PEC1: Introducción a las ómicas

Sergi Roig Soucase

Tabla de contenidos

1. Resumen

El presente estudio tiene como objetivo principal la obtención y análisis de datos provenientes de experimentos de metabolómica. Para ello, se seleccionó un estudio enfocado en la dinámica temporal de la concentración de metabolitos del metabolismo del ácido fólico en células de leucemia, expuestas a bajas concentraciones de ácido fólico. Utilizando el paquete *metabolomicsWorkbenchR* de Bioconductor, se importó el objeto *SummarizedExperiment*, que contiene tanto los datos como los metadatos del estudio. Posteriormente, se emplearon varios paquetes de R para extraer los datos de concentración de los metabolitos, los cuales fueron formateados y representados de diversas maneras para facilitar el análisis de los datos. Los resultados muestran que cinco de los seis metabolitos experimentan una disminución en su concentración, mientras que el metabolito restante no presenta la misma tendencia, posiblemente debido a su alta variabilidad entre réplicas. Finalmente, el código en R, los archivos generados y los metadatos se han depositado en un repositorio en GitHub para su consulta y reutilización.

2. Objetivos del estudio

Los objetivos principales del estudio son los siguientes:

- Búsqueda de estudios de metabolómica en la base de datos *Metabolomics Workbench*
- Obtención de datos de metabolómica mediante R, utilizando el paquete *metabolomicsWorkbenchR (Bioconductor)*

- Manejo de clases tipo *SummarizedExperiment* para visualizar los datos y metadatos
- Exploración de los datos utilizando diferentes funciones
- Exploración de los datos utilizando diferentes representaciones gráficas como gráfico de líneas, boxplot, heatmap y PCA
- Manejo de *GitHub* para subir todos los archivos y los metadatos generados del análisis del estudio

3. Materiales y Métodos

Origen y naturaleza de los datos

Los datos se corresponden al estudio "*Folate levels in K562 cells following folate depletion*" cuya referencia es ST002892 de acuerdo con la base de datos Metabolomics Workbench (https://www.metabolomicsworkbench.org/). Este dataset contiene cuantificaciones de concentración de 6 metabolitos derivados del ácido fólico durante 8 días en células de leucemia K562, las cuales han sido tratadas con una concentración constante de ácido fólico. Las cuantificaciones han sido realizadas con un espectrómetro de masas (MS) Orbitrap.

Herramientas informáticas

La descarga y procesamiento de los datos ha sido completamente realizado con R. Para la obtención de los datos se ha utilizado el paquete *metabolomicsWorkbenchR (Bioconductor)*. Para el manejo de la clase *SummarizedExperiment* se ha utilizado el paquete *SummarizedExperiment (Bioconductor)*. Para la representación de los datos, se ha utilizado el paquete *ggplot2*, requiriendo de los paquetes *tidyr*, *dplyr* y *reshape2* para el formateo de los datos. Para la elaboración del informe se ha utilizado Microsoft Word.

Procedimiento general del análisis

Primero se ha importado la clase *SummarizedExperiment* mediante la función *do_query()* del paquete *metabolomicsWorkbenchR*. El objeto contenedor con los datos se ha guardado en formato binario en ST002892_SE.Rda utilizando la función *save()*. Se ha accedido a las cuantificaciones de los metabolitos utilizando la función *assays()* del paquete *SummarizedExperiment*, los cuales se han guardado en un archivo ST002892_data.csv. Del mismo paquete, se ha accedido a los metadatos con la función *metadata()*. Posteriormente, se ha representado de maneras diferentes los datos utilizando los paquetes *ggplot2, tidyr, dplyr y reshape2. Por último, se ha volcado el código, los datos y metadatos en un repositorio de GitHub.*

4. Resultados

Primero se ha importado la clase *SummarizedExperiment* mediante el ID del estudio ST002892 utilizando la función *do_query()*:

```
library(metabolomicsWorkbenchR)

SE = do_query(
    context = 'study', input_item = 'study_id', input_value = 'ST002892',
    output_item = 'SummarizedExperiment')
SE

## class: SummarizedExperiment
## dim: 6 23
## metadata(8): data_source study_id ... description subject_type
## assays(1): ''
## rownames(6): ME754051 ME754046 ... ME754047 ME754049
## rowData names(3): metabolite_name metabolite_id refmet_name
## colnames(23): K562_100nM_Timecourse_Day0_Folate_1
##   K562_100nM_Timecourse_Day0_Folate_2 ...
##   K562_100nM_Timecourse_Day8_Folate_2
##   K562_100nM_Timecourse_Day8_Folate_3
## colData names(6): local_sample_id study_id ... raw_data
##   X100nM_Folic_Acid_Growth
```

