Análisis de Datos Ómicos - PEC1

Juan A. Martínez Velasco

Mayo 2020

Tabla de Contenidos

Resumen

Este informe presenta los resultados del estudio de los datos de microarrays obtenidos por el Ovarian Cancer Institute (Atlanta) a partir de pacientes de cáncer de ovario tratados con quimioterapia. El informe ha sido organizado de acuerdo con el procedimiento a seguir en un estudio de microarrays de un solo color. Los datos del estudio se encuentran en la base de datos Gene Expression Omnibus (GEO) con el código GSE7463. El resumen del estudio tal como aparece en **GEO** es el siguiente:

Gene expression profiles of malignant carcinomas surgically removed from ovarian cancer patients pre-treated with chemotherapy (neo-adjuvant) prior to surgery group into two distinct clusters. One group clusters with carcinomas from patients not pre-treated with chemotherapy prior to surgery (C-L) while the other clusters with non-malignant adenomas (A-L). Although the C-L cluster is preferentially associated with p53 loss-of-function (LOF) mutations, the C-L cluster cancer patients display a more favorable clinical response to chemotherapy as evidenced by enhanced long-term survivorships. A set of 43 ovarian tumors was obtained from the Ovarian Cancer Institute (Atlanta). Tissue was collected at the time of surgery and preserved in RNAlater (Ambion, Austin, TX) within one minute of collection. Labeled probe was hybridized to the Affymetrix HG-U95Av2 arrays.

Introducción

El Ovarian Cancer Institute (Atlanta) realizó un estudio de expresión de genes de carcinomas malignos extraídos con cirugía en pacientes de cáncer de ovario y tratamiento con quimioterapia antes de la cirugía. El estudio clasificó las pacientes antes de realizar cirugía en dos grupos: (i) el primero agrupaba las pacientes con carcinoma y sin tratamiento con quimioterapia antes de la cirugía (grupo C-L); (ii) el segundo agrupaba los pacientes con adenoma no maligno (A-L). El objetivo del estudio era la obtención de perfiles de expresión de genes. El trabajo fue presentado inicialmente en la referencia [1], y posteriormente ampliado en la referencia [2].

Este informe detalla el estudio realizado para obtener genes expresados diferencialmente y su significación biológica a partir de los datos obtenidos en un ensayo con microarrays y disponibles en **GEO**.

El procedimiento seguido consta de varios pasos comunes a la mayoría de estudios de expresión de genes con datos obtenidos en experimentos con microarrays. Para un

estudio detallado, ver referencia [3]. Este trabajo se ha basado fundamentalmente en los procedimientos implantados en [4] y [5].

La Figura 1 muestra un resumen del procedimiento seguido desde la descarga de los datos hasta la obtención de un significado biológico de esos datos.

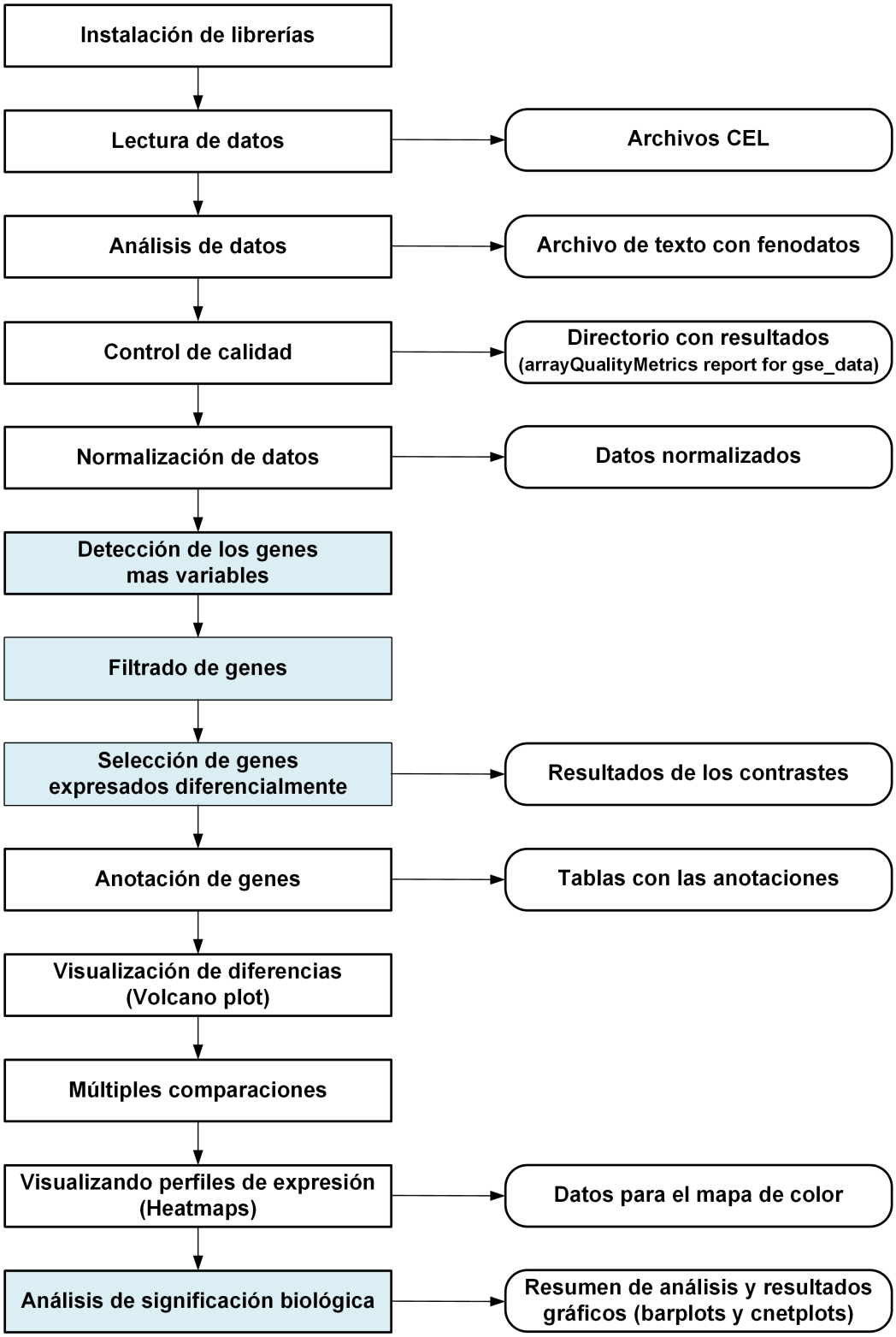

Figura 1. Procedimiento para análisis de datos de microarrays.

El procedimiento de la figura se puede resumir en los siguientes pasos:

1 - Instalación de librerías (**R** y **Bioconductor**).

2 - Lectura de datos con **GEOquery**.

3 - Interpretación y preparación de los datos (obtención de los archivos CEL).

4 - Control de calidad de los datos originales.

5 - Normalización de los datos.

6 - Filtrado de los datos.

7 - Selección de genes expresados diferencialmente.

8 - Anotación de genes.

9 - Visualización de las diferencias (*volcano plots*).

10 - Comparaciones.

11 - Visualización de los perfiles de expresión (uso de *heatmaps*).

14 - Significación biológica de los resultados.

NOTAS: En la elaboración de este informe se han teniendo en cuenta los siguientes aspectos:

a - El código empleado (con instrucciones de **R** y **Bioconductor**) está oculto.

b - Se muestran solo los resultados seleccionados. En el caso de las figuras se ha añadido título, que aparecerá numerado de forma consecutiva debajo de la figura.

c - Las referencias han sido numeradas de forma consecutiva conforme van siendo mencionadas en el informe. La lista completa aparece al final del informe.

d - El código empleado en la manipulación de datos y la realización de cálculos se presenta en un archivo adjunto.

Instalación de librerías

La instalación de las librerías que se han de utilizar en la preparación de este informe es el primer paso. La lista incluye las librerías de **Bioconductor** que se han empleado en este trabajo (p.e. **Biobase**, **GEOquery**, **limma**, etc) más algunas de **R** que se han necesitado para realizar otras tareas (p.e. **MVA**).

Lectura de datos

El repositorio GEO y la librería GEOquery

Organización de datos en un ExpressionSet

Gene Expression Omnibus (GEO) es un repositorio público que archiva y distribuye datos de expresión de genes de alto rendimiento. Los datos están organizados en estructuras normalizadas en clases de tipo *ExpressionSet*, que combina información de varios de fuentes y clases en una única estructura: un archivo *ExpressionSet* puede ser copiado, modificado, o servir de entrada/salida para muchas funciones implantadas en **Bioconductor**.

