PEC1 - Análisis de Datos Ómicos

Curso 2024-2025

Lidia Getino Álvarez

6 de noviembre, 2024

Contents

1. Introducción

1.1. Resumen

El presente informe recoge una propuesta de solución a la primera prueba de evaluación contínua (PEC) de la asignatura *Análisis de Datos Ómicos*.

A lo largo de este, se ha trabajado con un conjunto de datos de metabolómica, sobre el cual se han aplicado las herramientas bioinformáticas y estadísticas necesarias para llevar a cabo una exploración del mismo:

- En primer lugar, se ha mostrado el proceso de selección e importación de los datos en *RStudio*.
- A continuación, se ha realizado una exploración inicial del conjunto de datos con la finalidad de familiarizarse con su estructura e información.
- Finalmente, se ha planteado una pregunta de interés biológico y se ha resuelto aplicando las técnicas estadísticas que se han considerado más apropiadas.

1.2. Objetivos

El objetivo principal de este trabajo consiste en llevar a cabo un análisis exploratorio de un conjunto de datos metabolómicos, aplicando un contenedor de tipo *SummarizedExperiment*, perteneciente al conjunto de paquetes de *Bioconductor*.

Como objetivos secundarios se encuentran la familiarización con los datos de experimetos metabolómicos y el manejo de repositorios colaboratorivos con control de versiones (*github*).

Adicionalmente, como parte de la exploración de los datos, se planteará una pregunta biológica apropiada a la naturaleza de los datos con los que se trabaje y se intentará responder empleando los métodos estadísticos que se consideren más apropiados para ello.

2. Materiales y métodos

2.1. Conjunto de datos de trabajo

El conjunto de datos empleado se ha obtenido del repositorio de datos de metabolómica *Metabolomics Workbench* (https://www.metabolomicsworkbench.org/). Para la selección, se realizó una breve revisión de los estudios disponibles en la mencionada base de datos y se escogió uno en función de su temática y el interés del mismo: ***ST002787 - Metabolomic analysis of gut metabolites in colorectal cancer patients: correlation with disease development and outcome***, perteneciente al proyecto con ID: PR001737 [Xie, 2024].

En este estudio de la *Wuhan University of Science and Technology*, los investigadores llevaron a cabo un análisis exploratorio del perfil metabolómico de las muestras fecales de 35 pacientes de cáncer colorectal, 37 pacientes de adenoma colorrectal y 30 pacientes sanos, con el objetivo de determinar posibles biomarcadores metabolómicos que diferenciasen a los tres grupos.

La recolección de las muestras fecales se llevó a cabo en pacientes que se sometieron a una colonoscopia y a un examen histopatológico en el Hospital Tianyou (Wuhan, China). A continuación, estas se prepararon y se sometieron a un análisis de espectrometría de masas en tándem por cromatografía líquida (LC-MS/MS). Finalmente, para la busqueda de potenciales biomarcadores, aplicaron técnicas de análisis estadístico, como el *Orthogonal Partial Least Squares Discriminant Analysis* (OPLS-DA) o el *Receiver operating characteristic Analysis* (ROC).

2.2. Métodos y herramientas bioinformáticas utilizados

Para la realización de este trabajo, se ha empleado el entorno de desarrollo integrado (IDE) *RStudio* [RStudio Team, 2020], el cual proporciona una interfaz cómoda para la programación en lenguaje *R* y la redacción de infórmes dinámicos con *RMarkdown*.

La importación y exploración inicial del conjunto de datos se ha llevado a cabo empleando algunas de las librerías del proyecto Bioconductor [Huber et al., 2015], concretamente *metabolomicsWorkbenchR*, *SummarizedExperiment* y *mixOmmics*.

Finalmente, como parte del proceso de exploración, se ha planteado una pregunta biológica de interés y se ha resuelto aplicando técnicas de análisis estadístico como el *Principal components analysis* (PCA) o el método de *Partial Least Squares Discriminant Analysis* (PLS-DA), cuya relevancia y motivo de selección se especificará a lo largo del informe.

