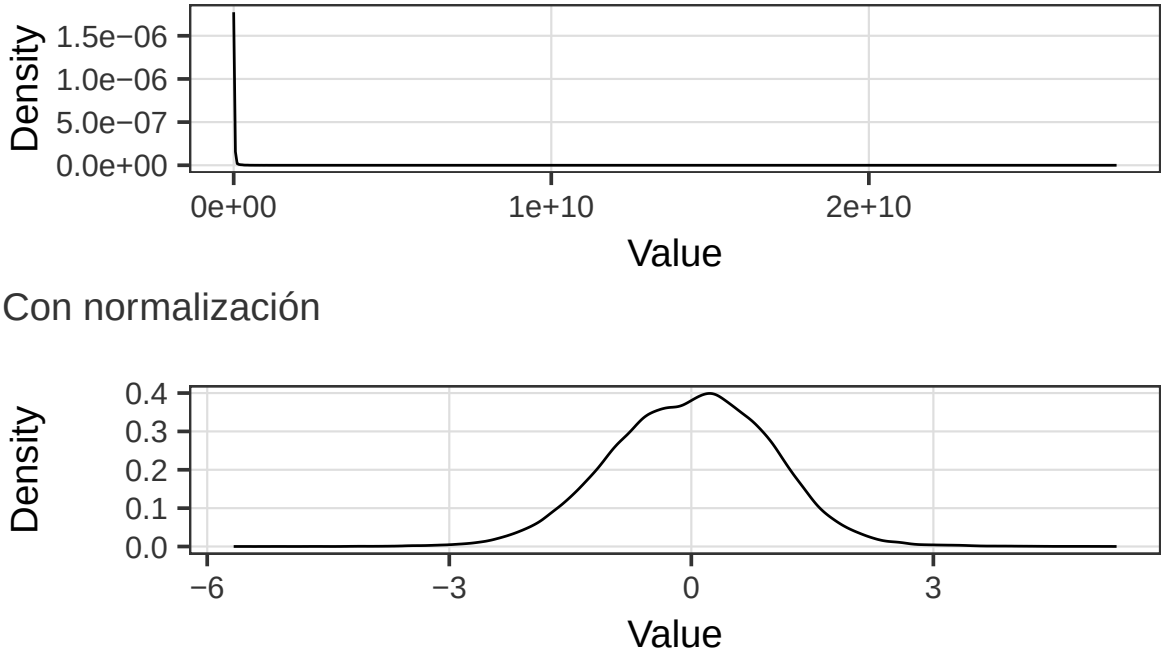

Figura 2: Curva de densidad de los datos de AN000464 antes y después de normalizar

Otro aspecto importante en la exploración del dataset es la detección de *outliers*. Para esto podemos usar la función `PomaOutliers`. Antes de la normalización, en ambos análisis experimentales podíamos encontrar muestras consideradas *outliers* (Cuadros 6 y 7). No obstante, tras normalizar los datos, dichas muestras dejaron de serlo.

Cuadro 6: Outliers detectados por el algoritmo de POMA para AN000464 sin normalizar

sample	groups	distance_to_centroid	limit_distance
a12	Urine after drinking apple juice	15900922501	14413980039

Cuadro 7: Outliers detectados por el algoritmo de POMA para AN000465 sin normalizar

sample	groups	distance_to_centroid	limit_distance
a2	Urine after drinking apple juice	4729800515	2970397106
b4	Baseline urine	4362395599	3807855112
c2	Urine after drinking cranberry juice	5363376444	3914154641

Finalmente, como último paso en la exploración de los datos, se llevó a cabo un análisis de componentes principales (PCA) para cada uno de los análisis experimentales, con el fin de visualizar la estructura de los datos normalizados y detectar posibles patrones o agrupaciones de las muestras según su condición (Figura 3). Como se observa en los gráficos, no se evidencia una agrupación clara de las muestras en función de su grupo experimental, lo que sugiere que, al menos con la información disponible, no existe una separación

evidente entre las condiciones evaluadas (orina basal, orina tras consumo de jugo de manzana u orina tras consumo de jugo de arándano).

