1 Resumen

2 Objetivos

3 Métodos

4 Resultados

5 Exploratory Data Analysis Report

6 Discusión

7 Conclusiones

8 Referencias.

PEC1 - Análisis de datos ómicos

Rodrigo Hortal García

2025-04-02

1 Resumen

Utilizaremos un conjunto de datos metabolómicos procedentes del siguiente artículo:

Hartman, A. L., Lough, D. M., Barupal, D. K., Fiehn, O., Fishbein, T., Zasloff, M., & Eisen, J. A. (2009). Human gut microbiome adopts an alternative state following small bowel transplantation. Proceedings of the National Academy of Sciences - PNAS, 106(40), 17187–17192. https://doi.org/10.1073/pnas.0904847106 (https://doi.org/10.1073/pnas.0904847106)

para realizar un análisis metabolómico para ver como afecta el transplante de intestino delgado a la flora intestinal.

2 Objetivos

Se quiere estudiar como afecta un transplante de intestino delgado a la flora intestinal. Para ello se analizan los datos metabolómicos de pacientes antes y después de la operación, con el propósito de comprender las alteraciones en la microbiota y los cambios metabólicos asociados.

3 Métodos

3.1 Origen y naturaleza de los datos.

Se utilizó el conjunto de datos `ST000002` disponible en la plataforma *Metabolomics Workbench*. Estos fueron descargados desde la web en formato de texto `.txt`.

Los datos incluyen:

- 12 muestras (6 antes y 6 después del trasplante)
- 142 metabolitos medidos
- 2 grupos experimentales (antes y después del trasplante)

Al abrir el archivo con un editor, se identificaron tres secciones bien diferenciadas:

- Metadatos del experimento (Filas 0-70): con información general sobre el estudio, incluyendo detalles sobre la institución responsable, el investigador principal y el proyecto.
- Datos de abundancia de metabolitos en muestras (Filas 71-216): esta sección contiene los valores de abundancia de distintos metabolitos en cada muestra:
    - La fila 72 presenta los nombres de las muestras.
    - La fila 73 incluye un factor que indica si la muestra fue tomada antes o después del trasplante.
    - Cada fila posterior representa un metabolito y cada columna una muestra, proporcionando datos cuantitativos sobre la concentración relativa de los metabolitos.
- Información sobre los metabolitos (Filas 217-363): proporciona detalles sobre los metabolitos analizados, incluyendo su nombre, identificadores en bases de datos y otras características relevantes.

3.2 Metodología

Para el análisis de los datos seguiremos los siguientes pasos:

- Carga y limpieza de los datos: importándolos en RStudio y eliminando filas y columnas irrelevantes.
- Tratamiento de valores faltantes.
- Normalización: aplicando normalización para reducir el impacto de variaciones técnicas.
- Análisis exploratorio: análisis de componentes principales (PCA) y heatmap para visualizar agrupaciones y tendencias en los datos.
- Análisis univariante: pruebas de t-test para identificar metabolitos significativamente alterados entre los grupos pre y post-trasplante.
- Análisis discriminante de mínimos cuadrados parciales (PLS-DA) para encontrar patrones que permitan diferenciar entre los dos grupos.

3.3 Herramientas estadísticas y bioinformáticas

- RStudio (R):
    - Biobase: paquete para gestionar datos genómicos y de expresión en R.
    - metabolomicsWorkbenchR: interfaz para trabajar con datos de la base de datos Metabolomics Workbench.
    - SummarizedExperiment: estructura de datos para almacenar y analizar resultados de experimentos biológicos.
    - POMA: herramienta para normalizar y analizar datos metabolómicos.
    - ggplot2: para visualización.
- Notepad++ para abrir y analizar el archivo de texto con los datos.

4 Resultados

Para facilitar el análisis de los datos comenzamos creando un objeto de clase *SummarizedExperiment*. Para ello necesitamos:

- Los datos de expresión (matriz con valores numéricos por muestra).
- Metadatos de filas (información sobre los metabolitos).
- Metadatos de columnas (información sobre las muestras).

```
# Comenzamos cargando los datos brutos
ruta_archivo <- "/home/hortal/Bioinformatica/Datos omicos/PEC1/Dataset_PEC1/ST00
0002_AN000002.txt"

datos_brutos <- readLines(ruta_archivo)
```

A continuación, procedemos a extraer toda la información que necesitamos para construir el objeto:

Comenzamos por los resultados numéricos del experimento, descartando para ello los metadatos del archivo:

```
datos <- read.table(ruta_archivo, skip = 71, nrows = 143, header = TRUE, sep =
"\t", quote = "", check.names = FALSE)
# Eliminamos los factores para dejar solo los datos numéricos
datos <- datos[-1,]
```