2.3. Repositorio en *github*

El presente informe, así como los archivos asociados al mismo, se alojarán en un repositorio público en el servicio web de repositorios con control de versiones, *github* [github, 2020]. Para ello, se ha creado el repositorio en la web de *github* y, a continuación, se ha abierto un proyecto con control de versiones en *RStudio*, al cual se le ha proporcionado la URL del repositorio previamente creado, quedando así enlazados.

3. Resultados

3.1. Obtención de los datos y creación del contenedor

Una vez seleccionado el conjunto de datos, se procede a su importación en *R*, empleando la función *do_query()* del paquete *metabolomicsWorkbenchR*. Esta función permite realizar una consulta directa en *Metabolomics Workbench* e importar los datos de un estudio como un objeto de clase *SummarizedExperiment*.

```
# Realizo una consulta para acceder a los datos del estudio ST002787
se_data <- do_query(
  # Tipo de busqueda: por estudio
  context = 'study',
  # Dato proporcionado para la búsqueda: ID del estudio
  input_item = 'study_id',
  # ID del estudio
  input_value = 'ST002787',
  # Tipo de datos de salida: SummarizedExperiment
  output_item = 'SummarizedExperiment'
)

# Observo el output de la consulta
se_data
```

```
## $AN004534
## class: SummarizedExperiment
## dim: 1074 102
## metadata(8): data_source study_id ... description subject_type
## assays(1): ''
## rownames(1074): ME723397 ME722671 ... ME723400 ME722714
## rowData names(3): metabolite_name metabolite_id refmet_name
## colnames(102): CRA001 CRA002 ... HC029 HC030
## colData names(7): local_sample_id study_id ... Group_type Sex
##
## $AN004535
## class: SummarizedExperiment
## dim: 567 102
## metadata(8): data_source study_id ... description subject_type
## assays(1): ''
## rownames(567): ME723973 ME723992 ... ME723784 ME723808
```

```
## rowData names(3): metabolite_name metabolite_id refmet_name
## colnames(102): CRA001 CRA002 ... HC029 HC030
## colData names(7): local_sample_id study_id ... Group_type Sex
```

A primera vista, se observa que los datos importados del estudio *ST002787* se conforman de dos objetos de la clase *SummarizedExperiment*, pertenecientes a dos análisis diferentes. Estos parecen compartir formato, aunque no dimensiones, pues uno de ellos (AN004534) consta de 1074 filas (o metabolitos detectados) y el otro (AN004535) de 567. En los siguientes apartados profundizaremos más en las carácterísticas de estos datos.

3.1.1. Ejemplo de creación de un contenedor *SummarizedExperiment* desde cero

Inicialmente, por una mala interpretación del enunciado del trabajo, realicé la importación de los datos directamente en formato *SummarizedExperiment* desde *Metabolomics Workbench*, como se ha demostrado previamente. Sin embargo, tras recibir una comunicación aclaratoria en el foro de la asignatura, entendí que parte del interés del trabajo consistía en transformar, nosotros mismos, unos datos de partida en formato matriz a un contenedor de tipo *SummarizedExperiment*.

Dado que en el momento de la recepción de dicho anuncio ya había avanzado con el resto de los ejercicios propuestos usando el conjunto de datos arriba mencionado, tomé la decisión de centrar mi trabajo en dicho conjunto. No obstante, muestro a continuación un ejemplo de creación de un contenedor *SummarizedExperiment* a partir de uno de los conjuntos de datos del repositorio de *github* proporcionado en el enunciado: *human_cachexia.csv*.