La Figura 2 muestra cómo se organiza un archivo de clase *ExpressionSet*. Se puede observar que esta clase de archivos combina los siguientes tipos de datos:

- *assayData*: Datos de expresión obtenidos mediante ensayos con microarrays; los datos de las sondas aparecen en las filas y los identificadores de las muestras en las columnas.
- *phenoData*: Combina los resultados del ensayo con los identificadores de las muestras en las filas y la información que describe el ensayo en las columnas.
- *featureData*: Datos sobre las características de la tecnología empleada en el experimento con las mismas filas que el contenedor assayData mientras que las columnas pueden mostrar diversos tipos de anotaciones (por ejemplo, anotaciones de genes extraídas de bases de datos biomédicas).

A esta lista hay que añadir otros tipos de datos (no mostrados en la figura):

- *experimentData*: Una estructura empleada para describir el experimento.
- *protocolData*: Información generada por el equipo sobre protocolos.
- *annotation*: Identificación de la plataforma con la que se realizó el ensayo.

Para más información sobre estos aspectos ver [6] y [7].

Aplicación de GEOquery

La descarga de datos desde GEO se realiza con la librería **GEOquery** y la función `getGEO`. Esta función analiza el formato del archivo a descargar para determinar cómo realizar la descarga de los datos y obtener una estructura que pueda interpretar **R**.

En este trabajo se descarga un archivo de datos con formato *GSE* y código GSE7463. Los datos del correspondiente ensayo fueron inicialmente cargados en **GEO** con fecha de 5 de abril de 2007; la última actualización fue realizada con fecha de 7 de junio de 2019.

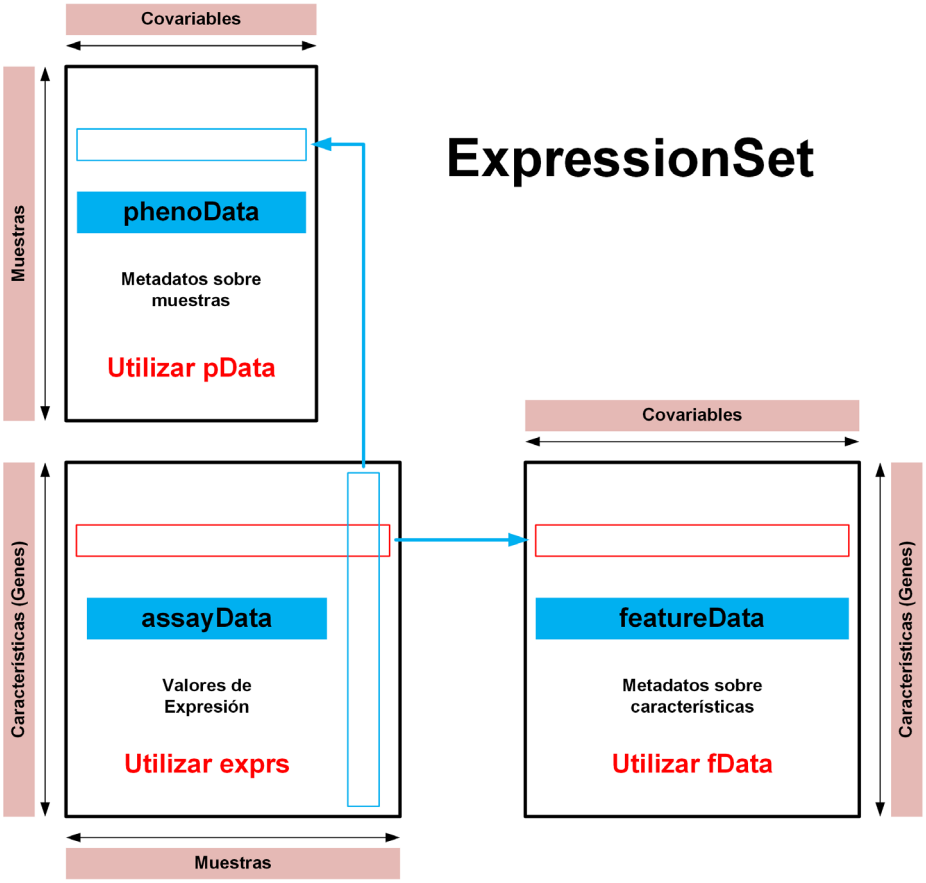

Figura 2. Organización de datos con expressioSet.

Análisis de los datos

Una vez descargado el archivo, se analiza su contenido; por ejemplo:

- el nombre del archivo descargado (en formato comprimido y con extensión *gz*)

```
[1] "GSE7463_series_matrix.txt.gz"
```

- el tipo de datos descargados

```
[1] "list"
```

- el tamaño o longitud

```
[1] 1
```

El archivo descargado contiene un único objeto cuya estructura se analizará con más detalle en los siguientes pasos. Como precaución se realiza una copia del objeto (no hay que confundir el objeto con el archivo general descargado).

Ahora se averigua la clase de datos que se van a estudiar y manipular.

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

Se presenta los primeros datos con la opción `head`.

```
ExpressionSet (storageMode: lockedEnvironment)
assayData: 1 features, 43 samples
  element names: exprs
protocolData: none
phenoData
  sampleNames: GSM180626 GSM180627 ... GSM180668 (43 total)
  varLabels: title geo_accession ... Stage:ch1 (41 total)
  varMetadata: labelDescription
featureData: none
experimentData: use 'experimentData(object)'
  pubMedIds: 17505532
19077237
Annotation: GPL8300
```

Se observa que puede haber los 6 tipos de datos mencionados anteriormente: assayData, protocolData, phenoData, featureData, experimentData, y Annotation. Por otro lado, se comprueba que en el experimento en estudio no se ha incluido información sobre protocolData y featureData.

Se empieza analizando experimentData.

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

Se trata de un objeto de clase *MIAME* (Minimum information about a microarray experiment), que es una norma creada para presentar resultados de experimentos con microarrays; ver [8].

Length	Class	Mode
1	MIAME	S4

El objeto está editado en modo S4, el utilizado actualmente en **Bioconductor** para programación orientada a objetos; ver [9].

Se puede ver que la información que hay en experimentData hace referencia al laboratorio o institución donde se realizó el ensayo, datos de la persona de contacto, abstract del trabajo presentado y basado en el experimento, tecnología empleada en el experimento, dirección del repositorio donde se hallan depositados los datos, etc. Se trata básicamente de información sobre el experimento, y que no será utilizada en el estudio posterior.

Se continua con los phenoData. Para examinar esta información, primero, se comprueba la dimensión. Para ello se utiliza la función `pData` que permite acceder a los fenodatos del dataset descargado y la opción `dim`.

```
Dimensiones de los datos

[1] 43 41
```

Se trata de una base de datos con 43 archivos que contienen las muestras (filas) y 41 variables (columnas) en cada uno de los archivos de las muestras.

Se utilizan las opciones

- `rownames` para saber el nombre de los archivos *GSM* comprimidos dentro del archivo *gz* descargado

```
 [1] "GSM180626" "GSM180627" "GSM180628" "GSM180629" "GSM180630" "GSM180631"
 [7] "GSM180632" "GSM180633" "GSM180634" "GSM180635" "GSM180636" "GSM180637"
[13] "GSM180638" "GSM180639" "GSM180640" "GSM180641" "GSM180642" "GSM180643"
[19] "GSM180644" "GSM180645" "GSM180646" "GSM180647" "GSM180648" "GSM180649"
[25] "GSM180650" "GSM180651" "GSM180652" "GSM180653" "GSM180654" "GSM180655"
[31] "GSM180656" "GSM180657" "GSM180658" "GSM180659" "GSM180660" "GSM180661"
[37] "GSM180662" "GSM180663" "GSM180664" "GSM180665" "GSM180666" "GSM180667"
[43] "GSM180668"
```

- y `colnames` para saber cómo está organizada la información disponible en cada archivo

```
 [1] "title"                   "geo_accession"
 [3] "status"                  "submission_date"
 [5] "last_update_date"        "type"
 [7] "channel_count"           "source_name_ch1"
 [9] "organism_ch1"            "characteristics_ch1"
[11] "characteristics_ch1.1"   "characteristics_ch1.2"
[13] "characteristics_ch1.3"   "treatment_protocol_ch1"
[15] "growth_protocol_ch1"     "molecule_ch1"
[17] "extract_protocol_ch1"    "label_ch1"
[19] "label_protocol_ch1"      "taxid_ch1"
[21] "hyb_protocol"            "scan_protocol"
[23] "description"             "data_processing"
[25] "platform_id"             "contact_name"
[27] "contact_email"           "contact_phone"
[29] "contact_laboratory"      "contact_department"
[31] "contact_institute"       "contact_address"
[33] "contact_city"            "contact_state"
[35] "contact_zip/postal_code" "contact_country"
[37] "supplementary_file"      "data_row_count"
[39] "Age at surgery:ch1"      "Histology:ch1"
[41] "Stage:ch1"
```

Se puede comprobar que existen 43 archivos (el nombre de las filas) tipo *GSM* y 41 variables (columnas) en cada archivo.

