Analisis estadistico

Elena Vicario Rubio

2025-04-01

Análisis estadístico

En la primera parte del análisis, vamos a cargar las librerías necesarias, los datos ya pre-procesados del SummarizedExperiment y una exploración básica para asegurarnos que los datos han sido correctamente cargados.

```
# Cargamos las librerías

library(SummarizedExperiment)
```

```
## Cargando paquete requerido: MatrixGenerics
```

```
## Cargando paquete requerido: matrixStats
```

```
## Warning: package 'matrixStats' was built under R version 4.4.3
```

```
##
## Adjuntando el paquete: 'MatrixGenerics'
```

```
## The following objects are masked from 'package:matrixStats':
##
##     colAlls, colAnyNAs, colAnys, colAvgsPerRowSet, colCollapse,
##     colCounts, colCummaxs, colCummins, colCumprods, colCumsums,
##     colDiffs, colIQRDiffs, colIQRs, colLogSumExps, colMadDiffs,
##     colMads, colMaxs, colMeans2, colMedians, colMins, colOrderStats,
##     colProds, colQuantiles, colRanges, colRanks, colSdDiffs, colSds,
##     colSums2, colTabulates, colVarDiffs, colVars, colWeightedMads,
##     colWeightedMeans, colWeightedMedians, colWeightedSds,
##     colWeightedVars, rowAlls, rowAnyNAs, rowAnys, rowAvgsPerColSet,
##     rowCollapse, rowCounts, rowCummaxs, rowCummins, rowCumprods,
##     rowCumsums, rowDiffs, rowIQRDiffs, rowIQRs, rowLogSumExps,
##     rowMadDiffs, rowMads, rowMaxs, rowMeans2, rowMedians, rowMins,
##     rowOrderStats, rowProds, rowQuantiles, rowRanges, rowRanks,
##     rowSdDiffs, rowSds, rowSums2, rowTabulates, rowVarDiffs, rowVars,
##     rowWeightedMads, rowWeightedMeans, rowWeightedMedians,
##     rowWeightedSds, rowWeightedVars
```

```
## Cargando paquete requerido: GenomicRanges
```

```
## Cargando paquete requerido: stats4
```

```
## Cargando paquete requerido: BiocGenerics
```

```
##
## Adjuntando el paquete: 'BiocGenerics'
```

```
## The following objects are masked from 'package:stats':
##
```

```
##     IQR, mad, sd, var, xtabs
## The following objects are masked from 'package:base':
##
##     anyDuplicated, aperm, append, as.data.frame, basename, cbind,
##     colnames, dirname, do.call, duplicated, eval, evalq, Filter, Find,
##     get, grep, grepl, intersect, is.unsorted, lapply, Map, mapply,
##     match, mget, order, paste, pmax, pmax.int, pmin, pmin.int,
##     Position, rank, rbind, Reduce, rownames, sapply, saveRDS, setdiff,
##     table, tapply, union, unique, unsplit, which.max, which.min
## Cargando paquete requerido: S4Vectors
##
## Adjuntando el paquete: 'S4Vectors'
## The following object is masked from 'package:utils':
##
##     findMatches
## The following objects are masked from 'package:base':
##
##     expand.grid, I, unname
## Cargando paquete requerido: IRanges
## Warning: package 'IRanges' was built under R version 4.4.2
##
## Adjuntando el paquete: 'IRanges'
## The following object is masked from 'package:grDevices':
##
##     windows
## Cargando paquete requerido: GenomeInfoDb
## Warning: package 'GenomeInfoDb' was built under R version 4.4.2
## Cargando paquete requerido: Biobase
## Welcome to Bioconductor
##
##     Vignettes contain introductory material; view with
##     'browseVignettes()'. To cite Bioconductor, see
##     'citation("Biobase")', and for packages 'citation("pkgname")'.
##
## Adjuntando el paquete: 'Biobase'
## The following object is masked from 'package:MatrixGenerics':
##
##     rowMedians
## The following objects are masked from 'package:matrixStats':
##
##     anyMissing, rowMedians
library(ggplot2)
library(dplyr)
##
## Adjuntando el paquete: 'dplyr'
```

```
## The following object is masked from 'package:Biobase':
##
##     combine

## The following objects are masked from 'package:GenomicRanges':
##
##     intersect, setdiff, union

## The following object is masked from 'package:GenomeInfoDb':
##
##     intersect

## The following objects are masked from 'package:IRanges':
##
##     collapse, desc, intersect, setdiff, slice, union

## The following objects are masked from 'package:S4Vectors':
##
##     first, intersect, rename, setdiff, setequal, union

## The following objects are masked from 'package:BiocGenerics':
##
##     combine, intersect, setdiff, union

## The following object is masked from 'package:matrixStats':
##
##     count

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(tidyr)
```

```
##
## Adjuntando el paquete: 'tidyr'

## The following object is masked from 'package:S4Vectors':
##
##     expand
```

```
library(matrixStats)
library(PCAtools)
```

```
## Cargando paquete requerido: ggrepel

## Warning: package 'ggrepel' was built under R version 4.4.2

##
## Adjuntando el paquete: 'PCAtools'

## The following objects are masked from 'package:stats':
##
##     biplot, screeplot
```

```
# Cargamos los datos
load("data/SummarizedExperiment_PEC1.Rda")
```

```r
# Exploración de los datos
dim(se)
```

```
## [1] 120 277
```

```r
head(se)
```

```
## class: SummarizedExperiment
## dim: 6 277
## metadata(0):
## assays(1): counts
## rownames(6): 1,3-Dihydroxyacetone 2' fucosyl lactose ...
##   2-Aminobutyrate 2-Hydroxybutyrate
## rowData names(3): High.Confidence.Annotation HMDB KEGG
## colnames(277): EP.2.1 EF.1.1 ... BF.14.2 BF.32.2
## colData names(11): Dataset Subject ... diet Description
```

```r
head(colData(se))
```

```
## DataFrame with 6 rows and 11 columns
##                      Dataset     Subject Study.Group       Age   Age.Units
##                  <character> <character> <character> <numeric> <character>
## EP.2.1  HE_INFANTS_MFGM_2019         EP2    Baseline         2      Months
## EF.1.1  HE_INFANTS_MFGM_2019         EF1    Baseline         2      Months
## EF.3.1a HE_INFANTS_MFGM_2019         EF3    Baseline         2      Months
## EF.7.1  HE_INFANTS_MFGM_2019         EF7    Baseline         2      Months
## EP.9.1  HE_INFANTS_MFGM_2019         EP9    Baseline         2      Months
## EP.11.1 HE_INFANTS_MFGM_2019        EP11    Baseline         2      Months
##              Gender                   DOI       Publication.Name
##         <character>           <character>            <character>
## EP.2.1         Male 10.1038/s41598-019-4.. Fecal microbiome and..
## EF.1.1       Female 10.1038/s41598-019-4.. Fecal microbiome and..
## EF.3.1a      Female 10.1038/s41598-019-4.. Fecal microbiome and..
## EF.7.1       Female 10.1038/s41598-019-4.. Fecal microbiome and..
## EP.9.1         Male 10.1038/s41598-019-4.. Fecal microbiome and..
## EP.11.1        Male 10.1038/s41598-019-4.. Fecal microbiome and..
##         BarcodeSequence                  diet Description
##             <character>           <character> <character>
## EP.2.1         AAGCGCTT Standard infant form..        SF_0
## EF.1.1         AAGCGGTA Standard infant form..        SF_0
## EF.3.1a        AAGCTACC Standard infant form..        SF_0
## EF.7.1         AAGCTTCG Experimental infant ..        EF_0
## EP.9.1         AAGCTTGC Experimental infant ..        EF_0
## EP.11.1        AAGGAAGG Standard infant form..        SF_0
```

```r
head(rowData(se))
```

```
## DataFrame with 6 rows and 3 columns
##                      High.Confidence.Annotation        HMDB        KEGG
##                                       <logical> <character> <character>
## 1,3-Dihydroxyacetone                        TRUE HMDB0001882      C00184
## 2' fucosyl lactose                          TRUE HMDB0002098          NA
## 2'-Deoxyinosine                             TRUE HMDB0000071      C05512
## 2'-Deoxyuridine                             TRUE HMDB0000012      C00526
## 2-Aminobutyrate                             TRUE HMDB0000452      C02356
## 2-Hydroxybutyrate                           TRUE HMDB0000008      C05984
```