Podemos ver que la dimensión del dataset es de 6 filas por 23 columnas. En las filas tenemos los 6 diferentes metabolitos cuantificados:

```
library(knitr)

kable(rowData(SE))
```

Tabla 1 *Metabolitos derivados del ácido fólico cuantificados en el estudio*

	metabolite_name	metabolite_id	refmet_name
ME754051	10-formyltetrahydrofolate	ME754051	N10-Formyl-THF
ME754046	5,10-Methylenetetrahydrofolate	ME754046	5,10-Methylene-THF
ME754050	5-formyltetrahydrofolate	ME754050	N5-Formyl-THF
ME754048	5-methyltetrahydrofolate	ME754048	5-Methyl-THF
ME754047	Folic Acid	ME754047	Folic acid
ME754049	tetrahydrofolate	ME754049	THF

En las columnas tenemos los diferentes días a los que se ha cuantificado (0, 1, 2, 4, 6 y 8) y las cuatro réplicas. A día 8 solo hay tres réplicas por lo que hay un total de 23 columnas.

Utilizando el paquete *SummarizedExperiment* se puede acceder a los metadatos utilizando la función *metadata()*:

```r
library(SummarizedExperiment)

metadata(SE)
```

```
## $data_source
## [1] "Metabolomics Workbench"
##
## $study_id
## [1] "ST002892"
##
## $analysis_id
## [1] "AN004751"
##
## $analysis_summary
## [1] "Reversed phase UNSPECIFIED ION MODE"
##
## $units
## [1] "Arbitrary Units"
##
## $name
## [1] "ST002892:AN004751"
##
## $description
## [1] "Folate levels in K562 cells following folate depletion"
##
## $subject_type
## [1] NA
```

Podemos apreciar que la fuente de los datos es de *Metabolomics Workbench*. También aparece el ID del estudio y del análisis, así como las unidades de las concentraciones, que son unidades arbitrarias. Además, aparecen otros datos que ya he comentado anteriormente.

La forma de guardar en un dataframe las cuantificaciones ha sido utilizando la función *assays() d*el paquete *SummarizedExperiment*:

```r
# generamos un data frame con los datos
datos <- assays(SE)[[1]]
# renombramos la columna de metabolitos
datos$Metabolito <- rowData(SE)[,3]
```

Una vez obtenido el dataframe resultante, el cual no mostraré debido a que es bastante extenso, he realizado una serie de representaciones gráficas para explorar los datos.

Lo primero ha sido ver la **tendencia de concentraciones** de los diferentes metabolitos del ácido fólico a lo largo de los días. Para ello, primero queremos tener un dataframe con cuatro columnas: Metabolito, Día, Réplica y Concentración.

```r
library(tidyr)
library(dplyr)

# Transformar los datos en formato:
# tipo de metabolito, día, réplica y concentración
```

```r
datos_long <- datos %>%  pivot_longer(
    cols = -Metabolito,  # seleccionamos todas las columnas excepto "Meta
bolito"
    names_to = c("Día", "Réplica"),  # nobmres para las nuevas columnas
    names_pattern = ".*Day(\\d+)_Folate_(\\d+)",  # extraemos el día y ré
plica del nombre original
    values_to = "Concentración"
  )

# Convertir "Día" y "Réplica" a variables numéricas
datos_long <- datos_long %>%
  mutate(Día = as.numeric(Día),
         Réplica = as.numeric(Réplica))
```

Una vez tenemos los datos en el formato correcto, representamos utilizando el paquete *ggplot2* de R:

```r
library(ggplot2)


ggplot(datos_long,
       aes(x = Día, y = Concentración, color = Metabolito, group = intera
ction(Metabolito, Réplica))) +
  geom_line(alpha = 0.6) +
  geom_point(size = 1) +
  labs(x = "Día",
       y = "Concentración") +
  theme_minimal()
```

En este caso hemos representado a lo largo de los días la concentración de los metabolitos (en unidades arbitrarias) de las cuatro réplicas de cada metabolito:

