Untitled

Alejandro_Segura_Alfaro

2025-03-31

Informe PEC 1 - Análisis de datos ómicos

Tabla de contenido

---

2. Abstract o Resumen

En este informe se describe el análisis exploratorio del dataset de metabolómica del estudio ST003557. Se ha creado un objeto `ExpressionSet` para manipular los datos y se ha aplicado un análisis de componentes principales (PCA) para identificar patrones subyacentes. Los datos sugieren un efecto del tratamiento únicamente en uno de los metabolitos y que la variedad representada en el análisis PCA indica que podría deberse a los tejidos muestreados y sus funciones. Además, se crearon diferentes gráficas para facilitar la interpretación de los resultados. El informe incluye un repositorio GitHub con el código utilizado para garantizar la transparencia y replicabilidad del análisis.

---

3. Objectivos

El objetivo principal de este informe es replicar el análisis exploratorio del conjunto de datos de metabolómica del estudio ST003557:

Como objetivos secundarios se podrían remarcar los siguientes: - Conseguir, a partir de herramientas bioinformáticas, el dataset del estudio en cuestion. - Estructurar los datos en un formato adecuado para su manipulación (ExpressionSet). - Explorar la distribución y variabilidad de los datos. - Aplicar análisis PCA, así como representar gráficamente, para identificar patrones. - Interpretar los resultados en un contexto biológico.

---

4. Métodos

Datos (Origen y tipología)

Entrando en el repositorio de la PEC1 se puede encontrar un enlace que aportaba información sobre cómo adquirir, a través de código y utilizando una API, la base de datos de un experimento concreto (bioconductor - metabolomicsWorkbenchR package).

Con ayuda de la documentación oficial sobre el paquete llamado 'metabolomicsWorkbenchR' no resulta muy complicado encontrar la forma de buscar y descargar databases. Lo primero que se debe hacer es instalar dicho paquete:

Existe una función llamada `do_query()` que permite realizar busquedas en la base de datos Metabolomics Workbench. Se decidió buscar estudios que en su título contuviera la palabra "cerebro":

Una vez obtenida la lista de estudios, se decidió que el estudio a replicar era el que tenía `ST003557` como `study_id` (después de una lectura entre líneas de los diferentes títulos).

Este mismo paquete tiene la opción de extraer el objeto `SummarizedExperiment` a partir de un `study_id`:

En el `SummarizedExperiment` se puede observar que hay dos elementos, los cuales se extrayeron de forma separada.

A partir de un objeto de clase `SummarizedExperiment` se puede generar un `ExpressionSet` con el paquete `SummarizedExperiment`:

En este punto se decidió fusionar ambos `expressionSet`, a pesar de provenir de diferentes elementos dentro de un mismo `SummarizedExperiment`.

```
expset_comb <- BiocGenerics::combine(expset1, expset2)
expr_matrix<- rbind(exprs_matrix1, exprs_matrix2)
#Combinamos las matrices tambien
```

Metodología de análisis

Para iniciar el análisis, primero se exploraron las dimensiones descriptivas de las matrices de expresión, así como una exploración de los diferentes factores en los que se clasificaban categóricamente cada muestra.

También se realizó un análisis por componentes principales (PCA). Todo esto acompañado de diferentes representaciones gráficas para un mayor entendimiento.

5. Resultats

Estadístiques descriptives

Para empezar, mostraremos un resumen de las variables del objeto creado `expset_comb` asi como la descripción extraída de los metadatos del objeto `se1`:

```
validObject(expset_comb) # Comprobación de que el objeto expressionSet es correcto
```

```
## [1] TRUE
```

```
dim(expset_comb) # Se exploran las dimensiones
```

```
## Features  Samples
##        6       39
```

```r
summary(pData(expset_comb)) #Se exploran variables
```

```
##  local_sample_id      study_id         sample_source      mb_sample_id
##  Length:39          Length:39          Length:39          Length:39
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##    raw_data               Genotype                          Treatment
##  Length:39          Aldh7a1-/- :39   N-methyl-arginine (80 mg/kg) :19
##  Class :character                    PBS                          :20
##  Mode  :character
##
##  Sample_source
##  Brain :10
##  Kidney: 9
##  Liver :10
##  Plasma:10
```

```r
summary(fData(expset_comb))
```

```
##  metabolite_name    metabolite_id      refmet_name
##  Length:6           Length:6           Length:6
##  Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character
```

```r
metadata(se1)$description #Se muestra el título del estudio
```

```
## [1] "Metabolomics analysis of kidney, brain, liver, and plasma from Aldh7a1-/- mice administered PBS
```

```r
expset_comb@featureData@data #Información de los metabolitos
```