En primer lugar, cargo el archivo de datos en *R* y visualizo su estructura:

```
# Cargo el archivo y guardo sus dimensiones
cachexia_data <- read.csv("human_cachexia.csv", stringsAsFactors = FALSE)
dim_cachexia <- dim(cachexia_data)

# Muestro de forma resumida la estructura del archivo
str(cachexia_data, list.len = 5)
```

```
## 'data.frame':    77 obs. of  65 variables:
##  $ Patient.ID                 : chr  "PIF_178" "PIF_087" "PIF_090" "NETL_005_V1" ...
##  $ Muscle.loss                : chr  "cachexic" "cachexic" "cachexic" "cachexic" ...
##  $ X1.6.Anhydro.beta.D.glucose: num  40.9 62.2 270.4 154.5 22.2 ...
##  $ X1.Methylnicotinamide      : num  65.4 340.4 64.7 53 73.7 ...
##  $ X2.Aminobutyrate           : num  18.7 24.3 12.2 172.4 15.6 ...
##   [list output truncated]
```

Vemos que el conjunto de datos consta de 77 observaciones, o pacientes, y 65 variables, entre las cuales tenemos el ID del paciente (*Patient.ID*), la variable de identificación del grupo del paciente en función de si padece o no cachexia (*Muscle.loss*) y una serie de metabolitos analizados.

Para estructurar el contenedor de tipo *SummarizedExperiment*, voy a crear los siguientes elementos propios de la clase:

- **Assay**: se compone de la matriz de datos principal, con los elementos biológicos de interés en las filas y las muestras, o pacientes, en las columnas. Por tanto, esta matriz la crearé trasponiendo el conjunto de datos original y eliminando las dos primeras columnas, para dejar así solo las columnas referentes a los metabolitos.

```r
# Antes de nada, asigno a las filas del conjunto el nombre del ID de paciente
rownames(cachexia_data) <- cachexia_data$Patient.ID

# Creo el assay con el conjunto original traspuesto, eliminado las dos primeras columnas
n_col <- dim_cachexia[2] #nº de columnas
cachexia_assay <- t(cachexia_data[,3:n_col])

# Muestro unos pocos registros del assay generado
kable(head(cachexia_assay[,1:5],4))
```

	PIF_178	PIF_087	PIF_090	NETL_005_V1	PIF_115
X1.6.Anhydro.beta.D.glucose	40.85	62.18	270.43	154.47	22.20
X1.Methylnicotinamide	65.37	340.36	64.72	52.98	73.70
X2.Aminobutyrate	18.73	24.29	12.18	172.43	15.64
X2.Hydroxyisobutyrate	26.05	41.68	65.37	74.44	83.93

- **colData**: consiste en los metadatos adicionales referentes a las muestras del experimento. En este caso, los metadatos de las muestras se compondrían de las dos primeras columnas del archivo de datos original, pues no disponemos de más información al respecto.

```r
# Creo el DataFrame de colData con las dos primeras columnas del conjunto original
cachexia_colData <- DataFrame(cachexia_data[,1:2])

# Muestro unos pocos registros del colData generado
kable(head(as.data.frame(cachexia_colData),4))
```

	Patient.ID	Muscle.loss
PIF_178	PIF_178	cachexic
PIF_087	PIF_087	cachexic
PIF_090	PIF_090	cachexic
NETL_005_V1	NETL_005_V1	cachexic

- **rowData**: consiste en los metadatos adicionales referentes a los elementos biológicos de interés. En este caso, los elementos biológicos de interés son los metabolitos, de los cuales solo conocemos el nombre. Por tanto, solo contendrá una columna con el nombre de los mismos.

```r
# Creo el DataFrame de rowData con los nombres de los metabolitos
met_names <- colnames(cachexia_data[,3:n_col])
cachexia_rowData <- DataFrame(Metabolites = met_names, row.names = met_names)

# Muestro unos pocos registros del colData generado
kable(head(as.data.frame(cachexia_rowData),4))
```

	Metabolites
X1.6.Anhydro.beta.D.glucose	X1.6.Anhydro.beta.D.glucose
X1.Methylnicotinamide	X1.Methylnicotinamide
X2.Aminobutyrate	X2.Aminobutyrate

