Análisis de datos ómicos - PEC 1

Liliana Cifuentes

Contents

1 Resumen **1**

2 Objetivos **1**

3 Métodos **2**

4 Resultados **2**

5 Discusión **7**

6 Conclusión **8**

7 Bibliografía **9**

8 Referencia - Repositorio **9**

9 Anexo **9**

9.1 Visualisación de la clase SummarizedExperiment 9
9.2 Medias y medianas por metabolito 11
9.3 Diagramas de barras de la variabilidad de los metabolitos 12
9.4 Mapa de calor de los metabolitos y muestras 14
9.5 Mapa de calor de la correlación entre los metabolitos 15
9.6 Gráficas de los análisis de componentes principales 16

1 Resumen

El síndrome metabólico de la caquexia es una condición importante en la salud pública al implicar catabolismo tisular, que se asocia a discapacidad y bajas tasas de supervivencia en los pacientes con una enfermedad subyacente. El `dataset human_cachexia` proporciona datos metabolómicos relevantes, pacientes con cáncer de colon o pulmón, con o sin el síndrome, que a través de un análisis y estudio de los mismos permite identificar las alteraciones y patrones en los metabolitos involucrados en el catabolismo muscular. Se implementó `SummarizedExperiment` por su flexibilidad y utilidad en el uso de datos experimentales, a través del cual se realizaron análisis descriptivos de los datos, encontrando que aquellos metabolitos vinculados con la pérdida muscular, tales como la Creatinina, la Glucosa y aquellos pertenecientes a las vías del metabolismo de aminoácidos, presentan variabilidades destacables, con tendencias a niveles mayores.

---

2 Objetivos

- Desarrollar habilidades en el manejo de GitHub y su integración con un proyecto en R, para la gestión de versiones.

- Analizar las diferencias clave entre las clases `ExpressionSet` y `SummarizedExperiment` del proyecto `Bioconductor`, para aplicarlo en el análisis de datos ómicos.
- Realizar un análisis exploratorio del `dataset` metabolómico `human_cachexia`, a partir de la clase `SummarizedExperiment`, con el fin de identificar patrones y relaciones relevantes entre las variables, para obtener una comprensión de los datos en el contexto de estudios sobre caquexia y la vía metabólica del catabolismo muscular en pacientes con cáncer.

---

3 Métodos

Se creó un repositorio en GitHub para la gestión de versiones. Se seleccionó el `dataset` `human_cachexia` por su relevancia en salud pública, y se leyó directamente, desde el repositorio `nutrimetabolomics` en GitHub, en lenguaje de programación R. Posteriormente, se creó la clase `SummarizedExperiment` a través de R, almacenando los datos de expresión de los metabolitos, los metadatos de los metabolitos, los metadatos de las muestras en el atributo correspondiente para cada uno de estos, y creando el atributo `metadata` con la información disponible en el artículo de Eisner et al., 2010 desde donde se extrajeron los datos del `dataset`.

Se realizó un análisis exploratorio de los datos, evaluando inicialmente la estructura de la base de datos y sus metadatos a través de los atributos assay, colData y rowData, verificando a su vez la ausencia de datos con las funciones sum e is.nan. Asimismo, se calculó un resumen estadístico por metabolito a través de la función summary. A partir de lo observado en este análisis, se calcularon las diferencias entre las medias y las medianas de los niveles de los metabolitos con el fin de verificar la simetría en la distribución de las muestras. También, se evaluó la variabilidad de los niveles de los metabolitos a través del cálculo de la desviación estándar para cada metabolito y se graficó en un diagrama de barras.

Se obtuvo un mapa de calor directamente de la matriz correspondiente a los niveles de los metabolitos, buscando identificar patrones en los mismos, y de igual forma se calculó la matriz de correlación entre los metabolitos, la cual fue visualizada a través de un mapa de calor; estos mapas se obtuvieron a través de la función pheatmap. Finalmente, se realizó un análisis de componentes principales (PCA) a través de la función prcomp, con el cual se redujo la dimensionalidad de la base de datos, y se pudieron identificar diferencias en la variabilidad de los datos de acuerdo con el grupo de pacientes (cachexia o control) o las categorías a las cuales se podían asociar los metabolitos de acuerdo con la literatura relacionada con las vías metabólicas, obtenida a partir de (METACYC, 2025).

---

4 Resultados

El `dataset` `human_cachexia` fue seleccionado debido a se trata de una base de datos públicamente disponible a través del repositorio `nutrimetabolomics` en GitHub de Sánchez Pla, A.; lo que permite su exploración y utilización. La información del Data_Catalog, en el mismo repositorio, describe que posee con 77 muestras de orina (47 pacientes con caquexia y 30 pacientes control) y 63 características; y dentro de su descripción se garantiza que las muestras no están emparejadas, que todos los valores de los datos son numéricos, y se detectaron 0 valores faltantes; lo que sugiere que se trata de una base de datos con buena calidad.

Esta base de datos se relaciona con un síndrome metabólico complejo, incapacitante y potencialmente mortal: la caquexia, que está asociada a enfermedades crónicas, como el cáncer, EPOC, CHF, etc.; y se caracteriza por alteraciones en el metabolismo que implican catabolismo tisular (pérdida de masa muscular), lo cual se traduce en pérdida de peso, fuerza y fatiga extrema. Este síndrome al estar relacionado con una enfermedad subyacente tiene un impacto significativo en el paciente al afectar la tolerancia hacia el tratamiento, ya que se asocia a la discapacidad y a bajas tasas de supervivencia, lo que la convierte en una condición relevante en la salud pública. Actualmente se cuenta con ciertos tratamientos, como la suplementación nutricional, sin embargo, estos no son capaces de ralentizar el metabolismo o revertir el síndrome (Pfizer, 2025)

Por lo tanto, esta base de datos metabolómicos en caquexia que incluye pacientes con cáncer de colon o pulmón, con o sin perdida muscular, a los cuales se les realizó un estudio de orina para identificar varios metabolitos excretados e involucrados en el catabolismo muscular, y por lo tanto relacionados con este síndrome (Eisner, R. et al, 2010), permitiría identificar alteraciones y patrones de dichos metabolitos; con el fin de comprender esta relación y como punto de partida para posteriores investigaciones que contribuyan a establecer tratamientos eficaces.

Para leer el `dataset` de metabolómica `human_cachexia`, del repositorio de GitHub proporcionado, se realizarón los siguientes pasos:

```
# Instalar paquete
#install.packages("readr")

# Llamar librerias
library(readr)
library(knitr)

# Definir la URL del archivo
url = "https://raw.githubusercontent.com/nutrimetabolomics/metaboData/refs/heads/main/Datasets/2024-Cach

# Leer el archivo CSV
data = read_csv(url)
```

```
## Rows: 77 Columns: 65
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr  (2): Patient ID, Muscle loss
## dbl (63): 1,6-Anhydro-beta-D-glucose, 1-Methylnicotinamide, 2-Aminobutyrate,...
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

Con la lectura del data frame se visualiza que se trata de una tabla con 77 filas y 65 columnas, con 2 características formato caracter (Patient ID, y Muscle loss) y 63 formato numerico. Luego, se creo el objeto clase `SummarizedExperiment` como se explica en el siguiente código:

```
# Instalar paquetes
#install.packages("BiocManager")
#BiocManager::install("SummarizedExperiment")

# Llamar libreria
library(SummarizedExperiment)

# Crear clase SummarizedExperiment

# Crear argumento assay (datos de expresion de metabolitos)
ensayo = t(as.matrix(data[,!(colnames(data) %in% c("Patient ID", "Muscle loss"))]))
colnames(ensayo) = data$`Patient ID` # Asignar nombre

# Crear argumento rowdata (metadatos de los metabolitos)
rowdata = data.frame(metabolito=tail(colnames(data),-2))

# Crear argumento coldata (metadatos de las muestras)
coldata = data.frame(grupo=data$`Muscle loss`, row.names=data$`Patient ID`)

# Añadir metadatos adicionales sobre el experimento
```

```
metadatos = list(experiment_date = "2010-08-01", protocol = "detección de metabolitos en orina a través

# Crear SummarizedExperiment
experimento = SummarizedExperiment(assay=list(ensayo=ensayo), rowData=rowdata, colData=coldata, metadata

# Guardar en formato .Rda
save(experimento, file = "SummarizedExperiment.Rda")
```

A continuación, se realizó un análisis exploratorio general del objeto clase `SummarizedExperiment` creado.

```
# Instalar paquetes
#install.packages("kableExtra")

# Llamar librerias
library(dplyr)

# Definir tabla con la informacion del atributo metadata
infoDF = data.frame(matrix(rep(NA, 2 * length(metadatos)), ncol = 2))
for (i in 1:length(metadatos)) {
    infoDF[i, 1] = names(metadatos)[i]
    infoDF[i, 2] = metadatos[i]
}
colnames(infoDF) = c("Campo", "Descripción")

# Imprimir
infoDF %>%
    kableExtra::kable() %>%
    kableExtra::kable_styling()
```

Campo	Descripción
experiment_date	2010-08-01
protocol	detección de metabolitos en orina a través del espectrómetro de RMN Varian INOVA de 600 MHz, son
researchers	Roman Eisner, Cynthia Stretch, Thomas Eastman, Jianguo Xia, David Hau, Sambasivarao Damaraju,
study_name	Learning to predict cancer-associated skeletal muscle wasting from 1H-NMR profiles of urinary metabo

Se verificó que la clase estuviera estructurada correctamente visualizando las primeras 6 características de la matriz `assay`, y las primeras y últimas 6 muestras con el respectivo grupo al que pertenecen, y las primeras 6 características (metabolitos); encontrando que se construyó correctamente (ver Anexo).

Se verificó que efectivamente la base de datos no tiene valores faltantes:

```
# Obtener el numero de valores faltantes en la matriz
sum(is.nan(assay(experimento)))
```

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

Se realizó un resumen estadístico por metabolito con el fin de identificar de forma más rápida si los valores encontrados en la muestra en general, corresponden a los rangos esperados para cada metabolito; encontrando que todos los metabolitos tienen valores mínimos y máximos extremos.