Toda la información está disponible si se tienen en cuenta las dimensiones y se solicita bien toda o una parte. Si se pide la información disponible en las columnas 1 (*title*) y 6 (*type*) de los archivos numerados como 5, 18 y 30 se obtiene lo siguiente:

	title	type
GSM180630	Adenoma patient 172	RNA
GSM180643	Carcinoma patient 66	RNA
GSM180655	Cancer Chemo patient 286	RNA

La información contenida en cualquier variable se puede obtener utilizando el nombre de las variables:

- data_processing (archivo GSM 1)

```
[1] The data were analyzed using GCRMA normalization with GeneTraffic Soft
ware (Iobion, La Jolla, CA).
Levels: The data were analyzed using GCRMA normalization with GeneTraffic
Software (Iobion, La Jolla, CA).
```

- description (Archivo GSM 11)

```
[1] Gene expression data from ovarian carcinoma.
3 Levels: Gene expression data from ovarian adenoma. ...
```

- scan_protocol (Archivo GSM 33)

```
[1] GeneChips were scanned using GeneArray Scanner Model G2500
Levels: GeneChips were scanned using GeneArray Scanner Model G2500
```

Toda esta información también se puede obtener indicando el nombre de la variable. Por ejemplo, si se consultan las variables *title* y *description* se obtiene

	title	description
GSM180626	Adenoma patient 125	Gene expression data from ovarian adenoma.
GSM180627	Adenoma patient 132	Gene expression data from ovarian adenoma.
GSM180628	Adenoma patient 146	Gene expression data from ovarian adenoma.
GSM180629	Adenoma patient 159	Gene expression data from ovarian adenoma.
GSM180630	Adenoma patient 172	Gene expression data from ovarian adenoma.

Datos de la expresión de los genes

En el siguiente paso se obtienen los valores numéricos que serán manipulados en el resto de este informe. Para ello se utiliza la opción `exprs`. Una información interesante que se puede derivar de este paso es la dimensión de la matriz de datos utilizando la opción `dim`. Con el presente estudio se tiene

```
[1] 12625    43
```

Es decir, los datos del ensayo incluyen la expresión de 12625 genes en cada una de las 43 muestras.

Dado el tamaño de esta información parece razonable que se analice (o se pueda analizar) por partes. Por ejemplo, los datos correspondientes a las 5 primeras filas (genes) y las 6 primeras columnas (muestras) sería la siguiente:

	GSM180626	GSM180627	GSM180628	GSM180629	GSM180630
1000_at	7.361102	7.640852	7.624269	7.229713	7.339200
1001_at	2.930170	3.430165	3.021592	2.875247	3.005486
1002_f_at	4.544077	4.661328	4.633082	4.405026	4.581405
1003_s_at	2.184544	2.279213	2.227368	2.073004	2.207774
1004_at	4.855048	4.974924	4.926967	4.690261	4.923349
1005_at	8.846703	5.968943	8.173467	7.420314	6.338455

También se puede obtener alguna estadística de las muestras. Por ejemplo, la correspondiente a las 3 primeras y las 2 últimas sería la siguiente:

	GSM180626	GSM180627	GSM180628	GSM180667	GSM180668
Min.	1.000000	1.000000	1.000000	1.000000	1.000000
1st Qu.	3.299012	3.427553	3.374706	3.773406	3.430084
Median	4.963983	4.950613	4.955331	5.096785	5.028503
Mean	5.379900	5.402553	5.391175	5.528004	5.463395
3rd Qu.	7.176073	7.144067	7.157236	7.009370	7.235717
Max.	14.960898	15.076636	14.986980	14.974531	15.110765

Obtención de archivos CEL

La información de cada muestra está contenida en archivos tipo CEL; ver [10]. Estos archivos se encuentran comprimidos en la base de datos descargada inicialmente, y pudo ser consultada aplicando la función `exprs`. El siguiente paso es la descompresión de los archivos CEL.

Primero, se confirma la existencia del archivo GSE descargado.

```
[1] "GSE7463"                          "GSE7463_series_matrix.txt.gz"
```

A continuación, se averigua que datos están comprimidos en el archivo *gz*.

```
[1] "CEL"             "GSE7463_RAW.tar"
```

Finalmente, se extraen los archivos CEL y se obtiene una lista de los mismos.

```
 [1] "GSE7463_SelectPhenoData.txt" "GSM180626.CEL.gz"
 [3] "GSM180627.CEL.gz"            "GSM180628.CEL.gz"
 [5] "GSM180629.CEL.gz"            "GSM180630.CEL.gz"
 [7] "GSM180631.CEL.gz"            "GSM180632.CEL.gz"
 [9] "GSM180633.CEL.gz"            "GSM180634.CEL.gz"
[11] "GSM180635.CEL.gz"            "GSM180636.CEL.gz"
[13] "GSM180637.CEL.gz"            "GSM180638.CEL.gz"
[15] "GSM180639.CEL.gz"            "GSM180640.CEL.gz"
[17] "GSM180641.CEL.gz"            "GSM180642.CEL.gz"
[19] "GSM180643.CEL.gz"            "GSM180644.CEL.gz"
[21] "GSM180645.CEL.gz"            "GSM180646.CEL.gz"
[23] "GSM180647.CEL.gz"            "GSM180648.CEL.gz"
[25] "GSM180649.CEL.gz"            "GSM180650.CEL.gz"
[27] "GSM180651.CEL.gz"            "GSM180652.CEL.gz"
[29] "GSM180653.CEL.gz"            "GSM180654.CEL.gz"
[31] "GSM180655.CEL.gz"            "GSM180656.CEL.gz"
[33] "GSM180657.CEL.gz"            "GSM180658.CEL.gz"
[35] "GSM180659.CEL.gz"            "GSM180660.CEL.gz"
[37] "GSM180661.CEL.gz"            "GSM180662.CEL.gz"
[39] "GSM180663.CEL.gz"            "GSM180664.CEL.gz"
[41] "GSM180665.CEL.gz"            "GSM180666.CEL.gz"
[43] "GSM180667.CEL.gz"            "GSM180668.CEL.gz"
```

La lista se guarda en el archivo *my.cels*.

Se observa que los nombres concuerdan con los de los archivos GSM, que ya fueron listados anteriormente.

Control de calidad de los datos originales

Aplicación de la librería ArrayQualityMetrics

Un paso muy útil es el control de la calidad de los datos, ya que una mala calidad puede ser la causa de ruido o defectos que la posterior normalización no pueda solucionar. La librería *ArrayQualityMetrics* realiza distintos estudios de calidad que presenta en forma gráfica (en forma de diagrama de cajas o análisis de componentes principales - Principal Component Analysis, PCA).

Por ejemplo, si la muestra correspondiente a un determinado (micro)array está por encima de un umbral definido en la función se marca con un asterisco como posible dato atípico (outlier), ver Figura 3; si el array se marca tres veces debería ser revisado cuidadosamente y quizás ser eliminado para mejorar la calidad del ensayo.

Los resultados del control de calidad se guardan en el subdirectorio *arrayQualityMetrics report for gse_data* que crea de forma automática la misma librería.

La Figura 3 presenta uno de los gráficos creados por la librería *ArrayQualityMetrics*.

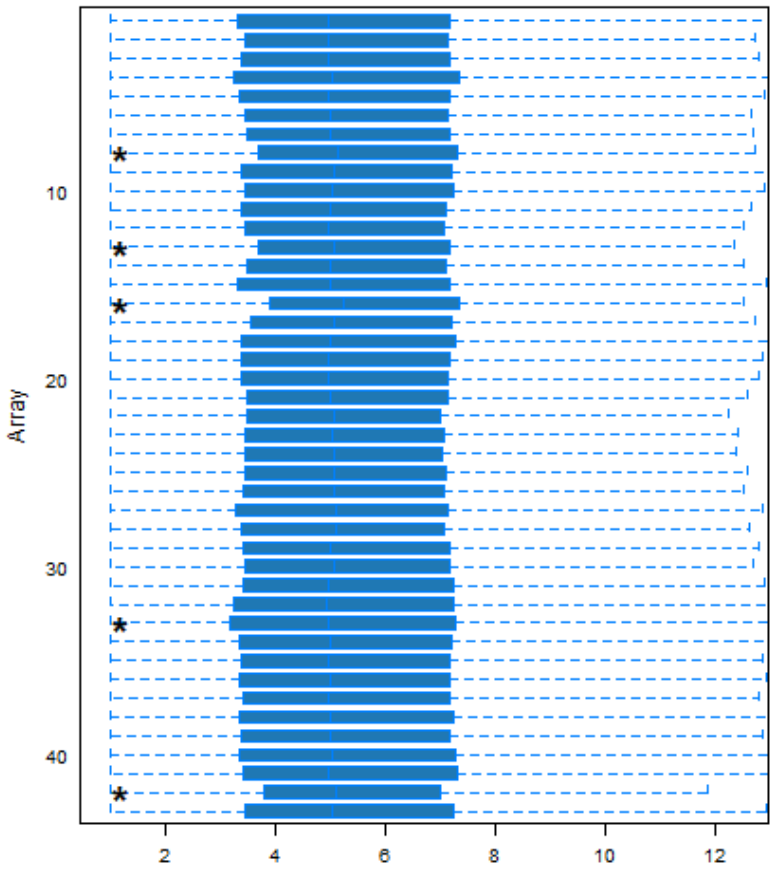

Figura 3. Control de calidad de los datos.