Separamos las filas con información sobre el experimento:

```
info_dataset <- read.table(ruta_archivo, skip = 1, nrows = 70, header = FALSE, s
ep = "\t", quote = "", check.names = FALSE)
```

Creamos tablas con información sobre las muestras y los metabolitos:

```
info_muestras <- read.table(ruta_archivo, skip = 71, nrows = 1, header = TRUE, s
ep = "\t", quote = "", check.names = FALSE)
# trasponemos la tabla y convertimos en factor
info_muestras <- as.factor(t(info_muestras[,-1]))

info_metabolitos <- read.table(ruta_archivo, skip = 218, nrows = 142, header = T
RUE, sep = "\t", quote = "", check.names = FALSE)
```

Por último, creamos el objeto *SummarizedExperiment*

```
se <- SummarizedExperiment(
  assays = list(counts = datos[, -1]),
  colData = info_muestras,
  rowData = info_metabolitos,
  metadata = info_dataset
)

colnames(colData(se))[colnames(colData(se)) == "X"] <- "Group"
se
```

```
## class: SummarizedExperiment
## dim: 142 12
## metadata(2): V1 V2
## assays(1): counts
## rownames(142): 2 3 ... 142 143
## rowData names(9): metabolite_name moverz_quant ... other_id
##   other_id_type
## colnames(12): LabF_684508 LabF_684512 ... LabF_684499 LabF_684503
## colData names(1): Group
```

Alternativamente, también podemos descargar el objeto de clase *SummarizedExperiment* directamente desde la base de datos de *Metabolomics Workbench*, utilizando el paquete *metabolomicsWorkbenchR* de *Bioconductor*.

```
SE = do_query(
    context = 'study',
    input_item = 'study_id',
    input_value = 'ST000002',
    output_item = 'SummarizedExperiment'
)

SE
```

```
## class: SummarizedExperiment
## dim: 142 12
## metadata(8): data_source study_id ... description subject_type
## assays(1): ''
## rownames(142): ME641269 ME641270 ... ME641409 ME641410
## rowData names(3): metabolite_name metabolite_id refmet_name
## colnames(12): LabF_684483 LabF_684487 ... LabF_684524 LabF_684528
## colData names(6): local_sample_id study_id ... raw_data Transplantation
```

La principal diferencia entre esta clase y *ExpressionSet* es su mayor flexibilidad en la información de las filas. Mientras que los *ExpressionSet* usan una estructura fija, los *SummarizedExperiment* permiten usar tanto rangos genómicos (`GRanges`) como `DataFrames` arbitrarios.

4.1 Procesamiento y normalización de los datos

Antes de analizar los datos, es importante comprobar y manejar los valores perdidos, ya que pueden afectar los resultados finales:

```
# Número total de valores faltantes
sum(is.na(assay(SE)))
```

```
## [1] 0
```

En este caso, no se ha encontrado ninguno.

