Informe de la PEC1

Ciencias ómicas

Guillermo Aguado Val

5 de noviembre, 2024

Índice

1. Abstract

En el presente trabajo se explora el dataset proveniente del estudio de Chan et al. (2016), en el que se estudia el cáncer gástrico a través de 149 metabolitos. Para ello, se utiliza el paquete *SummarizedExperiment* (Morgan et al., 2024) y *POMA* (Castellano-Escuder et al., 2021).Las exploraciones aplicadas han sido PCA, k-means y agrupaciones jerárquicas.

Los archivos generados se pueden encontrar en la siguiente dirección de github: https://github.com/Guillermo905/PAC1_omicas

2. Objetivos del estudio

En este trabajo se va a utilizar el dataset perteneciente al estudio realizado de Chan et al. (2016) en el que se estudia la metabolómica del cáncer gástrico. Los objetivos son:

- Utilizar el paquete *SummarizedExperiment* del repositorio Bioconductor
- Explorar y evaluar el dataset a través de PCA y análisis jerárquico.

En el dataset se recogen variables relacionadas con el cáncer gástrico. Las variables más desatacables son:

- **Name**: Es el número de referencia individual para cada metabolito. Están numerados desde M1 hasta M149.
- **Perc_missing**: Porcentaje de NAs para cada metabolito.
- **QC_RSD**: Medidor de calidad del metabolito a través de todas las muestras.
- **Class**: Indica el lugar de origen de la muestra para cada individuo. Las abreviaciones corresponden a: GC = Cáncer Gástrico, BN = Tumor Benigno, HE = Healthy Control, QC = Grupo control.
- **SampleType**: Indica si la muestra pertenece al grupo control. QC: grup control y Sample: muestra (puede ser GC, BN, o HE)

3. Material y métodos

El dataset está en formato *xlsx*, por lo tanto, se transforma a formato *csv*. A continuación, se carga en el programa R (R Core Team, 2024). De los datos se obtiene dos matrices, una de dimensiones 140, 152 con los datos de cada muestra y su concentración de metabolitos (dataset 0), y otra segunda matriz de dimensiones 149, 4 que contiene los metadatos de cada metabolito. (dataset 3)

Primero, se comprueba que el nombre de cada metabolito es compartido entre el que recoge la información de las muestras (dataset 0) con el que contiene los metadatos de los metabolitos (dataset 3). Luego, se divide el dataset 0 en dos. Uno contendrá la información de las muestras con los metabolitos (dataset 1) y otro las variables *sampleType* y *Class* (dataset 2).

Segundo, se carga la librería *SummarizedExperiment* (Morgan et al., 2024) del proyecto Bioconductor (Huber et al., 2015) para introducir los diferentes datasets. Es importante remarcar que la clase *SummarizedExperiment* necesita que las muestras estén en las columnas, haciendo necesario transponer la matriz de datos (dataset 1). A continuación, se genera un objeto a partir de la clase *SummarizedExperiment* que recoge los datsasets en los siguientes argumentos:

- *assays*: Dataset 1, recoge la concentración de metabolitos para cada muestra.
- *rowData* Dataset 3, corresponde a los metadatos de los metabolitos.
- *colData*: Dataset 2, contiene información sobre el origen de la muestra (variables *Class* y *SampleTipe*).

A continuación se muestra el objeto generado por la clase *SummarizedExperiment*:

```
## class: SummarizedExperiment
## dim: 149 140
## metadata(0):
## assays(1): counts
```

```
## rownames(149): M1 M2 ... M148 M149
## rowData names(4): Name Label Perc_missing QC_RSD
## colnames(140): sample_1 sample_2 ... sample_139 sample_140
## colData names(3): SampleID SampleType Class
```

Una vez cargados los datos se explora la cantidad de NAs que hay para cada individuo y se observa que el número máximo es de 25, el cual corresponde a la muestra 122.

Como se va a trabajar con el paquete *POMA* (Castellano-Escuder et al., 2021) es necesario eliminar los NAs para su correcto funcionamineto. En consecuencia, se va a utilizar el método k-nearest neighbor que substituye cada NA por un valor cercano en relación a cada como se situa espacialmente cada muestra. Además, en el caso que haya en una muestra con 21 o más NAs será eliminada porque significa que falta el 15 % de la información.

Luego, se procede a realizar una exploración univariable a través de boxplot con el package *POMA* (Castellano-Escuder et al., 2021). Además, se comparará visualmente los datos normalizados y sin normalizar debido a la alta heterogeneidad que presentan. En la figura 1 se observa que el rango de los datos es muy amplio y hay datos que pueden ser posibles outliers. Por lo tanto, se hace una transformación de los datos a través del método de Pareto con la finalidad de reducir la importancia de valores grandes, mantiene la estructura y centra los datos en el valor 0 (Van Den Berg et al., 2006).

