Code

Análisis de datos ómicos (M0-157) Primera prueba de evaluación continua.

Fecha de publicación del enunciado: 31/10/2022

Fecha límite para presentar la PEC: 16/11/2022[1]

1 Introducción

La solución de la PEC tiene dos objetivos principales: - Por un lado, debe mostrar un modelo de "resolución" de las preguntas planteadas que sirva de ejemplo de como se podía haber resuelto las preguntas y de comparación con vuestra propia solución. - Además debe contener explicaciones sobre el desarrollo de la PEC, que, normalmente no formaran parte de un informe, como es el caso de esta introducción. Para hacerlo sencillo he dejado los comentarios, es decir nuestras opiniones o valoraciones. en cursiva y lo que es la solución estricta, es decir lo que debe parecerse a vuestra resolución, en formato normal.

Tal como se os pedía, el código no aparecerá dentro del texto, aunque con fines didácticos lo podréis mostrar cuando así lo deseéis. Esto se puede hacer con un simple comando de Rmarkdown

En este documento se presenta una solución tipo, es decir no se valoran todos los casos sino que se resuelve uno. Eventualmente se añadirán comentarios sobre aquellos datasets que puedan haber dado problemas o que tengan especificidades concretas.

1.1 Objetivos

El objetivo principal de este trabajo es realizar un análisis exploratorio de unos datos de microarrays, descargados de la base de datos Gene Expression Omnibus (GEO) (https://www.ncbi.nlm.nih.gov/geo/) utilizando el programa estadístico R y las librerías para análisis de datos ómicos integradas en Bioconductor. En concreto se requiere que los datos descargados se almacenen en un objeto de clase `ExpressionSet` y que se acceda a ellos a través de este objeto.

2 Materiales y Métodos

Los datos con los que se trabajará se descargarán de GEO utilizando el paquete GEOquery (https://bioconductor.org/packages/release/bioc/html/GEOquery.html) de Bioconductor. Este paquete permite descargar los datos indicados (por un identificador GSE o GDS) y crear con ellos una estructura de datos del tipo `expressionSet` que contiene una matriz de datos preprocesados (habitualmente normalizados, así como una tabla con información sobre covariables y tras aspectos del experimento.

La exploración se llevará a cabo siguiendo la plantilla de un caso de estudio, Analisis_de_datos_omicos-Ejemplo_0-Microarrays (https://github.com/ASPteaching/Analisis_de_datos_omicos-Ejemplo_0-Microarrays) y por lo tanto está relativamente bien pautada. Básicamente dicha exploración consistirá en: - Análisis univariante de los datos, mediante boxplots y/o histogramas para estudiar la forma general de los mismos. - Análisis multivariante de los datos, mediante Análisis de Componentes Principales y Agrupamiento Jerárquico, para determinar si los grupos que aparezcan (en caso de hacerlo) parecen relacionarse con las fuentes de variabilidad del estudio o, si por el contrario, podrían haber otras fuentes de variabilidad como efectos batch.

Las técnicas mencionadas son muy habituales en estadística y bioinformática por lo que no nos entretendremos en explicarlas. Pueden verse resumidas en los materiales de la asignatura (documento “Módulo 1. Preliminares”).

2.1 Selección de los datos para el estudio

Los datos para este estudio se han tenido que escoger de entre una lista de estudios depositados en GEO. Aunque es posible escoger el que se desee, evitando datasets problemáticos, una forma “neutra” de hacerlo es sorteando el identificador.

El registro de la fila 27 da problemas al descargarlo por lo que se suprime

Code

Fijamos una semilla para el generador interno de números aleatorios, lo que nos asegura que, mientras no la cambiemos, se repetirá la asignación del estudio.

Code

Una vez escogida una fila de forma aleatoria es posible descargar los datos a partir del “DataSet” o de la “Serie”. - En el primer caso se obtiene una lista con un `expressionSet` por cada posible DataSet del estudio. - En el segundo se obtiene un objeto de clase `GDS` , que contiene campos adicionales con información sobre el experimento.

Para este estudio hemos seleccionado el resultado de la selección aleatoria, GSE3737.

3 Resultados

Este apartado contiene embebido el código de análisis utilizado. Esto permite seguir los principios de la “investigación reproducible” y “programación literata” en la que el informe se integra con el análisis. El código se mantiene oculto para facilitar la lectura por parte de personas no familiarizadas con R

3.1 Estructura de los datos y del estudio

3.1.1 Opción 1: Utilizar la clase `GDS` y trabajar con el resultado de descargar con el identificador `GDS`

Aunque podríamos trabajar únicamente con el objeto de clase “GDS” los utilizaremos ambos.