```
# Realizar resumen estadistico descriptivo de los datos por metabolito
summary(t(assay(experimento)))
```

```
##  1,6-Anhydro-beta-D-glucose 1-Methylnicotinamide 2-Aminobutyrate
##  Min.   :  4.71             Min.   :   6.42      Min.   :  1.28
```

```
##  1st Qu.: 28.79             1st Qu.:  15.80      1st Qu.:  5.26
##  Median : 45.60             Median :  36.60      Median : 10.49
##  Mean   :105.63             Mean   :  71.57      Mean   : 18.16
##  3rd Qu.:141.17             3rd Qu.:  73.70      3rd Qu.: 19.49
##  Max.   :685.40             Max.   :1032.77      Max.   :172.43
##  2-Hydroxyisobutyrate 2-Oxoglutarate    3-Aminoisobutyrate 3-Hydroxybutyrate
##  Min.   : 4.85        Min.   :   5.53   Min.   :   2.61    Min.   :  1.70
##  1st Qu.:15.80        1st Qu.:  22.42   1st Qu.:  11.70    1st Qu.:  5.99
##  Median :32.46        Median :  55.15   Median :  22.65    Median : 11.70
##  Mean   :37.25        Mean   : 145.09   Mean   :  76.76    Mean   : 21.72
##  3rd Qu.:54.60        3rd Qu.:  92.76   3rd Qu.:  56.26    3rd Qu.: 29.96
##  Max.   :93.69        Max.   :2465.13   Max.   :1480.30    Max.   :175.91
##  3-Hydroxyisovalerate 3-Indoxylsulfate  4-Hydroxyphenylacetate    Acetate
##  Min.   :  0.92       Min.   :  27.66   Min.   : 15.49         Min.   :  3.49
##  1st Qu.:  5.26       1st Qu.:  82.27   1st Qu.: 41.68         1st Qu.: 16.28
##  Median : 12.55       Median : 144.03   Median : 70.11         Median : 39.65
##  Mean   : 21.65       Mean   : 218.88   Mean   :112.02         Mean   : 66.14
##  3rd Qu.: 30.27       3rd Qu.: 333.62   3rd Qu.:145.47         3rd Qu.: 86.49
##  Max.   :164.02       Max.   :1043.15   Max.   :796.32         Max.   :411.58
##     Acetone          Adipate          Alanine          Asparagine
##  Min.   :  2.29   Min.   :  1.55   Min.   :  16.78   Min.   :  6.69
##  1st Qu.:  4.95   1st Qu.:  6.11   1st Qu.:  78.26   1st Qu.: 20.49
##  Median :  7.10   Median : 10.18   Median : 194.42   Median : 42.10
##  Mean   : 11.43   Mean   : 24.76   Mean   : 273.56   Mean   : 62.28
##  3rd Qu.: 10.49   3rd Qu.: 19.11   3rd Qu.: 399.41   3rd Qu.: 89.12
##  Max.   :206.44   Max.   :327.01   Max.   :1312.91   Max.   :273.14
##     Betaine         Carnitine         Citrate           Creatine
##  Min.   :  2.29   Min.   :  2.18   Min.   :   59.74   Min.   :   2.75
##  1st Qu.: 28.79   1st Qu.: 14.44   1st Qu.:  788.40   1st Qu.:  17.64
##  Median : 64.72   Median : 23.81   Median : 1790.05   Median :  44.26
##  Mean   : 90.32   Mean   : 52.09   Mean   : 2235.35   Mean   : 126.83
##  3rd Qu.:127.74   3rd Qu.: 60.95   3rd Qu.: 3071.74   3rd Qu.: 117.92
##  Max.   :391.51   Max.   :487.85   Max.   :13629.61   Max.   :1863.11
##    Creatinine    Dimethylamine      Ethanolamine        Formate
##  Min.   : 1002   Min.   :  41.26   Min.   :  16.12   Min.   :   6.42
##  1st Qu.: 3498   1st Qu.: 142.59   1st Qu.:  86.49   1st Qu.:  53.52
##  Median : 7631   Median : 304.90   Median : 204.38   Median :  95.58
##  Mean   : 8734   Mean   : 358.17   Mean   : 276.26   Mean   : 147.40
##  3rd Qu.:12333   3rd Qu.: 454.86   3rd Qu.: 407.48   3rd Qu.: 167.34
##  Max.   :33860   Max.   :1556.20   Max.   :1436.55   Max.   :1480.30
##     Fucose          Fumarate        Glucose          Glutamine
##  Min.   :  5.70   Min.   : 0.79   Min.   :  26.84   Min.   :  23.34
##  1st Qu.: 29.37   1st Qu.: 2.23   1st Qu.:  80.64   1st Qu.: 113.30
##  Median : 61.56   Median : 4.10   Median : 210.61   Median : 225.88
##  Mean   : 88.67   Mean   : 8.44   Mean   : 559.85   Mean   : 306.87
##  3rd Qu.:123.97   3rd Qu.: 7.85   3rd Qu.: 407.48   3rd Qu.: 445.86
##  Max.   :407.48   Max.   :96.54   Max.   :8690.62   Max.   :1685.81
##     Glycine          Glycolate      Guanidoacetate     Hippurate
##  Min.   :  38.09   Min.   :  5.42   Min.   :  7.03   Min.   :   92.76
##  1st Qu.: 262.43   1st Qu.: 50.91   1st Qu.: 33.78   1st Qu.:  492.75
##  Median : 528.48   Median :130.32   Median : 64.72   Median : 1224.15
##  Mean   : 880.72   Mean   :187.99   Mean   : 86.37   Mean   : 2286.84
##  3rd Qu.:1096.63   3rd Qu.:267.74   3rd Qu.:108.85   3rd Qu.: 2921.93
##  Max.   :5064.45   Max.   :720.54   Max.   :561.16   Max.   :19341.34
```

```
##    Histidine        Hypoxanthine      Isoleucine        Lactate
##  Min.   :  14.15   Min.   :  3.78   Min.   : 1.790   Min.   :   7.32
##  1st Qu.:  66.69   1st Qu.: 20.70   1st Qu.: 3.900   1st Qu.:  35.52
##  Median : 174.16   Median : 40.04   Median : 7.170   Median :  81.45
##  Mean   : 292.64   Mean   : 61.10   Mean   : 8.709   Mean   : 158.46
##  3rd Qu.: 419.89   3rd Qu.: 83.93   3rd Qu.:11.250   3rd Qu.: 139.77
##  Max.   :1863.11   Max.   :265.07   Max.   :40.040   Max.   :3640.95
##     Leucine           Lysine        Methylamine    Methylguanidine
##  Min.   :  2.51   Min.   : 10.49   Min.   : 1.51   Min.   :  1.70
##  1st Qu.:  9.12   1st Qu.: 30.27   1st Qu.: 5.26   1st Qu.:  4.26
##  Median : 19.11   Median : 69.41   Median :14.73   Median :  7.85
##  Mean   : 24.36   Mean   :108.79   Mean   :17.38   Mean   : 15.32
##  3rd Qu.: 31.19   3rd Qu.:121.51   3rd Qu.:24.05   3rd Qu.: 19.30
##  Max.   :103.54   Max.   :788.40   Max.   :52.46   Max.   :141.17
##  N,N-Dimethylglycine O-Acetylcarnitine  Pantothenate    Pyroglutamate
##  Min.   :  0.79      Min.   :  1.23    Min.   :  2.59   Min.   :  21.33
##  1st Qu.:  7.03      1st Qu.:  3.94    1st Qu.: 11.13   1st Qu.:  68.72
##  Median : 21.98      Median : 11.47    Median : 22.65   Median : 157.59
##  Mean   : 26.35      Mean   : 19.73    Mean   : 44.88   Mean   : 211.45
##  3rd Qu.: 40.04      3rd Qu.: 20.91    3rd Qu.: 41.26   3rd Qu.: 301.87
##  Max.   :120.30      Max.   :254.68    Max.   :692.29   Max.   :1064.22
##     Pyruvate       Quinolinate         Serine          Succinate
##  Min.   :  0.90   Min.   :  5.21   Min.   :  16.12   Min.   :  1.72
##  1st Qu.:  4.85   1st Qu.: 26.58   1st Qu.:  83.10   1st Qu.:  8.58
##  Median : 13.46   Median : 51.42   Median : 142.59   Median : 30.88
##  Mean   : 21.29   Mean   : 66.44   Mean   : 197.69   Mean   : 60.23
##  3rd Qu.: 29.08   3rd Qu.: 87.36   3rd Qu.: 270.43   3rd Qu.: 74.44
##  Max.   :184.93   Max.   :259.82   Max.   :1248.88   Max.   :589.93
##     Sucrose          Tartrate         Taurine          Threonine
##  Min.   :   6.49   Min.   :  2.20   Min.   :  17.81   Min.   :  8.25
##  1st Qu.:  19.30   1st Qu.:  6.89   1st Qu.:  99.48   1st Qu.: 31.82
##  Median :  40.85   Median : 12.94   Median : 249.64   Median : 64.07
##  Mean   : 113.23   Mean   : 40.00   Mean   : 525.12   Mean   : 95.36
##  3rd Qu.:  94.63   3rd Qu.: 25.79   3rd Qu.: 665.14   3rd Qu.:137.00
##  Max.   :2079.74   Max.   :837.15   Max.   :4272.69   Max.   :450.34
##   Trigonelline     Trimethylamine N-oxide   Tryptophan        Tyrosine
##  Min.   :  10.07   Min.   :  55.7         Min.   :  8.67   Min.   :  4.22
##  1st Qu.:  53.52   1st Qu.: 175.9         1st Qu.: 21.33   1st Qu.: 23.57
##  Median : 114.43   Median : 383.8         Median : 46.99   Median : 60.34
##  Mean   : 270.44   Mean   : 652.2         Mean   : 66.24   Mean   : 81.76
##  3rd Qu.: 340.36   3rd Qu.: 735.1         3rd Qu.: 96.54   3rd Qu.:113.30
##  Max.   :2252.96   Max.   :5486.2         Max.   :259.82   Max.   :539.15
##      Uracil           Valine           Xylose        cis-Aconitate
##  Min.   :  3.10   Min.   :  4.10   Min.   :  10.07   Min.   :  12.94
##  1st Qu.: 11.94   1st Qu.: 12.18   1st Qu.:  29.96   1st Qu.:  36.23
##  Median : 27.39   Median : 33.12   Median :  50.40   Median : 129.02
##  Mean   : 35.56   Mean   : 35.67   Mean   : 100.93   Mean   : 204.22
##  3rd Qu.: 44.26   3rd Qu.: 50.40   3rd Qu.:  89.12   3rd Qu.: 254.68
##  Max.   :179.47   Max.   :160.77   Max.   :2164.62   Max.   :1863.11
##   myo-Inositol    trans-Aconitate  pi-Methylhistidine tau-Methylhistidine
##  Min.   : 11.59   Min.   :  4.90   Min.   :  11.36    Min.   :  8.00
##  1st Qu.: 30.27   1st Qu.: 12.43   1st Qu.:  67.36    1st Qu.: 27.39
##  Median : 78.26   Median : 26.84   Median : 162.39    Median : 68.72
##  Mean   :135.40   Mean   : 40.63   Mean   : 370.29    Mean   : 89.69
```