Análisis de componentes principales

Un análisis de componentes principales también puede ser útil para entender los resultados del ensayo con microarrays. En el análisis de control realizado con la librería

arrayQualityMetrics se presenta un primer resultado. Ahora se va a realizar otro análisis utilizando las prestaciones de la librería *MVA*; en concreto la función `princomp`.

Los resultados del análisis se pueden consultar utilizando varias opciones disponibles en **R**, como `str` o `summary`. Si se utiliza esta última se obtiene la siguiente salida:

```
Importance of components:
                          Comp.1     Comp.2      Comp.3     Comp.4      Comp.5
Standard deviation     6.2951114 1.05181235 0.577642033 0.50363177 0.407837189
Proportion of Variance 0.9215913 0.02572812 0.007759775 0.00589872 0.003868167
Cumulative Proportion  0.9215913 0.94731945 0.955079229 0.96097795 0.964846116
                            Comp.6      Comp.7      Comp.8      Comp.9
Standard deviation     0.354750830 0.344176655 0.327862164 0.316071318
Proportion of Variance 0.002926701 0.002754827 0.002499851 0.002323281
Cumulative Proportion  0.967772817 0.970527644 0.973027495 0.975350776
                           Comp.10     Comp.11     Comp.12     Comp.13
Standard deviation     0.306814344 0.285053960 0.263543122 0.252748141
Proportion of Variance 0.002189187 0.001889669 0.001615232 0.001485619
Cumulative Proportion  0.977539963 0.979429632 0.981044864 0.982530483
                           Comp.14     Comp.15     Comp.16     Comp.17
Standard deviation     0.243869265 0.240865633 0.230263288 0.216627568
Proportion of Variance 0.001383075 0.001349215 0.001233051 0.001091337
Cumulative Proportion  0.983913558 0.985262773 0.986495824 0.987587161
                           Comp.18      Comp.19      Comp.20     Comp.21
Standard deviation     0.207898212 0.2011728141 0.1994905165 0.192328569
Proportion of Variance 0.001005155 0.0009411744 0.0009254992 0.000860239
Cumulative Proportion  0.988592316 0.9895334908 0.9904589900 0.991319229
                            Comp.22      Comp.23      Comp.24      Comp.25
Standard deviation     0.1840896086 0.1831638874 0.1760793951 0.1698322703
Proportion of Variance 0.0007881159 0.0007802095 0.0007210222 0.0006707674
Cumulative Proportion  0.9921073450 0.9928875545 0.9936085766 0.9942793441
                            Comp.26    Comp.27      Comp.28      Comp.29
Standard deviation     0.1627972540 0.15139591 0.1487472546 0.1430190418
Proportion of Variance 0.0006163476 0.00053304 0.0005145522 0.0004756848
Cumulative Proportion  0.9948956917 0.99542873 0.9959432839 0.9964189687
                            Comp.30      Comp.31      Comp.32      Comp.33
Standard deviation     0.1358616615 0.1234635386 0.1221676277 0.1202991711
Proportion of Variance 0.0004292649 0.0003544941 0.0003470914 0.0003365556
Cumulative Proportion  0.9968482337 0.9972027277 0.9975498191 0.9978863747
                            Comp.34      Comp.35      Comp.36      Comp.37
Standard deviation     0.1189355706 0.1135338857 0.1013803885 0.0986842068
Proportion of Variance 0.0003289691 0.0002997661 0.0002390229 0.0002264784
Cumulative Proportion  0.9982153438 0.9985151099 0.9987541328 0.9989806112
                            Comp.38      Comp.39      Comp.40      Comp.41
Standard deviation     0.0985271011 0.0943600802 0.0846687355 0.0820276387
Proportion of Variance 0.0002257579 0.0002070657 0.0001667162 0.0001564775
Cumulative Proportion  0.9992063691 0.9994134348 0.9995801509 0.9997366285
                            Comp.42      Comp.43
Standard deviation     0.0773091842 0.0731318411
Proportion of Variance 0.0001389933 0.0001243783
Cumulative Proportion  0.9998756217 1.0000000000
```

Se observa que la primera componente puede servir para representar más del 90% de la variabilidad. Esto se puede entender mejor si se realiza el cálculo de valores propios de la matriz de varianzas-covarianzas.

```
Valores propios de la matriz de datos
```

```
 [1] 246.06964104   6.84579538   2.02534779   1.52236258   1.02310524
 [6]   0.75954401   0.73068309   0.66296490   0.62922078   0.57110817
[11]   0.50003972   0.42510659   0.39269786   0.36754365   0.35061911
[16]   0.32890178   0.28575465   0.27245158   0.24914948   0.24332723
[21]   0.23330510   0.21490799   0.21194886   0.19271352   0.18121863
[26]   0.16113850   0.14148317   0.14064481   0.12544379   0.11149130
[31]   0.09669941   0.09264560   0.09204100   0.08895086   0.08145206
[36]   0.06511296   0.06124224   0.05982100   0.05555086   0.04441951
[41]   0.04107385   0.03804823   0.03360229
```

Se confirma que el primer valor es mucho más grande que los restantes valores.

Para obtener información del análisis de componentes principales se puede utilizar la opción `names`, que proporciona una lista de los distintos resultados obtenidos con el análisis. No es igual la información que proporciona el resultado del análisis que la que proporciona el resumen (`summary`) del análisis; la lista de objetos es distinta.

```
Información del análisis de componentes principales

[1] "sdev"     "loadings" "center"   "scale"    "n.obs"    "scores"   "cal
l"

Información del resumen

[1] "sdev"           "loadings"       "center"         "scale"
[5] "n.obs"          "scores"         "call"           "cutoff"
[9] "print.loadings"
```

Si se piden los primeros resultados de las 6 primeras componentes (*loadings*) se obtiene la siguiente salida:

```
Primeras componentes-Primeras muestras
```

	Comp.1	Comp.2	Comp.3	Comp.4	Comp.5	Comp.6
GSM180626	0.155	0.182	0.009	0.019	0.101	0.002
GSM180627	0.155	0.167	-0.049	0.044	0.056	0.122
GSM180628	0.155	0.176	-0.038	0.053	0.043	0.057
GSM180629	0.154	0.135	0.082	-0.151	-0.158	0.094
GSM180630	0.155	0.141	0.003	-0.036	-0.066	0.045

Del análisis de componentes principales se pueden obtener varios resultados gráficos. La Figura 4 presenta los primeros componentes.

Se observa que el primer componente puede ser suficiente para explicar la variabilidad del ensayo en estudio.

Se visualizan los resultados con las dos primeras componentes aplicando un cálculo distinto; ahora se usa la función `prcomp`, sin escalado; ver Figura 5.

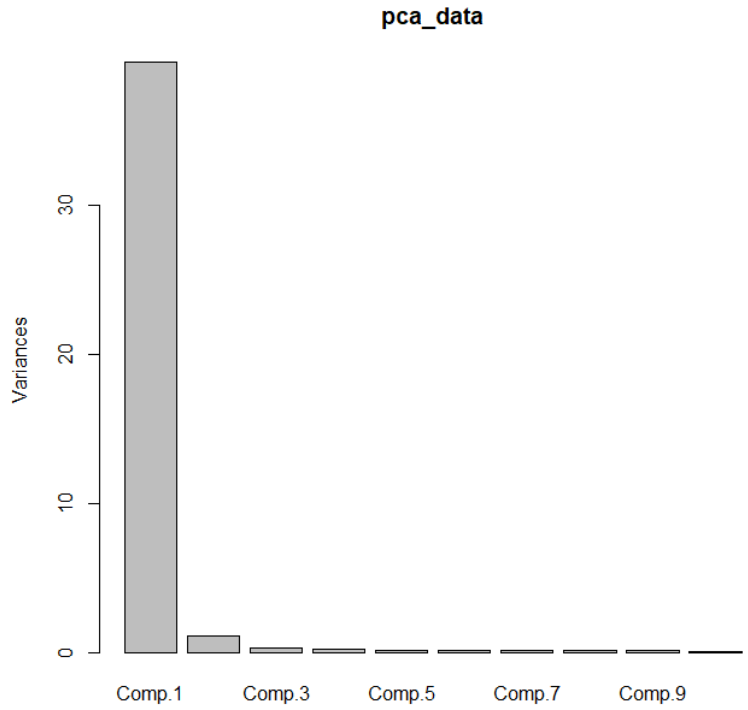

Figura 4. Análisis de componentes principales - 1.