Para este ejercicio utilizaremos el `ExpressionSet` obtenido a partir de la serie.

Code

La clase ‘GDS’ contiene - información sobre el estudio en forma de metadatos que podemos utilizar para crear una tabla resumen de la información del estudio. - Una tabla con información sobre grupos y muestras en cada grupo - La matriz de expresión, codificada de forma no evidente, aunque puede extraerse creando un `expressionSet`

Code

```
$channel_count
[1] "1"

$dataset_id
[1] "GDS1736" "GDS1736"

$description
[1] "Analysis of PC-3 prostate cancer cells incubated with arachidonic acid
(AA).  AA is an omega-6 fatty acid shown to induce cancer cell proliferatio
n. Results suggest AA plays an important role in stimulation of growth-relat
ed genes and proliferation via phosphatidylinositol 3-kinase (PI3K) signalin
g."
[2] "control"
[3] "arachidonic acid"

$email
[1] "geo@ncbi.nlm.nih.gov"

$feature_count
[1] "22283"

$institute
[1] "NCBI NLM NIH"

$name
[1] "Gene Expression Omnibus (GEO)"

$order
[1] "none"

$platform
[1] "GPL96"

$platform_organism
[1] "Homo sapiens"

$platform_technology_type
[1] "in situ oligonucleotide"

$pubmed_id
[1] "16452198"

$ref
[1] "Nucleic Acids Res. 2005 Jan 1;33 Database Issue:D562-6"

$reference_series
[1] "GSE3737"

$sample_count
[1] "8"
```

```
$sample_id
[1] "GSM86079,GSM86080,GSM86081,GSM86082" "GSM86083,GSM86084,GSM86085,GSM860
86"

$sample_organism
[1] "Homo sapiens"

$sample_type
[1] "RNA"

$title
[1] "Arachidonic acid effect on prostate cancer cells"

$type
[1] "Expression profiling by array" "agent"
[3] "agent"

$update_date
[1] "May 31 2006"

$value_type
[1] "count"

$web_link
[1] "http://www.ncbi.nlm.nih.gov/geo"
```

Code

```
    sample            agent
1 GSM86079          control
2 GSM86080          control
3 GSM86081          control
4 GSM86082          control
5 GSM86083 arachidonic acid
6 GSM86084 arachidonic acid
7 GSM86085 arachidonic acid
8 GSM86086 arachidonic acid
                                                        description
1               Value for GSM86079: 1 PC3 0hr; src: PC-3, untreated
2              Value for GSM86080: 1 PC3 0hra; src: PC-3, untreated
3               Value for GSM86081: 2 PC3 0hr; src: PC-3, untreated
4        Value for GSM86082: B final PC3 ctrl; src: PC-3, untreated
5          Value for GSM86083: 3 PC3 2hr; src: PC-3, AA treated 2hr
6          Value for GSM86084: 4 PC3 2hr; src: PC-3, AA treated 2hr
7 Value for GSM86085: C final PC3 AA 2hr; src: PC-3, AA treated 2hr
8 Value for GSM86086: D final PC3 AA 2hr; src: PC-3, AA treated 2hr
```

Code

```
     ID_REF IDENTIFIER GSM86079 GSM86080 GSM86081 GSM86082 GSM86083 GSM86084
1 1007_s_at    MIR4640    738.9    620.1    706.6    985.2    664.3    557.2
2   1053_at       RFC2     96.1    312.8    129.9    341.1    109.1    143.4
3    117_at      HSPA6    100.8     83.1     83.5    245.5    102.3     66.2
4    121_at       PAX8    823.4    646.2    805.9   1122.7    635.1    598.1
5 1255_g_at     GUCA1A     38.3     23.2     13.0     51.3     29.9     23.7
6   1294_at    MIR5193    239.7    100.6    180.3    198.9    249.8    150.6
7   1316_at       THRA     84.6     49.6     62.3    180.0     35.8     33.9
  GSM86085 GSM86086
1    991.6    873.5
2    420.5    439.2
3    161.5    185.2
4    991.9    809.9
5     47.9     56.9
6    106.0     90.7
7    113.7     86.6
 [ reached 'max' / getOption("max.print") -- omitted 22276 rows ]
```

Con algo de trabajo es posible crear una tabla con la información de los Metadatos.

