Máster Universitarioen bioinformática y bioestadística

Análisis de Datos Ómicos

Universitat Oberta de Catalunya

Prueba de Evaluación Continua 1

Aaron Peñas Cruz

06-11-2024

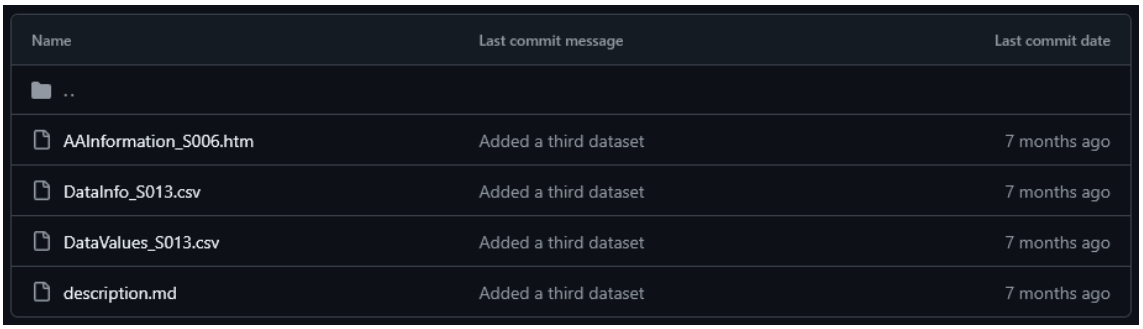
Contenido

Abstract

Seleccionar un dataset de metabolómica

En primer lugar, accedemos al repositorio de GitHub https://github.com/nutrimetabolomics/metaboData/. En mi caso, he optado por basar mi análisis en los datos de la carpeta “2018-MetabotypingPaper”. Dado que actualmente estoy trabajando como bioinformático en un proyecto similar, considero que me puede aportar unas habilidades directamente aplicables a las funciones que estoy desarrollando actualmente.

El proceso para descargar los datos es rápido e intuitivo, gracias a la interfaz de GitHub.



Name	Last commit message	Last commit date
..		
AAInformation_S006.htm	Added a third dataset	7 months ago
DataInfo_S013.csv	Added a third dataset	7 months ago
DataValues_S013.csv	Added a third dataset	7 months ago
description.md	Added a third dataset	7 months ago

La descripción indica que estos datos han sido extraídos de un artículo titulado "*Metabotypes of response to bariatric surgery independent of the magnitude of weight loss*". Como suele ser habitual, al publicar dicho artículo también se incluyen los datos empleados.

El archivo “*DataInfo_S013.csv*” contiene los metadatos, que nos aportan información sobre el contexto, la estructura y las características básicas de los datos.

Finalmente, disponemos del archivo “DataValues_S013.csv”, que contiene la información clínica y metabólica de 39 pacientes en 5 etapas temporales distintas.

Aclarar, que aunque en la descripción se menciona que el archivo “AAInformation_S600.csv” corresponde al formato *.csv*, dicho archivo se encuentra en formato *.htm*.

Cread un contenedor del tipo SummarizedExperiment

A continuación, declaramos dos nuevas variables: dataInfo y dataValues. En la primera de ellas almacenamos los metadatos, mientras que en la segunda guardamos los datos experimentales.

Una vez instalamos y cargamos correctamente el paquete SummarizedExperiment, disponemos de todas las herramientas necesarias para crear el contenedor en el formato deseado.

Objetivos del estudio

A priori, el requerimiento para la resolución de esta actividad es que se realice una exploración de los datos. Por lo tanto, en este informe voy a responder a diferentes preguntas que puede resultar de interés para un posterior análisis del que se pueda extraer diferentes conclusiones.