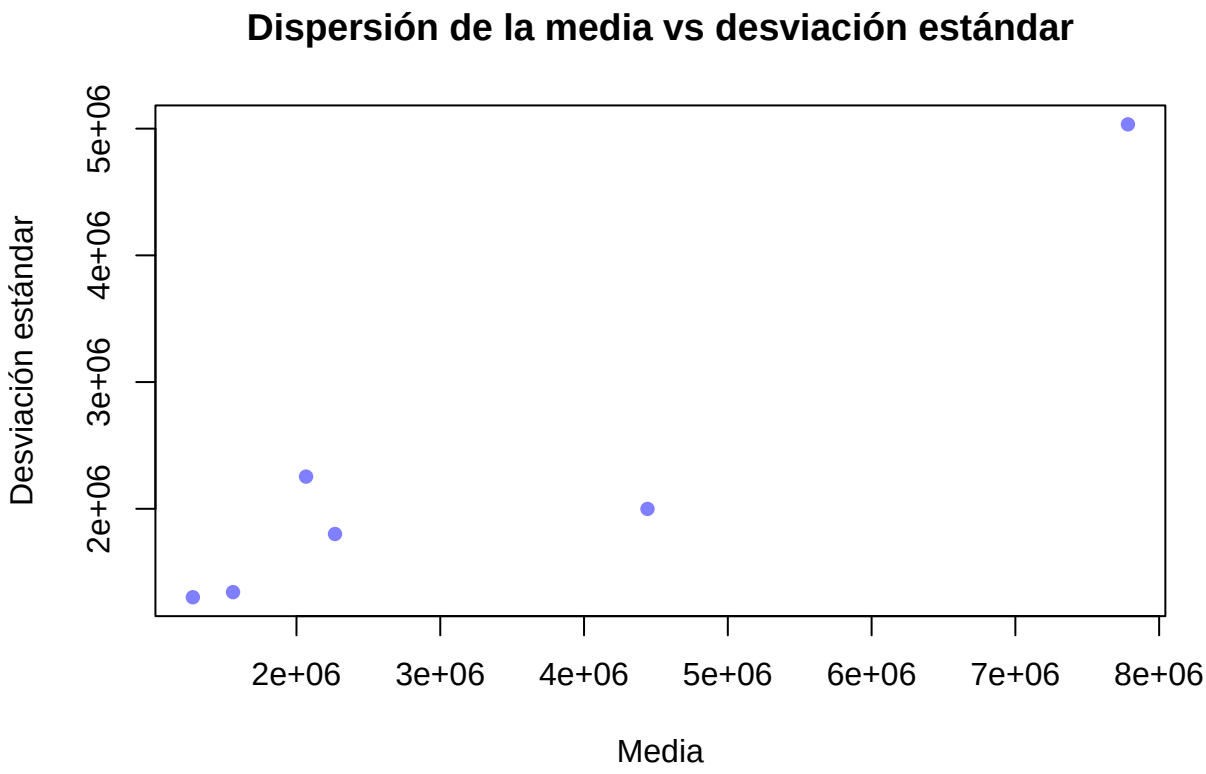
```
##                          metabolite_name metabolite_id      refmet_name
## ME927344                     AMINOADIPATE      ME927344 Aminoadipic acid
## ME927341                           LYSINE      ME927341           Lysine
## ME927345               NG-METHYL-ARGININE      ME927345        Targinine
## ME927343 PIPERIDEINE 6-CARBOXYLATE (P6C)      ME927343
## ME927342                     SACCHAROPINE      ME927342     Saccharopine
## ME927346                       PIPECOLATE      ME927346   Pipecolic acid
```