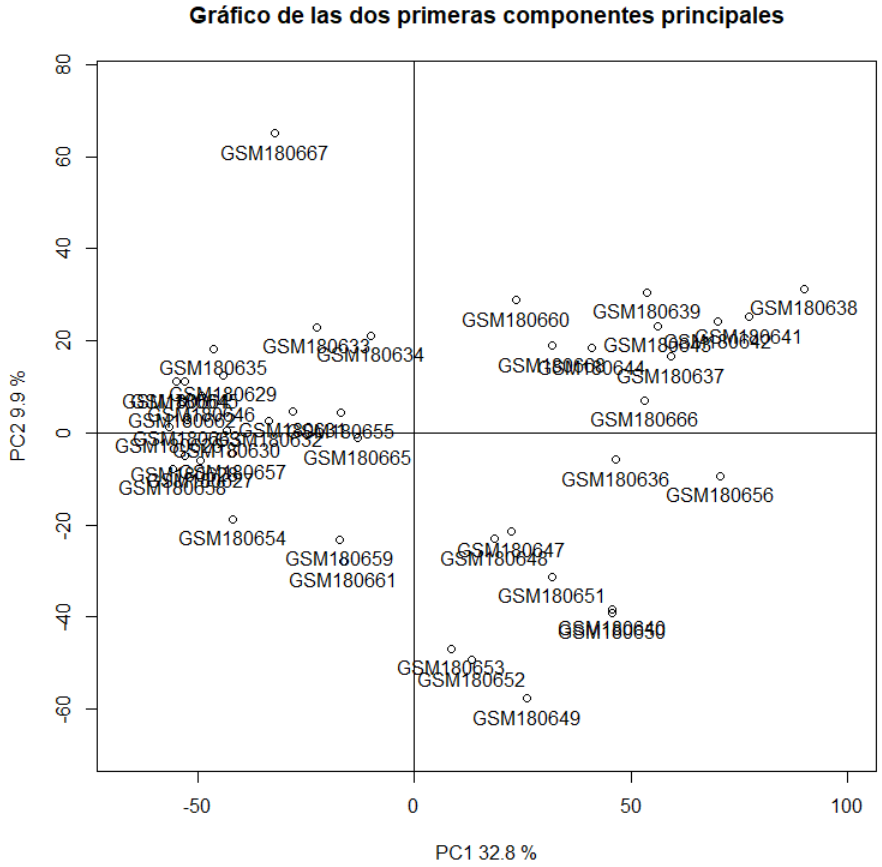

Figura 5. Análisis de componentes principales - 2.

Se observa que hay diferencias entre los dos resultados; con este nuevo estudio, las dos primeras componentes explican poco más del 60% de variabilidad.

Finalmente, se presenta un diagrama de cajas de todas las muestras antes de realizar la normalización, ver Figura 6.

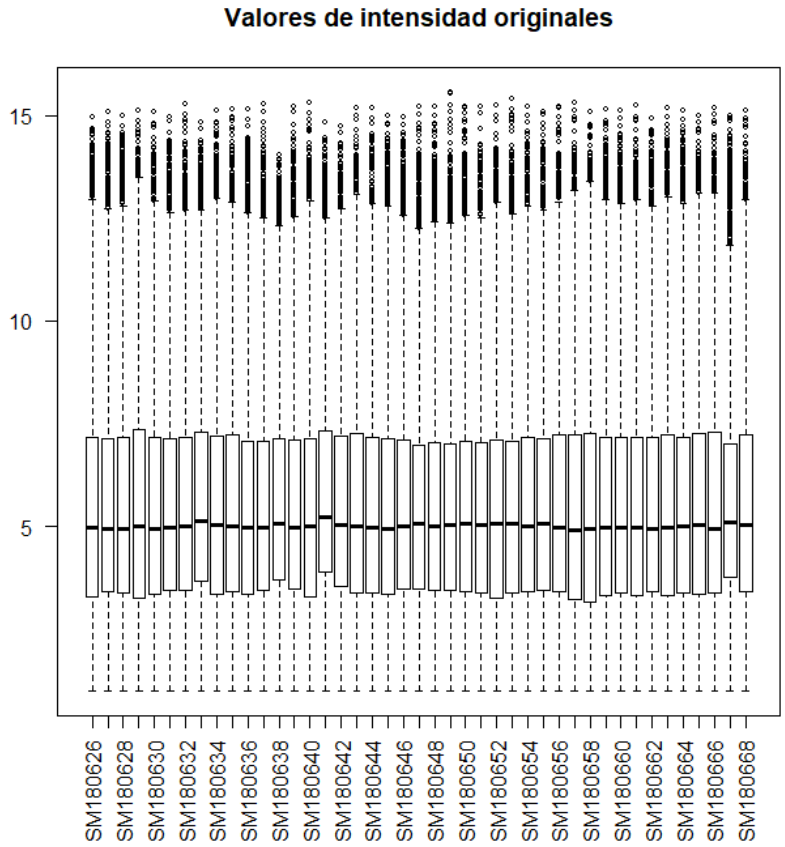

Figura 6. Diagrama de cajas con intensidades originales.

Este resultado se puede comparar con los diagramas de cajas que resultarán de aplicar la función `rma` con datos sin normalizar y normalizados.

Procesamiento de datos

Preparación de fenodatos - Archivos CEL

Se leen los archivos CEL, y se obtiene un resumen y el tipo de información que contienen (clase, dimensiones, etc). Este paso se realiza con la librería **affy** ya que los ensayos se han realizado con tecnología Affymetrix de un solo color.

Se genera un archivo en formato texto GSE7463_SelectPhenoData.txt con la información de las muestras (archivos *GSM*).

```
Datos de Affymetrix

AffyBatch object
size of arrays=640x640 features (79 kb)
cdf=HG_U95Av2 (12625 affyids)
number of samples=43
number of genes=12625
annotation=hgu95av2
notes=
```

Se presentan los nombres de las filas y columnas.

```
[1] "100_g_at"  "1000_at"   "1001_at"   "1002_f_at" "1003_s_at"

[1] "GSM180626" "GSM180627" "GSM180628" "GSM180629" "GSM180630"
```

Se comprueba, como ya se sabía, que hay 43 muestras de clase *AffyBatch*. El número de genes en cada array es 12625 y las anotaciones se han realizado con el formato `hgu95av2`. Esta será la dimensión de la matriz de datos obtenida con la opción `exprs`.

Las dimensiones son las mismas que se tenían con el archivo *gse_data*.

```
Dimensiones de los fenodatos

[1] 43 41
```

Se consultan los fenodatos disponibles en el archivo de clase AffyBatch con la función `pData`.

```
> class(pData(my.affy))

[1] "data.frame"
```

Se presenta el nombre de las filas y columnas.

```
> rownames(pData(my.affy))[1:5]

[1] "GSM180626" "GSM180627" "GSM180628" "GSM180629" "GSM180630"

> colnames(pData(my.affy))[1:5]

[1] "title"            "geo_accession"    "status"           "submission_d
ate"
[5] "last_update_date"
```

Los datos se pueden seleccionar o consultar indicando filas y columnas. Así, para las 6 primeras filas y las 3 primeras columnas, se tiene

En ese caso, puesto que se trata de un data frame, se puede obtener información por columnas indicando el nombre de la columna. Por ejemplo, con la columna *title* se obtiene el siguiente resultado (que muestra 5 filas de cada uno de los grupos):

```
 [1] "Adenoma patient 125"      "Adenoma patient 132"
 [3] "Adenoma patient 146"      "Adenoma patient 159"
 [5] "Adenoma patient 172"      "Carcinoma patient 183"
 [7] "Carcinoma patient 196"    "Carcinoma patient 2"
 [9] "Carcinoma patient 204"    "Carcinoma patient 212"
[11] "Cancer Chemo patient 150" "Cancer Chemo patient 184"
[13] "Cancer Chemo patient 187" "Cancer Chemo patient 199"
[15] "Cancer Chemo patient 253" "Cancer Chemo patient 255"
```

Agrupamiento de datos

Del último resultado se deduce que hay tres tipos de muestras que se pueden clasificar en los siguientes grupos: *Adenoma*, *Carcinoma*, *Quimio*.

```
Número de muestras - Grupo Adenoma    = 10
```

```
Número de muestras - Grupo Carcinoma = 9

Número de muestras - Grupo Quimio    = 24
```

Se crea una tabla con un nombre para cada grupo y el número de muestras.

Adenoma	Carcinoma	Quimio
10	9	24

Y otra tabla más completa en la que se presentan las muestras según esta clasificación. Se muestran las mismas filas que se mostraron anteriormente.

/	Adenoma	Carcinoma	Quimio
Adenoma patient 125	1	0	0
Adenoma patient 132	1	0	0
Adenoma patient 146	1	0	0
Adenoma patient 159	1	0	0
Adenoma patient 172	1	0	0
Cancer Chemo patient 150	0	0	1
Cancer Chemo patient 184	0	0	1
Cancer Chemo patient 187	0	0	1
Cancer Chemo patient 199	0	0	1
Cancer Chemo patient 253	0	0	1
Cancer Chemo patient 255	0	0	1
Carcinoma patient 183	0	1	0
Carcinoma patient 196	0	1	0
Carcinoma patient 2	0	1	0
Carcinoma patient 204	0	1	0
Carcinoma patient 212	0	1	0

Esto será posteriormente utilizado para obtener la matriz de diseño.