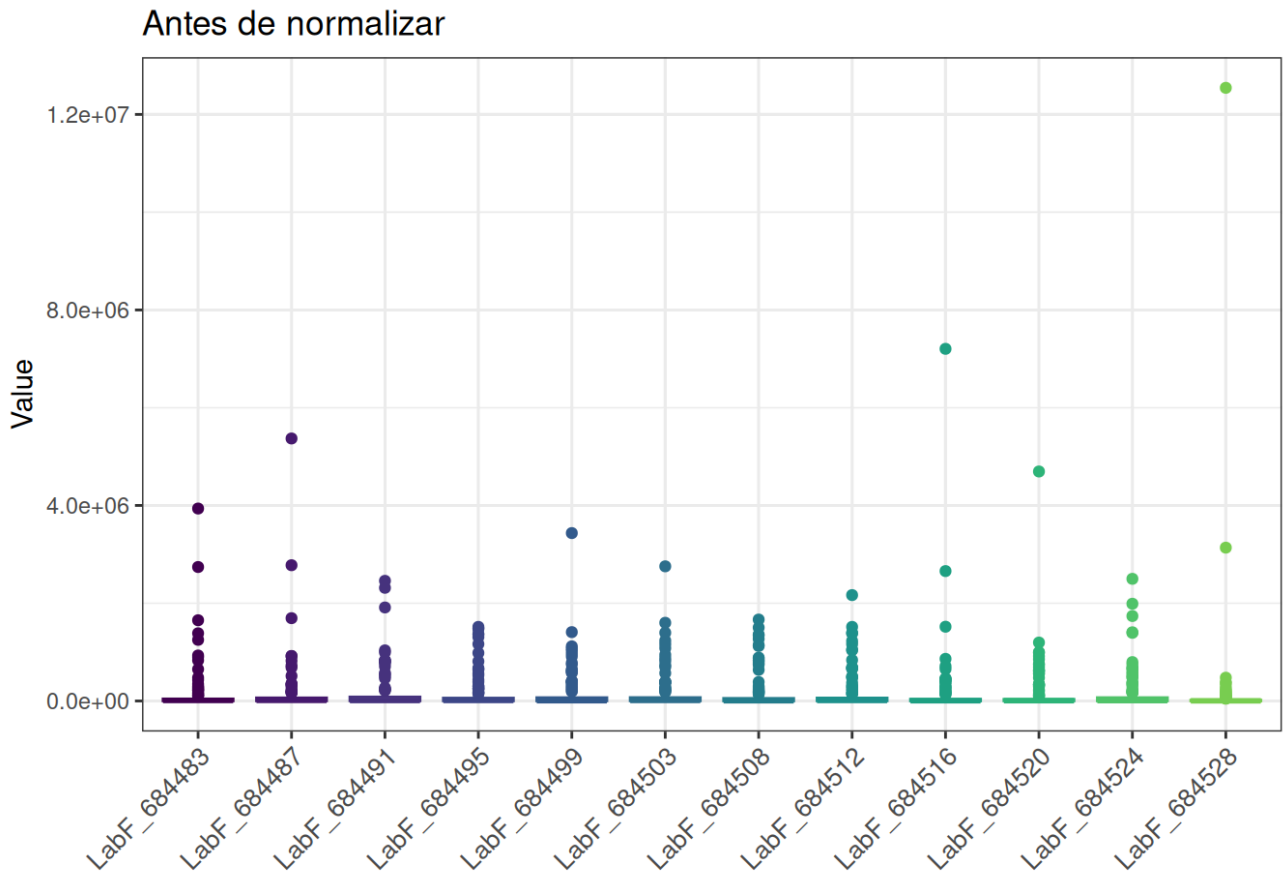
A continuación, para reducir la variabilidad técnica sin reducir la variabilidad biológica, y para facilitar la comparación entre muestras, procedemos con la normalización de los datos. En esta ocasión utilizamos el método de *log_pareto* del paquete `POMA` :

```
SE_norm <- PomaNorm(SE, method = "log_pareto")
# Mantenemos el nombre de los metabolitos
rownames(SE_norm) <- rownames(SE)
SE_norm
```

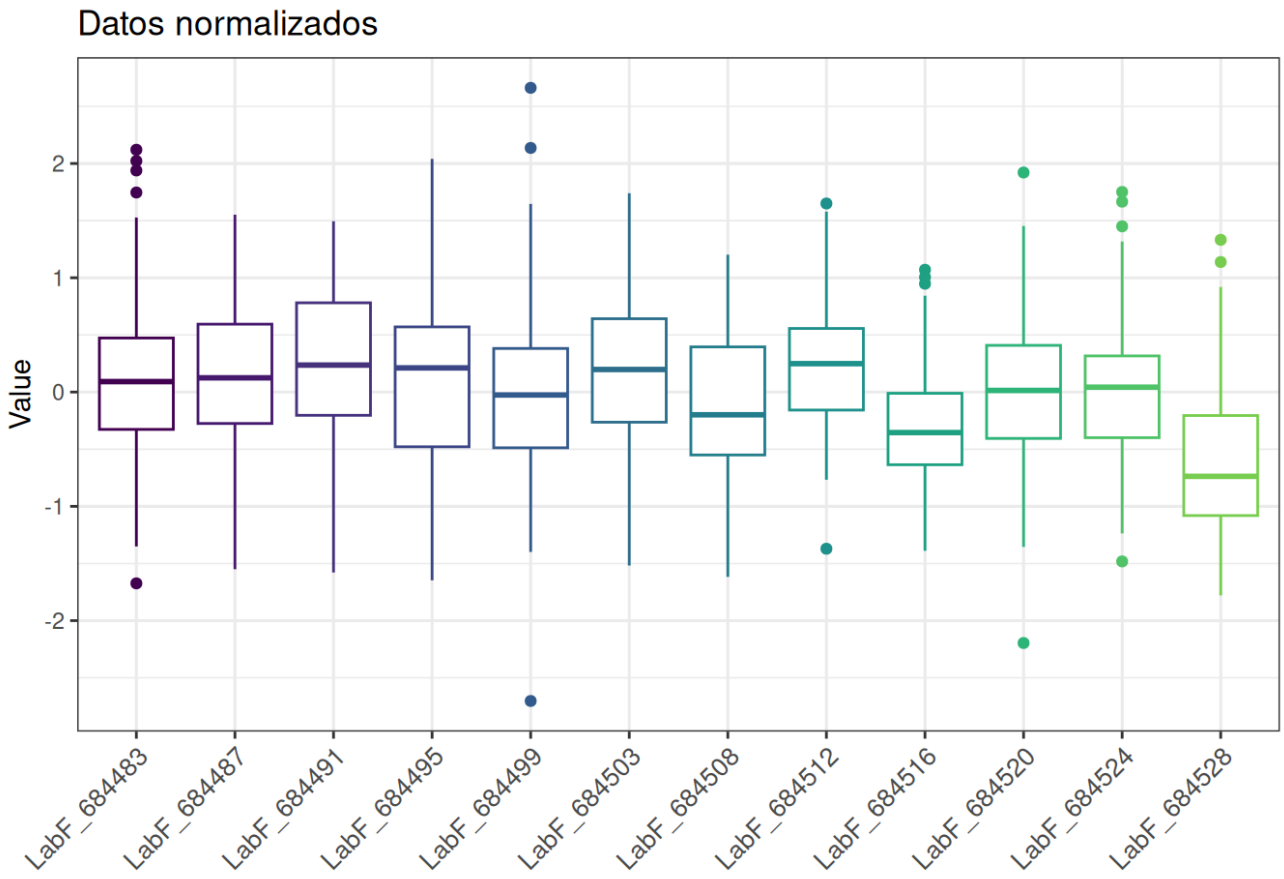
```
## class: SummarizedExperiment
## dim: 142 12
## metadata(0):
## assays(1): ''
## rownames(142): ME641269 ME641270 ... ME641409 ME641410
## rowData names(0):
## colnames(12): LabF_684483 LabF_684487 ... LabF_684524 LabF_684528
## colData names(6): local_sample_id study_id ... raw_data Transplantation
```