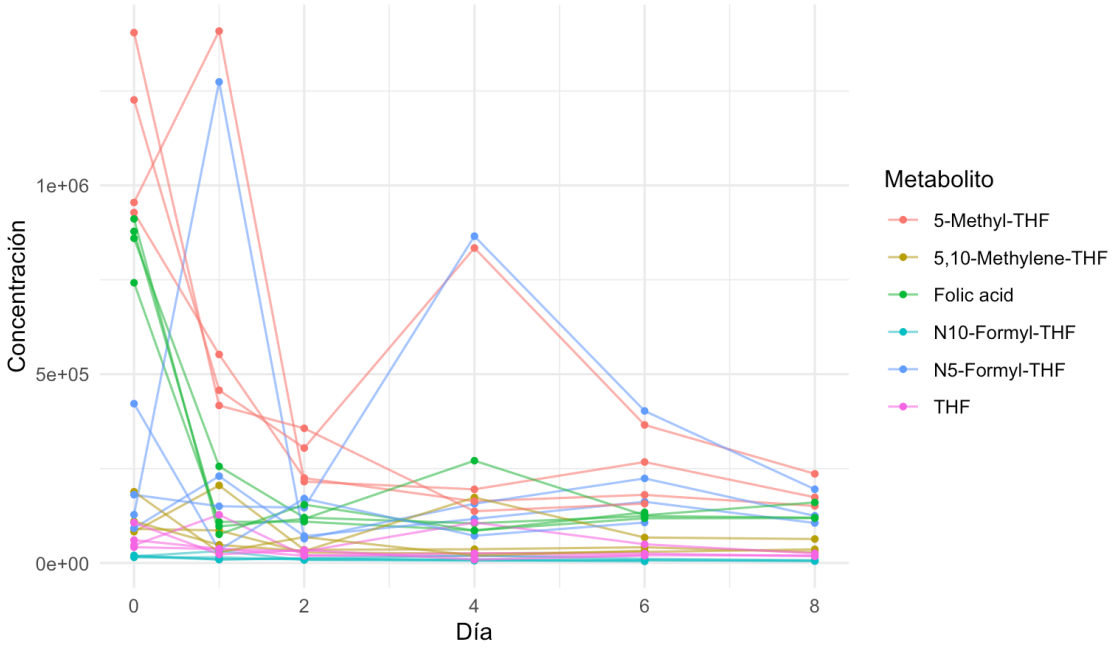

Figura 1 *Tendencias de Concentración de Metabolitos a lo Largo del Tiempo*

Uno de los problemas que vemos es que hay metabolitos cuya concentración es baja y algunas réplicas tienen poca correlación entre ellas. Por ello, podemos separar por metabolito mediante un diagrama de cajas (***boxplot***) para apreciar la posible variabilidad entre réplicas y si existen valores anómalos:

```
ggplot(datos_long, aes(x = factor(Día), y = Concentración, fill = Metabol
ito)) +
  geom_boxplot() +
  facet_wrap(~ Metabolito, scales = "free_y") +
  labs(x = "Día",
       y = "Concentración") +
  theme_minimal()
```