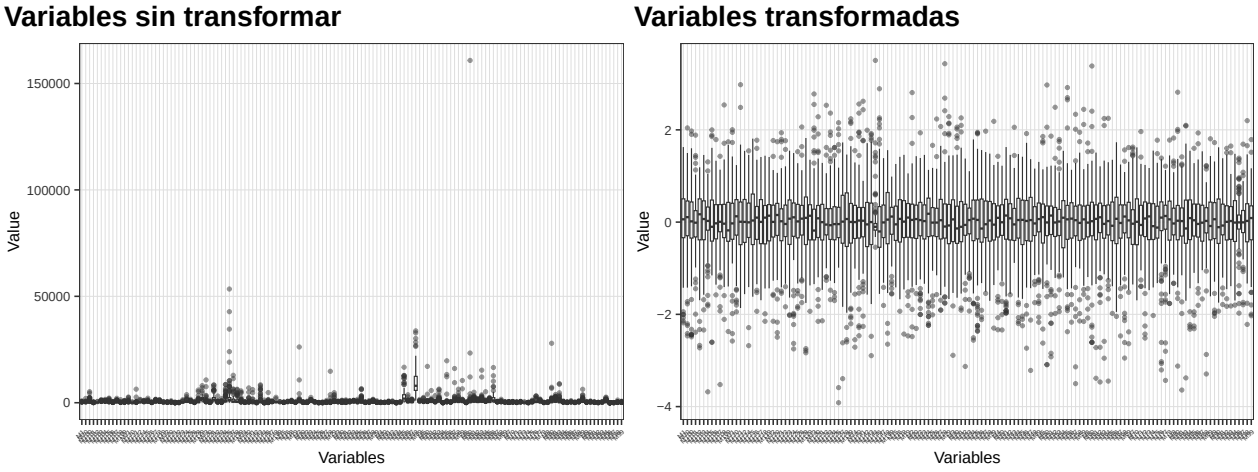

Figura 1: Boxplot de las variables sin transformar y transformadas.

Como en el gráfico de cajas se observan posibles valores extremos o outliers se decide eliminarlos de los datos a través de la función *PomaOutliers* que utiliza el método MDS. Eliminando del dataset un total de 5 muestras.

En la exploración de los datos se realiza una correlación entre las diferentes variables para conocer el nivel de correlación a través de la función *PomaCorr* del paquete Castellano-Escuder et al. (2021) utilizando el método de Spearman porque es más robusto delante de outliers o valores grandes.

A posteriori, se realizará una PCA a través del paquete Castellano-Escuder et al. (2021) en el que se obvia la opción de centrar los datos, porque ya se centraron durante el análisis anterior, y se diferencia a los individuos en función de la clase a la que pertenecen. El número de componentes principales incluidos corresponde al número de variables.

A nivel de técnicias de agrupación se van a realizar dos técnicas no supervisadas. La primera es k-means a través del paquete Castellano-Escuder et al. (2021), en la que se visualizará la agrupación de individuos en 3 y 4 núcleos. Ya que esto coincide con las diferentes clases de muestras que contiene el dataset. La otra técnica va a ser la clasificación a través de un dendograma o cluster jerárquico a través del paquete *stats* de R Core Team (2024), que recoge la función *hclust*. Se utilitzará como matriz de distancias la generada para las correlaciones.

4. Resultados

En la exploración de la correlación de las diferentes variables se observa que la mayoría de metabolitos tienen una correlación positiva entre ellos (Figura 2). Si se explora los *p-value* se observa que hay 8616 pares de metabolitos con una correlación significativa entre ellos. En cambio, 1114 pares de metabolitos no tienen un valor significativo.