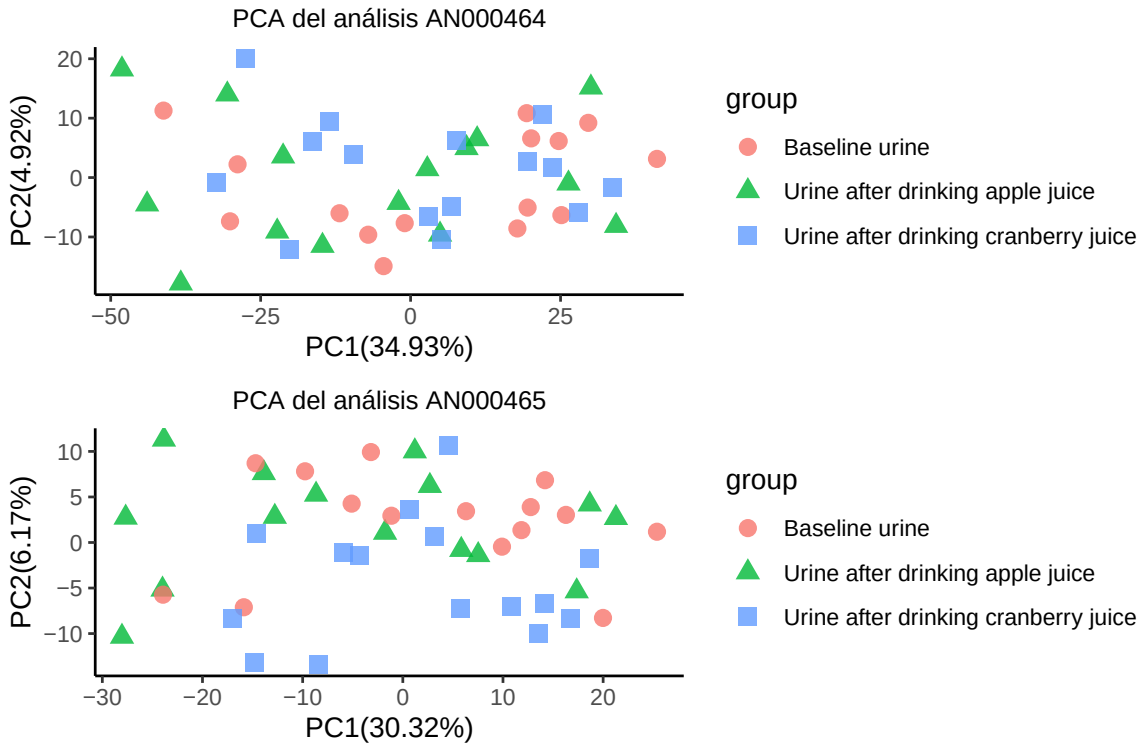

Figura 3: PCA de los datos normalizados de AN000464 y AN000465

5. Discusión

El presente estudio se centró en la exploración de un dataset metabolómico proveniente del *Metabolomics Workbench*, en el que se evaluaron las diferencias en el perfil metabolómico urinario de mujeres jóvenes tras el consumo de jugo de manzana y jugo de arándano. Para ello se utilizó un objeto de la clase `SummarizedExperiment`, lo que permitió integrar de forma eficiente los datos experimentales, metadatos, datos de metabolitos y la información relativa a las muestras.

Interpretación Biológica de los Resultados

El análisis exploratorio mostró que los datos presentan una alta variabilidad, la cual se redujo considerablemente tras la normalización mediante *log-scaling*. Este proceso fue esencial para estabilizar la varianza y facilitar la detección de patrones en el dataset. Sin embargo, el análisis de componentes principales (PCA) realizado sobre los datos normalizados no evidenció una separación clara entre las condiciones experimentales (orina basal, orina tras consumo de jugo de manzana y orina tras consumo de jugo de arándano). Desde el punto de vista biológico, este hallazgo puede interpretarse de dos maneras:

- **Ausencia de cambios metabólicos significativos:** Es posible que el consumo de jugo de manzana o de arándano no induzca cambios suficientemente pronunciados en el perfil metabolómico urinario

para ser detectados mediante un análisis multivariado como el PCA. Esto sugiere que, a nivel del metaboloma urinario, las alteraciones metabólicas asociadas a estos tratamientos son sutiles o se manifiestan en otros biomarcadores o vías metabólicas que no han sido capturadas en este estudio.

- **Limitaciones en el diseño del estudio:** La falta de separación clara podría deberse también a aspectos del diseño experimental, tales como el número limitado de muestras o la ventana temporal en la que se realizaron las tomas post consumo, lo que podría haber afectado la detección de diferencias entre grupos.

Limitaciones del Estudio

Si bien el trabajo realizado aportó importantes aprendizajes en el manejo y análisis de datos metabolómicos mediante `SummarizedExperiment`, es relevante destacar algunas limitaciones:

1. **Tamaño muestral reducido:** Con un total de 45 muestras distribuidas en tres condiciones experimentales, la potencia estadística para detectar diferencias sutiles se ve limitada. Un mayor número de réplicas podría ayudar a revelar patrones que en este estudio quedaron indetectables.
2. **Proceso de imputación de datos:** La utilización de la función `PomaImpute` permitió manejar los valores faltantes, pero este proceso puede introducir sesgos, especialmente en aquellos metabolitos con un alto porcentaje de datos ausentes. Dicho sesgo puede afectar la interpretación de la variabilidad y la distribución de los datos.
3. **Normalización mediante *log-scaling*:** Aunque el *log-scaling* resultó efectivo para reducir la dispersión de los datos, es posible que otras técnicas de normalización o transformaciones complementarias pudieran aportar una visión más detallada de la estructura subyacente del dataset.
4. **Análisis multivariado limitado:** La aplicación del PCA permitió visualizar la estructura global de los datos, pero la ausencia de agrupaciones claras sugiere que sería pertinente aplicar técnicas estadísticas adicionales que puedan detectar mejor distinciones menos evidentes.