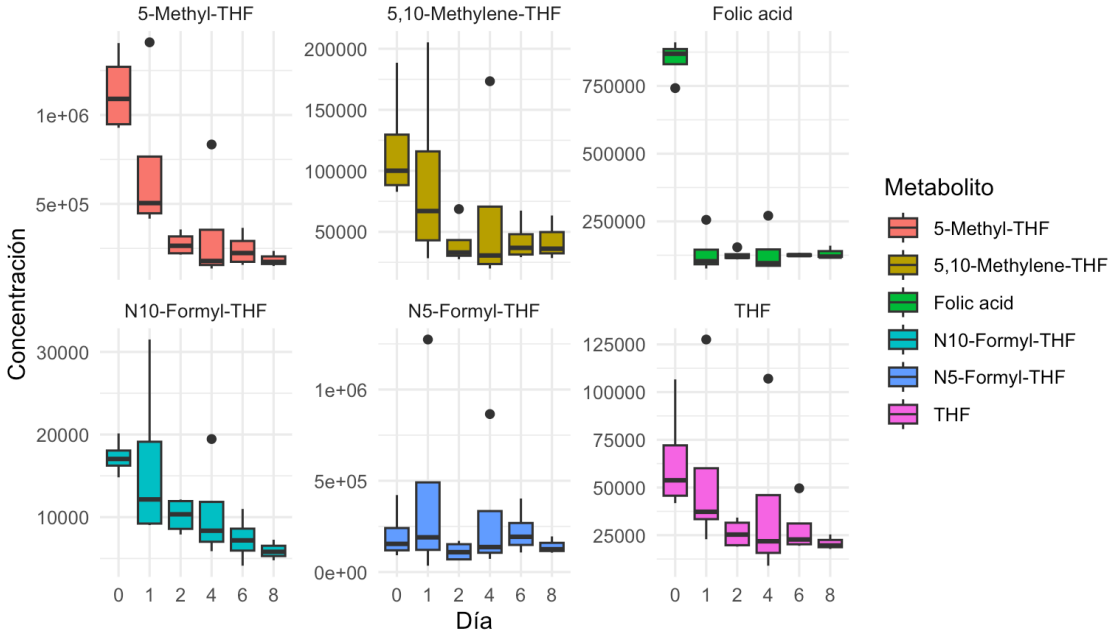

Figura 2 *Distribución de Concentraciones de Metabolitos por Día mediante boxplot*

De manera alternativa a las anteriores dos formas de representación, para detectar patrones de concentraciones de los diferentes metabolitos a lo largo del tiempo, se puede realizar un gráfico de calor o ***heatmap***.

```
library(reshape2)

# cambiamos formato de los datos para el heatmap
# calculamos la media de concentración de cada metabolito
datos_wide <- dcast(datos_long, Metabolito ~ Día, value.var = "Concentrac
ión", fun.aggregate = mean)
# creamos el heatmap
ggplot(melt(datos_wide), aes(x = variable, y = Metabolito, fill = value))
+
  geom_tile() +
  scale_fill_gradient(low = "blue", high = "red") +
  labs(x = "Día",
```

```
       y = "Metabolito") +
  theme_minimal()

## Using Metabolito as id variables
```

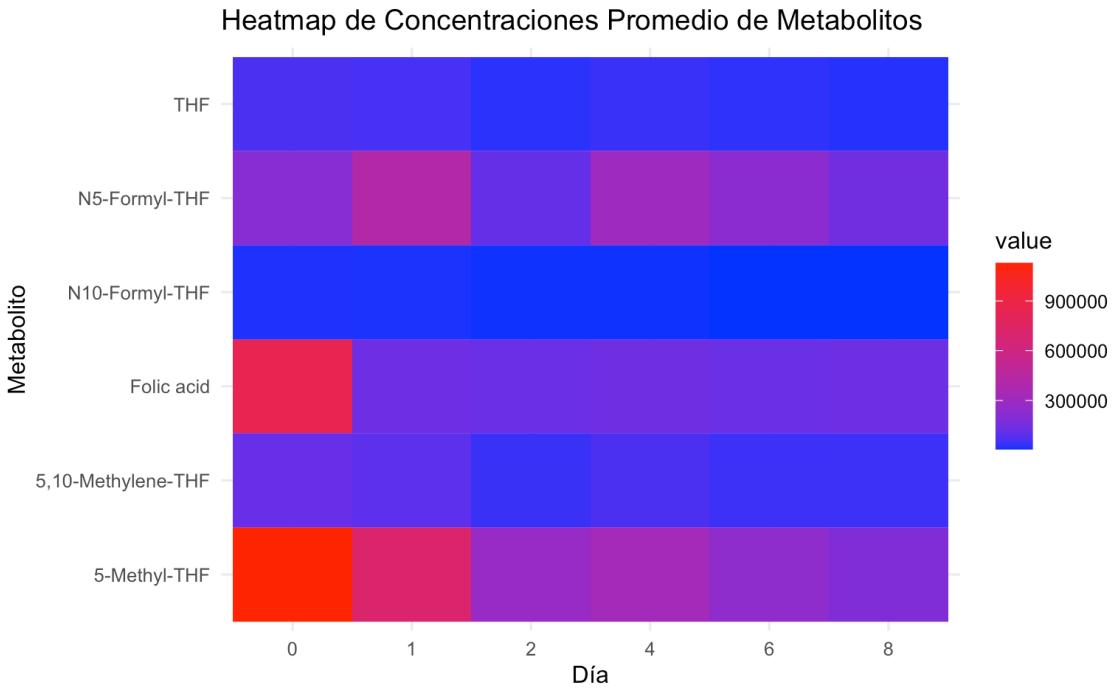

Figura 3 *Heatmap de Concentraciones Promedio de Metabolitos*

Finalmente, he realizado un **análisis de los componentes principales (PCA)**. El PCA nos ayuda para observar patrones entre muestras e identificar variaciones principales en los datos. Para ello necesitamos un dataframe con las 23 columnas de las cuantificaciones y dos columnas más con el día y el número de la réplica:

```
datos_wide <- datos_long %>%
  pivot_wider(names_from = Metabolito, values_from = Concentración) %>%
  arrange(Día, Réplica)  # ordenamos por día y réplica
# seleccionamos únicamente las columnas de metabolitos para el PCA
datos_pca <- datos_wide %>%
  select(-Día, -Réplica)  # excluímos las columnas "Día" y "Réplica"
```