Code

Campo	Descripción
channel_count	1
dataset_id	GDS1736
description	Analysis of PC-3 prostate cancer cells incubated with arachidonic acid (AA). AA is an omega-6 fatty acid shown to induce cancer cell proliferation. Results suggest AA plays an important role in stimulation of growth-related genes and proliferation via phosphatidylinositol 3-kinase (PI3K) signaling.
email	geo@ncbi.nlm.nih.gov (mailto:geo@ncbi.nlm.nih.gov)
feature_count	22283
institute	NCBI NLM NIH
name	Gene Expression Omnibus (GEO)
order	none
platform	GPL96
platform_organism	Homo sapiens
platform_technology_type	in situ oligonucleotide

Campo	Descripción
pubmed_id	16452198
ref	Nucleic Acids Res. 2005 Jan 1;33 Database Issue:D562-6
reference_series	GSE3737
sample_count	8
sample_id	GSM86079,GSM86080,GSM86081,GSM86082
sample_organism	Homo sapiens
sample_type	RNA
title	Arachidonic acid effect on prostate cancer cells
type	Expression profiling by array
update_date	May 31 2006
value_type	count
web_link	http://www.ncbi.nlm.nih.gov/geo (http://www.ncbi.nlm.nih.gov/geo)

Para disponer de la tabla de grupos, así como de los valores de expresión basta con convertir el objeto `myGDS` a un objeto de clase `ExpressionSet`.

Code

	sample	agent	description
GSM86079	GSM86079	control	Value for GSM86079: 1 PC3 0hr; src: PC-3, untreated
GSM86080	GSM86080	control	Value for GSM86080: 1 PC3 0hra; src: PC-3, untreated
GSM86081	GSM86081	control	Value for GSM86081: 2 PC3 0hr; src: PC-3, untreated
GSM86082	GSM86082	control	Value for GSM86082: B final PC3 ctrl; src: PC-3, untreated
GSM86083	GSM86083	arachidonic acid	Value for GSM86083: 3 PC3 2hr; src: PC-3, AA treated 2hr
GSM86084	GSM86084	arachidonic acid	Value for GSM86084: 4 PC3 2hr; src: PC-3, AA treated 2hr

	sample	agent	description
GSM86085	GSM86085	arachidonic acid	Value for GSM86085: C final PC3 AA 2hr; src: PC-3, AA treated 2hr
GSM86086	GSM86086	arachidonic acid	Value for GSM86086: D final PC3 AA 2hr; src: PC-3, AA treated 2hr

3.1.2 Opción 2: Trabajar con el resultado de descargar con el identificador `GSE`

Antes de empezar a trabajar con los datos estos se extraen de la lista que los contiene (en forma de objeto de clase `ExpressionSet`).

El objeto contiene la matriz de expresión, las posibles covariables e información del estudio almacenada de forma poco intuitiva, puesto que se repite para cada muestra.

La matriz de expresión y la información fenotípica pueden extraerse con las funciones `exprs` y `pData` respectivamente.

Code

```
[1] "ExpressionSet"
attr(,"package")
[1] "Biobase"
```

En este caso el objeto no dispone de un tabla como la del caso anterior por lo que debemos elaborarla a partir de una **inspección detallada** del objeto `pData(eSet)`. Como puede verse parece que las únicas columnas con información sobre los grupos experimentales parecen ser la 2 y la 8

Code

```
                    title      source_name_ch1
GSM86079          1 PC3 0hr      PC-3, untreated
GSM86080         1 PC3 0hra      PC-3, untreated
GSM86081          2 PC3 0hr      PC-3, untreated
GSM86082   B final PC3 ctrl      PC-3, untreated
GSM86083          3 PC3 2hr PC-3, AA treated 2hr
GSM86084          4 PC3 2hr PC-3, AA treated 2hr
GSM86085 C final PC3 AA 2hr PC-3, AA treated 2hr
GSM86086 D final PC3 AA 2hr PC-3, AA treated 2hr
```

- La columna “source_name_ch1” sugiere que hay dos grupos “treated/untreated”.
- La columna “title” no es muy explicativa y sugiere la existencia de otra clasificación superpuesta PC3/final.