Esta información, que ya se tenía disponible, se puede mostrar cómo se aparean las muestras, de acuerdo con su nombre, y la clasificación recientemente seleccionada. Aquí se muestran todas las filas.

GSM180626	Adenoma patient 125	Adenoma
GSM180627	Adenoma patient 132	Adenoma
GSM180628	Adenoma patient 146	Adenoma
GSM180629	Adenoma patient 159	Adenoma
GSM180630	Adenoma patient 172	Adenoma
GSM180631	Adenoma patient 221	Adenoma
GSM180632	Adenoma patient 300	Adenoma
GSM180633	Adenoma patient 64	Adenoma
GSM180634	Adenoma patient 77A	Adenoma

GSM180635	Adenoma patient 97	Adenoma
GSM180636	Carcinoma patient 183	Carcinoma
GSM180637	Carcinoma patient 196	Carcinoma
GSM180638	Carcinoma patient 2	Carcinoma
GSM180639	Carcinoma patient 204	Carcinoma
GSM180640	Carcinoma patient 212	Carcinoma
GSM180641	Carcinoma patient 23	Carcinoma
GSM180642	Carcinoma patient 4	Carcinoma
GSM180643	Carcinoma patient 66	Carcinoma
GSM180644	Carcinoma patient 99	Carcinoma
GSM180645	Cancer Chemo patient 150	Quimio
GSM180646	Cancer Chemo patient 184	Quimio
GSM180647	Cancer Chemo patient 187	Quimio
GSM180648	Cancer Chemo patient 199	Quimio
GSM180649	Cancer Chemo patient 253	Quimio
GSM180650	Cancer Chemo patient 255	Quimio
GSM180651	Cancer Chemo patient 259	Quimio
GSM180652	Cancer Chemo patient 269	Quimio
GSM180653	Cancer Chemo patient 272	Quimio
GSM180654	Cancer Chemo patient 279	Quimio
GSM180655	Cancer Chemo patient 286	Quimio
GSM180656	Cancer Chemo patient 29	Quimio
GSM180657	Cancer Chemo patient 303	Quimio
GSM180658	Cancer Chemo patient 310	Quimio
GSM180659	Cancer Chemo patient 311	Quimio
GSM180660	Cancer Chemo patient 312	Quimio
GSM180661	Cancer Chemo patient 314	Quimio
GSM180662	Cancer Chemo patient 325	Quimio
GSM180663	Cancer Chemo patient 326	Quimio
GSM180664	Cancer Chemo patient 338	Quimio
GSM180665	Cancer Chemo patient 36	Quimio
GSM180666	Cancer Chemo patient 76	Quimio
GSM180667	Cancer Chemo patient 9	Quimio
GSM180668	Cancer Chemo patient 94	Quimio

Normalización de datos

Un paso previo al estudio de expresión diferencial es la normalización de datos [11]. Con este proceso se pretende eliminar la variabilidad causada por factores no biológicos o malfuncionamiento de los escáneres y microarrays, y conseguir que las diferencias de

intensidad en los arrays reflejen la expresión diferencial entre genes. El método más popular de normalización es el conocido como *Robust Multichip Analysis* (RMA) [12].

Con este proceso los datos disponibles en el archivo creado con la librería **affy**, de clase *AffyBatch*, se convierten en un *ExpressiónSet* utilizando la función `rma` disponible en **affy**.

Primero, se realiza el proceso con datos sin normalizar; a continuación, otro con los datos normalizados.

Datos sin normalizar

El paso se realiza con la función `rma` y la opción `normalize=FALSE`.

```
Calculating Expression
```

El tipo de archivo creado, con nombre *my.rma0*, y sus dimensiones se muestran a continuación:

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

Features  Samples
   12625       43
```

mientras que las del data frame con fenodatos son:

```
Dimensiones de los feno datos sin normalizar

[1] 43 42
```

Se genera una lista con los objetos creados por `rma`.

```
 [1] "title"                   "geo_accession"
 [3] "status"                  "submission_date"
 [5] "last_update_date"        "type"
 [7] "channel_count"           "source_name_ch1"
 [9] "organism_ch1"            "characteristics_ch1"
[11] "characteristics_ch1.1"   "characteristics_ch1.2"
[13] "characteristics_ch1.3"   "treatment_protocol_ch1"
[15] "growth_protocol_ch1"     "molecule_ch1"
[17] "extract_protocol_ch1"    "label_ch1"
[19] "label_protocol_ch1"      "taxid_ch1"
[21] "hyb_protocol"            "scan_protocol"
[23] "description"             "data_processing"
[25] "platform_id"             "contact_name"
[27] "contact_email"           "contact_phone"
[29] "contact_laboratory"      "contact_department"
[31] "contact_institute"       "contact_address"
[33] "contact_city"            "contact_state"
[35] "contact_zip.postal_code" "contact_country"
[37] "supplementary_file"      "data_row_count"
[39] "Age.at.surgery.ch1"      "Histology.ch1"
[41] "Stage.ch1"               "sample.levels"
```

Los valores de la matriz de datos de este ExpressionSet se pueden obtener utilizando la opción `exprs` y seleccionando filas y columnas. Por ejemplo, la información disponible en las primeras 5 filas y 5 columnas será

	GSM180626	GSM180627	GSM180628	GSM180629	GSM180630
100_g_at	8.862609	9.117964	9.250954	8.630628	9.294633
1000_at	9.286057	9.353794	9.583127	8.699458	9.505831
1001_at	7.570453	7.728362	7.562745	6.945264	7.746467
1002_f_at	6.833398	6.846095	7.018486	6.361166	7.103694
1003_s_at	7.889211	7.905227	8.079750	7.145236	8.020635

El procedimiento a seguir es similar con los fenodatos.

	title	geo_accession	status
GSM180626	Adenoma patient 125	GSM180626	Public on May 31 2007
GSM180627	Adenoma patient 132	GSM180627	Public on May 31 2007
GSM180628	Adenoma patient 146	GSM180628	Public on May 31 2007
GSM180629	Adenoma patient 159	GSM180629	Public on May 31 2007
GSM180630	Adenoma patient 172	GSM180630	Public on May 31 2007
GSM180631	Adenoma patient 221	GSM180631	Public on May 31 2007

Se comprueban los factores o niveles del archivo recién creado con la función `rma` (*my.rma0*).

```
Factores

[1] "Adenoma"   "Carcinoma" "Quimio"
```

Se visualiza la distribución de intensidades utilizando un diagrama de cajas y curvas de densidad de probabilidad; ver Figuras 7 y 8. Los resultados en ambas figuras se han clasificado en función de los tres grupos o factores establecidos anteriormente.

Datos normalizados

Para distinguir los datos creados en el paso anterior de los nuevos datos, se nombra el nuevo archivo como *my.rma1*.