El estudio elegido es un experimento cuyo objetivo era describir a nivel metabólico muestas de plasma, higado, cerebro y riñón en ratones Aldh7a1-/- a los que se les había subministrado PBS o N-metil-arginina a través de una inyección intraperitoneal. Este estudio se basa en 39 muestras (10 de cada órgano estudiado, excepto para riñón, que son 9). La mitad de los ratones fueron tratados con PBS y la otra mitad con 80mg/kg de N-metil-arginina. Se estudió la expresión de 6 metabolitos en cada muestra.

Para ver si hay algún metabolito cuya expresión varie mucho dependiendo de la situación, se realizó un gráfica de dispersión enfrentando la media con su desviación estándar:

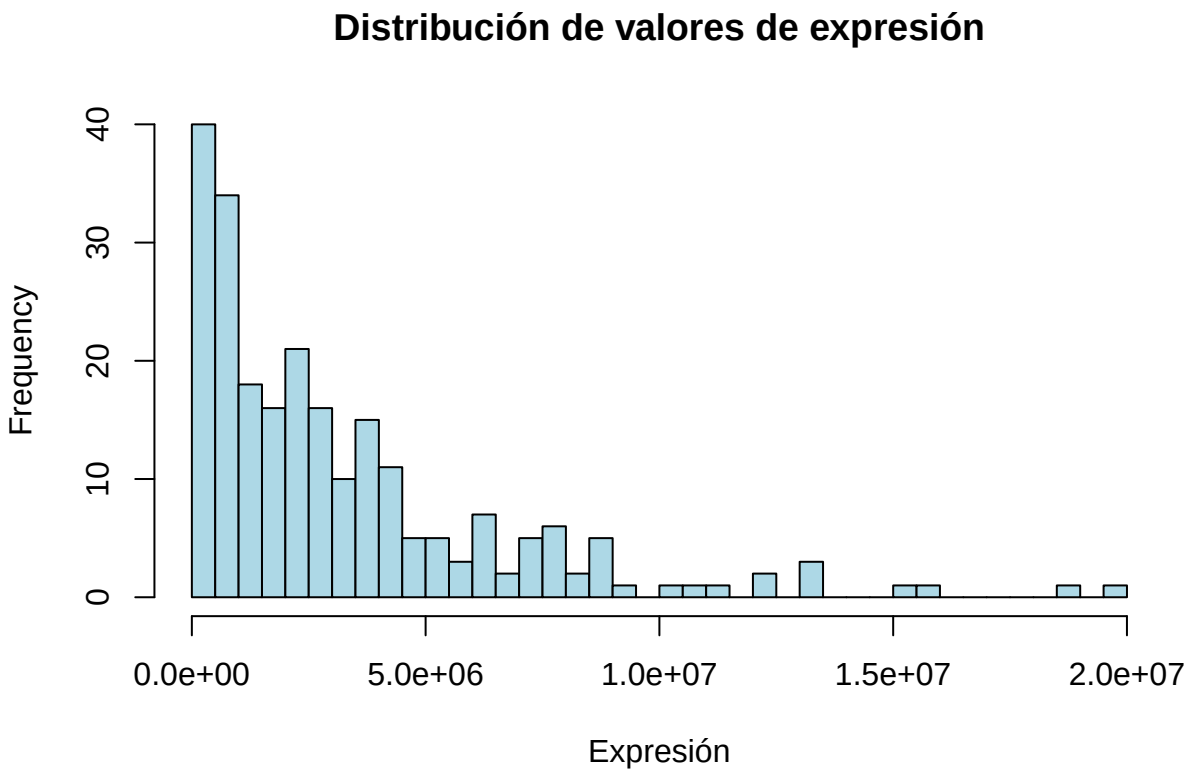
```r
metabolite_means <- apply(expr_matrix, 1, mean) # Calculamos la media
metabolite_sd <- apply(expr_matrix, 1, sd) # Calculamos la desviación

# Graficar media vs desviación estándar
plot(metabolite_means, metabolite_sd, xlab="Media", ylab="Desviación estándar", main="Dispersión de la
```