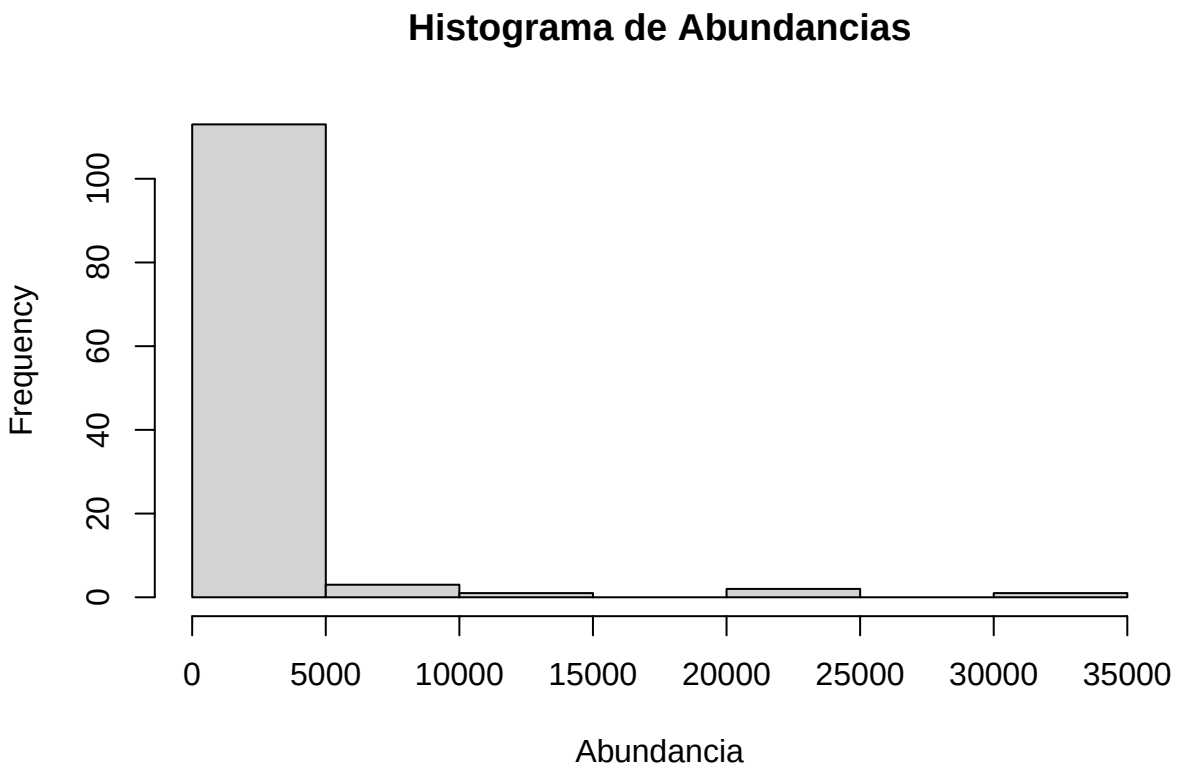
```r
assayNames(se)
```

```
## [1] "counts"
```