	Metabolites
X2.Hydroxyisobutyrate	X2.Hydroxyisobutyrate

A continuación, procedo a la constitución del contenedor a partir de los elementos recién creados:

```
# Aplico la función SummarizedExperiment()
cachexia_se <- SummarizedExperiment(
  assays = list(counts = cachexia_assay),
  rowData = cachexia_rowData,
  colData = cachexia_colData
)

cachexia_se
```

```
## class: SummarizedExperiment
## dim: 63 77
## metadata(0):
## assays(1): counts
## rownames(63): X1.6.Anhydro.beta.D.glucose X1.Methylnicotinamide ...
##   pi.Methylhistidine tau.Methylhistidine
## rowData names(1): Metabolites
## colnames(77): PIF_178 PIF_087 ... NETL_003_V1 NETL_003_V2
## colData names(2): Patient.ID Muscle.loss
```

Como se puede observar, el conjunto de datos *human_cachexia.csv* está ahora correctamente formateado en forma de objeto *SummarizedExperiment*.

Visto este ejemplo, el resto del informe, como mencioné anteriormente, está realizado con el conjunto de datos importado directamente desde *Metabolomics Workbench*.

3.2. Exploración de los datos

3.2.1. Observación de los metadatos

Como punto de partida, se explora la información experimental de cada análisis del conjunto de datos con el fin de determinar las diferencias entre ambos. Esta información es accesible mediante la función *metadata()* del paquete *SummarizedExperiment*.

```
# Creo un data frame para visualizar los metadatos en formato tabla, con las
# categorías en filas y los 2 análisis en columnas
metadata_table <- function(metadata) {
  df <- data.frame(t(data.frame(metadata)))
  colnames(df) <- metadata$analysis_id
  return(df)
}

# Aplico la función a los dos análisis
meta_data <- cbind(metadata_table(metadata(se_data$AN004534)),
                   metadata_table(metadata(se_data$AN004535)))
```

	AN004534	AN004535
data_source	Metabolomics Workbench	Metabolomics Workbench
study_id	ST002787	ST002787
analysis_id	AN004534	AN004535
analysis_summary	Reversed phase POSITIVE ION MODE	Reversed phase NEGATIVE ION MODE
units	Peak Area	Peak area
name	ST002787:AN004534	ST002787:AN004535
description	Metabolomic analysis of gut metabolites in colorectal cancer patients: correlation with disease development and outcome	Metabolomic analysis of gut metabolites in colorectal cancer patients: correlation with disease development and outcome
subject_type	NA	NA

Como vemos, los dos análisis se diferencian únicamente en el modo de ionización (positiva/negativa) de las muestras a la hora de entrar al espectrómetro de masas. Esto se debe a que en el estudio, se empleó el método ESI (*electrospray ionization*) para la ionización de las muestras, el cual consta de dos modos de polaridad: iones positivos o negativos [Agarwal]. Realizar el análisis en ambos modos de ionización aumenta la sensibilidad de la técnica, permitiendo una mejor detección de los diversos metabolitos presentes en la muestra.

3.2.2. Estructura de los datos y diseño de experimento

Ahora que se conoce la diferencia entre los dos análisis, se analiza la estructura de cada uno de ellos:

```
# Extraigo la matriz de datos de AN004534
AN004534_data <- assay(se_data$AN004534)

# Muestro unos pocos registros de AN004534
kable(head(AN004534_data[,1:5],4),)
```

CRA001	CRA002	CRA003	CRA004	CRA005
28072	60481	8978.5	338360	359550.0
18767	40098	101140.0	24963	8489.2
71243	60693	72576.0	60035	97100.0
517600	319350	105920.0	695690	458780.0

```
# Extraigo la matriz de datos de AN004535
AN004535_data <- assay(se_data$AN004535)

# Muestro unos pocos registros de AN004535
kable(head(AN004535_data[,1:5],4))
```

CRA001	CRA002	CRA003	CRA004	CRA005
16205	33631	34256	3693.9	26133
81808	181930	50529	18550.0	52277
16561000	4373200	2240900	10866000.0	600140

CRA001	CRA002	CRA003	CRA004	CRA005
9521900	780990	6977000	1211700.0	343540

- El análisis AN004534, realizado con el modo de ionización positiva, consta de 1074 metabolitos identificados y 102 muestras.
- El análisis AN004535, realizado con el modo de ionización negativa, consta de 567 metabolitos identificados y 102 muestras.