De momento se crea una etiqueta para cada muestra que contenga ambas informaciones

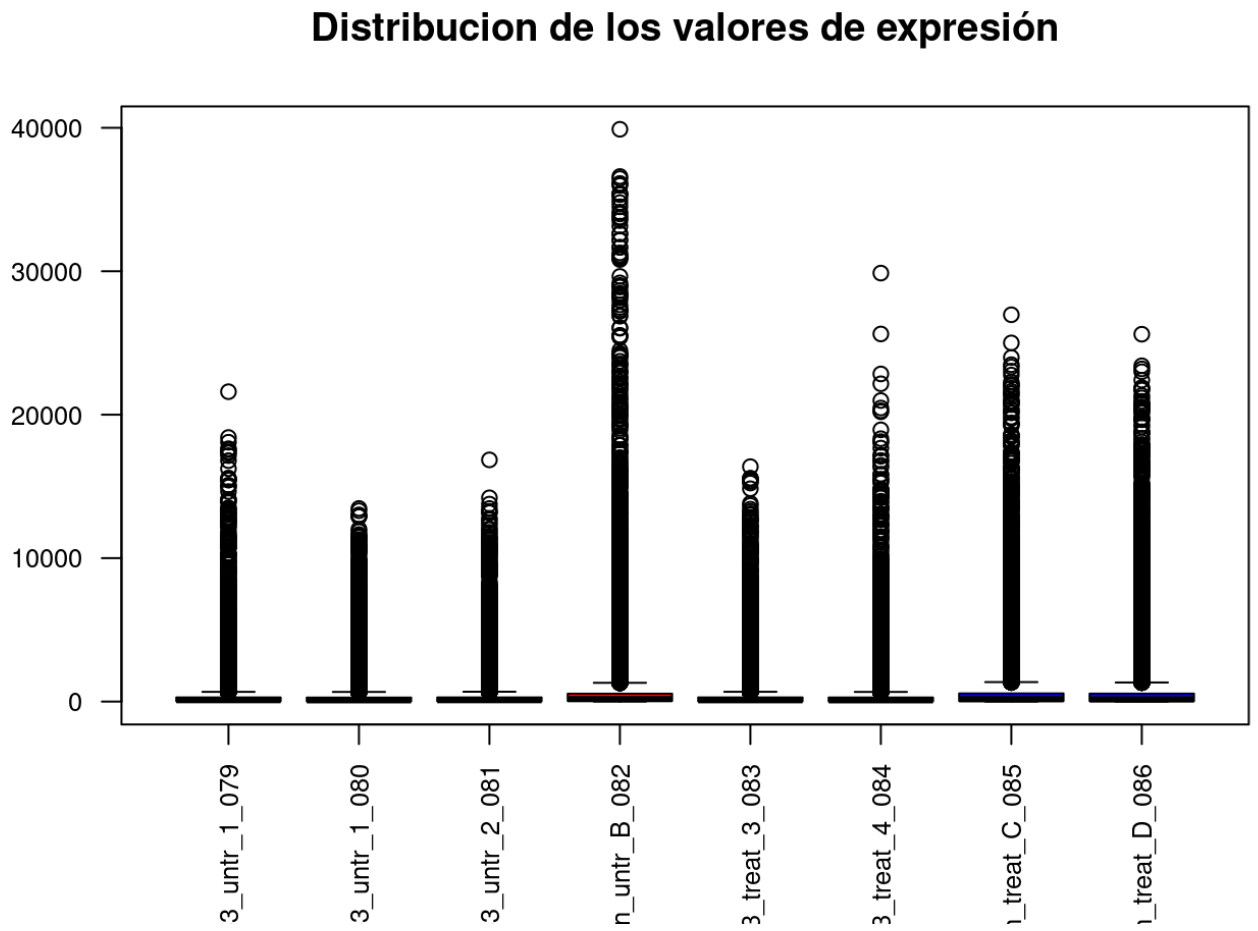
Code

A pesar de que parece más intuitivo trabajar con la información extraída del objeto de clase "GDS", en lo que resta del ejercicio se utilizara la información extraída de la lista generada al descargar el objeto desde el código de serie ("GSExxxx"). Obviamente ambos dan lugar a los mismos resultados.