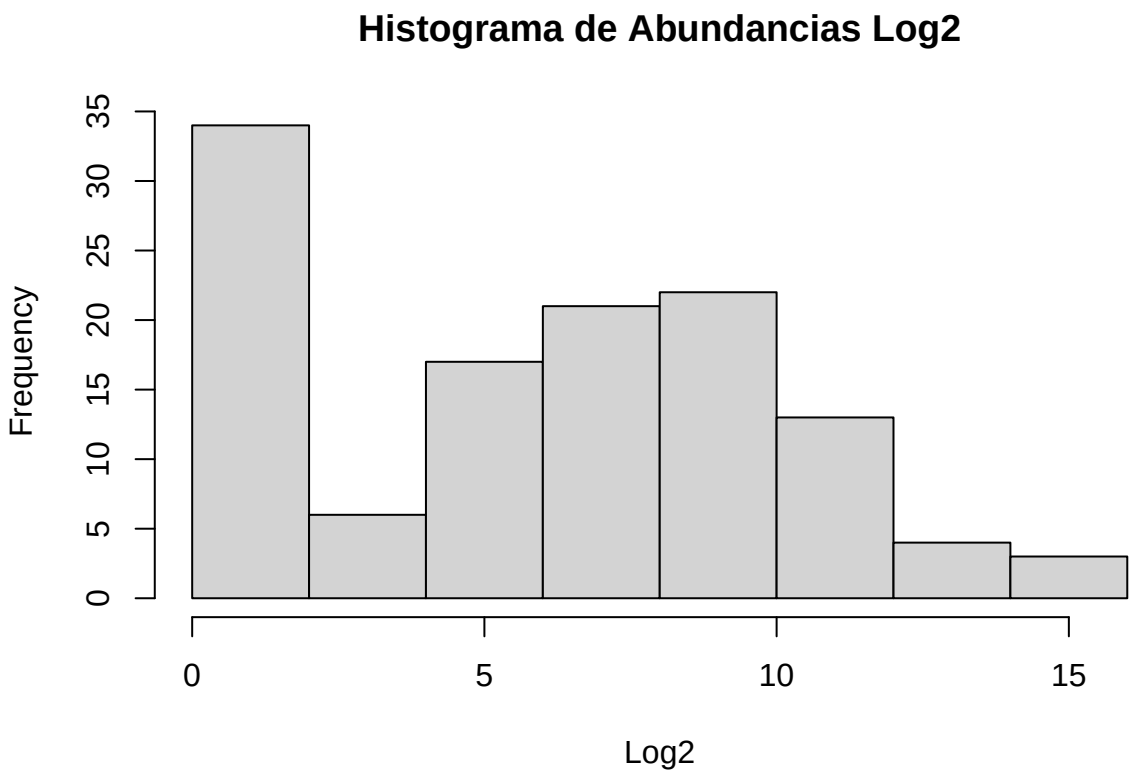
Reducción de dimensional (PCA)

A continuación, preparamos las matrices de datos para su análisis. Esta preparación consiste en una transformación logarítmica para reducir el sesgo y hacer comparables las varianzas, como hace el articulo original (He et al., 2019) y el escalado de datos para centrar la matriz.

```r
# Preparamos los datos
## Extraemos la matriz de datos originales
assay_matriz_original <- assay(se, "counts")

## Comprobamos las distribuciones
### Histograma de las abundancias
hist(assay_matriz_original[, 1], main = "Histograma de Abundancias", xlab = "Abundancia")
```




```r
## Hacemos una transformación logarítmica de los datos
assay_matriz_log <- log2(assay_matriz_original + 1)

### Comprobamos la distribución
hist(assay_matriz_log[, 1], main = "Histograma de Abundancias Log2", xlab = "Log2")
```