Gracias a tratarse de objetos de la clase *SummarizedExperiment*, se puede visualizar información adicional acerca de las muestras y de los metábolitos detectados accediendo a esta mediante las funciones *colData()* y *rowData()*, respectivamente.

- **Muestras - colData:**

A continuación, se explora la información referente a los muestras empleadas en cada análisis, accediendo mediante la función *colData()*:

```
# Extraigo el data frame de las muestras del análisis AN004534
AN004534_samples <- as.data.frame(colData(se_data$AN004534))

# Extraigo el data frame de las muestras del análisis AN004535
AN004535_samples <- as.data.frame(colData(se_data$AN004535))

# Compruebo si es coincidente
identical(AN004534_samples, AN004535_samples)
```

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

Como cabía de esperar por la naturaleza del estudio, el conjunto de muestras empleado para cada uno es el mismo. A continuación, se muestra un par de registros de muestras por cada uno de los grupos de estudio:

	local_sample_id	study_id	sample_source	mb_sample_id	raw_data	Group_type	Sex
CRA001	CRA001	ST002787	Feces	SA299023	CRA001	Colorectal adenoma	Male
CRA002	CRA002	ST002787	Feces	SA298988	CRA002	Colorectal adenoma	Female
CRC001	CRC001	ST002787	Feces	SA299027	CRC001	Colorectal cancer	Female
CRC002	CRC002	ST002787	Feces	SA299044	CRC002	Colorectal cancer	Male
HC001	HC001	ST002787	Feces	SA299070	HC001	Heathy control	Female
HC002	HC002	ST002787	Feces	SA299069	HC002	Heathy control	Female

A partir de la tabla mostrada, se puede extraer que la nomenclatura empleada para codificar las muestras depende del grupo: los pacientes de **adenoma colorrectal** se nombran con las letras **"CRA"** seguidas de una clave numérica consecutiva de tres cifras, de igual modo, los pacientes de **cáncer colorrectal** se indican con las letras **"CRC"** y, por último, los pacientes **control** reciben las letras **"HC"**.

Otra información relevante que se puede extraer de esta tabla sería el origen de la muestra (aunque en todos los casos es fecal) o si el paciente es hombre o mujer:

	Colorectal adenoma	Colorectal cancer	Heathy control	Totals
Female	14	16	18	48
Male	23	19	12	54
Totals	37	35	30	102

Se observa que el diseño experimental es muy equilibrado en cuanto a la representación de los distintos grupos, así como en la representación de hombres y mujeres en cada uno de ellos.

- **Metabolitos detectados - rowData:**

A continuación, se explora la información referente a los metabolitos detectados en cada análisis, accediendo mediante la función *rowData()*:

```
# Extraigo el data frame de losmetabolitos del análisis AN004534
AN004534_met <- as.data.frame(rowData(se_data$AN004534))

# Extraigo el data frame de las muestras del análisis AN004535
AN004535_met <- as.data.frame(rowData(se_data$AN004535))

# Chequeo si hay metabolitos detectados en ambos análisis
common_met  <- AN004534_met$metabolite_id %in% AN004535_met$metabolite_id
table(common_met)
```

```
## common_met
## FALSE
##  1074
```

Como era de esperar por la naturaleza de cada análisis, los metabolitos detectados en cada uno son completamente diferentes. Esto pone de manifiesto la utilidad de haber realizado el análisis con ambos modos de ionización, pues de haberse limitado a solo un modo, muchos metabolitos presentes en las muestras no se habrían identificado.