Ahora se realiza el PCA utilizando la función *prcomp()*:

```
# estandarizamos los datos antes del PCA
datos_pca <- scale(datos_pca)
# realizamos el PCA
pca <- prcomp(datos_pca, center = TRUE, scale. = TRUE)


# vemos el del PCA para ver la varianza explicada
summary(pca)
```

```
## Importance of components:
##                           PC1    PC2     PC3     PC4     PC5     PC6
## Standard deviation     2.1861 1.0070 0.35479 0.20486 0.15419 0.12279
## Proportion of Variance 0.7965 0.1690 0.02098 0.00699 0.00396 0.00251
## Cumulative Proportion  0.7965 0.9656 0.98653 0.99352 0.99749 1.00000

# convertimos resultados del PCA en un data frame para representarlo medi
ante ggplot2
pca_data <- as.data.frame(pca$x)
pca_data$Día <- datos_wide$Día  # agregamos la información de día
pca_data$Réplica <- datos_wide$Réplica  # agregamos la información de rép
lica
```

Vemos que el primer componente (PC1) explica el 79.7% de la varianza y el segundo (PC2) el 16.9%. Representamos los datos del PCA:

```
ggplot(pca_data, aes(x = PC1, y = PC2, color = factor(Día))) +
  geom_point(size = 3, alpha = 0.7) +
  labs(title = "PCA de Concentraciones de Metabolitos",
       x = paste0("PC1 (", round(pca$sdev[1]^2 / sum(pca$sdev^2) * 100, 1
), "% varianza)"),
       y = paste0("PC2 (", round(pca$sdev[2]^2 / sum(pca$sdev^2) * 100, 1
), "% varianza)"),
       color = "Día") +
  theme_minimal()
```

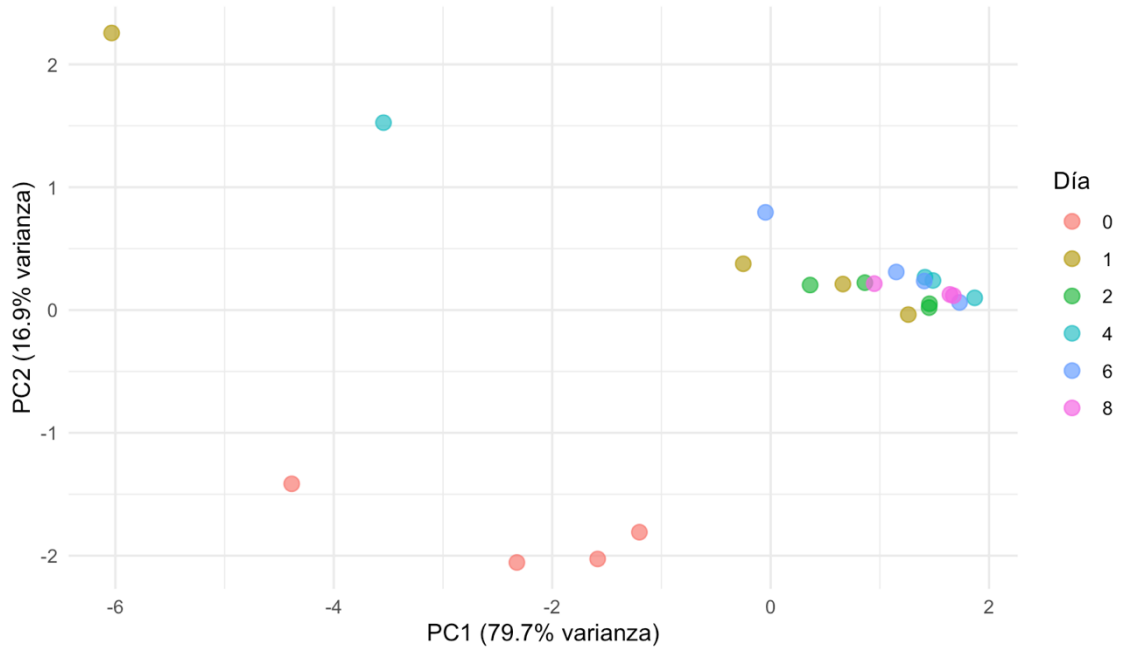

Figura 4 *PCA de Concentraciones de Metabolitos*

Se puede ver que las 4 réplicas a día 0 se agrupan juntas y a los siguientes días se alejan con respecto a las de día 0. Vemos también que hay una réplica a día 1 y a día 4 que se aleja bastante de las demás. Esto nos da una idea de la consistencia de las réplicas.