Reflexiones sobre el Trabajo Realizado

El análisis llevado a cabo demuestra la utilidad de trabajar con objetos `SummarizedExperiment` para integrar y manipular datos complejos provenientes de experimentos metabolómicos. Se destaca la importancia de seguir una serie de pasos rigurosos: desde la inspección de los metadatos y matrices de datos, pasando por la imputación de valores faltantes y la normalización, hasta la aplicación de métodos multivariados como el PCA. Cada uno de estos pasos contribuyó a garantizar la calidad del análisis.

Desde una perspectiva biológica, aunque no se evidenciaron, a priori, cambios drásticos en el perfil metabolómico urinario tras el consumo de jugos, esto solo ha sido un análisis muy somero. Se requiere la utilización de otras técnicas estadísticas más avanzadas y específicas para poner de manifiesto las posibles diferencias que pueda haber en la toma de uno u otro jugo.

6. Conclusiones

El estudio ha permitido no solo familiarizarse con el manejo de datos metabolómicos a través de `SummarizedExperiment`, sino también identificar las limitaciones y desafíos asociados a la integración y análisis de este tipo de datos. La ausencia de agrupaciones claras en el PCA sugiere que, en este caso, los efectos metabólicos del consumo de jugo de manzana y arándano no son tan evidentes y requieren enfoques analíticos más profundos o estudios complementarios para ser interpretados de manera concluyente en el contexto de la salud humana.

7. Referencias (incluye el repositorio de GitHub)

Castillo Vargas, E. (2025). *Castillo-Vargas-Erick-Marcos-PEC1* [Repositorio en GitHub]. GitHub. https://github.com/ErickCastilloVargas/Castillo-Vargas-Erick-Marcos-PEC1.git

Castellano-Escuder, P. (2025). *Package 'POMA'* (1.16.0) [Software]. Bioconductor. https://www.bioconductor.org/packages/release/bioc/manuals/POMA/man/POMA.pdf

Castellano-Escuder, P. (2024, 29 octubre). *Get started*. Bioconductor. https://www.bioconductor.org/packages/release/bioc/vignettes/POMA/inst/doc/POMA-workflow.html

Castellano-Escuder, P. (2022, 1 noviembre). *POMA workflow*. Bioconductor. http://bioconductor.jp/packages/3.16/bioc/vignettes/POMA/inst/doc/POMA-demo.html

Castellano-Escuder, P. (2024, 29 noviembre). *Use case: LC-MS Based Approaches to Investigate Metabolomic Differences in the Urine of Young Women after Drinking Cranberry Juice or Apple Juice*. Bioconductor. https://www.bioconductor.org/packages/devel/bioc//vignettes/fobitools/inst/doc/MW_ST000291_enrichment.html

Morgan, M., Obenchain, V., & Hester, J., Pagès, H. (2025). *Package 'SummarizedExperiment'* (1.37.0) [Software]. Bioconductor. https://www.bioconductor.org/packages/release/bioc/manuals/metabolomicsWorkbenchR/man/metabolomicsWorkbenchR.pdf

Morgan, M., Obenchain, V., & Hester, J., Pagès, H. (2018, 3 mayo). *SummarizedExperiment for Coordinating Experimental Assays, Samples, and Regions of Interest*. Bioconductor. https://bioconductor.statistik.tu-dortmund.de/packages/3.8/bioc/vignettes/SummarizedExperiment/inst/doc/SummarizedExperiment.html

Liu, H. (2016). *LC-MS Based Approaches to Investigate Metabolomic Differences in the Urine of Young Women after Drinking Cranberry Juice or Apple Juice*. [Conjunto de datos; ST000291]. Metabolomics Workbench. https://doi.org/10.21228/M8J590

Liu, H., Garrett, T. J., Su, Z., Khoo, C., & Gu, L. (2017). UHPLC-Q-Orbitrap-HRMS-based global metabolomics reveal metabolome modifications in plasma of young women after cranberry juice consumption. *The Journal Of Nutritional Biochemistry, 45*, 67-76. https://doi.org/10.1016/j.jnutbio.2017.03.007

Rhys Lloyd, G., & Maria Weber, R. J. (2025). *Package 'metabolomicsWorkbenchR'* (1.16.0) [Software]. Bioconductor. https://www.bioconductor.org/packages/release/bioc/manuals/metabolomicsWorkbenchR/man/metabolomicsWorkbenchR.pdf

Rhys Lloyd, G., & Maria Weber, R. J. (2024, 29 octubre). *Introduction to metabolomicsWorkbenchR*. Bioconductor. https://www.bioconductor.org/packages/release/bioc/vignettes/metabolomicsWorkbenchR/inst/doc/Introduction_to_metabolomicsWorkbenchR.html