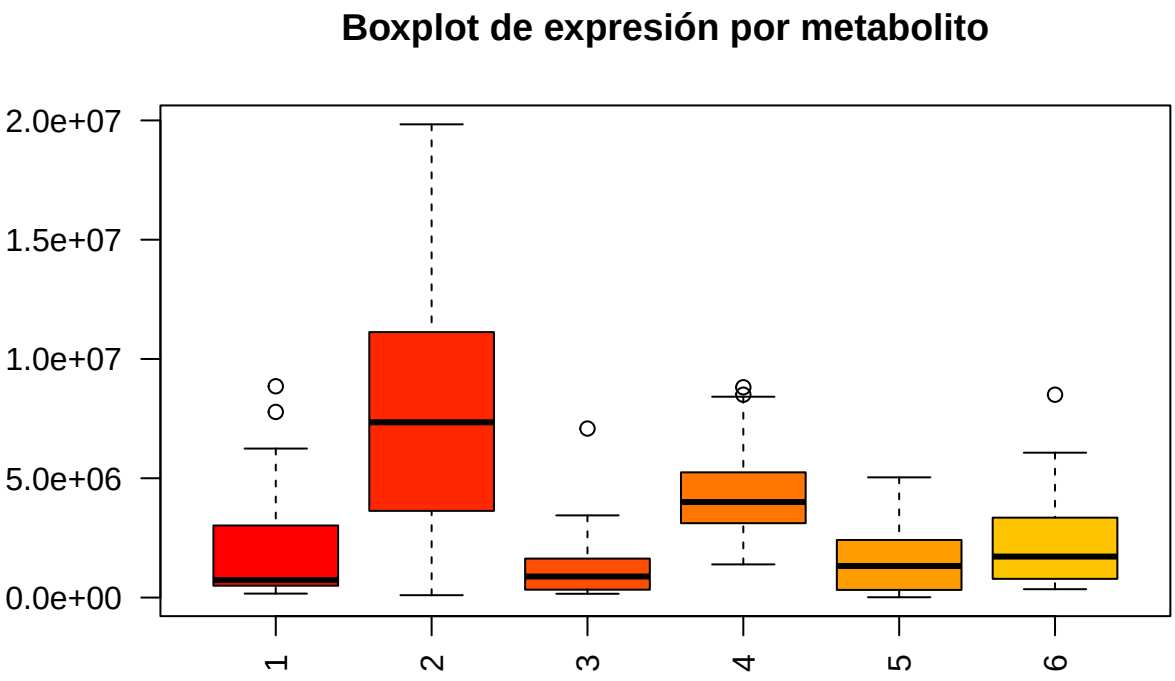


Se puede observar que, a excepción de un metabolito, la media suele estar por encima de la desviación estándar, es decir, hay baja variabilidad relativa en comparación con su nivel medio de expresión.

```
hist(expr_matrix, breaks = 50, main = "Distribución de valores de expresión", xlab = "Expresión", col =
```




```
boxplot(t(expr_matrix), main = "Boxplot de expresión por metabolito", col = rainbow(ncol(expr_matrix)),
```



Con la distribución de valores de expresión, se puede ver que la mayoría de valores están por debajo de 1.0e+07. Parecería que hay algún efecto/causa que podría hacer que algunos datos se mostraran por encima de ese umbral. Tras ver el boxplot, se puede ver que el metabolito ME927341, lisina, es el único metabolito con gran varianza entre tejidos y tratamiento. Los demás, no presentan gran diferencia entre ellos.