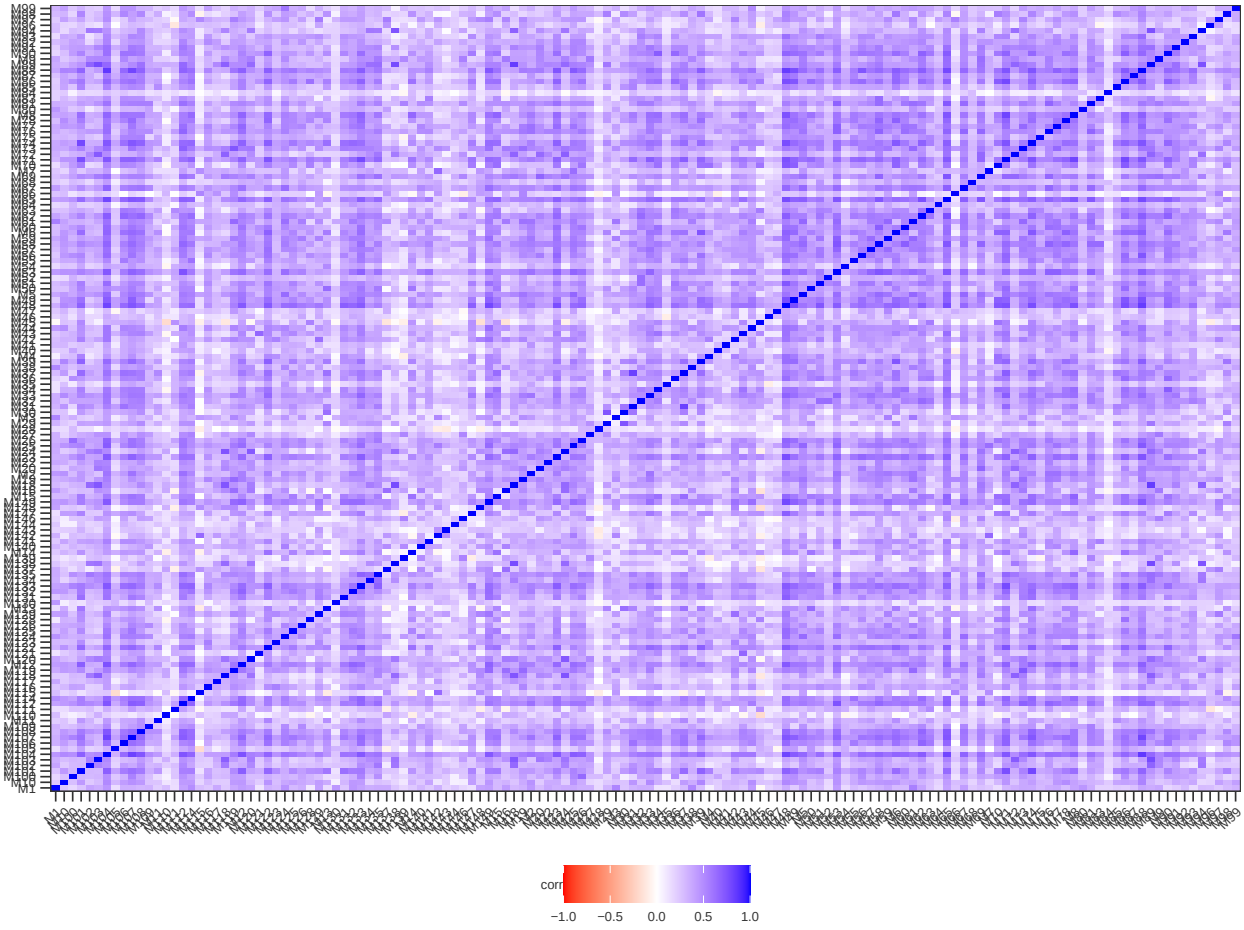

Figura 2: Heatmap que correlaciona los diferentes metabolitos.

La PCA muestra que la mayoría de la variabilidad está retenida en las dos primeras componentes principales (Figura 3), a pesar de que solo recogen el 39.25 % de la varianza explicada.

Al analizar conjuntamente las dos primeras componentes principales se observa en el centro el grupo de control, mientras que el resto de la muestra está más distribuido en el espacio de las dos primeras componentes principales. Sin embargo, en el primer componente parece que las muestras analizadas con tumor benigno se sitúen a la derecha de la gráfica mientras que el resto de casos ocupan el resto del eje. Por otro lado, en el segundo componente se diferencie ligeramente el grupo de individuos sanos (situados mayoritariamente en la mitad inferior de la gráfica) respecto al grupo que presenta un tumor, ya sea maligno o benigno (mayoritariamente en la mitad superior de la gráfica).

Al explorar las posibles agrupaciones a través del método de k-means se observa la diferenciación de tres grupos (número óptimo de agrupaciones, Figura 8), pero no parece que ofrezca información relevante para el estudio que se está realizando porque no se observa ninguna asociación con los datos obtenidos a través de la PCA. En cambio, si se realiza la agrupación con 4 núcleos el grupo control está bastante bien identificado, mientras que la diferenciación entre grupos comentada anteriormente en el eje del primer componente