Con las funciones *PomaBoxplots* y *PomaDensity* visualizamos cómo ha afectado la normalización a los datos. Los *boxplots* del antes y el después muestran cómo cambia la dispersión y los valores atípicos entre las muestras, mientras que las *densidades* nos permiten observar el cambio en las distribuciones de los metabolitos.

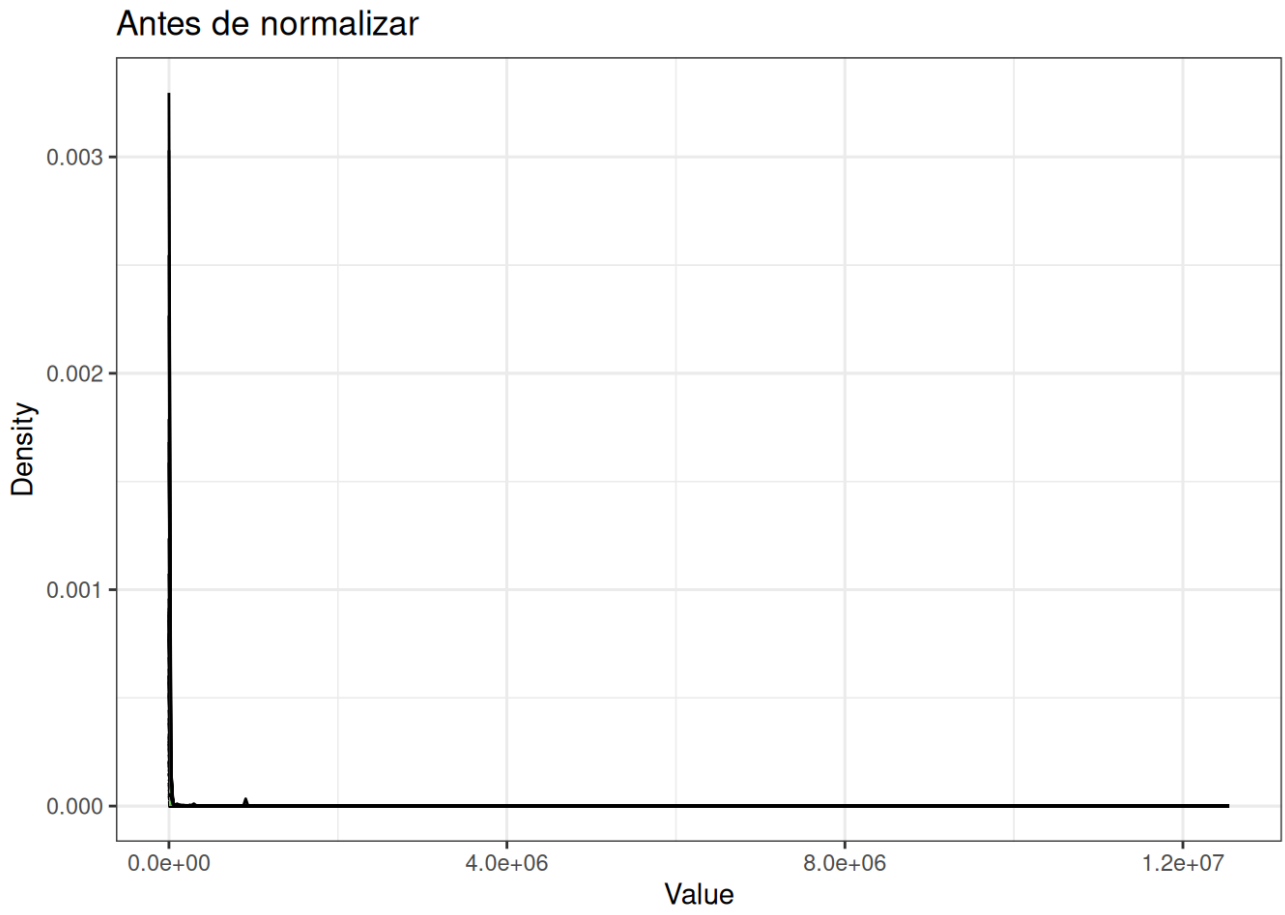
```
# Datos antes de normalizar
PomaBoxplots(SE, group = "samples",
             jitter = FALSE,
             legend_position = "none") +
  ggplot2::ggtitle("Antes de normalizar")
```