```
Background correcting
Normalizing
Calculating Expression
```

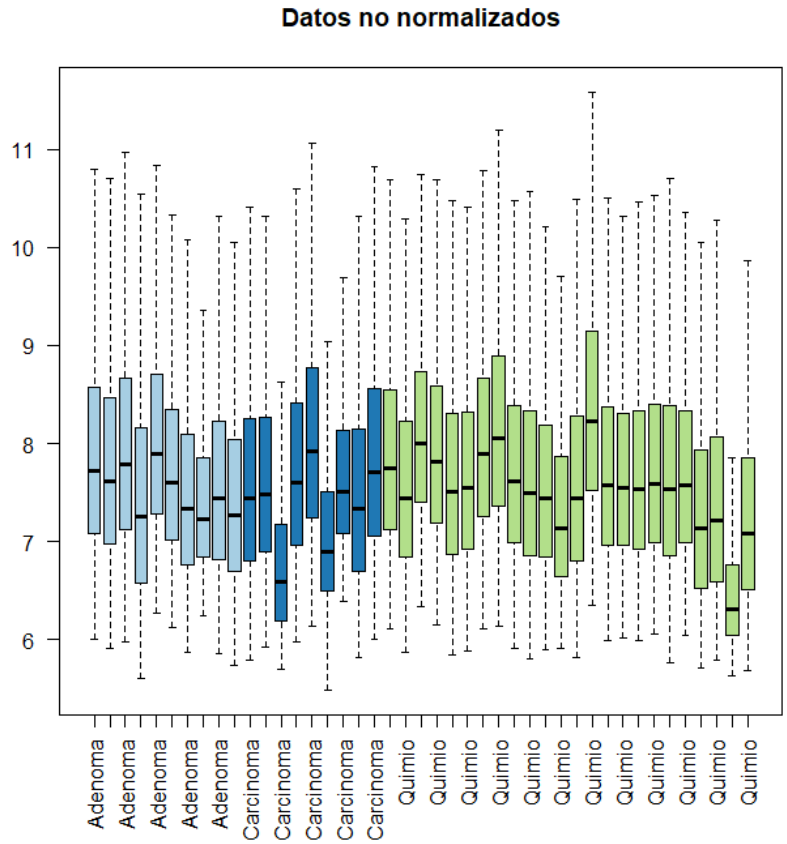

Figura 7. Intensidades de microrrays - Datos no normalizados.

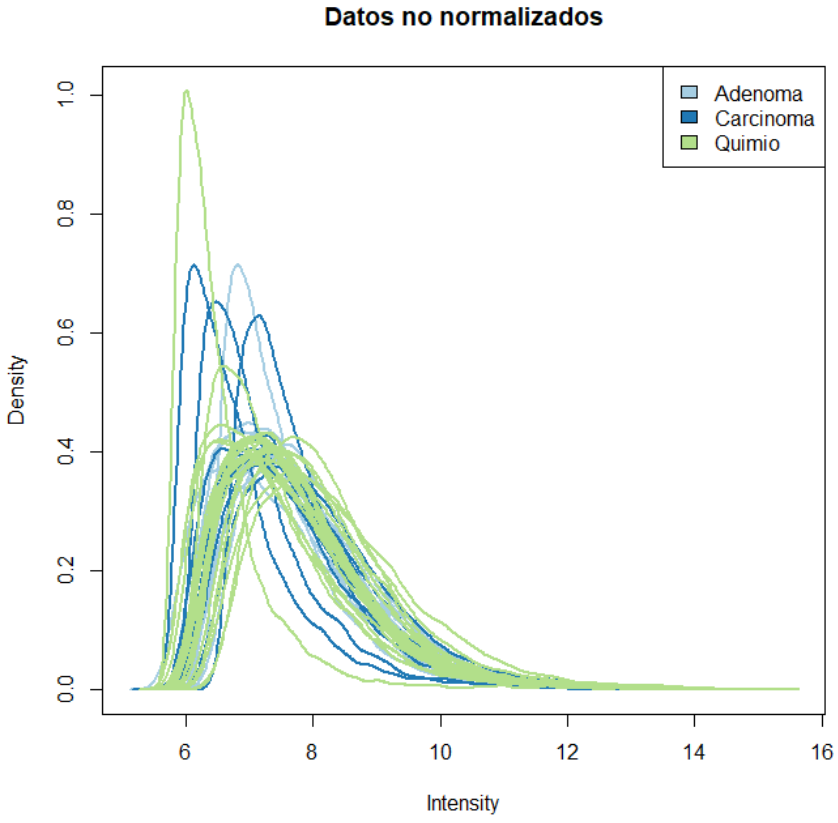

Figura 8. Densidades de probabilidad de las intensidades - Datos no normalizados.

Los nuevos objetos creados por `rma` tienen los mismos nombres que en los datos sin normalizar. El tipo de datos también es el mismo, así como el procedimiento a seguir para visualizar los nuevos resultados.

```
 [1] "Adenoma"   "Adenoma"   "Adenoma"   "Adenoma"   "Adenoma"   "Adenoma"
 [7] "Adenoma"   "Adenoma"   "Adenoma"   "Adenoma"   "Carcinoma" "Carcinom
a"
[13] "Carcinoma" "Carcinoma" "Carcinoma" "Carcinoma" "Carcinoma" "Carcinom
a"
[19] "Carcinoma" "Quimio"    "Quimio"    "Quimio"    "Quimio"    "Quimio"
[25] "Quimio"    "Quimio"    "Quimio"    "Quimio"    "Quimio"    "Quimio"
[31] "Quimio"    "Quimio"    "Quimio"    "Quimio"    "Quimio"    "Quimio"
[37] "Quimio"    "Quimio"    "Quimio"    "Quimio"    "Quimio"    "Quimio"
[43] "Quimio"
```

Se confirma que hay tres niveles o factores.

Las Figuras 9 y 10 presentan estos resultados.

Con la normalización se pretende obtener una distribución de todas las muestras con valores iguales o muy similares. Se puede comprobar qué tanto en uno como en otro gráfico, todas las muestras tienen un aspecto muy similar. Las dos figuras sugieren que la normalización ha funcionado correctamente.

Finalmente, los datos normalizados se guardan en un archivo de texto.

Detección de los genes más variables

La selección de genes diferencialmente expresados depende del número de genes que se desean analizar: cuantos más genes, más elevado será el valor p a utilizar, lo que puede causar el descarte de muchos genes. Si hay un gen diferencialmente expresado, su varianza será más grande que la de aquellos que no lo están. Presentar una variabilidad general puede ser útil para decidir qué porcentaje de genes muestra variabilidad que no pueda ser atribuida más que a causas biológicas. La Figura 11 muestra la desviación típica de todos los genes ordenada de menor a mayor, y los genes más variables cuya desviación estándar está en la franja 90-95% de todas las desviaciones estándar.

Filtrado de genes

El filtrado de genes, cuya variabilidad puede ser atribuida a causas aleatorias, puede ser muy útil para reducir el número posterior de pruebas a realizar. Un problema a tener en cuenta es el denominado error estadístico de Tipo I: falsos positivos. Esto se puede evitar reduciendo el número de pruebas estadísticas independientes. En el análisis de genes expresados diferencialmente, se deben eliminar las sondas de genes que no están expresados.

Se han propuesto e implantado (en **Bioconductor**) varios procedimientos para filtrar genes. Aquí se presentan dos; el resto del estudio se realizará con los resultados del primer filtrado.

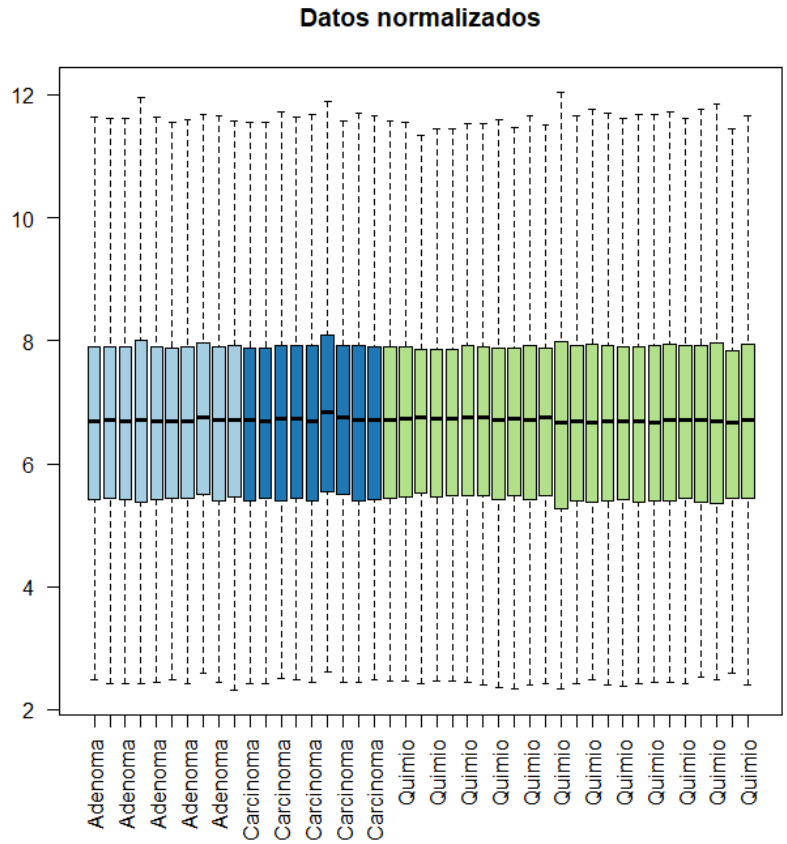

Figura 9. Intensidades de microrrays - Datos normalizados.

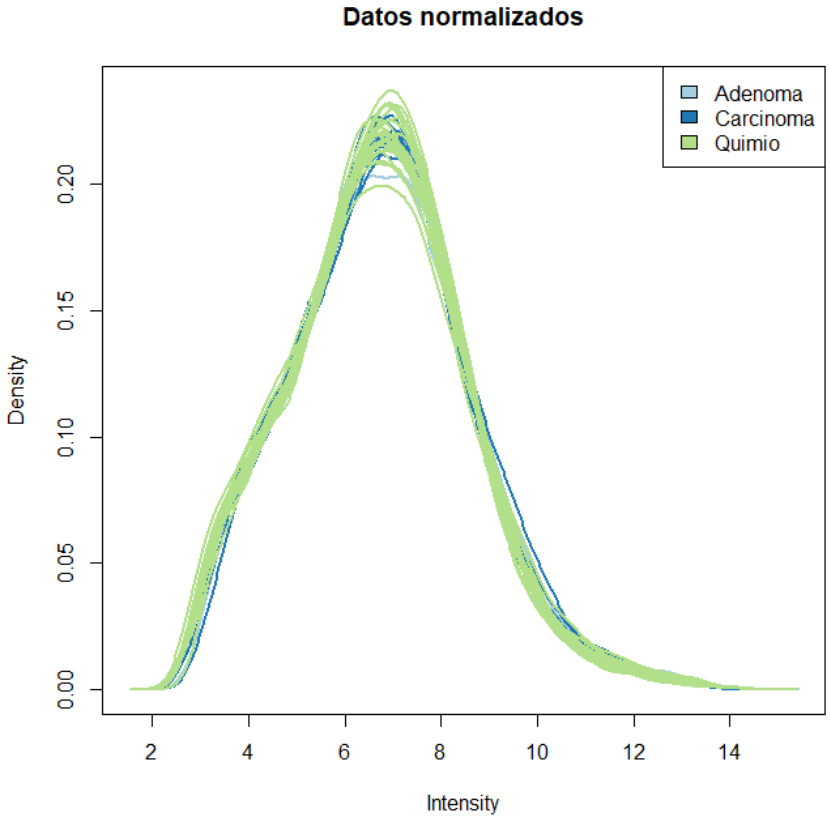

Figura 10. Densidades de probabilidad de las intensidades - Datos normalizados.

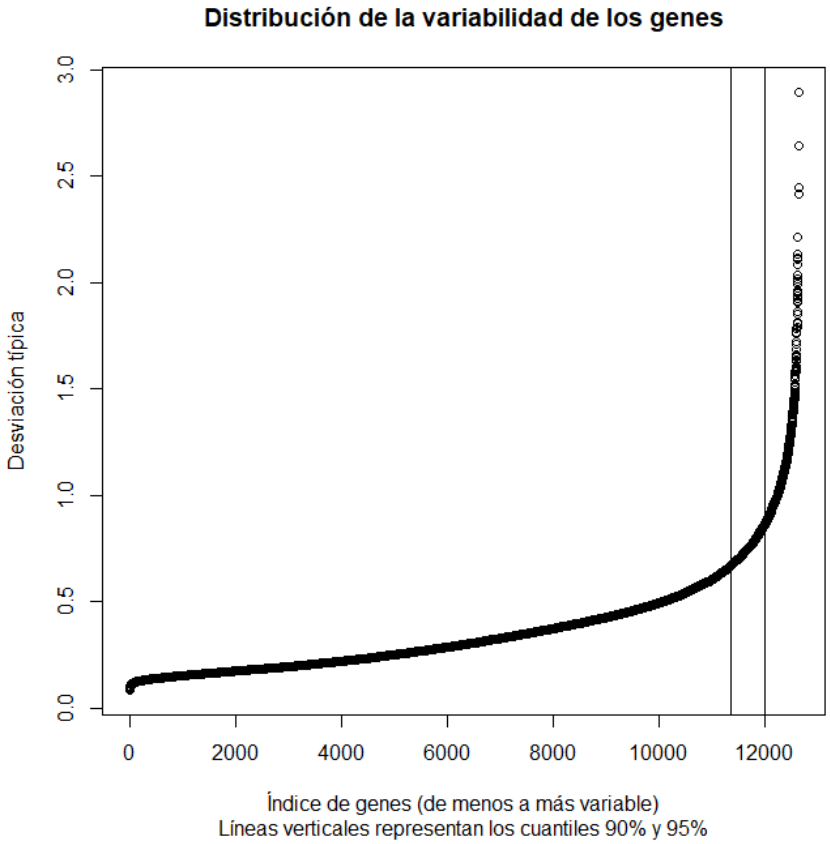

Figura 11. Desviaciones típicas con valores ordenados de menor a mayor.

1 - La función `mas5calls` disponible en la librería **affy** realiza un test de Wilcoxon con objetos AffyBatch y genera una clasificación de genes en tres niveles: P = present, M = marginal, and A = absent [13], [14].

Esta clasificación se basa en los siguientes criterios:

- P si $p - value < \alpha_1$
- M si $\alpha_1 <= p - value < \alpha_2$
- A si $\alpha_2 <= p - value$.

Los valores por defecto son: $\alpha_1 = 0.04$, $\alpha_2 = 0.06$.

Valores pequeños de p indican presencia de transcritos, mientras que valores grandes indican ausencia 14.

2 - La función `nsFilter`, disponible en la librería *genefilter*, funciona de manera similar; es decir, seleccionando genes a partir de un valor umbral.

Se comprueba el tipo de datos a manipular.