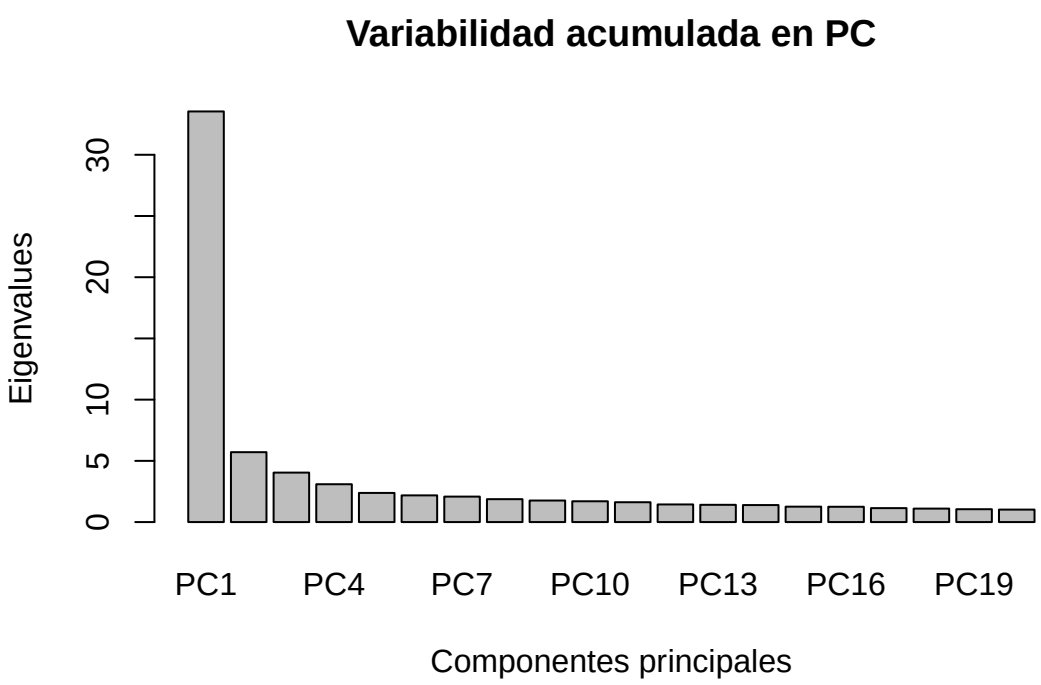

Figura 3: Gráfico de barras con la acumulación de la varianza de cada componente principal.

principal parece que este aquí más definida (Figura 5) .

A través de los dendogramas también se puede analizar cómo se agrupan los datos. En el caso de la agrupación a través de las muestras (Figura 6) se observan 4 grupos que se podrían corresponder a las diferentes clases de muestras que hay. Por lo tanto, el número de grupos podría coincidir con el número de clases que hay. En cambio, en el caso de los metabolitos (Figura 7), se observa la formación de 4 o 5 grupos.

5. Discusión

El estudio del metaboloma significa la inclusión de varios datasets que tiene que funcionar de manera sincronizada a medida que se vayan eliminando variables o individuos de la muestra. En esto, el package *SummarizedExperiment* (Morgan et al., 2024) juega un papel importante al no permitir que los datos se desordenen. Sin embargo, el manejo del programa para el usuario no es muy amigable y dificulta algunos análisis al guardar en las columnas las muestras. Por otro lado, el paquete *POMA* (Castellano-Escuder et al., 2021) permite un manejo muy fácil de los datos para hacer los diferentes análisis y transformación de los datos. Sin embargo, la perdida de las variables guardadas en el atributo *rowdata* podría perjudicar análisis diferentes a los hechos en este trabajo. Por otro lado, a pesar de la facilidad que presenta, tiene el inconveniente que solo puede clasificar individuos si la primera columna de los metadatos (dataset 2) es de tipo factor, no permitiendo al usuario escoger otras columnas de interés. Esto hace necesario modificar el orden de las variables cada vez que se quiera valorar una varaible cualitativa diferente. Igualmente, en la actualización del 2 de noviembre de 2024 de la librería *POMA* aparece el argumento *outcome* que permite marcar la variable de interés en la versión 3.20 de Bioconductor. En cambio, la versión instalada para la realización del trabajo es la versión 3.19 y, en consecuncia, no se ha podido utilizar esta opción.

En relación con los resultados obtenidos en las diferentes exploraciones, es importante tener en cuenta que parte de los resultados obtenidos depende de las transformaciones aplicadas en los datos. En consecuencia, hubiese sido conveniente probar otras transformaciones para ver si los resultados obtenidos variaban. Además,