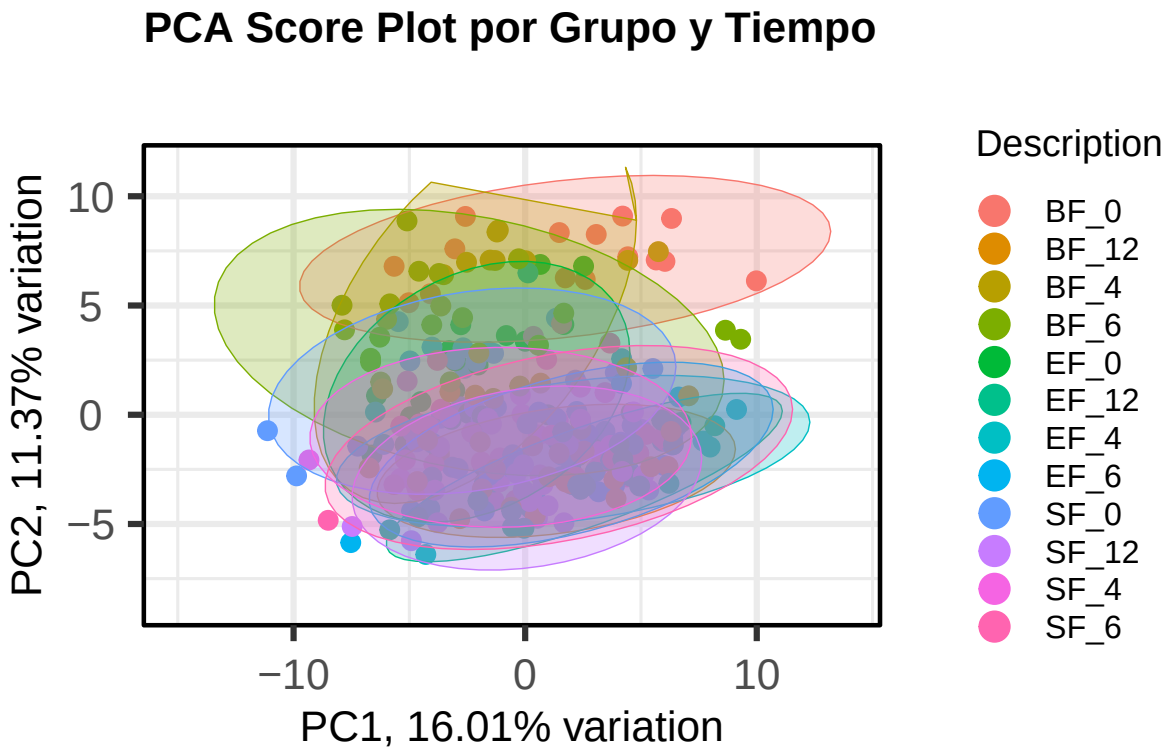



```
## Escalamos los datos
assay_matriz_scal <- t(scale(t(assay_matriz_log), center = TRUE, scale = TRUE))

## Creamos el factor AgeFactor
colData(se)$AgeFactor <- factor(colData(se)$Age)

## Hacemos el análisis PCA
pca_def <- pca(assay_matriz_scal, metadata = colData(se), removeVar = 0.1)
```

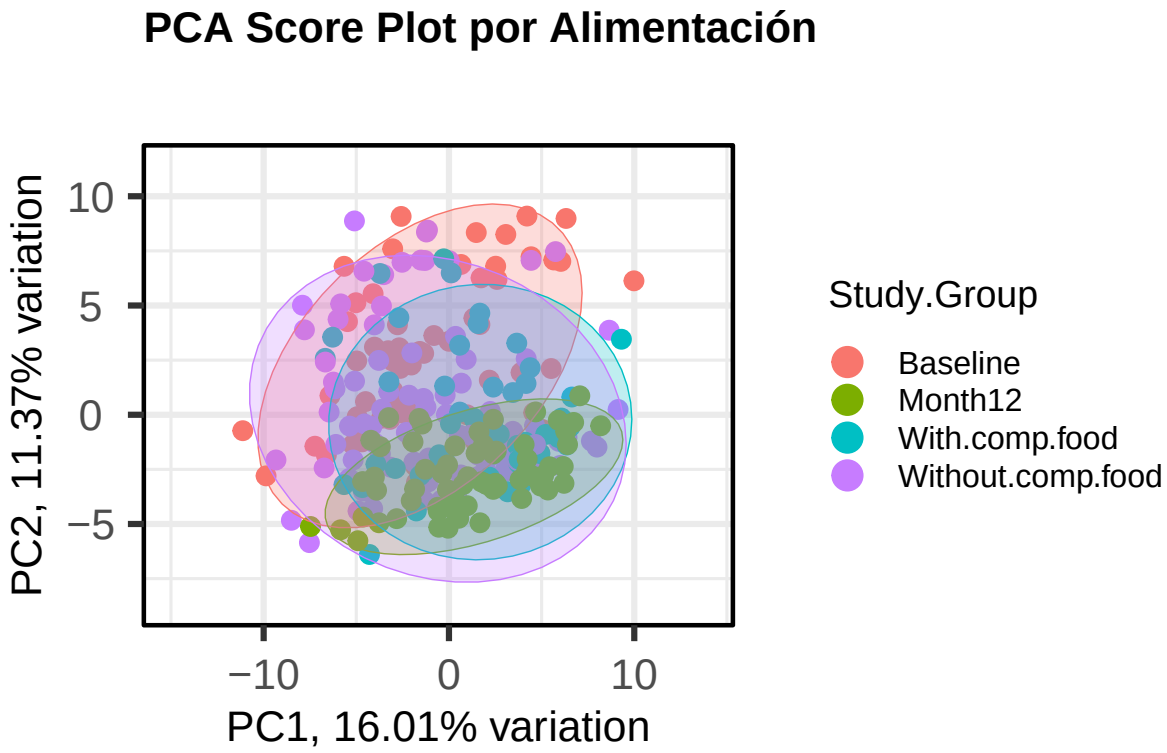
```
## -- removing the lower 10% of variables based on variance
```