Por último, he subido los archivos generados (tanto el .csv como el .Rda), el código R, el RMarkdown y el informe utilizando el paquete *usethis*. Después de configurarlo, se utiliza la siguiente función para conectar RStudio y GitHub y crear un nuevo repositorio con el proyecto entero:

```
library(usethis)


usethis::use_github()
```

No muestro la configuración entera pero se encuentra el código en R. Cada vez que actualizo una versión nueva de los archivos, hago *Git --> Commit* y después *Push* para actualizarlo en GitHub.

5. Discusión, conclusiones y limitaciones del estudio

Con este pequeño estudio, hemos buscado en la base de datos *Metabolomics Workbench* un estudio en el cual cuantifican diferentes metabolitos derivados del ácido fólico (**Tabla 1**) en células de leucemia durante 8 días. Las células K562 han sido tratadas con una concentración constante de 100 nM. Una vez he obtenido los datos del estudio utilizando el paquete de Bioconductor denominado metabolomicsWorkbenchR, se ha creado un contenedor de tipo *SummarizedExperiment,* el cual he visto la estructura de los datos y he accedido a los datos y metadatos utilizando el paquete *SummarizedExperiment* de Bioconductor. Además, he guardado el objeto contenedor con los datos y los metadatos en formato binario. Las cuantificaciones de los diferentes metabolitos los he guardado en un archivo separado por comas y he creado un dataframe, para realizar posteriormente las diferentes representaciones.

Primero he representado mediante un gráfico de líneas las concentraciones de los metabolitos a lo largo de los días. Como para cada metabolito hay 4 réplicas diferentes, las he representado de manera separada. Se aprecia en la **Figura 1** que la concentración de los metabolitos 5-methyl-THF y ácido fólico disminuye a lo largo del tiempo en todas las réplicas. Además, el N5-formyl-THF muestra variabilidad entre algunas réplicas.

Como en los demás metabolitos no se aprecian grandes diferencias, he realizado un diagrama de cajas o boxplot (**Figura 2**). Se observa que en todos los casos hay una disminución de la concentración de los metabolitos, excepto el N5-formyl-THF, que es posible que sea debido a su elevada variabilidad en las réplicas a día 1. De manera alternativa, se puede ver si hay algún patrón en la disminución de la concentración de los metabolitos utilizando un mapa de calor o heatmap (**Figura 3**). Se identifica que los grandes cambios ocurren, como he comentado anteriormente, en las concentraciones de los metabolitos de 5-methyl-THF y ácido fólico.

La última gráfica que he hecho para explorar los datos ha consistido en un PCA o análisis de componentes principales con el fin de estudiar si existe una agrupación de los diferentes metabolitos a lo largo de los días por réplicas. Se puede apreciar en la **Figura 4** que las cuatro réplicas están bastante juntas a día 0 y se dispersan a lo largo de los días. Además, hay una réplica a día 1 y a día 4 que está bastante alejada de los demás puntos. Esto puede deberse a la alta presencia de valores anómalos que aparecen en los boxplots.

Se puede concluir de este proyecto que he aprendido a manejar datos reales de estudios de metabolómica utilizando R y paquetes de Bioconductor, y he podido explorar los datos para analizar cambios de concentraciones de los metabolitos analizados y la variabilidad de las réplicas.

En cuanto a las limitaciones de este trabajo, no he encontrado grandes dificultades para el análisis del estudio metabolómico. La única limitación es la falta de sentido biológico, ya que este estudio está hecho con un dataset de un proyecto, por lo que faltaría analizar el estudio de las células sin tratar, y hacer comparaciones entre tratadas y sin tratadas. Entiendo que este trabajo es meramente exploratorio de los datos, y que el siguiente paso será realizar comparaciones entre experimentos y estudios.

Hay que destacar que todos los archivos y los metadatos generados del análisis del estudio, así como el código R empleado se puede mediante el siguiente enlace de GitHub:

https://github.com/sergiroigso/Roig-Soucase-Sergi-PEC1/