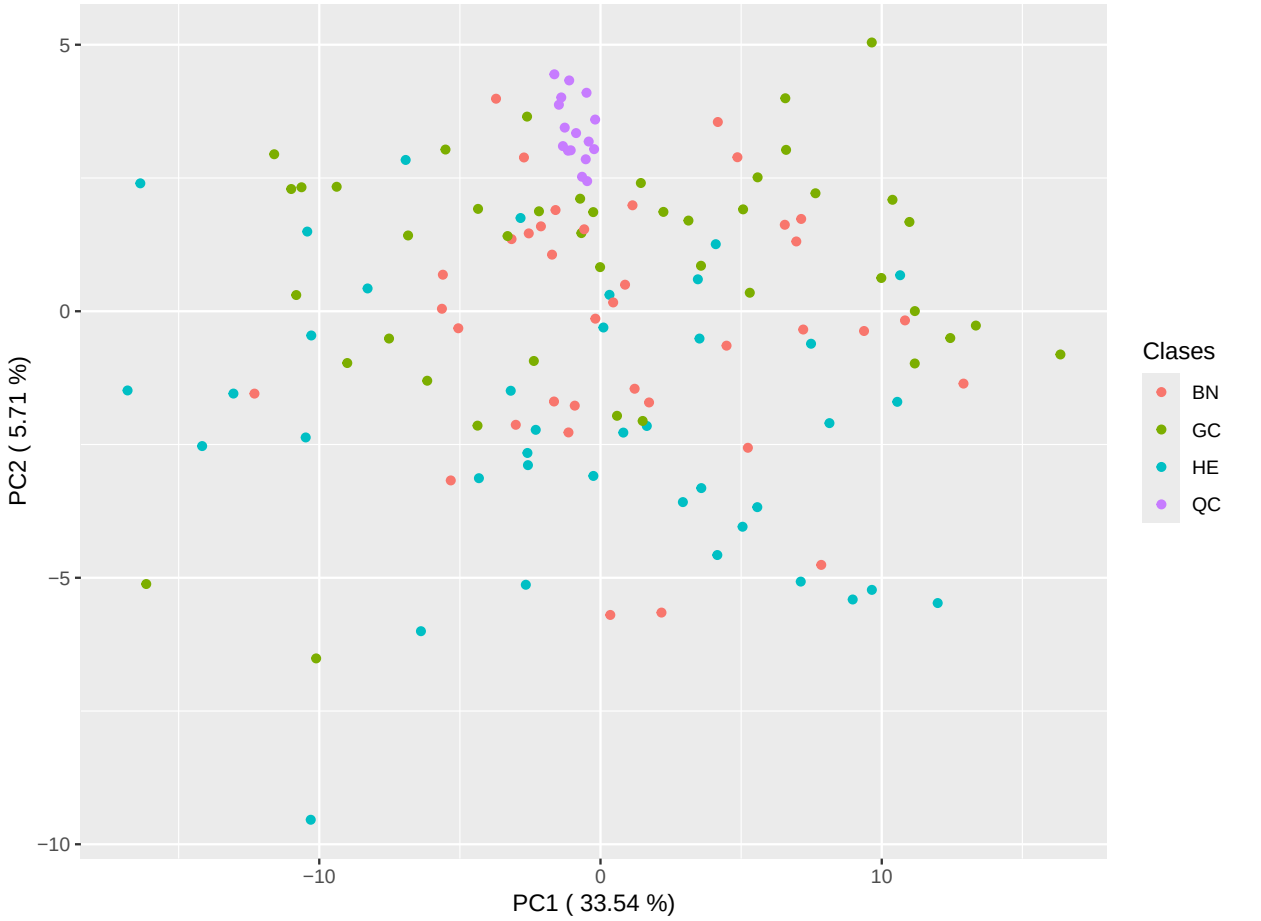

Figura 4: PCA de los dos primeros componentes principales. En color los diferentes grupos que se han incluido en el estudio: GC = Cáncer Gástrico, BN = Tumor Benigno, HE = Healthy Control, QC = Grupo control.