Visualización Gráfica

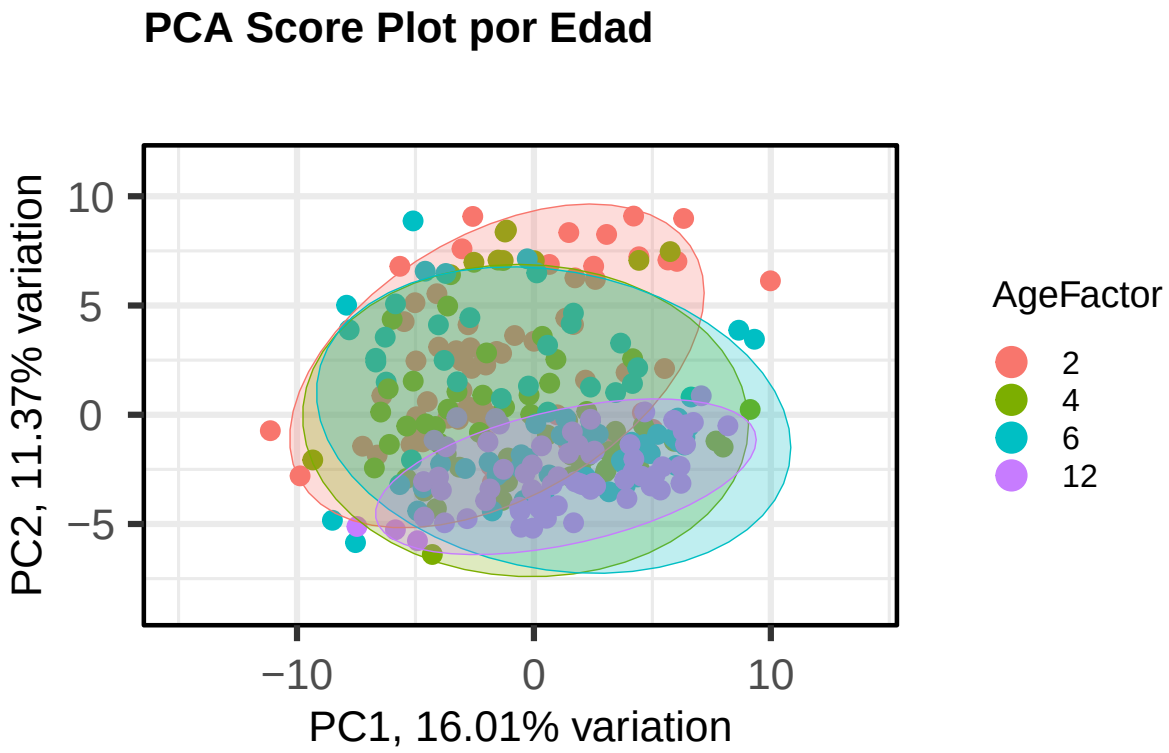
```
# Hacemos los gráficos para la visualización
## PCA Score Plots
biplot(pca_def,
       colby = 'Description',
       ellipse = TRUE,
       ellipseLevel = 0.95,
       pointSize = 3,
       lab = NULL,
       legendPosition = 'right',
       title = 'PCA Score Plot por Grupo y Tiempo')
```




```
biplot(pca_def,
       colby = 'Study.Group',
       ellipse = TRUE,
       ellipseLevel = 0.95,
       pointSize = 3,
       lab = NULL,
       legendPosition = 'right',
       title = 'PCA Score Plot por Alimentación')
```

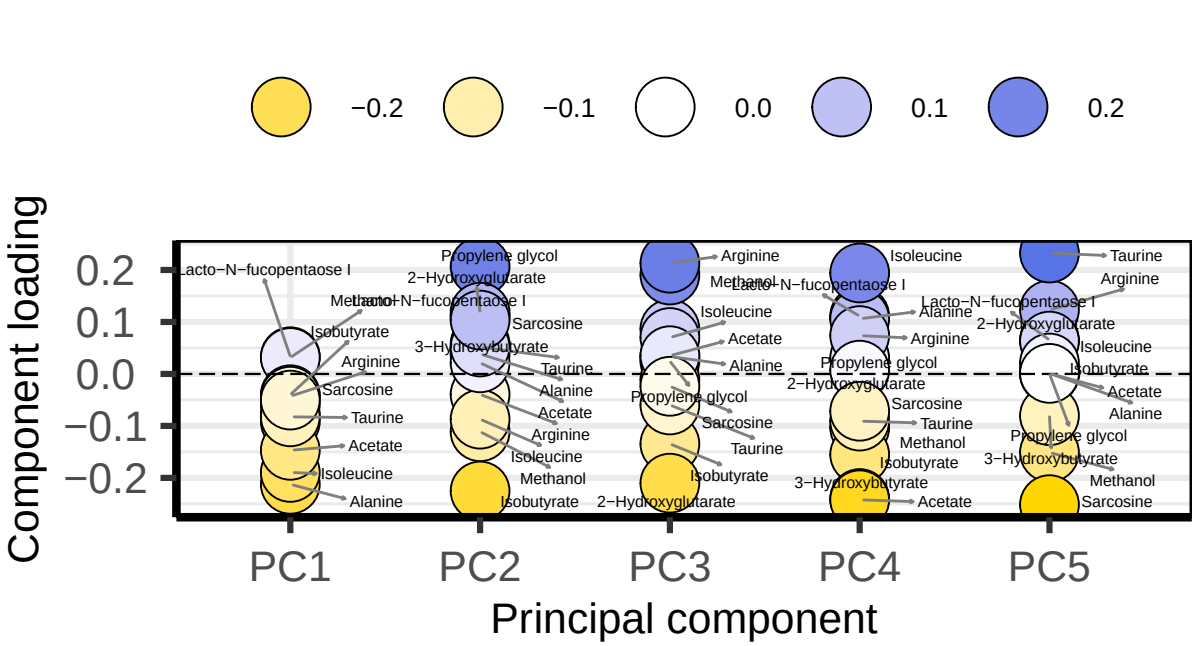



```
colData(se)$AgeFactor <- factor(colData(se)$Age)
biplot(pca_def,
       colby = 'AgeFactor',
       ellipse = TRUE,
       ellipseLevel = 0.95,
       pointSize = 3,
       lab = NULL,
       legendPosition = 'right',
       title = 'PCA Score Plot por Edad')
```




```
## Hacemos un loadings plot
plotloadings(pca_def,
             rangeRetain = 0.01,
             labSize = 2,
             title = 'PCA Loadings Plot')
```

```
## -- variables retained:
## Lacto-N-fucopentaose I, Methanol, Alanine, Propylene glycol, Isobutyrate, Arginine, 2-Hydroxyglutara
## Warning: ggrepel: 5 unlabeled data points (too many overlaps). Consider
## increasing max.overlaps
```