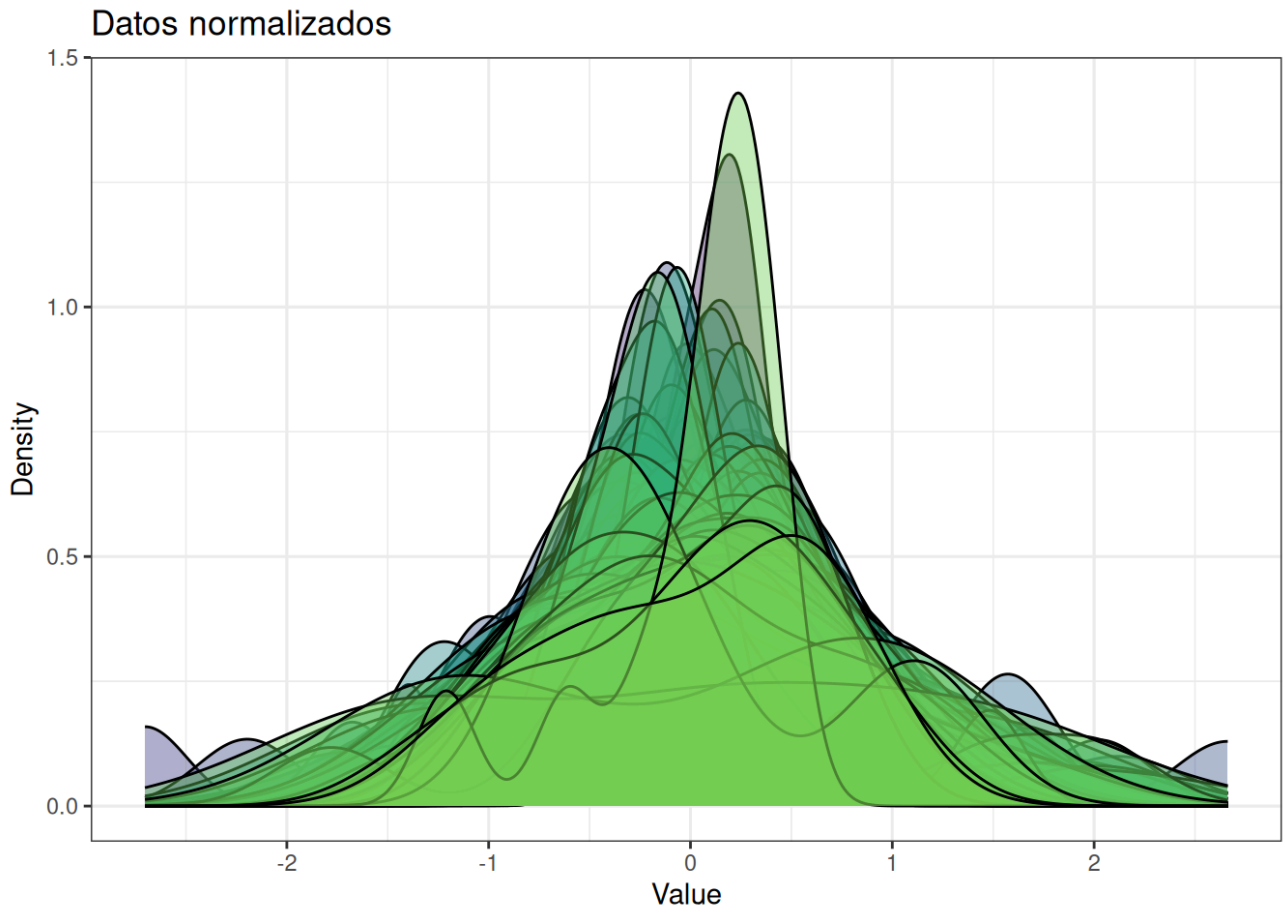

```
# Datos normalizados
PomaBoxplots(SE_norm,
             group = "samples",
             jitter = FALSE,
             legend_position = "none") +
  ggplot2::ggtitle("Datos normalizados")
```




```
# Datos antes de normalizar
PomaDensity(SE,
            group = "features",
            legend_position = "none") +
  ggplot2::ggtitle("Antes de normalizar")
```




```
# Datos normalizados
PomaDensity(SE_norm,
            group = "features",
            legend_position = "none") +
  ggplot2::ggtitle("Datos normalizados")
```