3.2 Análisis exploratorio de los datos

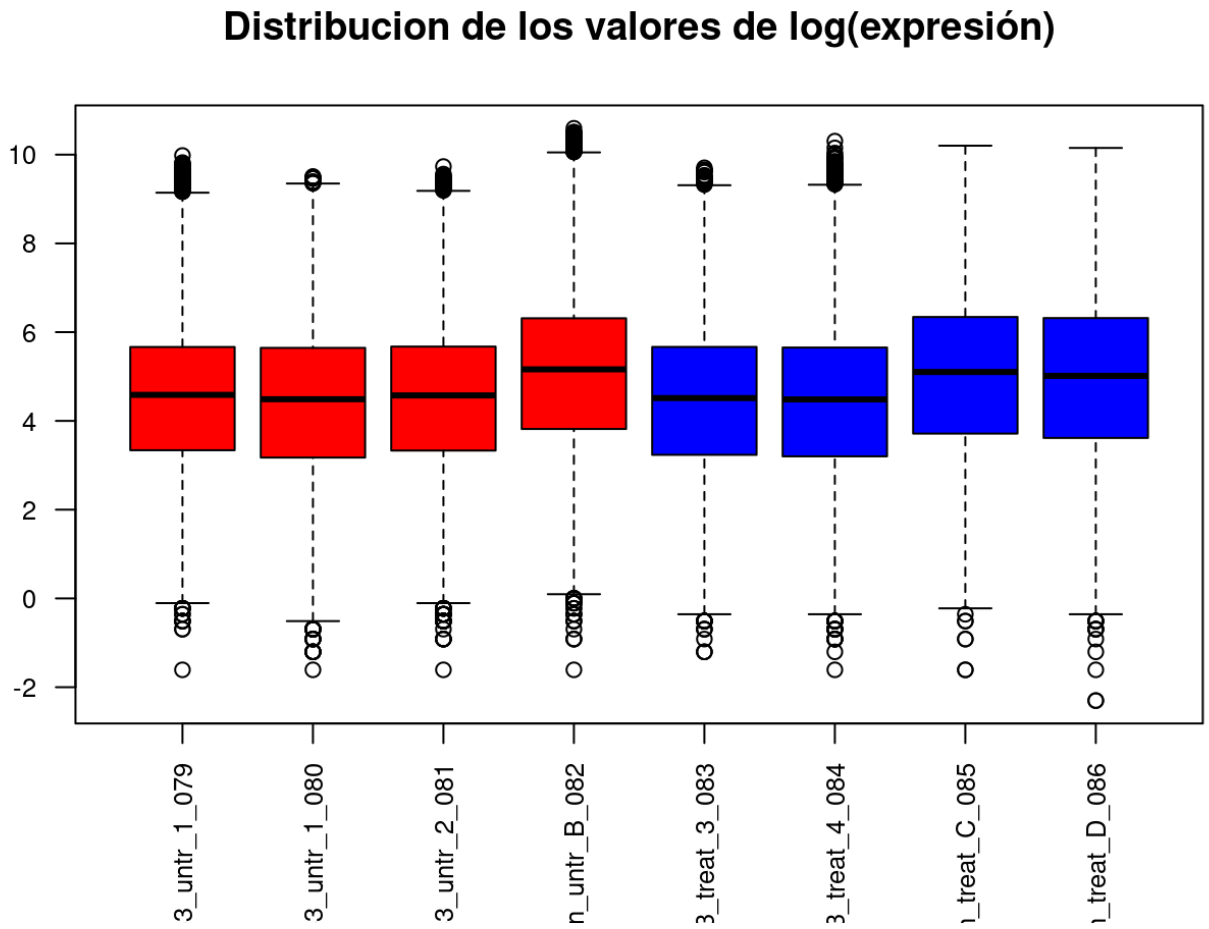
Una vez extraídos los datos y la información podemos proceder a realizar una exploración básica, similar a la del caso de estudio.

Code



Los datos son claramente asimétricos, lo que sugiere que puede tener sentido trabajar con los mismos datos en escala logarítmica.

Code



Claramente, a la vista del segundo gráfico, **es mejor trabajar con los datos transformados logarítmicamente**

3.2.1 Exploración multivariante

Un análisis en componentes principales puede facilitar la visualización de los datos en dimensión reducida y, sobretodo, detectar posibles patrones que no se detecten a simple vista.

El PCA transforma las variables originales de forma que las nuevas componentes (las variables transformadas) resultan tener dos propiedades muy interesantes: - Son independientes entre ellas, es decir explican propiedades distintas de los datos). - Cada componente explica un porcentaje de variabilidad mayor que la anterior, con lo que suele bastar con las dos o tres primeras componentes para obtener una visualización de los datos en dimensión reducida.