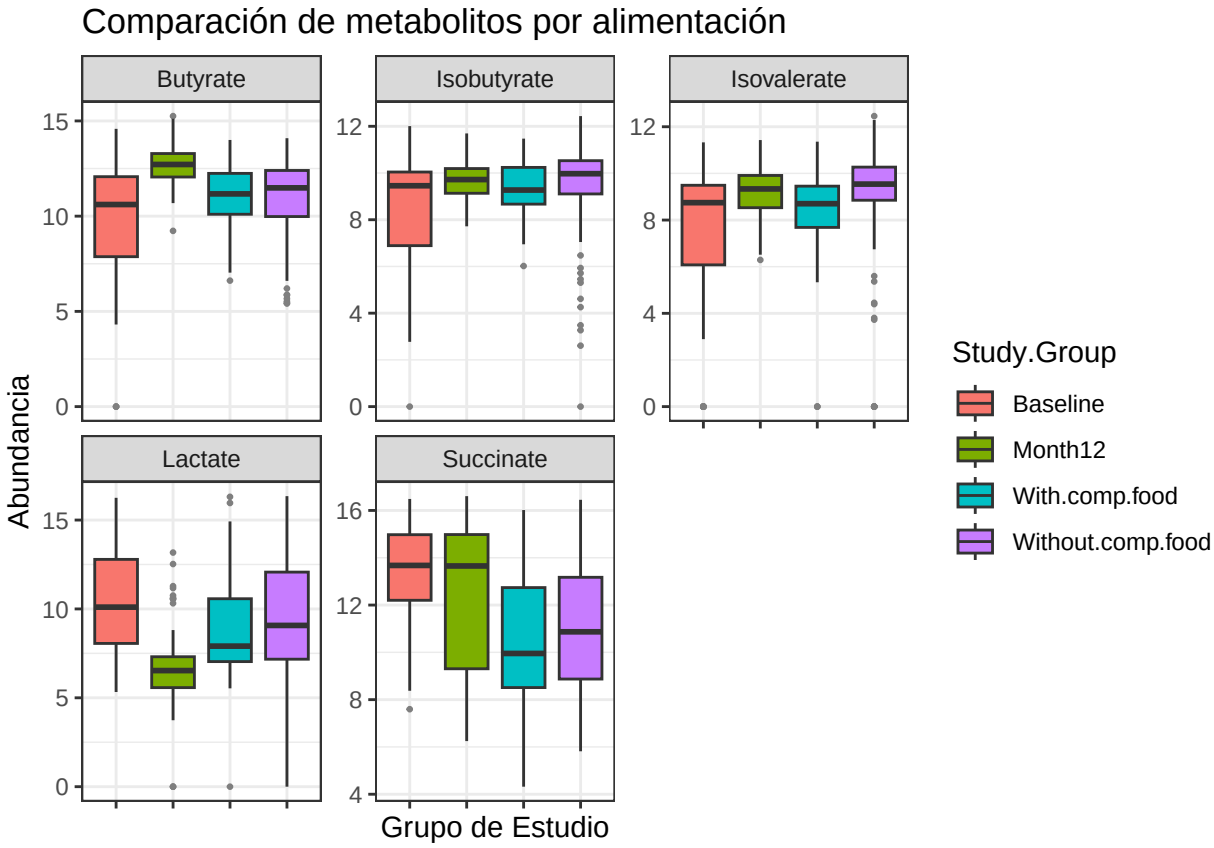


Profundizamos en los metabolitos por edad y por grupo de estudio a través de los gráficos boxplot. Vamos a usar la matriz logarítmica para poder ver las diferencias entre el mismo nivel. Preparando los datos para las gráficas combinando los metadatos del estudio con los datos para poder extraer analizar la información relevante.