Primero vamos a describir, contextualizar y filtrar los datos originales, para garantizar que en el procesamiento de estos no haya errores. Por ello, debemos comprender su naturaleza, identificar posibles outliers, comprobar la distribución de las variables más importantes y comprobar si hay valores nulos. Algunas de las preguntas que podemos plantear son las siguientes:

- ¿Cómo están estructurados los datos? ¿Cuántas filas y columnas hay?
- ¿Qué representa el dataset? ¿ Cómo se han obtenido los valores?
- ¿Las variables presentan missing values?
- ¿Presencia de outliers?

Considero que debemos responder a estas preguntas, con la idea de llevar a cabo un proceso exploratorio de los datos.

Materiales y métodos

Origen de los datos

Los datos provienen de un estudio en el que se obtienen muestras de 39 pacientes con obesidad mórbida. Todas ellas, fueron recolectadas en el Hospital Universitario de la Victoria, en la provincia de Málaga.

El análisis de los metabolitos consiste en estudiar las diferencias presentes en diferentes momentos de tiempo, antes y después de que los sujetos se sometieran a una intervención bariátrica.

Naturaleza

Los valores obtenidos corresponden a datos metabolómicos extraídos de suero sanguíneo. Se específica la presencia de varios metabolitos: acilcarnitinas, fosfatidicolinas, esfingomielinas, aminoácidos, aminas biogénicas y hexosas.

Respecto a la obtención de las muestras, se indica que se empleó espectrometría de masas: LC-MS/MS y FIA-ESI-MS/MS. Posteriormente, se procesaron los valores con la herramienta MetIDQ.

Herramientas

En lo que respecta a este informe, las herramientas empleadas han sido entorno de programación R Studio y la plataforma GitHub. En el análisis se utiliza el lenguaje R, con paquetes propios del proyecto de código abierto: *Bioconductor*. Concretamente, el paquete *SummarizedExperiment*, que ha permitido crear un contenedor en el que se fusionan datos y metadatos en una sola variable.

Procedimiento general

Comenzamos con la selección y la carga del dataset correspondiente, a partir del repositorio de GitHub indicado en las instrucciones de esta tarea. A continuación, filtramos las variables que no corresponden a las concentraciones de los metabolitos. También transponemos el dataframe, para poder crear correctamente el contenedor en formato SummariedExperiment. Verificamos la estructura de dicho contenedor, comprobando que las columnas corresponden a los pacientes y las filas a las concentraciones de metabolitos.

Finalmente, iniciamos el análisis exploratorio de los datos, en el que utilizamos la matriz resultante de aplicar la función assay() en el contenedor anteriormente mencionado. Se grafica un heatmap que nos muestras las asociaciones entre metabolitos y muestras, inspeccionamos de forma visual uno de los metabolitos mediante un boxplot y concluimos con un análisis de componentes principales (PCA) para obtener la distribución de las muestras y valorar si existe algún patrón.

Resultados

Creación del objeto SummarizedExperiment

En primer lugar, creamos el objeto de tipo SummarizedExperiment. Para ello, debemos excluir aquellas variables que no aportan los información sobre los valores de concentración de los metabolitos. En nuestro caso, eliminamos las variables: SURGERY, AGE, GENDER y Group y las guardamos como metadatos en una nueva variable llamada meta_values.

```
data_values <- read.csv("2018-MetabotypingPaper/DataValues_S013.csv", row.names = 1)
meta_data <- data_values[, 1:5]
meta_data <- as.data.frame(meta_data)
grupo <- data_values$Group

data_values <- data_values[, -(1:5)]
data_values <- as.data.frame(t(data_values))
```

Además, debemos transponer los datos antes de crear el contenedor, ya que las filas de un objeto de tipo SummarizedExperiment han de corresponder a las variables de los metabolitos, mientras que las columnas han de ser los pacientes.

```
se <- SummarizedExperiment(
       assays = list(counts = as.matrix(data_values)),
)
```

Una vez hemos creado el objeto, podemos proceder con el análisis exploratorio de los datos. A continuación, mostramos por pantalla la estructura principal del objeto en sí mismo:

```
> se
class: SummarizedExperiment
dim: 690 39
metadata(0):
assays(1): counts
rownames(690): MEDDM_T0 MEDCOL_T0 ... SM.C24.0_T5 SM.C24.1_T5
rowData names(0):
colnames(39): 1 2 ... 38 39
colData names(5): SUBJECTS SURGERY AGE GENDER Group
```

Se puede comprobar que la estructura es correcta, ya que las columnas corresponden a los 39 pacientes, mientras que las 690 filas representan las diferentes concentraciones de metabolitos.