```
##  3rd Qu.:167.34   3rd Qu.: 57.40   3rd Qu.: 387.61    3rd Qu.:130.32
##  Max.   :854.06   Max.   :217.02   Max.   :2697.28    Max.   :317.35
```

Se establecieron las diferencias entre la media y la mediana de los metabolitos, para definir la simetría en la distribución de los niveles de los metabolitos, y se encontró que en todos los casos la media era mayor que la mediana (ver Anexo).

Se realizó un análisis de variabilidad para establecer aquellos metabolitos que más cambios evidencian entre las muestras, observando que hay una variabilidad relativamente baja entre todos los metabolitos, a excepción del Citrato, Creatinina, Glucosa, Glicina, Hipurato, Lactato, y la trimetilamina N-oxido, donde se pueden destacar el Citrato con aproximadamente 2000, el Hipurato con aproximadamente 3000 y la Creatinina con más de 6000 (ver Anexo).

También se hizo un mapa de calor de los niveles de metabolítos, donde no se observaron patrones entre las muestras. Asimismo, se realizó un análisis de correlación entre los metabolitos, donde se identificaron tres grupos entre los niveles evidenciados (ver Anexo).

Se realizó un análisis de componentes principales, para reducir la dimensionalidad de los datos, y poder identificar patrones en la variabilidad de los niveles de metabolitos, basado en: los grupos de pacientes (donde caquexia presenta una mayor dispersión que el grupo de control, cuyos datos se encuentran más agrupados) y las categorías de metabolitos (encontrando mayor dispersión en las categorías metabolismo de carbohidratos, de aminoácidos, vías de detoxificación y microbioma, y biomarcadores de estrés y energía) (ver Anexo).

---

5 Discusión

Los reultados de la creación del objeto `SummarizedExperiment` permite confirmar que fue creado correctamente, con los datos de expresión de los metabolitos almacenados en `assay`, los metadatos de los metabolitos en `rowData`, y los metadatos de las muestras en `colData`. Además, coinciden las dimensiones con lo realizado: 63 metabolitos en filas, y 77 muestras en columnas.

De acuerdo con los resultados de los estadísticos descriptivos, se evidencia que todos los metabolitos tienen valores mínimos y máximos extremos, indicando una alta variabilidad, algunos con más de 1000 de diferencia. Evaluando las medidas de tendencia central, se encontró que en todos los casos la media era mayor que la mediana, indicando una tendencia hacia una mayor concentración de los metabolitos, en especial Glucosa, Glicina, Citrato, Hipurato y Creatinina.

En general se observa que hay una variabilidad relativamente baja entre todos los metabolitos, a excepción del Citrato, Creatinina, Glucosa, Glicina, Hipurato, Lactato, y la trimetilamina N-oxido, donde se pueden destacar el Citrato con aproximadamente 2000, el Hipurato con aproximadamente 3000 y la Creatinina con más de 6000, indicando que las muestras poseen una alta variabilidad para estos metabolitos específicos; estos últimos son característicos en la pérdida de masa muscular y en la disminución en la capacidad energética de las células, comunes en la caquexia.

En el mapa de calor se observa que no se evidencian patrones destacables en los niveles de los metabolitos. Sin embargo, al realizar el análisis de correlación entre los metabolitos, se evidencian patrones en tres grupos, que pueden estar asociados con las rutas metabólicas en las cuales están involucrados estos metabolitos.

Posteriormente, al aplicar el análisis de componentes principales, teniendo en cuenta los grupos de pacientes, se observa que, en los tres primeros componentes, el grupo de caquexia presenta una mayor dispersión que el grupo de control, cuyos datos se encuentran más agrupados. Por su parte, entre las categorías de metabolitos, definidos de acuerdo con su pertenencia a alguna vía metabólica, se encontró que, en metabolismo de carbohidratos, de aminoácidos, vías de detoxificación y microbioma, y biomarcadores de estrés y energía presentan mayor dispersión que las demás categorías. Esto concuerda con que el metabolismo de carbohidratos y aminoácidos es fundamental para la disponibilidad de nutrientes, y al verse afectado en pacientes por cáncer, estos nutrientes son secuestrados por las células tumorales, que termina provocando el catabolismo muscular; por otro lado, una alteración en las vías de detoxificación y microbioma puede afectar la respuesta inmune

e inflamatoria de los tejidos, lo cual está ampliamente evidenciado en cáncer de colon y que se encuentra siendo estudiado en cáncer de pulmón (American Association for Cancer Research, 2020). Además, los biomarcadores de estrés y energía son indicadores del estrés metabólico y el estado energético de las células, asociado directamente con el desarrollo del cáncer y de la caquexia.

Para destacar, se evidenció que la Creatinina tiende a tener mayor variabilidad y asimetría hacia valores altos, la cual está relacionada con el catabolismo muscular, al ser un producto de la degradación de la creatina, una molécula específica de la energía muscular, lo cual coincide con lo encontrado en Eisner et al. (2010). Asimismo, la Glucosa tuvo un comportamiento similar a la Creatinina, y también se destaca por estar relacionada con la pérdida muscular de acuerdo con lo encontrado por los autores.

Esto también está vinculado con el hallazgo de que los metabolitos asociados con las vías del metabolismo de aminoácidos presentaron mayor variabilidad, como lo expresan los autores, al estar relacionado con el metabolismo de proteínas, posiblemente involucradas con el tejido muscular.

Teniendo en cuenta lo observado anteriormente, se puede destacar que las diferencias clave entre la clase `ExpressionSet` y `SummarizedExperiment` consisten en:

- **Modulo:** Ambas clases hacen parte del proyecto `Bioconductor`, por un lado, `ExpressionSet` hace parte del paquete Biobase; mientras que `SummarizedExperiment` hace parte del paquete con el mismo nombre `SummarizedExperiment`.
- **Estructura:** `ExpressionSet` almacena los datos de expresión en una matriz `assayData`, además es posible disponer de los metadatos de la característica de interés (genes, exones...) en el data frame `featureData` y/o los metadatos de las muestras a través del data frame `phenoData`. Por su parte la clase `SummarizedExperiment` almacena los datos de expresión en una matriz `assay`, los metadatos de la característica de interés están en el data frame `rowData`, y es posible disponer los metadatos de las muestras a través del data frame `colData`.
- **Flexibilidad en las filas de los metadatos de la característica de interés:** `SummarizedExperiment` permite manejar la información de las filas de manera flexible, ya sea de forma convencional a través del data frame (descrito en la viñeta anterior) o integrando el objeto `GRanges` a través de su subclase `RangedSummarizedExperiment` para rangos genómicos; por otro lado el `ExpressionSet` tiene una estructura predeterminada y establecida donde siempre se almacena la información de la característica de interés a través del data frame (descrito en la viñeta anterior).
- **Coordinación de los metadatos y los datos de expresión para manipular subconjuntos:** Para eliminar subconjuntos con información no deseada, a pesar de que el `ExpressionSet` también trabaja con metadatos, esta clase no tiene una coordinación automática con los datos de expresión, por lo que es necesario una verificación manual; mientras que `SummarizedExperiment` permite realizarlo de forma sincronizada ya que al coordinar estas dos fuentes de información, es posible eliminar subconjuntos completos sin error.

(Morgan, M., 2023), (Irizarry, R., S.F.), (Morgan, M., 2020), (Falcon, S., 2007), (Davis, S., 2014), y (R Documentation, S.F.)

---

6 Conclusión

El `SummarizedExperiment` es una clase más actualizada y flexible para el adecuado procesamiento y análisis de datos ómicos ya que sincroniza la información de los datos de expresión con los metadatos garantizando estudios sin errores en la manipulación de subconjuntos de información; además, posibilita la integración con otros objetos como `GRanges` para potenciar su utilidad.

También, a partir de los datos fue posible evidenciar los patrones en los niveles de metabolitos relacionados con la caquexia, de acuerdo con los resultados esperados con base en lo evidenciado en el artículo del estudio del dataset. De esta forma, se observó como aquellos metabolitos vinculados con la pérdida muscular, tales

como la Creatinina, la Glucosa y aquellos pertenecientes a las vías del metabolismo de aminoácidos, presentan variabilidades destacables, con tendencias a niveles mayores.

Finalmente, se evidenció la utilidad de gestionar proyectos de R con el apoyo del control de versiones con Git, ya que permite realizar un seguimiento a los cambios realizados en código y documentación, y mantener y almacenar los proyectos en un entorno seguro.

---

7 Bibliografía

- Pfizer. (2025) Caquexia. Pfizer. Recuperado de: https://www.pfizer.com/disease-and-conditions/cachexia
- Eisner et al. (2010) Learning to predict cancer-associated skeletal muscle wasting from 1H-NMR profiles of urinary metabolites. Metabolomics 7:25-34. https://doi.org/10.1007/s11306-010-0232-9
- Morgan, M. et al. (2023). SummarizedExperiment for Coordinating Experimental Assays, Samples, and Regions of Interest. Bioconductor. Recuperado de: https://bioconductor.org/packages/release/bioc/vignettes/SummarizedExperiment/inst/doc/SummarizedExperiment.html
- Irizarry, R. & Love, M. (S.F.) SummarizedExperiment class in Depth. GitHub. Recuperado de: https://genomicsclass.github.io/book/pages/bioc1_summex.html
- Morgan, M. et al. (2020). RangedSummarizedExperiment-class: RangedSummarizedExperiment objects. R Documentation. Recuperado de: https://rdrr.io/bioc/SummarizedExperiment/man/RangedSummarizedExperiment-class.html
- Falcon, S. et al. (2007). An Introduction to Bioconductor's ExpressionSet Class. Bioconductor. Recuperado de: https://www.bioconductor.org/packages/devel/bioc/vignettes/Biobase/inst/doc/ExpressionSetIntroduction.pdf
- Davis, S. (2014). Introduction to the ExpressionSet. GitHub. Recuperado de: https://seandavi.github.io/BiocBrazil2014/vignettes/ExpressionSetSlides.pdf
- R Documentation. (S.F.). ExpressionSet: Class to Contain and Describe High-Throughput Expression Level Assays. R Documentation. Recuperado de: https://www.rdocumentation.org/packages/Biobase/versions/2.32.0/topics/ExpressionSet
- METACYC. (2025). Database. METACYC. Recuperado de: https://metacyc.org/
- American Association for Cancer Research. (2020). El microbioma pulmonar puede afectar la patogénesis y el pronóstico del cáncer de pulmón. American Association for Cancer Research. Recuperado de: https://ecancer.org/es/news/19061-el-microbioma-pulmonar-puede-afectar-la-patogenesis-y-el-pronostico-del-cancer-de-pulmon

---

8 Referencia - Repositorio

Dirección del repositorio: https://github.com/L-Cifuentes-Y/Cifuentes-Yepes-LilianaCarolina-PEC1

9 Anexo

9.1 Visualisación de la clase SummarizedExperiment