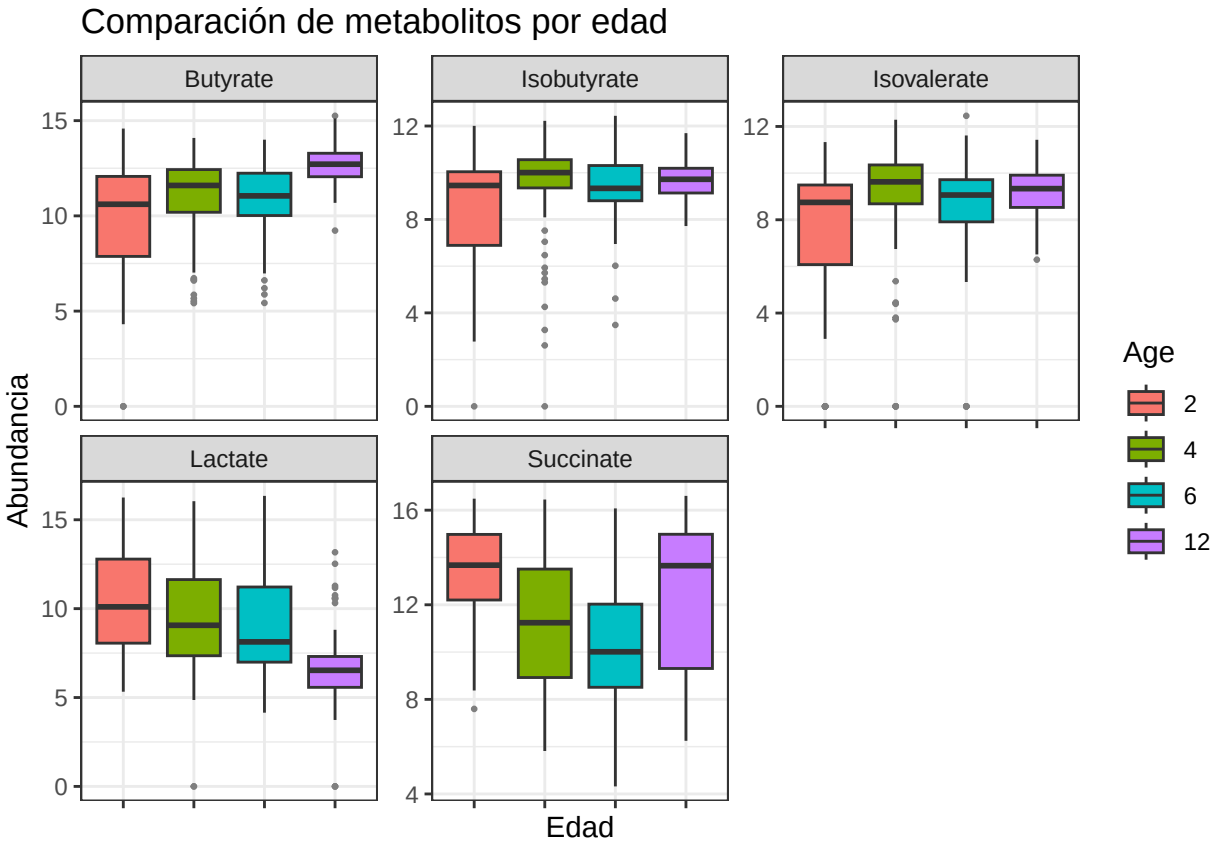
Convertimos los datos a formato largo para que puedan ser compatibles con ggplot2

```
# Seleccionamos los metabolitos de interés
metabolitos_graficos <- c("Lactate", "Succinate", "Butyrate", "Isovalerate", "Isobutyrate")

# Preparamos los datos los gráficos
data_graficos <- as.data.frame(colData(se)[, c("Study.Group", "Age")]) %>%
  cbind(t(assay_matriz_log[metabolitos_graficos, , drop = FALSE]))

# Convertimos a formato largo
data_graficos_long <- data_graficos %>%
  pivot_longer(cols = all_of(metabolitos_graficos),
               names_to = "Metabolite",
               values_to = "Log2Abundance")

# Convertimos la columna 'Age' a factor para facilitar la visualización de los datos
data_graficos_long$Age <- factor(data_graficos_long$Age)

# Creamos los gráficos de caja y bigotes
plot1 <- ggplot(data_graficos_long, aes(x = Study.Group, y = Log2Abundance, fill = Study.Group)) +
  geom_boxplot(outlier.colour = "grey50", outlier.size = 0.5) +
  facet_wrap(~ Metabolite, scales = "free_y", ncol = 3) +
  labs(title = "Comparación de metabolitos por alimentación",
       x = "Grupo de Estudio",
       y = "Abundancia") +
  theme_bw() +
```

```
  theme(axis.text.x = element_blank())

print(plot1)
```




```
plot2 <- ggplot(data_graficos_long, aes(x = Age, y = Log2Abundance, fill = Age)) +
  geom_boxplot(outlier.colour = "grey50", outlier.size = 0.5) +
  facet_wrap(~ Metabolite, scales = "free_y", ncol = 3) +
  labs(title = "Comparación de metabolitos por edad",
       x = "Edad",
       y = "Abundancia") +
  theme_bw()+
  theme(axis.text.x = element_blank())

print(plot2)
```



Interpretación biológica

PCA Score Plot por Grupo y Tiempo En este gráfico se muestra como se distribuyen las muestras según su perfil metabolómico. En el se muestra un solapamiento de casi todas las elipses, siendo relevante la distribución de los grupos de estudio. El grupo BF podemos observarlo en la parte superior del gráfico, mientras que los grupos EF y SF se agrupan en la parte inferior del gráfico, muy mezclado.