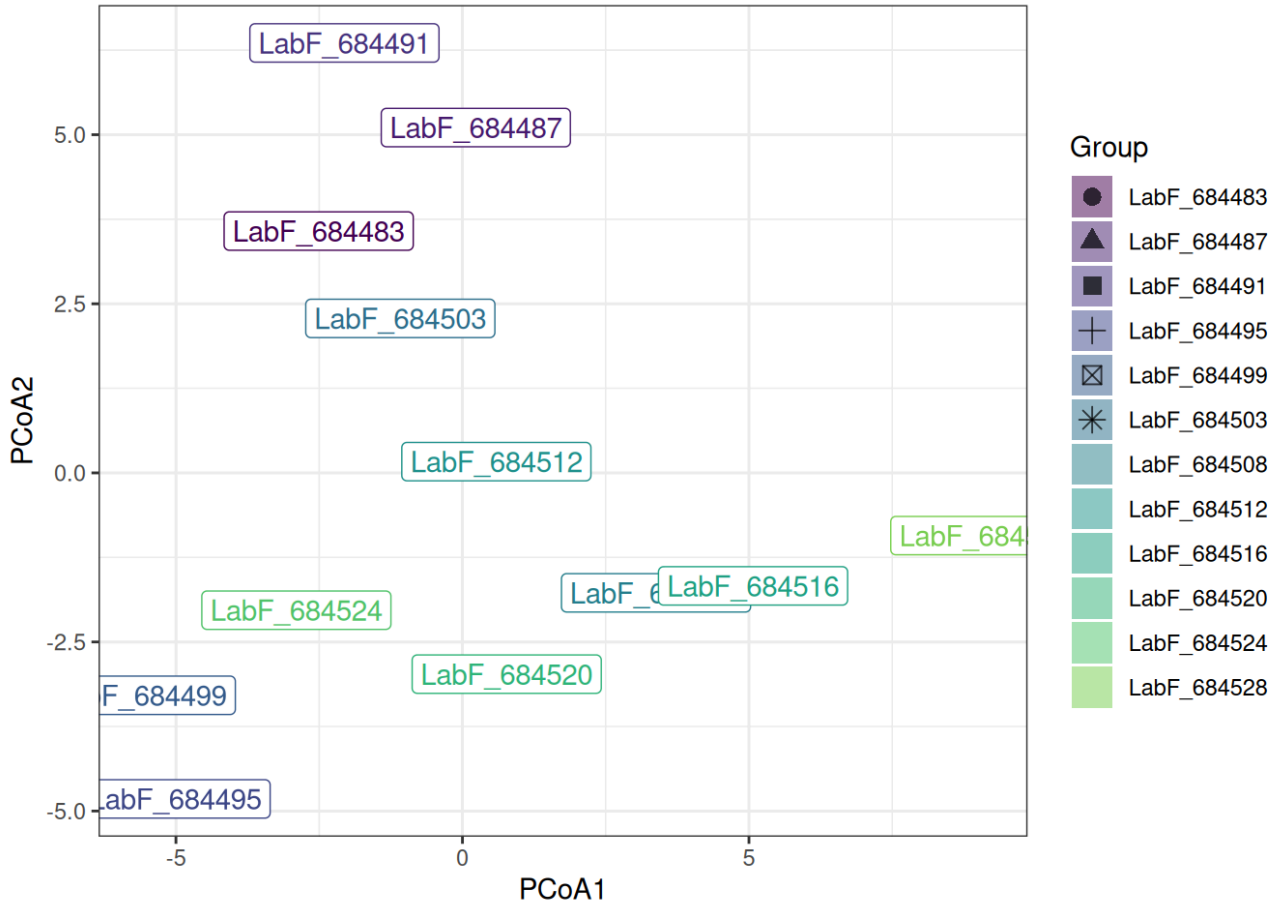
Estos análisis visuales nos muestran que la normalización ha logrado reducir la variabilidad no deseada y mejorar la comparabilidad entre las muestras.