```
# Imprimir primeras 6 filas de atributo assay
head(assay(experimento)) %>%
```

```
  kableExtra::kable() %>%
  kableExtra::kable_styling()
```

	PIF_178	PIF_087	PIF_090	NETL_005_V1	PIF_115	PIF_110	NETL_019_V
1,6-Anhydro-beta-D-glucose	40.85	62.18	270.43	154.47	22.20	212.72	151.4
1-Methylnicotinamide	65.37	340.36	64.72	52.98	73.70	31.82	36.6
2-Aminobutyrate	18.73	24.29	12.18	172.43	15.64	18.36	8.6
2-Hydroxyisobutyrate	26.05	41.68	65.37	74.44	83.93	80.64	42.5
2-Oxoglutarate	71.52	67.36	23.81	1199.91	33.12	47.94	223.6
3-Aminoisobutyrate	1480.30	116.75	14.30	555.57	29.67	17.46	56.2

```
# Imprimir primeras 6 filas de atributo colData
head(colData(experimento)) %>%
  kableExtra::kable() %>%
  kableExtra::kable_styling()
```

	grupo
PIF_178	cachexic
PIF_087	cachexic
PIF_090	cachexic
NETL_005_V1	cachexic
PIF_115	cachexic
PIF_110	cachexic

```
# Imprimir ultimass 6 filas de atributo colData
tail(colData(experimento)) %>%
  kableExtra::kable() %>%
  kableExtra::kable_styling()
```

	grupo
PIF_112	control
NETCR_019_V2	control
NETL_012_V1	control
NETL_012_V2	control
NETL_003_V1	control
NETL_003_V2	control

```
# Imprimir primeras 6 filas de atributo rowData
head(rowData(experimento)) %>%
  kableExtra::kable() %>%
  kableExtra::kable_styling()
```

	metabolito
1,6-Anhydro-beta-D-glucose	1,6-Anhydro-beta-D-glucose
1-Methylnicotinamide	1-Methylnicotinamide
2-Aminobutyrate	2-Aminobutyrate
2-Hydroxyisobutyrate	2-Hydroxyisobutyrate

2-Oxoglutarate	2-Oxoglutarate
3-Aminoisobutyrate	3-Aminoisobutyrate

9.2 Medias y medianas por metabolito

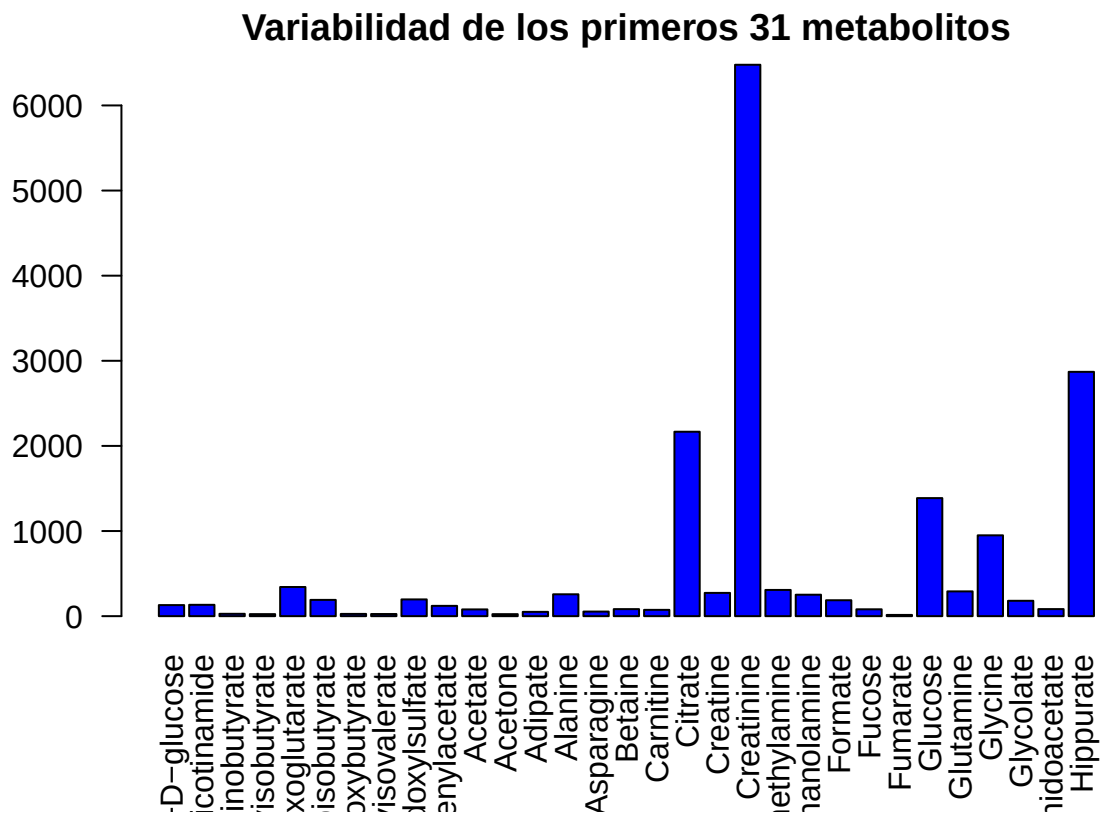
```
# Obtener medias y medianas para cada metabolito
medias = apply(t(assay(experimento)),2,mean)
medianas = apply(t(assay(experimento)),2,median)