Esto puede demostrar que los bebéss alimentados con leche materna y los alimentados con fórmula a a pesar de que la fórmula este suplementada, en los primeros meses tienen un metaboloma fecal muy diferente, y con el paso de los mese pareciéndose mas entre sí.

PCA Score Plot por Alimentación En este gráfico se muestran como se distribuyen as muestras según su alimentación. Podemos observar como todas las elipses están solapadas entre sí, diferenciando el tamaño de la elipse roja (2 meses), la cual es grande, mientras que la elipse de 12 meses es más compacta y pequeña, lo que nos puede decir que la población es mas homogénea. Esto concuerda con el articulo ya que los bebés a los 12 meses no tienen grandes diferencias en el microbioma intestinal.

PCA Score Plot por Edad En este gráfico se muestran como se distribuyen según la edad de los bebéss agrupando todas las dietas, en el se puede observar como todas las elipses están solapadas entre sí, destacando como cambia el perfil metabolomico a lo largo del primer año de vida. Viendo que con 2 meses los datos se agrupan a la derecha del gráfico, ocupando gran parte del dentro, mientras que a los 12 meses, los datos se agrupan mas hacia abajo del gráfico y están mas juntos.

Metabolitos que más contribuyen a PC1 y PC2 En este gráfico se muestran cuales son los principales metabolitos que más contribuyen en cada Componente Principal. Dichos metabolitos están relacionados con la dieta del bebé a lo largo de su primer año de vida.

Comparación de metabolitos por grupo de estudio y por edad En estos gráficos comparamos las cantidades de aminoacidos en las distintas dietas y edad que se muestran en el estudio.

En ellos se puede interpretar como el butirato, el isobutirato y el isovalerato tienen una tendencia positiva con la edad, siendo el nivel más bajo a los 2 meses y creciendo hasta alcanzar su maximo a los 12 meses. En este grupo de metabolitos tambien podemos observar como el paso del tiempo es mucho mas relevante para el aumento de sus oncentraciones que el tipo de suplementacion con comida de la fórmula o la leche materna en los meses intermedios (4-6 meses). Esto concuerda con la maduración del microbioma.

Mientras que, en metabolitos como el lactato y el succinato se muestra una tendencia descendente con la edad, siendo los niveles más altos a los 2 meses y los niveles más bajos a los 12 meses. En el caso del succinato si que se aprecia nivel un poco mas bajos con la suplementacion con comida de la fórmula o la leche materna en los meses intermedios (4-6 meses). Esto tambien confirma el descenso de los procesos de fermentación temprana, para dar paso a mejores digestiones de los bebés.

Referencias:

- Alboukadel. (2018, noviembre 12). How to Customize GGPLot Axis Ticks for Great Visualization. *Datanovia*. https://www.datanovia.com/en/blog/ggplot-axis-ticks-set-and-rotate-text-labels/
- Bakker, J. D. (2024). *PCA*. https://uw.pressbooks.pub/appliedmultivariatestatistics/chapter/pca/
- Bobbitt, Z. (2022, marzo 23). *How to Use pivot_longer() in R*. Statology. https://www.statology.org/pivot_longer-in-r/
- Carmona, A. S. y F. (2024, octubre 23). *Casos y Ejemplos de Análisis Multivariante con R*. https://aspteaching.github.io/AMVCasos/#an%C3%A1lisis-de-componentes-principales
- Gabriel, K. R. (1971). The biplot graphic display of matrices with application to principal component analysis. *Biometrika*, *58*(3), 453-467. https://doi.org/10.1093/biomet/58.3.453
- He, X., Parenti, M., Grip, T., Lönnerdal, B., Timby, N., Domellöf, M., Hernell, O., & Slupsky, C. M. (2019). Fecal microbiome and metabolome of infants fed bovine MFGM supplemented formula or standard formula with breast-fed infants as reference: A randomized controlled trial. *Scientific Reports*, *9*, 11589. https://doi.org/10.1038/s41598-019-47953-4
- *PCAtools*. (s. f.). Bioconductor. Recuperado 2 de abril de 2025, de http://bioconductor.org/packages/PCAtools/
- *PCAtools: Everything Principal Component Analysis*. (s. f.). Recuperado 2 de abril de 2025, de https://www.bioconductor.org/packages/release/bioc/vignettes/PCAtools/inst/doc/PCAtools.html
- *PCAtools package—RDocumentation*. (s. f.). Recuperado 2 de abril de 2025, de https://www.rdocumentation.org/packages/PCAtools/versions/1.99.4
- *Principal Component Analysis*. (s. f.). Recuperado 2 de abril de 2025, de https://masedki.github.io/enseignements/pca.html
- *Rbind() and cbind() functions in R*. (2023, noviembre 19). RCODER. https://r-coder.com/rbind-cbind-r/
- Río, F. M. del. (s. f.). *Tema 8 Tipos de datos: Data frames | Programación con R*. Recuperado 2 de abril de 2025, de https://www4.ujaen.es/~fmartin/R/tipos-de-datos-data-frames.html
- *RPubs—Series de tiempo con el paquete ggplot2*. (s. f.). Recuperado 2 de abril de 2025, de https://rpubs.com/profe_ferro/1048518
- *screeplot function—RDocumentation*. (s. f.). Recuperado 2 de abril de 2025, de https://www.rdocumentation.org/packages/PCAtools/versions/2.5.13/topics/screeplot