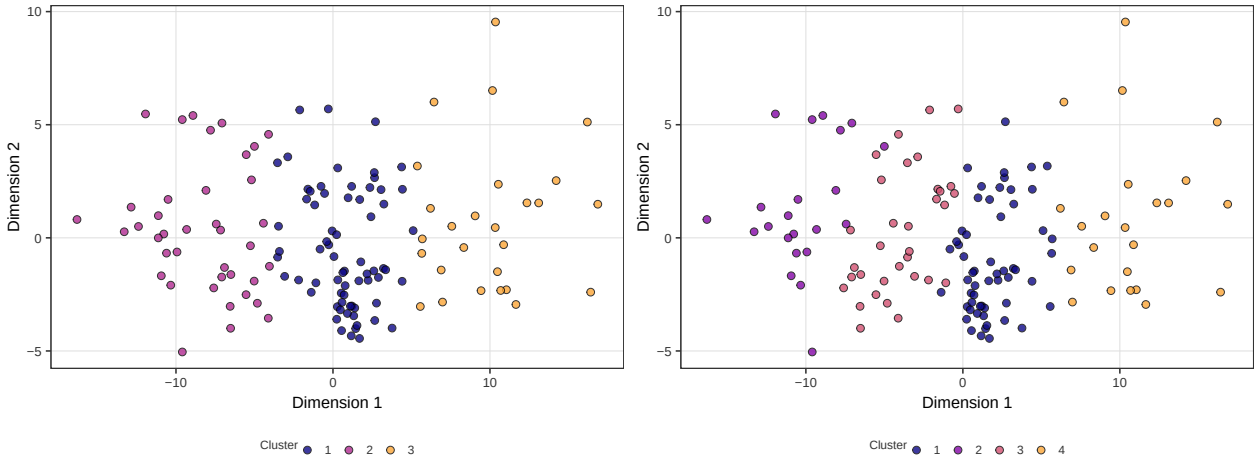

Figura 5: Agrupación a través del método k-means con 3 y 4 núcleos

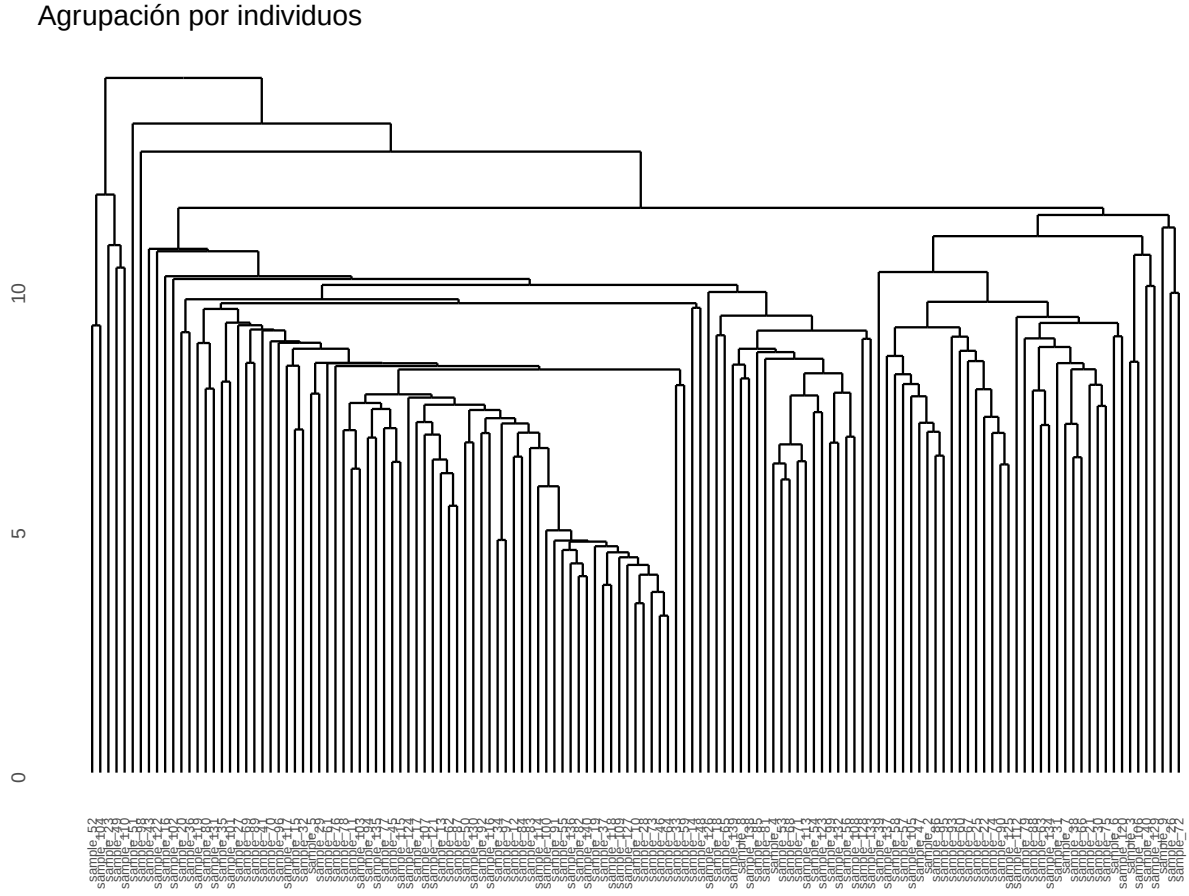

Figura 6: Dendograma en el que se agrupan las diferentes muestras

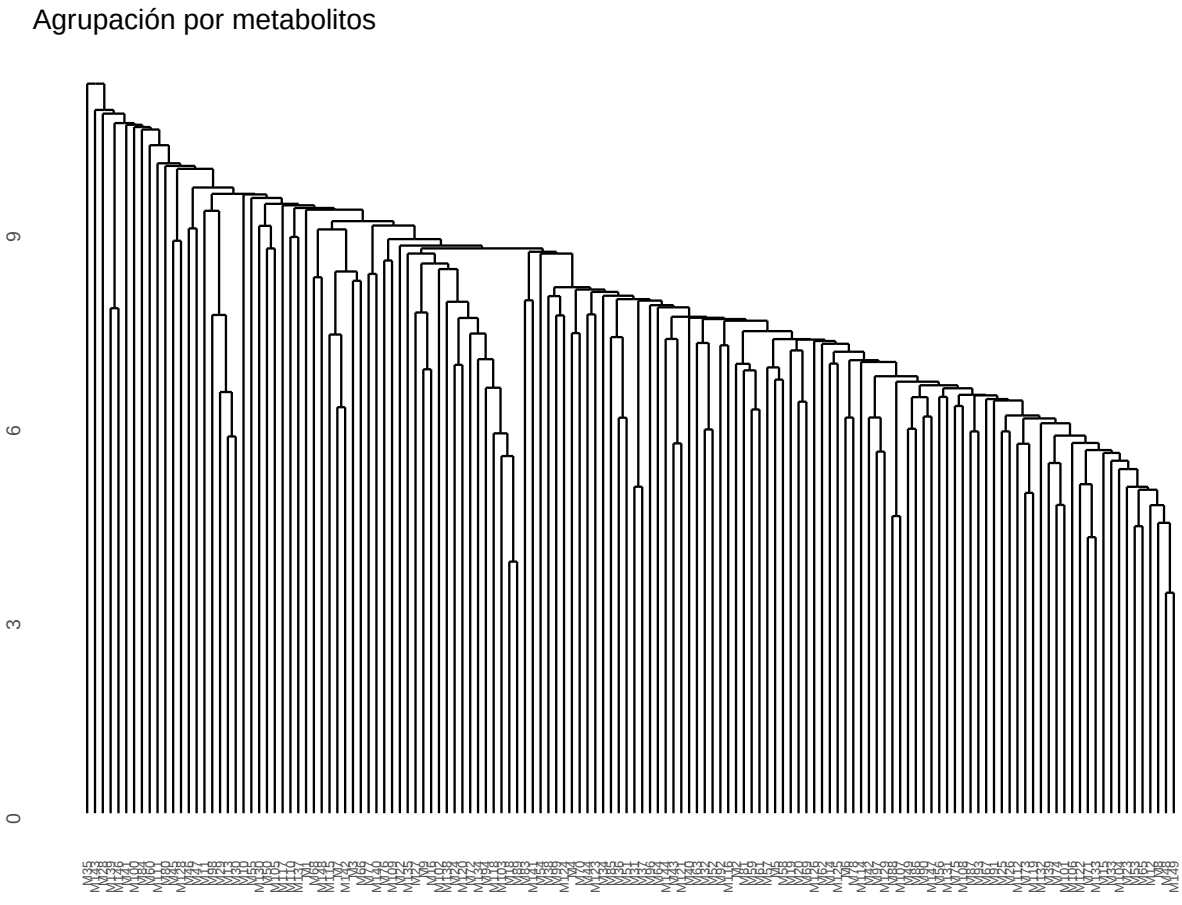

Figura 7: Dendograma en el que se agrupan las diferentes metabolitos

en este caso, a pesar de que se decidió eliminar los outliers a través del método MDS, hubiese sido conveniente analizar los individuos que tengan mucho peso o influencia en la muestra.

Por otro lado, tampoco se evaluó si las diferentes variables mostraban patrón de normalidad. Punto que puede modificar los resultados de las exploraciones. Por este motivo, en la correlación se decidió aplicar la correlación de Spearman en lugar de la de Pearson porque es más robusta, tanto en la presencia de outliers como en la posible falta de normalidad.

En las diferentes exploraciones realizadas, tanto en k-means como en los dendrogramas, se han podido observar diferentes agrupaciones de los datos. Para poder evidenciar la diferencia entre los grupos hubiese sido necesario aplicar algún test multivariante a los datos para ver si hay diferencias estadisticas entre las clases. También, hubiese sido una buena opción comprobar hasta que punto las diferentes clasificaciones coincidían con las clases que hay en el dataset.

Para acabar, faltaría estudiar que metabolitos son más relevantes o muestran una mayor asociación con el cáncer gástrico. Para hacer esto sería necesario ver que correlación muestran los diferentes metabolitos con la presencia de la enfermedad. O mejor aún, la combinación de ellos. En consecuencia, sería necesario utilizar otro tipo de técnicas multivariantes que reduzcan el número de variables a parte de la PCA. Esto, por ejemplo, se podría examinar a través de una regresión lineal cualitativa utilizando como predictor la presencia de cáncer mientras se combina con técnicas que reduzcan el número de variables, por ejemplo: LASSO, Partial Least Squares (PLS), etc.