# Calcular las diferencias
diferencias = medias-medianas

# Ordenar de menor a mayor
diferencias = diferencias[order(diferencias)]

# Imprimir
kable(diferencias)
```

	x
Isoleucine	1.539091
Valine	2.547013
Methylamine	2.646234
Acetone	4.327013
Fumarate	4.340130
N,N-Dimethylglycine	4.369610
2-Hydroxyisobutyrate	4.790649
Leucine	5.253636
Methylguanidine	7.474545
2-Aminobutyrate	7.669740
Pyruvate	7.834416
Uracil	8.167662
O-Acetylcarnitine	8.263377
3-Hydroxyisovalerate	9.097792
3-Hydroxybutyrate	10.017013
trans-Aconitate	13.790390
Adipate	14.576364
Quinolinate	15.019481
Tryptophan	19.253117
Asparagine	20.183636
tau-Methylhistidine	20.966883
Hypoxanthine	21.057662
Tyrosine	21.417273
Guanidoacetate	21.650520
Pantothenate	22.233766
Betaine	25.604675
Acetate	26.491429
Tartrate	27.064026
Fucose	27.108831
Carnitine	28.275065
Succinate	29.349091
Threonine	31.287403
1-Methylnicotinamide	34.973636

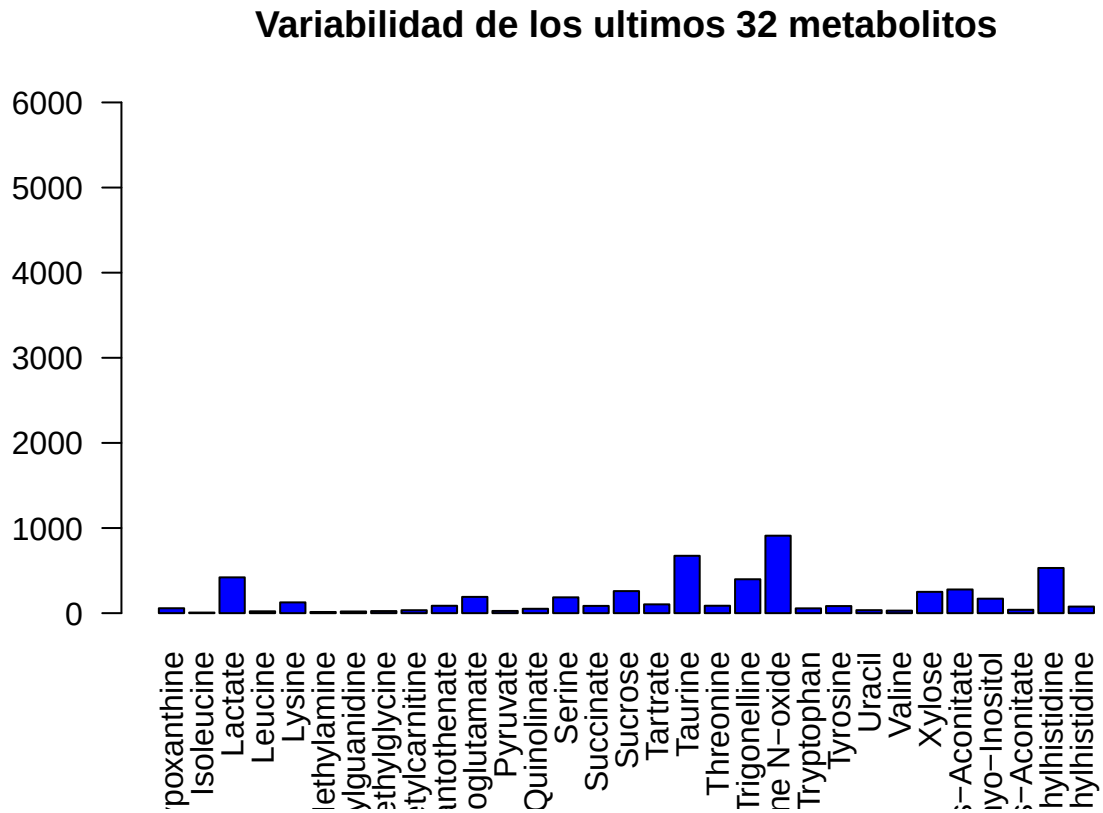
	x
Lysine	39.384156
4-Hydroxyphenylacetate	41.911039
Xylose	50.533377
Formate	51.822987
Dimethylamine	53.266104
Pyroglutamate	53.857792
3-Aminoisobutyrate	54.106364
Serine	55.096883
myo-Inositol	57.137532
Glycolate	57.669351
1,6-Anhydro-beta-D-glucose	60.030390
Ethanolamine	71.880390
Sucrose	72.377792
3-Indoxylsulfate	74.849221
cis-Aconitate	75.199740
Lactate	77.006494
Alanine	79.142338
Glutamine	80.991558
Creatine	82.571948
2-Oxoglutarate	89.937143
Histidine	118.477532
Trigonelline	156.006104
pi-Methylhistidine	207.898312
Trimethylamine N-oxide	268.406883
Taurine	275.483506
Glucose	349.234546
Glycine	352.237403
Citrate	445.295974
Hippurate	1062.687662
Creatinine	1102.771818

9.3 Diagramas de barras de la variabilidad de los metabolitos

```
# Calcular la desviacion estandar para cada metabolito
variabilidad = apply(t(assay(experimento)), 2, sd)

# Graficar diagrama de barras de la variabilidad de los primeros 31 metabolitos
barplot(head(variabilidad,31), main="Variabilidad de los primeros 31 metabolitos", col="blue", las=2, yl
```

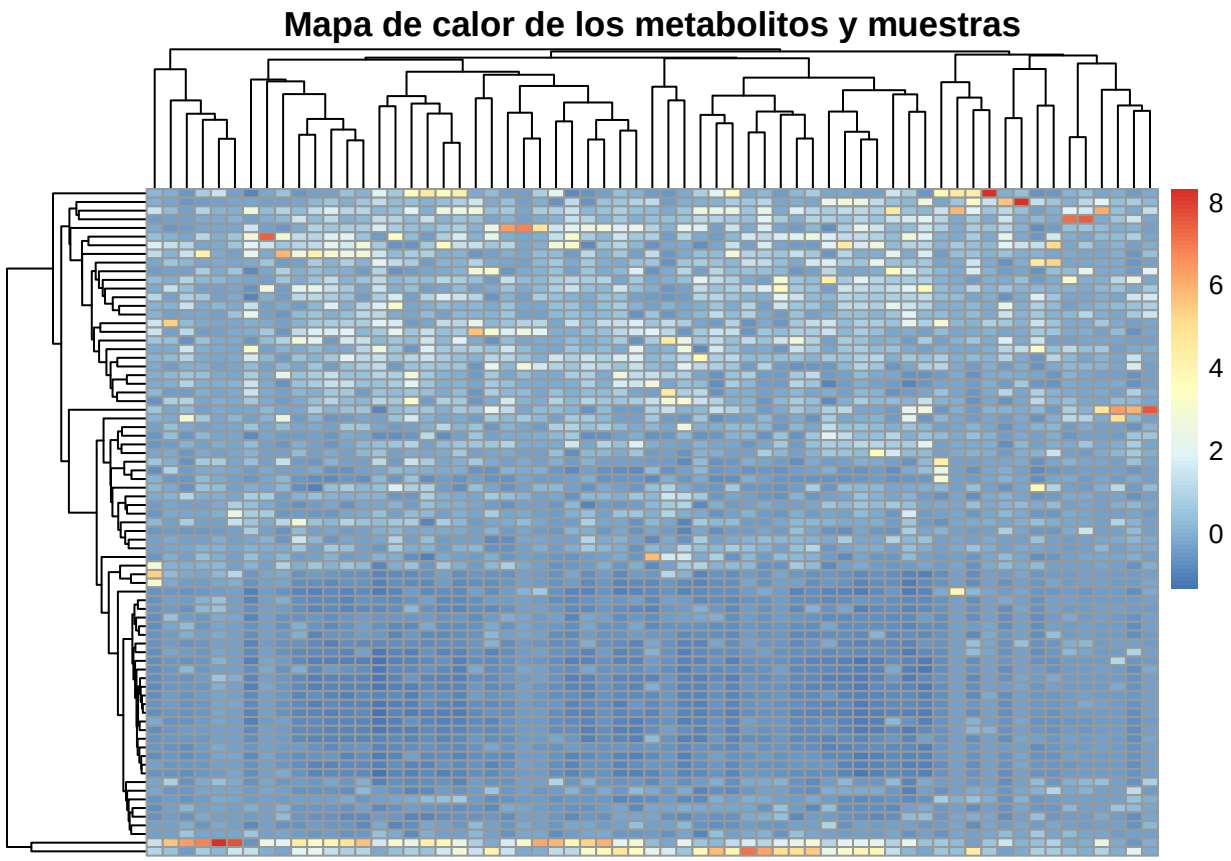