A continuación, muestro unos pocos registros de la tabla de metabolitos de uno de los análisis, como ejemplo:

```
# Análisis AN004534
kable(head(AN004534_met,4))
```

	metabolite_name	metabolite_id	refmet_name
ME723397	10-Formyl-Thf	ME723397	
ME722671	1-(2,4-Dihydroxyphenyl)-2-(4-hydroxyphenyl)propan-1-one	ME722671	O-Desmethylangolensin
ME722692	1,3-Dicyclohexylurea	ME722692	1,3-Dicyclohexylurea
ME723382	1,4-Dihydro-1-Methyl-4-Oxo-3-Pyridinecarboxamide	ME723382	

Como se observa, cada fila se corresponde a uno de los metabolitos detectados. Como información adjunta se proporciona el nombre del metabolito, su "id" y su nomenclatura RefMet (propia de *Metabolomics Workbench*).

3.2.3. Análisis estadístico: PLS-DA

Dependiendo de la pregunta biológica que se quiera responder mediante el análisis de los datos de los que se dispone, serán de mayor utilidad unos métodos estadísticos u otros. En este caso, una pregunta que podríamos plantear sería: ***¿Se puede diferenciar a los distintos grupos de pacientes en función de su perfil metabólico?***

Con el objetivo de responder a dicha pregunta, se plantea, a continuación, un análisis de *Partial Least Squares Discriminant Analysis* (PLS-DA), método supervisado que permite detectar patrones en los datos, partiendo de un conocimiento previo de los grupos presentes en los mismos. Por tanto, se trata de una herramienta muy valiosa para casos como este.

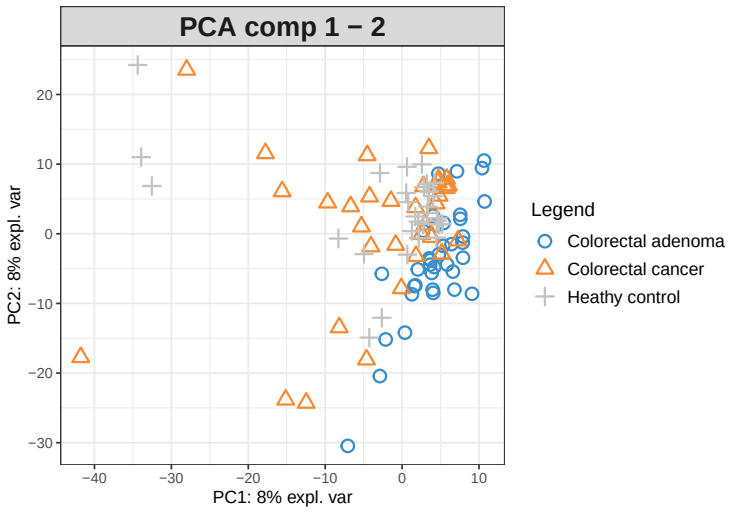
A lo largo de este apartado, se hará uso de las funciones del paquete *mixOmics* de *Bioconductor*. Para no extender excesivamente el informe, se ha decidido llevar a cabo el análisis estadístico sobre uno sólo de los dos *assay*: AN004534.

En primer lugar, se recogen los datos en dos variables "X" e "Y", de manera que "X" contiene la matriz de datos o *assay* traspuesta, es decir, con las muestras en las filas y los metabolitos en las columnas. Por su parte, "Y" será un vector con la información de la variable "Group_type", de la matriz de metadatos de las muestras:

```
# Guardo el contenido de la matriz de datos traspuesta (assay) en X
X <- as.data.frame(t(AN004534_data))
colnames(X) <- rownames(AN004534_met)

# Guardo la variable de interés "Group_type" en Y
Y <- AN004534_samples$Group_type
```