6. Apéndice 1. Selección del número adecuado de grupos en la clasificación k-means

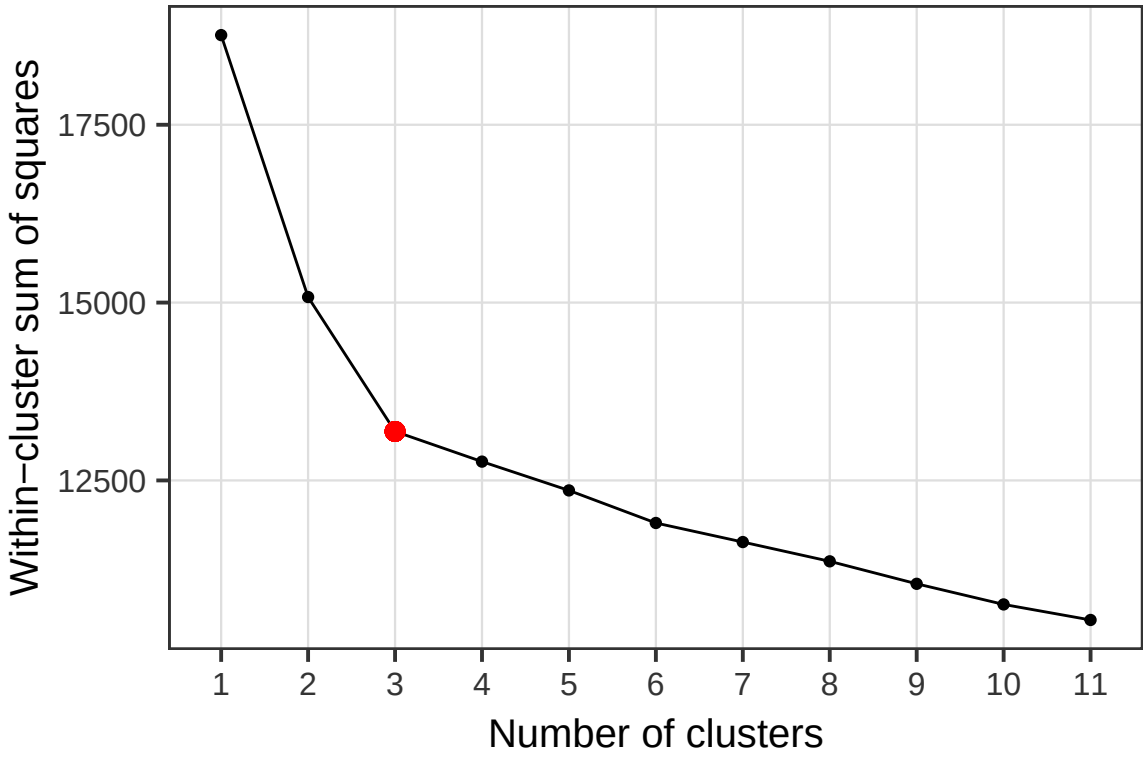

Figura 8: Número óptimo de núcleos para k-means

Bibliografía

Castellano-Escuder, P., González-Domínguez, R., Carmona-Pontaque, F., Andrés-Lacueva, C., & Sánchez-Pla, A. (2021). POMAShiny: A user-friendly web-based workflow for metabolomics and proteomics data analysis. *PLOS Computational Biology*, *17*(7), 1–15. https://doi.org/10.1371/journal.pcbi.1009148

Chan, A. W., Mercier, P., Schiller, D., Bailey, R., Robbins, S., Eurich, D. T., Sawyer, M. B., & Broadhurst, D. (2016). 1H-NMR urinary metabolomic profiling for diagnosis of gastric cancer. *British Journal of Cancer*, *114*(1), 59–62. https://doi.org/10.1038/bjc.2015.414

Huber, W., Carey, V. J., Gentleman, R., Anders, S., Carlson, M., Carvalho, B. S., Bravo, H. C., Davis, S., Gatto, L., Girke, T., Gottardo, R., Hahne, F., Hansen, K. D., Irizarry, R. A., Lawrence, M., Love, M. I., MacDonald, J., Obenchain, V., Ole's, A. K., … Morgan, M. (2015). Orchestrating high-throughput genomic analysis with Bioconductor. *Nature Methods*, *12*(2), 115–121. http://www.nature.com/nmeth/journal/v12/n2/full/nmeth.3252.html

Morgan, M., Obenchain, V., Hester, J., & Pagès, H. (2024). *SummarizedExperiment: SummarizedExperiment container*. https://doi.org/10.18129/B9.bioc.SummarizedExperiment

R Core Team. (2024). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. https://www.R-project.org/

Van Den Berg, R. A., Hoefsloot, H. C., Westerhuis, J. A., Smilde, A. K., & Van Der Werf, M. J. (2006). Centering, scaling, and transformations: Improving the biological information content of metabolomics data. *BMC Genomics*, *7*(1), 142. https://doi.org/10.1186/1471-2164-7-142