En primer lugar se realiza el cálculo de las componentes principales:

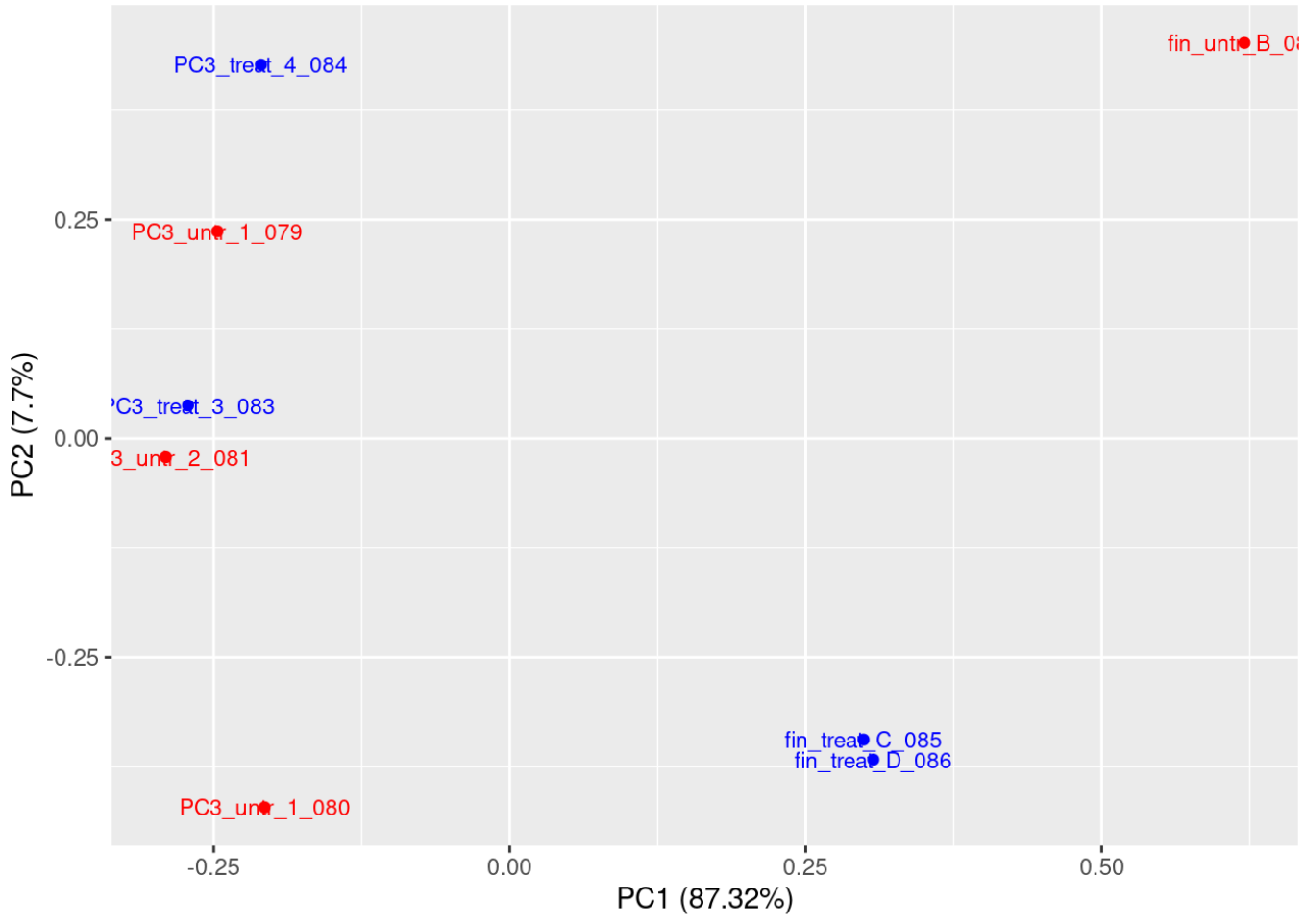
Code

```
Importance of components:
                            PC1       PC2       PC3       PC4       PC5
Standard deviation     9.607e+04 2.852e+04 1.760e+04 8.858e+03 8.195e+03
Proportion of Variance 8.732e-01 7.695e-02 2.929e-02 7.420e-03 6.350e-03
Cumulative Proportion  8.732e-01 9.502e-01 9.795e-01 9.869e-01 9.933e-01
                            PC6       PC7       PC8
Standard deviation     6.068e+03 5.864e+03 1.225e-10
Proportion of Variance 3.480e-03 3.250e-03 0.000e+00
Cumulative Proportion  9.968e-01 1.000e+00 1.000e+00
```

Las dos primeras componentes explican un 94% de la variabilidad de los datos, con lo que no precisaremos de ninguna otra.