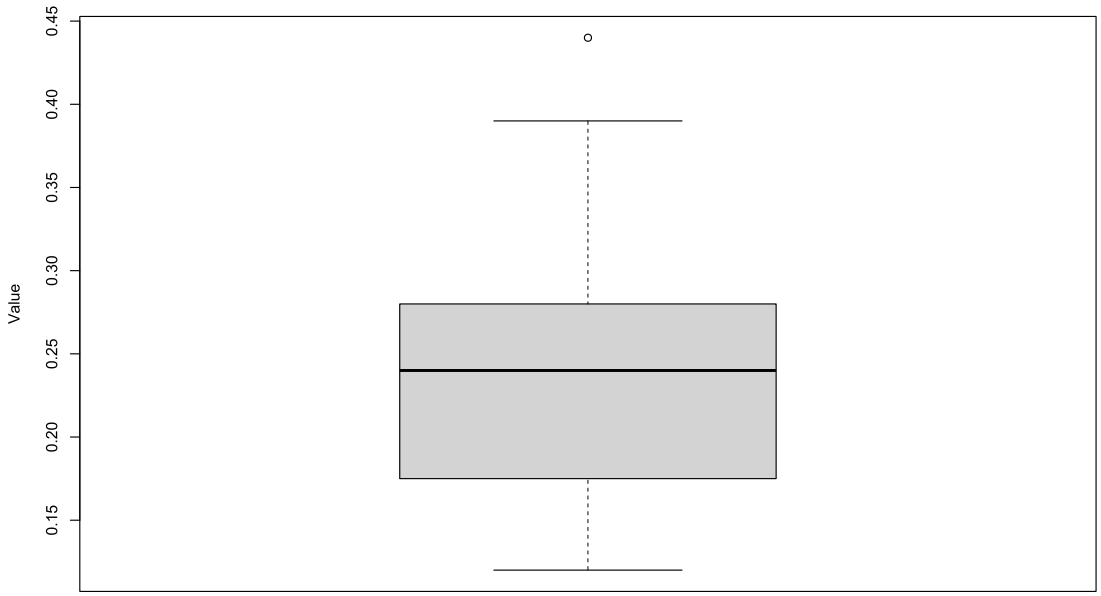
Descripción de los datos

Hay muchas funciones disponibles para utilizar en el paquete SummarizedExperiment y una de ellas es assay(), que nos permite acceder a la matriz que contiene los valores de nuestras medidas para cada uno de los pacientes. A continuación, incluyo una captura de pantalla en la que se muestra la matriz recortada, ya que hay una gran cantidad de datos y abarca demasiado como para incluirla en su totalidad.

```
> assay_se <-  assay(se, "counts")
> print(assay_se)
```