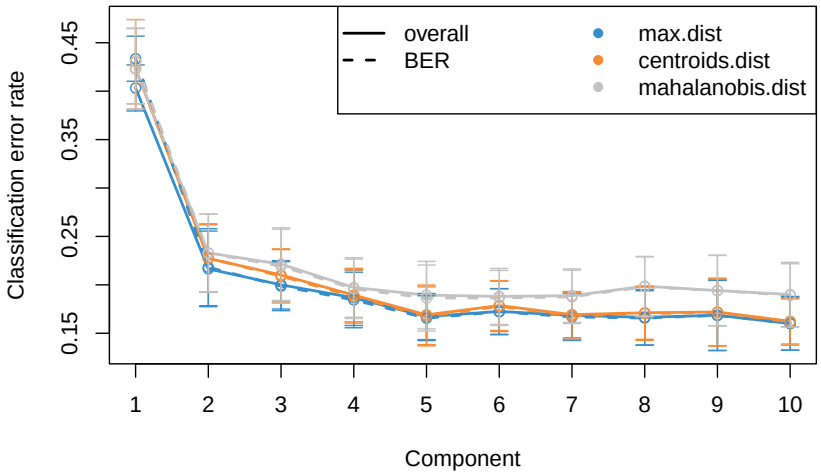
A continuación, se procede a hacer una evaluación preliminar de la agrupación natural de los datos sin tener en cuenta los grupos de los pacientes, aplicando un análisis de componentes principales o *Principal Component Analysis* (PCA):

```
# Aplico PCA escalando las variables para hacerlas comparables
pca <- pca(X, scale = TRUE)

# Ploteo los resultados
plotIndiv(pca, group = Y, ind.names = FALSE,
          legend = TRUE,
          title = "PCA comp 1 - 2")
```



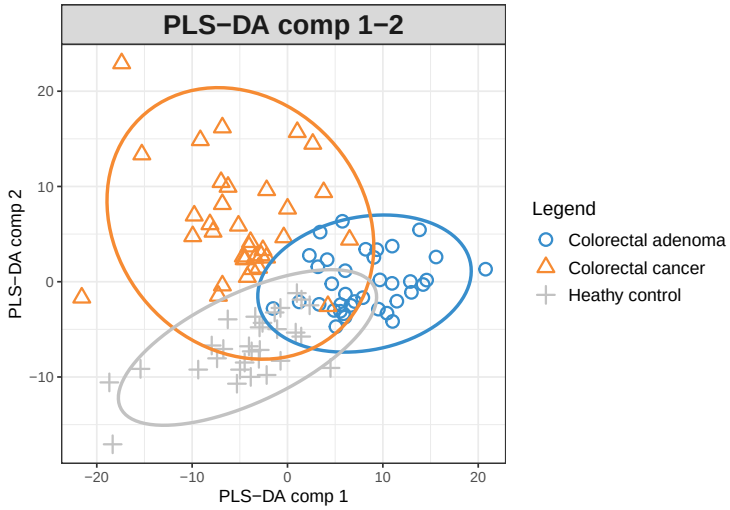
Como se puede observar, a penas existe separación entre los distintos grupos y el porcentaje de variabilidad explicado por las dos primeras componentes es bastante bajo. Esto pone de manifiesto que, a priori, no hay diferencias metabólicas claras entre los distintos grupos de pacientes. Sin embargo, se procede con el PLS-DA que, al ser un método supervisado, permitirá una mayor discriminación de los grupos por su perfil metabólico.

```
# Aplico PLS-DA partiendo de un numero elevado de componentes, 10, para luego
# realizar una validación del número óptimo de componentes
plsda <- plsda(X,Y, ncomp = 10)

# Llevo a cabo la validación aplicando el método de validación cruzada k-fold
set.seed(123)
perf.plsda <- perf(plsda, validation = 'Mfold', folds = 3,
                   nrepeat = 50)

plot(perf.plsda, sd = TRUE, legend.position = "horizontal")
```