Para la visualización de los resultados podemos utilizar el paquete `ggfortify` que genera un gráfico en `ggplot` en función del objeto que se le proporcione.

Code



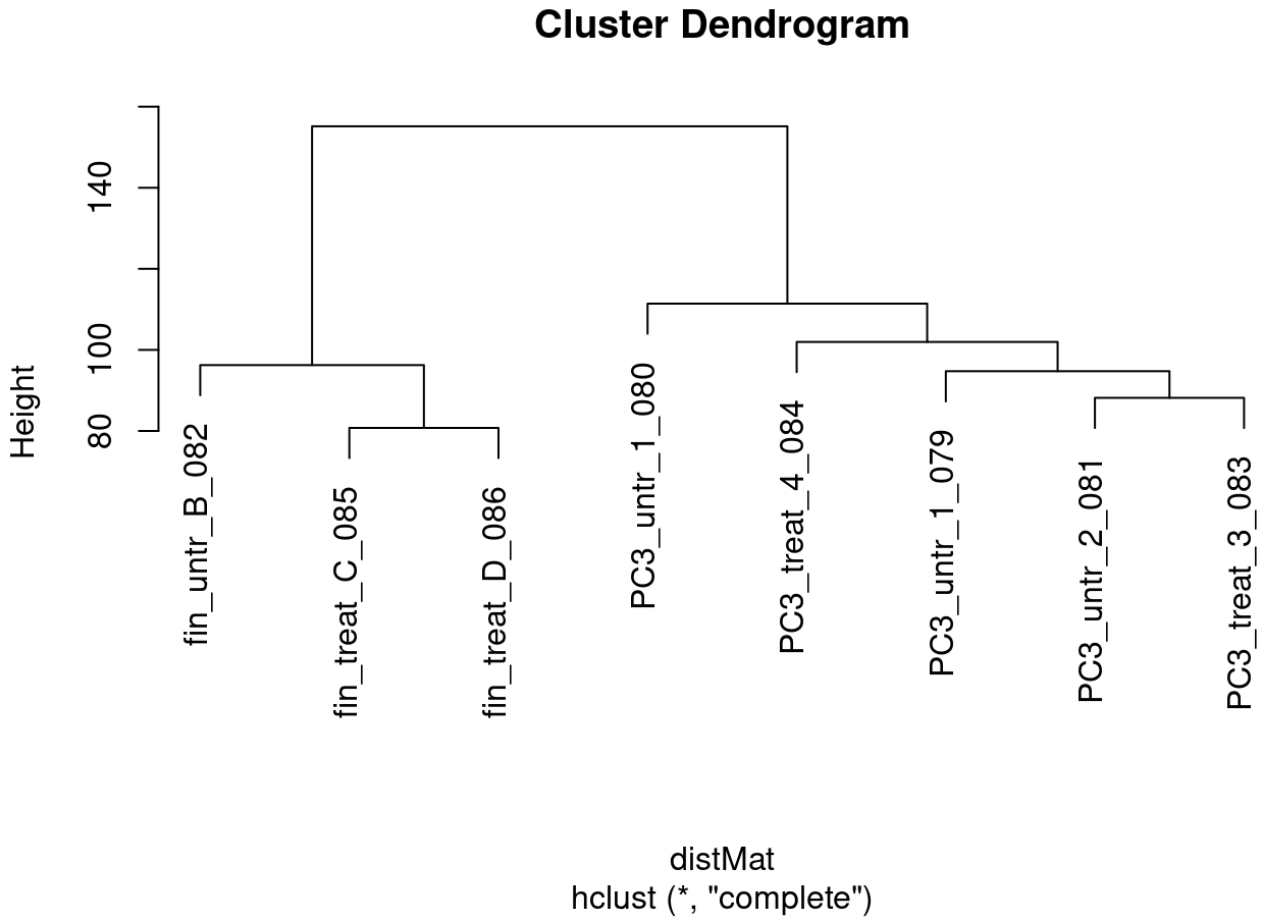
El resultado muestra que, en contra de lo que uno podría esperar, lo que diferencia dos grupos es la característica “fin/PC3”, que se encuentra claramente asociada con la primera componente.

La diferencia entre tratados y no tratados no se asocia, al menos en el grupo PC3, con la segunda.

El diseño, además está desbalanceado, con lo que, en caso de querer realizar un estudio comparativo con los dos factores no se podría realizar en condiciones óptimas.