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
MEDDM_T0	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
MEDCOL_T0	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
MEDINF_T0	0.000	0.000	0.000	0.000	0.000	0.000	0.000	1.000	0.000	0.000	1.000	0.000	0.000	0.000	0.000	0.000	0.000
MEDHTA_T0	1.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	1.000	1.000	1.000	1.000	0.000
GLU_T0	85.000	78.000	75.000	71.000	82.000	71.000	80.000	90.000	92.000	84.000	75.000	108.000	101.000	105.000	139.000	106.000	159.000
INS_T0	11.400	12.100	8.410	12.800	6.010	9.880	9.200	3.400	5.430	6.980	13.300	16.800	17.100	21.300	36.600	20.000	17.600
HOMA_T0	2.400	2.320	1.560	2.250	1.220	1.730	1.820	0.760	1.230	1.450	2.470	4.470	4.260	5.530	12.600	5.240	6.910
HBA1C_T0	NA	NA	5.400	5.100	5.600	5.100	5.600	5.500	5.700	5.500	5.700	NA	NA	NA	NA	5.800	NA
HBA1C.mmol.mol_T0	NA	NA	35.510	32.230	37.690	32.230	37.690	36.600	38.780	36.600	38.780	NA	NA	NA	NA	39.880	NA
PESO_T0	151.000	139.000	84.000	136.000	121.000	148.000	109.000	109.000	114.000	120.000	171.000	135.000	124.000	119.000	154.000	162.000	146.000
bmi_T0	62.900	47.000	29.800	53.100	46.600	48.800	43.700	41.800	44.000	40.600	54.400	55.500	52.300	45.900	62.500	61.700	49.900
CC_T0	0.700	NA	0.700	1.000	0.900	0.700	0.900	0.900	0.900	NA	1.100	0.900	0.900	0.900	1.000	0.900	0.900
CINT_T0	116.000	NA	90.000	157.000	123.000	110.000	122.000	124.000	136.000	NA	166.000	147.000	132.000	128.000	156.000	147.000	104.000
CAD_T0	167.000	NA	126.000	162.000	132.000	148.000	141.000	136.000	148.000	NA	149.000	162.000	151.000	143.000	156.000	164.000	110.000
TAD_T0	125.000	NA	79.000	73.000	84.000	74.000	65.000	75.000	91.000	NA	88.000	95.000	85.000	NA	90.000	102.000	103.000
TAS_T0	174.000	NA	111.000	127.000	122.000	131.000	121.000	128.000	128.000	NA	148.000	140.000	150.000	NA	180.000	144.000	159.000
TG_T0	147.000	150.000	45.000	109.000	30.000	61.000	75.000	53.000	164.000	145.000	79.000	268.000	204.000	225.000	145.000	143.000	171.000
COL_T0	256.000	180.000	211.000	205.000	102.000	121.000	192.000	181.000	154.000	220.000	144.000	261.000	181.000	207.000	254.000	232.000	295.000
LDL_T0	167.000	94.000	114.000	146.000	24.000	60.800	96.000	99.400	67.200	160.000	83.200	162.000	99.200	97.000	163.000	162.000	154.000
HDL_T0	60.000	56.000	88.000	37.000	72.000	48.000	81.000	71.000	54.000	31.000	45.000	45.000	41.000	65.000	62.000	41.000	107.000
VLDL_T0	29.400	30.000	9.000	21.800	6.000	12.200	15.000	10.600	32.800	29.000	15.800	53.600	40.800	45.000	29.000	28.600	34.200
PCR_T0	10.200	9.000	3.050	8.890	NA	1.600	4.520	NA	NA	NA	0.000	5.330	4.100	21.900	6.820	5.630	29.000
LEP_T0	155.000	84.000	27.000	46.000	NA	38.000	61.000	NA	NA	NA	NA	61.500	59.000	62.300	75.600	99.000	65.800