En la gráfica del rendimiento de clasificación, se observa que la tasa de error disminuye notablemente entre la primera y la segunda componente. A partir de ese punto, disminuye ligeramente alcanzando un mínimo en la quinta componente. Sin embargo, se seleccionan las dos primeras componentes para evitar un posible error de sobreajuste.

```
# Re-aplico el PLS-DA con el número de componentes seleccionado
plsda <- plsda(X,Y, ncomp = 2)

# Ploteo los resultados
plotIndiv(plsda,
          ind.names = FALSE,
          legend = TRUE,
          ellipse = TRUE,
          title = "PLS-DA comp 1-2",
          X.label = 'PLS-DA comp 1', Y.label = 'PLS-DA comp 2')
```



Tras la aplicación de un método supervisado, como es el PLS-DA, se observa una mejor agrupación de los datos en función del grupo al que pertenecen los pacientes. Sin embargo, se sigue observando bastante solapamiento entre los grupos, sugiriendo que, aunque hay perfiles metabólicos distintivos, estos no proporcionan una separación clara.

```
# Aplico de nuevo el método de validación cruzada k-fold sobre el PLS-DA final
set.seed(123)
perf.plsda <- perf(plsda, validation = 'Mfold', folds = 3,
                   nrepeat = 50)

# Muestro la tasa de error por clase en cada componente
perf.plsda$error.rate.class$max.dist
```

```
##                         comp1     comp2
## Colorectal adenoma  0.0600000 0.1578378
## Colorectal cancer   0.3405714 0.2668571
## Heathy control      0.9000000 0.2300000
```

A partir de este resultado, se puede observar que la primera componente tiende a clasificar muy adecuadamente a los pacientes de adenoma colorectal, mientras que tiene una elevada tasa de error a la hora de clasificar a los pacientes control. Al incluir la segunda componente, sin embargo, observamos como las tasas de error se equilibran notablemente entre los tres grupos, siendo más adecuado para el objetivo deseado.

Las tasas de error obtenidas con la segunda componente se encuentran en todos los casos por debajo del 30%, reforzando lo que se observaba en el gráfico previo: se podría considerar que sí existen perfiles metabólicos distintivos entre los tres grupos, aunque estos no proporcionan una separación perfecta de los mismos.

4. Discusión y conclusiones

El conjunto de datos seleccionado ha resultado de gran versatilidad a la hora de realizar los ejercicios propuestos. La presencia de varios ensayos (*assay*), así como de sus respectivos metadatos, han permitido enriquecer el proceso de exploración e interacción con los datos.

A la hora de realizar el análisis estadístico con PLS-DA, se tomó la decisión de emplear sólo uno de los dos ensayos disponibles, con el fin de mantener el informe conciso. Sin embargo, esto ha limitado la capacidad

de responder a la pregunta biológica planteada. Sería interesante ampliar el estudio realizando también el análisis estadístico sobre los datos del segundo ensayo e, incluso, se podría valorar la opción de unificar ambos conjuntos, teniendo en cuenta que los metabolitos detectados en cada uno de los ensayos son completamente diferentes entre sí y no habría riesgo de duplicados. Esta integración podría proporcionar una imagen más completa de los perfiles metabólicos de cada grupo y de las potenciales diferencias entre ellos.

5. URL del respositorio de *github*

La **URL** del respositorio público de *github* en el que se pueden localizar tanto el presente informe, como todos los archivos asociados a este, es: https://github.com/lidiaga/Getino-Alvarez-Lidia-PEC1

6. Bibliografía

Vipin Agarwal. Positive and negative mode in mass spectroscopy: Lc-ms/ms method. URL https://www.nebiolab.com/positive-and-negative-mode-in-mass-spectroscopy/.

github. Github, 2020. URL https://github.com/.

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

RStudio Team. *RStudio: Integrated Development Environment for R.* RStudio, PBC., Boston, MA, 2020. URL http://www.rstudio.com/.

Zhufu Xie. Pr001737, 2024. URL http://dx.doi.org/10.21228/M85X48.