Análisis por componentes principales

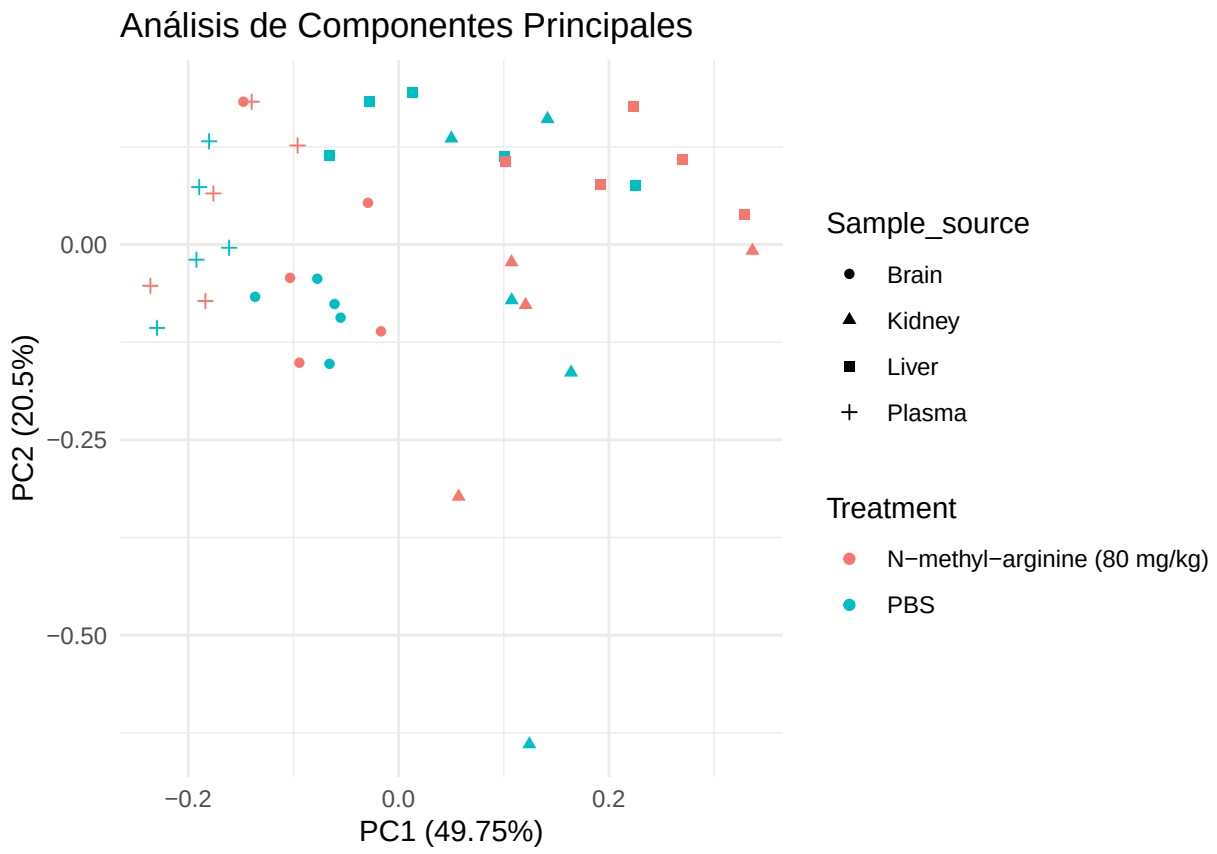
Para hacerse una idea del significado de estos datos, se procedió con un análisis por componentes principales:

```
library(ggplot2)
library(ggfortify)
```

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

```
pca_result <- prcomp(t(expr_matrix), scale. = TRUE) # Aplicamos PCA

autoplot(pca_result, data = pData(expset_comb),
         colour = "Treatment",
         shape = "Sample_source"
         ) + ggtitle("Análisis de Componentes Principales") + theme_minimal()
```




```
#Graficamos diferenciando tratamiento y tejido
summary(pca_result)
```

```
## Importance of components:
##                           PC1    PC2    PC3     PC4     PC5     PC6
## Standard deviation     1.7276 1.1091 0.9340 0.71446 0.51629 0.36863
## Proportion of Variance 0.4975 0.2050 0.1454 0.08508 0.04443 0.02265
## Cumulative Proportion  0.4975 0.7025 0.8478 0.93293 0.97735 1.00000
```

Con este gráfico, que enfrenta las dos componentes principales que mayor aporte tienen a la variabilidad de los datos (70,25 % en proporción acumulada), permite agrupar los datos en una simplificación de dimensionalidad. Se entiende que la variabilidad esta bien explicada cuando se supera el 80% de ella, por lo que se debe de tener en cuenta la tercera componente principal para tener una variabilidad correctamente explicada.