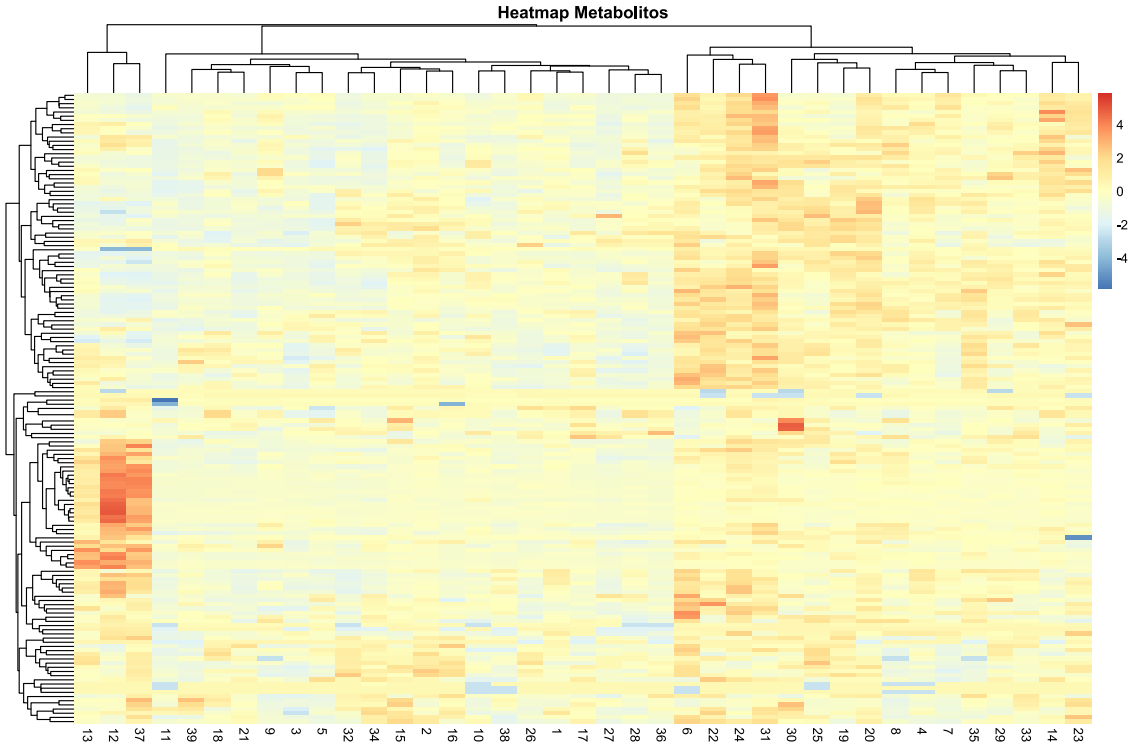
Boxplot



Gracias a esta función, podemos visualizar rápidamente la variabilidad entre muestras de un metabolito en específico, como en el siguiente ejemplo. Podemos observar un diagrama de cajas correspondiente al metabolito 98. La media se sitúa cerca del valor 0.24, habiendo la existencia de un valor extremo próximo a 0.45 unidades.

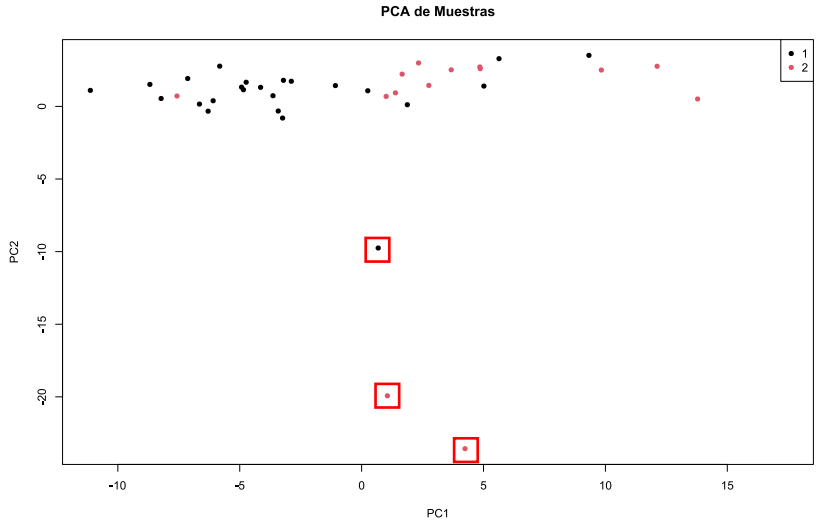
Heatmap

También podemos generar un heatmap, en el que se pueden ver diferentes asociaciones de forma clara y muy visual. La contrapartida es que no se pueden incluir los nombres de todas las variables estudiadas, ya que impide la correcta visualización de este.