Por ultimo detectamos y eliminamos los valores atipicos, ayuda de `PomaOutliers'

```
PomaOutliers(SE_norm, do = "analyze")$polygon_plot
```

```
## Warning: The shape palette can deal with a maximum of 6 discrete values becau
se more
## than 6 becomes difficult to discriminate
## i you have requested 12 values. Consider specifying shapes manually if you ne
ed
##   that many have them.
```

```
## Warning: Removed 6 rows containing missing values or values outside the scale
range
## (`geom_point()`).
```




```
SE_procesado <- PomaOutliers(SE_norm, do = "clean")
SE_procesado
```

```
## class: SummarizedExperiment
## dim: 142 12
## metadata(0):
## assays(1): ''
## rownames(142): ME641269 ME641270 ... ME641409 ME641410
## rowData names(0):
## colnames(12): LabF_684483 LabF_684487 ... LabF_684524 LabF_684528
## colData names(6): local_sample_id study_id ... raw_data Transplantation
```

Por algún problema en la configuración del *SummarizedExperiment* los grupos no son reconocidos correctamente. Posiblemente por esta razón no es posible ejecutar el resto de test:

4.2 Análisis univariante

El análisis univariante utiliza la función *PomaUnivariate()* para realizar un test t y detectar metabolitos con diferencias significativas entre grupos. Por otro lado, *PomaVolcano()* genera un gráfico de volcán que permite visualizar las diferencias significativas y la magnitud del cambio entre los grupos.

```
PomaUnivariate(SE_procesado, method = "ttest")
PomaVolcano(SE_procesado, pval = "adjusted")
```

4.3 Análisis multivariante

4.3.1 PCA

Se usa para simplificar los datos y encontrar patrones o agrupamientos entre las muestras, facilitando la visualización de diferencias entre grupos.

```
poma_pca <- PomaMultivariate(SE_procesado, method = "pca")
poma_pca$scoresplot +
  ggplot2::ggtitle("Scores Plot")
```

4.3.2 PLS-DA

Ayuda a identificar qué variables distinguen mejor entre grupos, combinando técnicas de regresión y clasificación, y es útil cuando queremos predecir a qué grupo pertenece una muestra.

```
poma_plsda <- PomaMultivariate(SE_procesado, method = "plsda")
poma_plsda$scoresplot +
  ggplot2::ggtitle("Scores Plot")

poma_plsda$errors_plsda_plot +
  ggplot2::ggtitle("Error Plot")
```