```
# Graficar diagrama de barras de la variabilidad de los ultimos 32 metabolitos
barplot(tail(variabilidad,-32), main="Variabilidad de los ultimos 32 metabolitos", col="blue", las=2, y
```



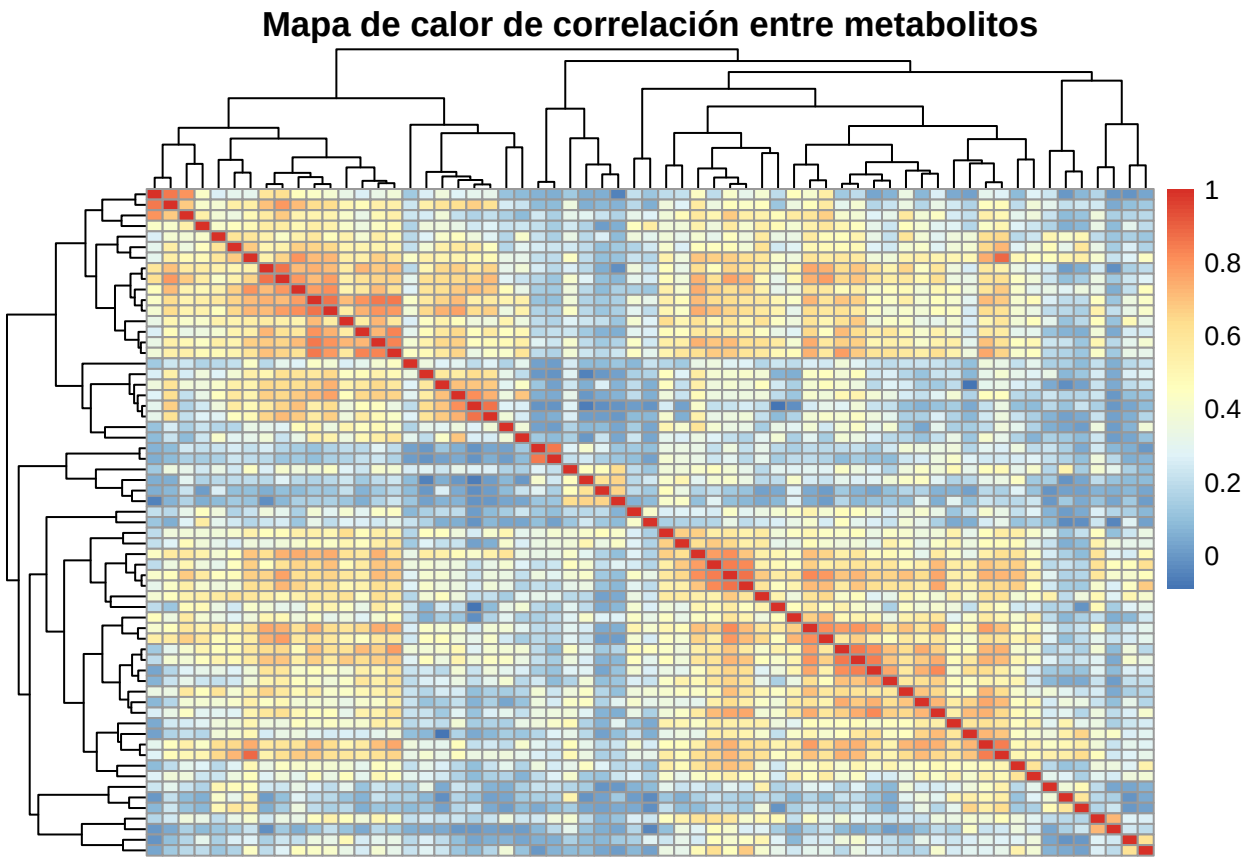
9.4 Mapa de calor de los metabolitos y muestras

```r
# Instalar paquete
# install.packages("pheatmap")

# Llamar libreria
library(pheatmap)

# Normalizar los datos
data_norm = scale(t(assay(experimento)))

# Obtener mapa de calor
pheatmap(data_norm, main="Mapa de calor de los metabolitos y muestras", show_rownames=FALSE, show_colnar
```



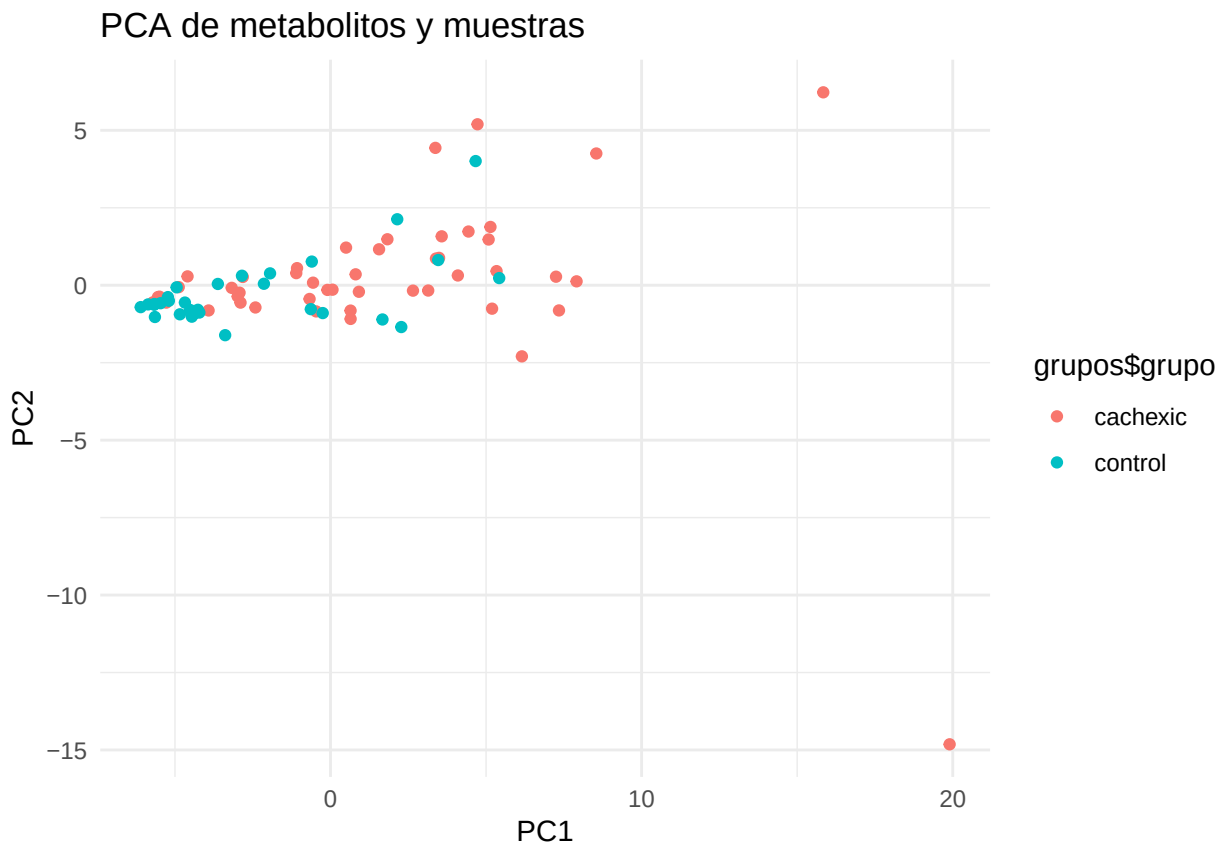
9.5 Mapa de calor de la correlación entre los metabolitos

```
# Obtener la matriz de correlacion
cor_matrix = cor(t(assay(experimento)))