Finalmente, también podemos utilizar un análisis de componentes principales (PCA). En el siguiente gráfico se ha incluido la variable *Group*, que probablemente corresponde al batch effect de las muestras. Podemos observar cómo, a priori, las muestras del lote 2 se sitúan en la zona derecha, mientras que las del lote 1 se ubican en la zona izquierda. Tal vez, deberíamos de realizar una corrección por batch effect en el caso hipotético de que llevásemos a cabo un análisis de significancia estadística. También he señalado manualmente tres puntos que podrían ser outliers, debido a la distancia que presentan respecto al resto.

Principal Analysis Component



Hay otras muchas funciones que nos sirven para visualizar el contenido del objeto, que nos retornan los metadatos, las dimensiones, los nombres de las filas y columnas...

Además de otras funciones que nos permiten modificar dicho contenedor, como subset() para filtrar según diferentes criterios o la función de inicialización SummarizedExperiment, que además de crear nuevos objetos nos permite modificar los existentes.

Discusión y limitaciones del estudio

A lo largo de este informe, hemos podido observar como algunos metabolitos presentan valores extremos además de una variabilidad alta entre los pacientes. Para un análisis de significancia, sería interesante incluir como covariables la edad o el género de los pacientes.

A partir del análisis de componentes principales (PCA), hemos identificado algunos posibles outliers alejados del grupo en los que los puntos se hayan distribuidos uniformemente. Al etiquetar los puntos presentes en el gráfico agrupando por la variable Group, vemos una clara diferenciación entre las muestras correspondientes al grupo 1 y 2. Esta enumeración se aplica habitualmente para indicar el lote del cual han formado parte al procesarse las muestras, ya que cuando hay muchas de ellas no se pueden incluir simultáneamente en las máquinas de extracción. Con lo que se puede inducir ruido, denominado efecto lote (batch effect). Para solucionarlo, se puede emplear un paquete concreto para este tipo de datos, conozco la función ComBat del paquete SVA, que se suele utilizar para corregir este tipo de efecto en análisis de niveles de expresión.

Las limitaciones que puedo mencionar son las que se suelen encontrar en este tipo de estudios. En primer lugar, el tamaño de la muestra nos puede llegar a limitar a la hora de sacar conclusiones significativas.

Además, la presencia de valores faltantes y outliers puede suponer un problema. La primera de ellas se suele obviar, ya que en muchas de las funciones de R se excluyen por defecto los valores faltantes y, en caso contrario, podemos utilizar la función na.omit() para depurar la base de datos. También podemos identificar la naturaleza de esos valores faltantes y buscar el método de imputación que mejor se ajuste a dicha naturaleza.

Respecto a los outliers, debemos tener en cuenta que existe la posibilidad de que los outliers sean fruto de una variabilidad real y significativa, su comportamiento no siempre es debido a errores en la obtención y procesado de las muestras. Por ello, debemos contemplar diferentes vías: sin filtrar pero aplicando una transformación logarítmica que reduzca su impacto o bien filtrando a partir de la media y 3 desviaciones estándar (en caso de que se siga una distribución normal). Debo mencionar que en un análisis de significancia, deberíamos de incluir como covariables algunas de las variables relacionadas con factores clínicos y ambientales que han sido excluidas en primera instancia.

Como conclusión, en un posterior procesamiento de estos datos, deberíamos tener en cuenta todos los puntos mencionados con anterioridad, garantizando un buen tratamiento de outliers y missing values, una normalización de los datos crudos y necesaria corrección del batch effect que parece haber mostrado la PCA realizada.

Repositorio GitHub

A continuación dejo el enlace de acceso a mi repositorio:

https://github.com/AaronPenas2024/Penas-Cruz-Aaron-PEC1