Podemos ver que al simplificar en dos componentes, la variabilidad en grupos tratamiento parece no tener una relación con ninguna de las componentes principales, por lo que el tratamiento podría no tener un efecto a considerar a la hora de entender nuestros datos. En cambio, se puede ver como hay grupos marcados según tejidos, concretamente, en plasma y cerebro en la esquina superior izquierda. Al parecer, la variabilidad entre higado y riñon no parece predecirse correctamente con la aportación de ambas componentes principales.

Discusión

Los ratones con mutación en ALDH7A1 con pérdida de función generan una acumulación tóxica de metabolitos de lisina, uno de los metabolitos estudiados en este experimento. De acuerdo con el boxplot de expresión

de metabolitos, los datos mostraban una gran variabilidad en la expresión de lisina, lo que concuerda con la aportación de la variabilidad de aquellos que no han sido tratados (valores altos de expresión de lisina). A la vez, se podría pensar que, según aquella pequeña proporción de valores de expresión altos en la gráfica de distribución, era debido a esa acumulación de lisina.

Se podría pensar que el posible efecto del tratamiento únicamente afecta a la lisina, pero no a la cadena de metabolitos y quizas por eso no se observa que las componentes principales no lo tienen en cuenta al darse en un solo metabolito.

La variación en la expresión general de los metabolitos, en principio y según el análisis exploratorio realizado, no tendría porque relacionarse con el tratamiento realizado a los ratones. Esa varianza vista en la expresión vendría a estar más relacionada con el tejido y las funciones de ese tejido.

Respecto al preprocesado, es posible que se hubiera añadido ruido a nuestros datos por el hecho de fusionar diferentes `expressionSet` en uno, ya que la metodología para obtener dichos datos es ligeramente diferentes (HELIC positive ion mode vs. HELIC negative ion mode). Quizas se debería de haber visto que en un principio ambos `SummarizedExperiment` provenian de forma separada durante la descarga de datos.

Para finalizar, cabe mencionar la necesidad de validar los resultados com métodos complementarios, como una ANOVA multifactorial.

---

7. Conclusiones

- El PCA ha permitido identificar ciertos patrones en tejidos y sus funciones.
- El análisis exploratorio no ha demostrado que los tratamientos afecten de forma general a la expresion de los metabolitos estudiados, unicamente en lisina.
- El tratamiento podría no actuar en todos los metabolitos, sino únicamente actuaría en un eslabón de la cadena metabólica.

---

8. Referencias

- Enlace al repositorio de GitHub
- Johal, A. S., Al-Shekaili, H. H., Abedrabbo, M., Kehinde, A. Z., Towriss, M., Koe, J. C. & Parker, S. J. (2024). Restricting lysine normalizes toxic catabolites associated with ALDH7A1 deficiency in cells and mice. Cell Reports, 43(12).
- Al-Shekaili, H. H., Petkau, T. L., Pena, I., Lengyell, T. C., Verhoeven-Duif, N. M., Ciapaite, J. & Leavitt, B. R. (2020). A novel mouse model for pyridoxine-dependent epilepsy due to antiquitin deficiency. Human Molecular Genetics, 29(19), 3266-3284.
- https://www.metabolomicsworkbench.org/data/DRCCMetadata.php?Mode=Study&DataMode=AllData&StudyID=ST003557&StudyType=MS&ResultType=1#DataTabs
- https://aspteaching.github.io/AMVCasos/
- https://aspteaching.github.io/Analisis_de_datos_omicos-Ejemplo_0-Microarrays/ExploreArrays.html

---