Esta visualización se confirma al realizar un cluster jerárquico y visualizarlo con un dendrograma

Code



3.3 Discusión

El análisis exploratorio, realizado tras tomar logaritmo sobre los datos, ha puesto de manifiesto que existen dos fuentes de variación distintas. La mayor, relacionada con el origen de las muestras, no estaba especificada en el diseño del estudio. El efecto del tratamiento no aparece como la principal fuente de variabilidad lo que puede, probablemente, atribuirse a la confusión entre ambas fuentes.

Aparte de este problema inesperado, los datos no presentan evidencias de problemas: las distribuciones de las muestras son similares y no hay valores faltantes o cero, lo que se habría evidenciado al tomar logaritmos, por lo que en una situación real se procedería a investigar si es posible eliminar lo que aparenta ser un efecto batch y llevar a cabo el análisis.

4 Apéndice: Datasets

GEO Dataset	GEO Serie	Especie	Título
GDS1251	GSE2401	Rattus norvegicus	Acute hypotension effect on kidneys
GDS1736	GSE3737	Homo sapiens	Arachidonic acid effect on prostate cancer cells

GEO Dataset	GEO Serie	Especie	Título
GDS2107	GSE3311	Rattus norvegicus	Long-term ethanol consumption effect on pancreas
GDS2153	GSE5370	Homo sapiens	Dermatomyositis
GDS2294	GSE5583	Mus musculus	Histone deacetylase 1 deficient embryonic stem cells
GDS2300	GSE5668	Mus musculus	Germinal vesicle stage and metaphase II stage oocyte comparison
GDS2406	GSE6077	Mus musculus	Proto-oncogene Nmyc overexpression effect on the embryonic lung
GDS2629	GSE7381	Mus musculus	Epithelial transcription factor Get-1 deficiency effect on embryonic skin
GDS2637	GSE6299	Rattus norvegicus	Keratinocyte cell line response to a (56)Fe ion beam
GDS2646	GSE6868	Gallus gallus	Homocysteine effect on cardiac neural crest cells
GDS2648	GSE6766	Mus musculus	Palmitate effect on myoblast cell line
GDS2686	GSE6376	Mus musculus	MyD88 deficient macrophage response to zymosan
GDS2698	GSE6461	Mus musculus	Synovial sarcoma model
GDS2699	GSE6383	Mus musculus	Mesenchymal and epithelial compartments of the developing intestine
GDS2707	GSE4936	Mus musculus	Neuralized embryoid bodies response to Hedgehog agonist
GDS2727	GSE7196	Mus musculus	Estrogen-related receptor alpha deficiency effect on the heart
GDS2744	GSE7765	Homo sapiens	Dioxin effect on breast cancer cell line (HG-U133A)

GEO Dataset	GEO Serie	Especie	Título
GDS2766	GSE6850	Mus musculus	Dominant negative cJun effect on apolipoprotein E deficient livers
GDS2855	GSE3307	Homo sapiens	Various muscle diseases (HG-U133B)
GDS2917	GSE4734	Mus musculus	Various brain regions of several inbred strains
GDS2922	GSE5180	Homo sapiens	Ascending aortic aneurysms
GDS2935	GSE6281	Homo sapiens	Allergic contact dermatitis: time course
GDS2936	GSE8972	Mus musculus	Neural retina leucine zipper deficiency effect on retinas: time course
GDS4843	GSE42806	Homo sapiens	Skeletal muscle disuse atrophy
GDS3221	GSE6297	Mus musculus	Liver response to human and chimpanzee diets
GDS3223	GSE8853	Homo sapiens	Interleukin-13 effect on esophageal epithelial cells from eosinophilic esophagitis patients
GDS781	GSE1746	Homo sapiens	CD14 cells from granulocyte colony stimulating factor mobilized peripheral blood: expression profile
GDS885	GSE1417	Homo sapiens	Tumor cell response to topoisomerase poison camptothecin
GDS2613	GSE6955	Homo sapiens	Rett syndrome: brain frontal cortex

5 Apendice: Código R

Para extraer el código R se ha utilizado la instrucción `knitr::purl("UOC-MU-AD0-2022-23-S1-PEC1-Solucion.Rmd")` *que ha creado el archivo "UOC-MU-AD0-2022-23-S1-PEC1-Solución.R". Dicho archivo se incluye automáticamente en un ultimo "chiunnk" de código. Para evitar que se ejecute, se asigna el valor FALSE a la opción "eval" del código*

Code

1. La fecha de entrega es la que se indica en el enunciado de la PEC. En caso de no coincidir con la indicada en el aula, ésta será la que predomine.↩