5 Exploratory Data Analysis Report

Como algunos análasis han fallado, utilizamos PomaEDA para generar un report de exploración de datos automático:

```
# Utilizamos el objeto creado manualmente
PomaEDA(se)
```

```
##
##
## processing file: POMA_EDA_report.Rmd
```

```
##   |                                                                      |
|   0%  |                                                                     |..
|   3%                     |
|...                                                  |   6% [unnamed-chunk-32]  |
|.....                                                |   9%                     |
|......                                               |  12% [unnamed-chunk-33]  |
|........                                             |  15%                     |
|.........                                            |  18% [unnamed-chunk-34]  |
|...........                                          |  21%                     |
|............                                         |  24% [unnamed-chunk-35]  |
|..............                                       |  27%                     |
|................                                     |  30% [unnamed-chunk-36]  |
|..................                                   |  33%                     |
|...................                                  |  36% [unnamed-chunk-37]  |
|.....................                                |  39%                     |
|......................                               |  42% [unnamed-chunk-38]
```

```
##   |
|.....................                                 |  45%                    |
|.......................                               |  48% [unnamed-chunk-39]
```

```
##   |
|.........................                             |  52%                    |
|...........................                           |  55% [unnamed-chunk-40]
```

```
##   |
|.............................                         |  58%                    |
|...............................                       |  61% [unnamed-chunk-41]
```

```
##   |
|.................................                     |  64%                    |
|...................................                   |  67% [unnamed-chunk-42]  |
|.....................................                 |  70%                    |
|.......................................               |  73% [unnamed-chunk-43]  |
|.........................................             |  76%                    |
|...........................................           |  79% [unnamed-chunk-44]
```

```
##   |
|.............................................         |  82%                    |
|...............................................       |  85% [unnamed-chunk-45]
```

```
##   |
|.................................................     |  88%                    |
|...................................................   |  91% [unnamed-chunk-46]
```

```
##   |
|.....................................................  |  94%                    |
|....................................................... |  97% [unnamed-chunk-47]
```

```
##   |
|.....................................................| 100%
```

```
## output file: POMA_EDA_report.knit.md
```

```
## /snap/rstudio/15/resources/app/bin/quarto/bin/tools/x86_64/pandoc +RTS -K512m
-RTS POMA_EDA_report.knit.md --to html4 --from markdown+autolink_bare_uris+tex_m
ath_single_backslash --output POMA_EDA_report.html --lua-filter /home/hortal/R/x
86_64-pc-linux-gnu-library/4.3/rmarkdown/rmarkdown/lua/pagebreak.lua --lua-filte
r /home/hortal/R/x86_64-pc-linux-gnu-library/4.3/rmarkdown/rmarkdown/lua/latex-d
iv.lua --embed-resources --standalone --variable bs3=TRUE --section-divs --table
-of-contents --toc-depth 3 --template /home/hortal/R/x86_64-pc-linux-gnu-librar
y/4.3/rmarkdown/rmd/h/default.html --no-highlight --variable highlightjs=1 --num
ber-sections --variable theme=bootstrap --mathjax --variable 'mathjax-url=http
s://mathjax.rstudio.com/latest/MathJax.js?config=TeX-AMS-MML_HTMLorMML' --includ
e-in-header /tmp/Rtmp404Q1I/rmarkdown-str2caef59e0d241.html
```

```
##
## Output created: POMA_EDA_report.html
```

5.1 PCA

En este caso se ve separación entre los dos grupos de estudio. El primer componente es capaz de explicar aproximadamente el 25% de la variabilidad de los datos y no observamos ningún valor atípico. Muestra que la microbiota ha cambiado después de la operación.

5.2 Análisis de clusters. Heatmap

Nos muestra la misma información que la PCA, observamos cierta separación entre las muestras de los dos grupos, sin embargo algunas tienen comportamiento anómalo y no se agrupan bien.

6 Discusión

En este estudio hemos analizado cómo el trasplante de intestino delgado afecta la microbiota intestinal. Aunque encontramos diferencias claras entre las muestras antes y después del trasplante, hay algunas limitaciones a tener en cuenta; la muestra es pequeña (solo 12 pacientes), lo que limita la validez de los resultados, y factores externos podrían haber influido.

7 Conclusiones

Los resultados parecen mostrar que el trasplante de intestino delgado puede alterar la microbiota intestinal. Aunque es necesario un análisis en profundidad de los datos para confirmar estos efectos y entender mejor los mecanismos implicados.

8 Referencias.

1. Repositorio de GitHub: Código utilizado para el análisis. Disponible en: https://github.com/rhortalg/Hortal_Garcia_Rodrigo_PEC1.git (https://github.com/rhortalg/Hortal_Garcia_Rodrigo_PEC1.git)
2. BIOCONDUCTOR. *POMA Workflow*. Disponible en: http://bioconductor.jp/packages/3.16/bioc/vignettes/POMA/inst/doc/POMA-demo.html#univariate-analysis (http://bioconductor.jp/packages/3.16/bioc/vignettes/POMA/inst/doc/POMA-demo.html#univariate-analysis)
3. BIOCONDUCTOR. *POMA Demo*. Disponible en: http://bioconductor.jp/packages/3.16/bioc/vignettes/POMA/inst/doc/POMA-demo.html (http://bioconductor.jp/packages/3.16/bioc/vignettes/POMA/inst/doc/POMA-demo.html)
4. BIOCONDUCTOR. *PomaOutliers Function Documentation*. Disponible en: https://rdrr.io/bioc/POMA/man/PomaOutliers.html (https://rdrr.io/bioc/POMA/man/PomaOutliers.html)
5. BIOCONDUCTOR. *POMA EDA*. Disponible en: https://bioconductor.statistik.tu-dortmund.de/packages/3.18/bioc/vignettes/POMA/inst/doc/POMA-eda.html#principal-component-analysis (https://bioconductor.statistik.tu-dortmund.de/packages/3.18/bioc/vignettes/POMA/inst/doc/POMA-eda.html#principal-component-analysis)
6. METABOLOMICS WORKBENCH. *MWRestAPI v1.0*. Disponible en: https://www.metabolomicsworkbench.org/tools/MWRestAPIv1.0.pdf (https://www.metabolomicsworkbench.org/tools/MWRestAPIv1.0.pdf)