# Crear mapa de calor de las correlaciones
pheatmap(cor_matrix, main="Mapa de calor de correlación entre metabolitos", clustering_distance_rows="c
```

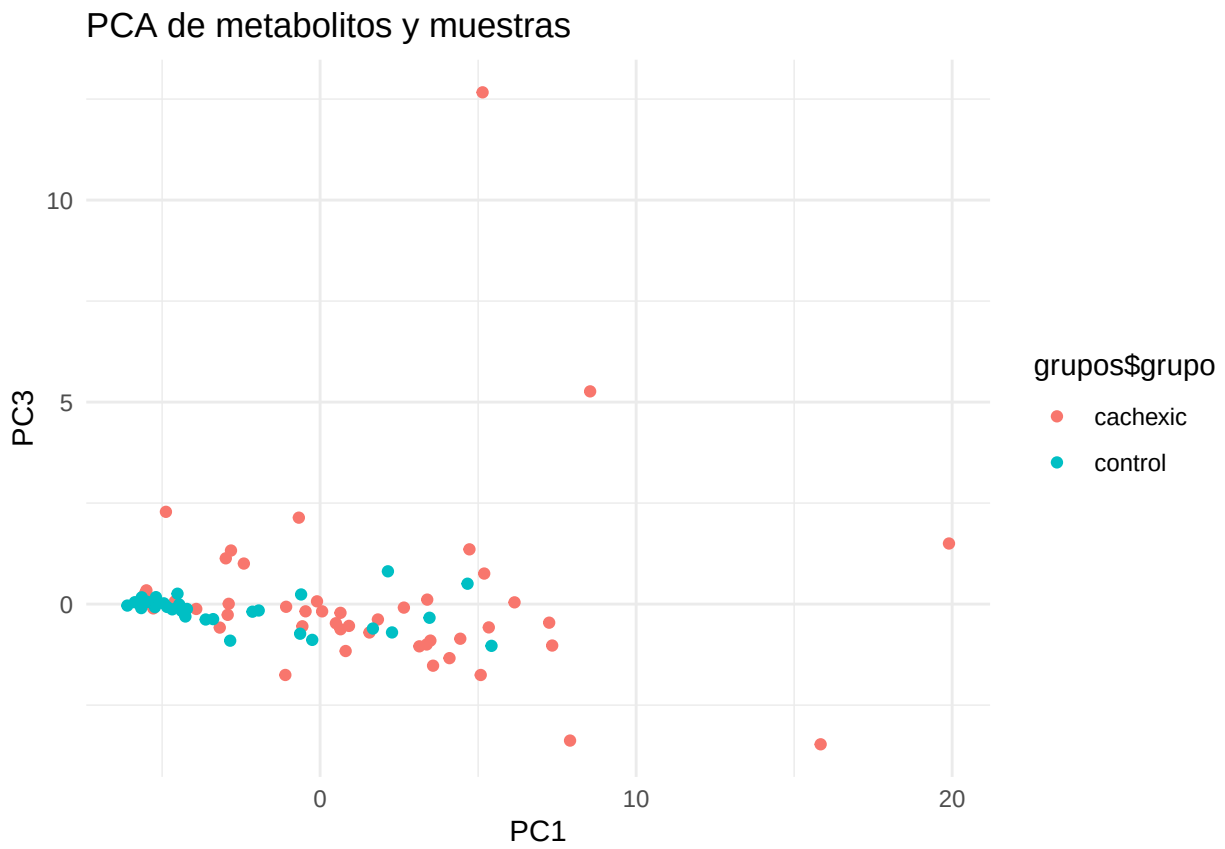


9.6 Gráficas de los análisis de componentes principales

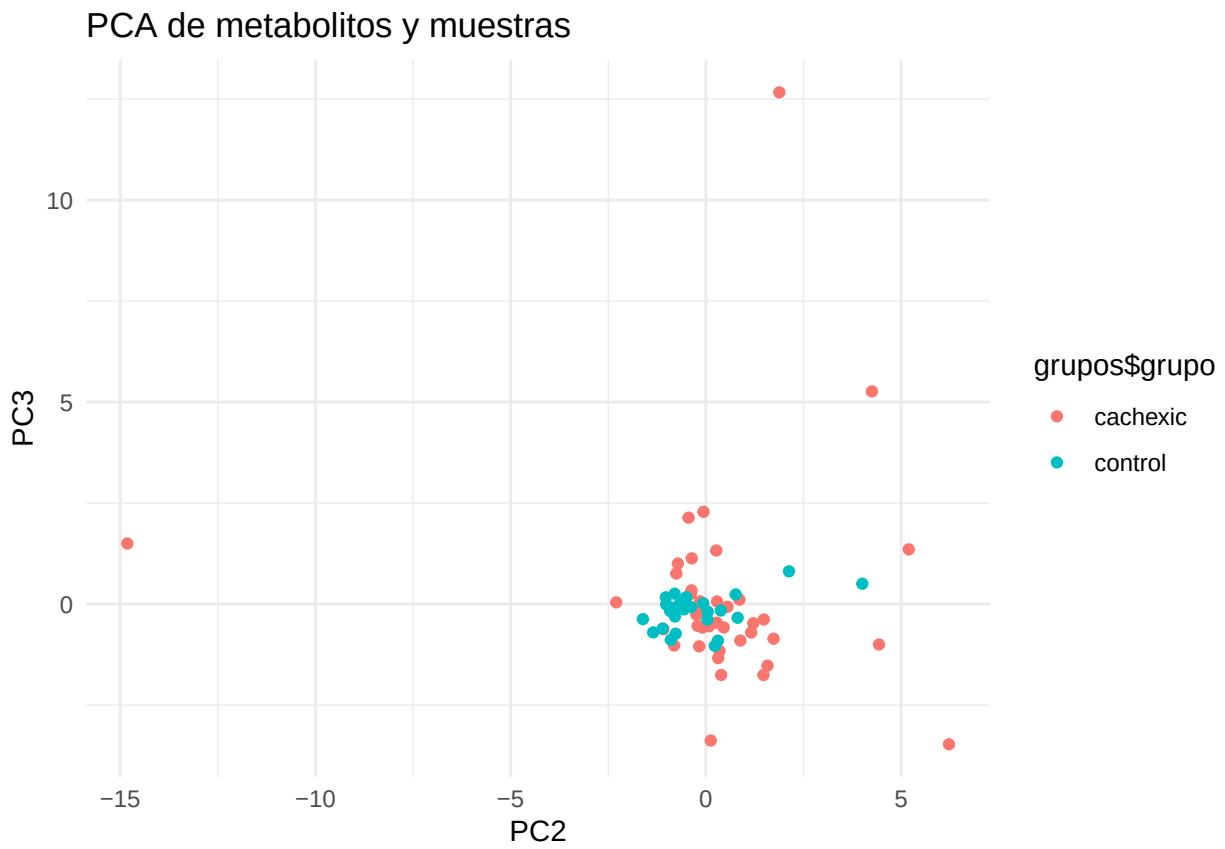
```
# Llamar libreria
library(ggplot2)

# Obtener los grupos de muestras
grupos = colData(experimento)

# Realizar analisis de componentes principales
pca = prcomp(t(assay(experimento)), scale.=TRUE)

# Obtener datos del PCA en data frame
pca_df = data.frame(pca$x)

# Graficar PCA (primeros dos componentes)
p1 = ggplot(pca_df, aes(PC1, PC2, color=grupos$grupo)) +
  geom_point() +
  labs(title="PCA de metabolitos y muestras", x="PC1", y="PC2") +
  theme_minimal()
# Imprimir
print(p1)
```




```
# Graficar PCA (primero y tercer componente)
p2 = ggplot(pca_df, aes(PC1, PC3, color=grupos$grupo)) +
  geom_point() +
  labs(title="PCA de metabolitos y muestras", x="PC1", y="PC3") +
  theme_minimal()
# Imprimir
print(p2)
```

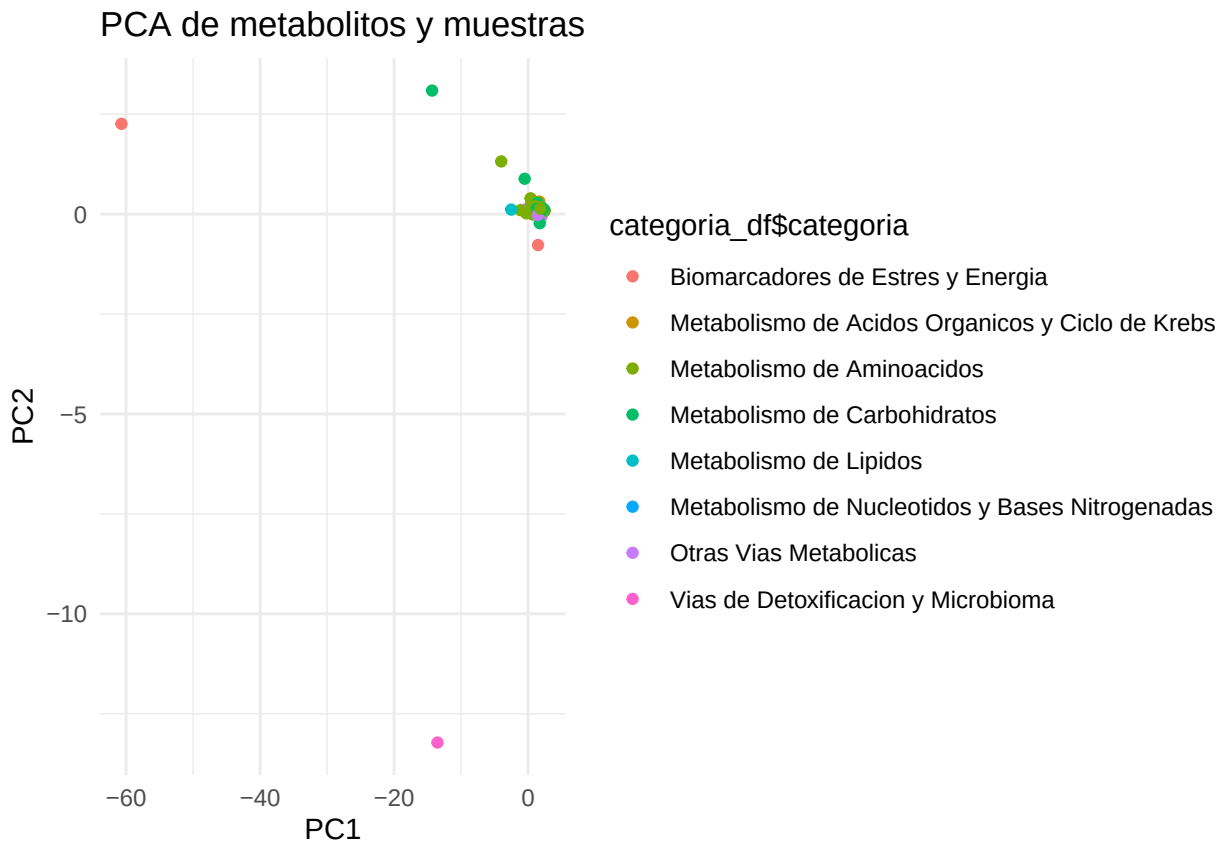



```
# Graficar PCA (segundo y tercer componente)
p3 = ggplot(pca_df, aes(PC2, PC3, color=grupos$grupo)) +
  geom_point() +
  labs(title="PCA de metabolitos y muestras", x="PC2", y="PC3") +
  theme_minimal()
# Imprimir
print(p3)
```

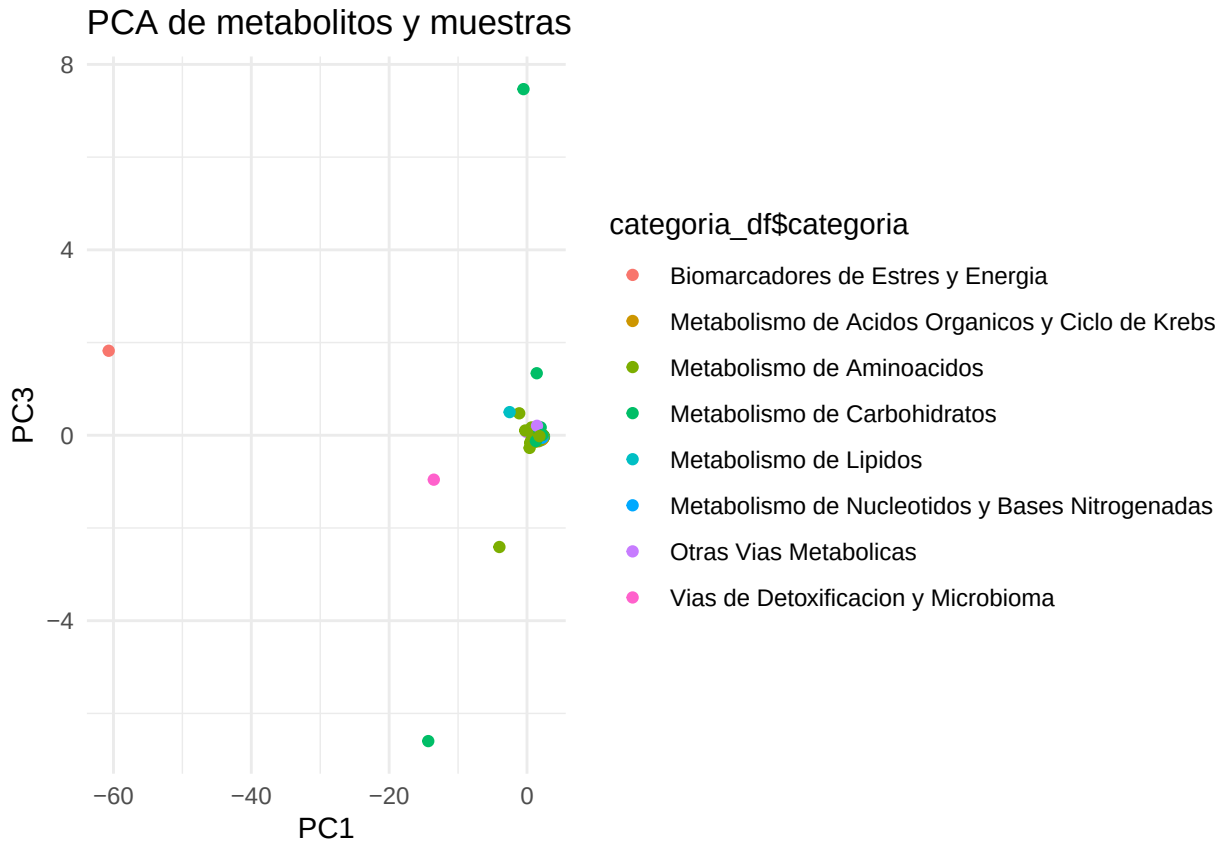



```
# Obtener los nombres de los metabolitos
metabolitos = rownames(rowData(experimento))

# Definir categorias de metabolitos
categorias = list(
  "Metabolismo de Carbohidratos" = c("Glucose", "Pyruvate", "Lactate", "Acetate", "Citrate", "Fumarate"

  "Metabolismo de Aminoacidos" = c("Alanine", "Glutamine", "Glycine", "Serine", "Leucine", "Isoleucine"

  "Metabolismo de Lipidos" = c("Carnitine", "Acetone", "Adipate", "Trimethylamine N-oxide"),

  "Metabolismo de Acidos Organicos y Ciclo de Krebs" = c("Succinate", "Citrate", "Fumarate", "2-Oxoglut

  "Metabolismo de Nucleotidos y Bases Nitrogenadas" = c("Hypoxanthine", "Uracil", "Quinolinate"),

  "Vias de Detoxificacion y Microbioma" = c("Indoxyl sulfate", "Hippurate", "3-Indoxylsulfate", "4-Hydr

  "Otras Vias Metabolicas" = c("1,6-Anhydro-beta-D-glucose", "Dimethylamine", "Methylamine", "Pantothena

  "Biomarcadores de Estres y Energia" = c("Creatine", "Creatinine", "3-Hydroxybutyrate")
)

# Encontrar la categoría correspondiente para cada metabolito
categoria_df = data.frame(
  metabolito=metabolitos,
  categoria=sapply(metabolitos, function(metabolito) {
```

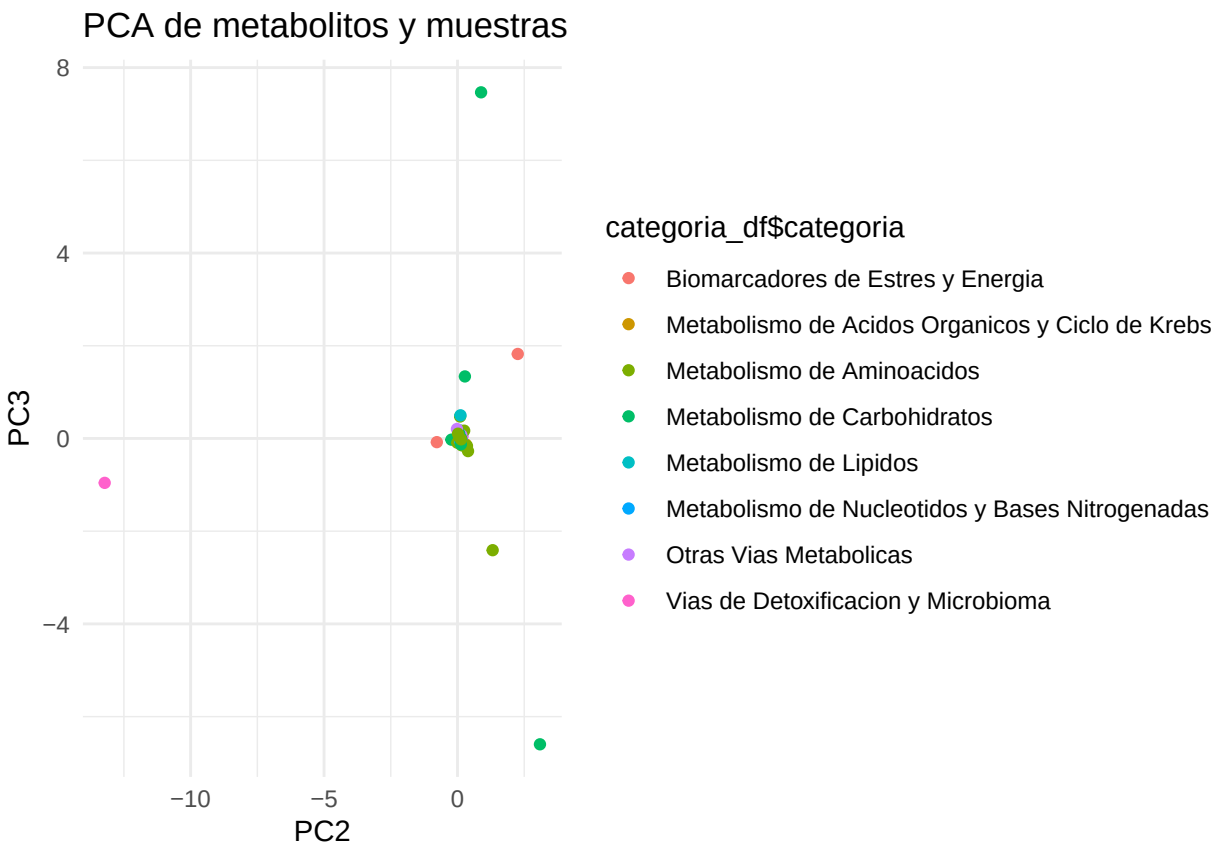
```
    category=NA
    for (cat in names(categorias)) {
      if (metabolito %in% categorias[[cat]]) {
        category=cat
        break
      }
    }
    return(category)
  })
)

# Realizar analisis de componentes principales
pca_metabolitos = prcomp(assay(experimento), scale.=TRUE)

# Obtener datos del PCA en data frame
pca_df_metabolitos = data.frame(pca_metabolitos$x)

# Graficar PCA (primeros dos componentes)
p1 = ggplot(pca_df_metabolitos, aes(PC1, PC2, color=categoria_df$categoria)) +
  geom_point() +
  labs(title="PCA de metabolitos y muestras", x="PC1", y="PC2") +
  theme_minimal()
# Imprimir
print(p1)
```



```r
# Graficar PCA (primer y tercer componente)
p2 = ggplot(pca_df_metabolitos, aes(PC1, PC3, color=categoria_df$categoria)) +
  geom_point() +
  labs(title="PCA de metabolitos y muestras", x="PC1", y="PC3") +
  theme_minimal()
# Imprimir
print(p2)
```




```r
# Graficar PCA (segundo y tercer componente)
p3 = ggplot(pca_df_metabolitos, aes(PC2, PC3, color=categoria_df$categoria)) +
  geom_point() +
  labs(title="PCA de metabolitos y muestras", x="PC2", y="PC3") +
  theme_minimal()
# Imprimir
print(p3)
```


PCA de metabolitos y muestras
8
4
0
−4
PC3
−10
−5
0
PC2
categoria_df$categoria
Biomarcadores de Estres y Energia
Metabolismo de Acidos Organicos y Ciclo de Krebs
Metabolismo de Aminoacidos
Metabolismo de Carbohidratos
Metabolismo de Lipidos
Metabolismo de Nucleotidos y Bases Nitrogenadas
Otras Vias Metabolicas
Vias de Detoxificacion y Microbioma