```
[1] "AffyBatch"
attr(,"package")
[1] "affy"

[1] "hgu95av2"
```

El objeto con los datos es de clase *AffyBatch* y utiliza la librería de anotaciones *hgu95av2.db*.

Y ahora para el objeto con los datos normalizados.

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

Se comprueba que es un objeto de clase *ExpressionSet*.

Es necesario tener en cuenta esta información para poder filtrar los datos.

Primero, se aplica la función `mas5calls` al archivo de clase *AffyBatch*.

```
Getting probe level data...
Computing p-values
Making P/M/A Calls
```

Se comprueba la clase del objeto creado con la función `mas5calls`.

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

El archivo resultante puede ser analizado utilizando la función `exprs`. El tamaño del nuevo archivo tiene tantos genes (filas) y muestras (columnas) como el archivo original.

```
Dimensiones de la matriz de datos filtrados

[1] 12625    43
```

Los resultados correspondientes a las 5 primeras filas y columnas es el siguiente:

	GSM180626	GSM180627	GSM180628	GSM180629	GSM180630
100_g_at	A	A	A	P	A
1000_at	P	P	P	P	P
1001_at	A	P	A	M	A
1002_f_at	A	A	A	A	A
1003_s_at	A	A	A	A	A

El resultado se puede resumir en forma de tabla.

A	M	P
5778	357	6490

Se comprueba que la suma de los genes clasificados en las tres categorías es el que corresponde.

```
Suma de genes= 12625
```

Se puede calcular el número de veces que un gen ha quedado dentro del mismo rango de $p - valores$. El código a utilizar para cada grupo es: 1 = P, 2 = M, 3 = A.

El resultado con el número de veces que los primeros genes han obtenido un valor dentro de cada nivel:

```
Nivel P

 100_g_at   1000_at   1001_at 1002_f_at 1003_s_at   1004_at
       18        43         9         0         0         0

Nivel M

 100_g_at   1000_at   1001_at 1002_f_at 1003_s_at   1004_at
        9         0         4         0         0         0

Nivel A

 100_g_at   1000_at   1001_at 1002_f_at 1003_s_at   1004_at
       16         0        30        43        43        43
```

Se puede observar que la suma de veces que cada gen aparece en los tres niveles es igual al número de muestras, 43.

Un resumen del filtrado de genes se presenta en forma de tabla. Primero, se averigua el número de niveles, y después el número de genes en cada nivel (aquí solo se muestran los primeros). Finalmente, se comprueba que todos los genes suman la cantidad esperada.

```
Número de niveles

[1] 44

Genes por nivel-Grupo P
```

0	1	2	3	4	5	6	7	8	9
3036	547	335	255	199	166	143	146	143	126

```
Número total de genes-Grupo P

[1] 12625
```

Se observa que hay 44 niveles y todos suman los 12625 genes que hay en este estudio.

Este resultado se puede presentar en forma de proporciones; las 10 primeras serían las siguientes:

```
Proporción de genes por nivel
```

0	1	2	3	4	5	6	7	8	9	10	11
24.05	4.33	2.65	2.02	1.58	1.31	1.13	1.16	1.13	1.00	0.84	0.97

El filtrado se puede realizar fijando un valor umbral para el nivel de $p - valor$. Por ejemplo, si se fija en 5 se tienen los siguientes valores:

```
Valor igual o superior a 5
```

FALSE	TRUE
4372	8253

```
Valor inferior a 5
```

```
 FALSE  TRUE
  8253  4372
```

Para entender mejor como funciona este valor umbral se compara con la selección que resulta de escoger un valor 15:

```
Valor igual o superior a 15
```

```
 FALSE  TRUE
  5642  6983
```

```
Valor inferior a 15
```

```
 FALSE  TRUE
  6983  5642
```

Al aumentar el valor umbral, aumenta el número de genes que quedan por debajo de este valor.

Si se escoge como umbral para filtrar genes el valor 10, se obtienen los resultados gráficos que muestran las Figuras 12 y 13.

Se comprueba que la función de densidad de la mayoría de los genes no es la misma por debajo que por encima del valor umbral seleccionado.

El paso final consistirá en obtener la lista de genes eliminando los genes de control (con la extensión $AFFX$). Se escoge el valor umbral 15.

```
Número de genes seleccionado = 5642
```

Las dimensiones provisionales de la matriz de datos serán, por tanto, las siguientes:

```
[1] 5642   43
```

El número de genes de control (con código $AFFX$) es en este estudio el siguiente:

```
Número de genes de control = 67
```

Con lo que la dimensión final de la matriz será:

Se comprueba la clase y las dimensiones del objeto con los datos filtrados.

```
[1] "matrix"
```

```
[1] 5575   43
```

Se trata de una matriz con 5575 filas y 43 columas.

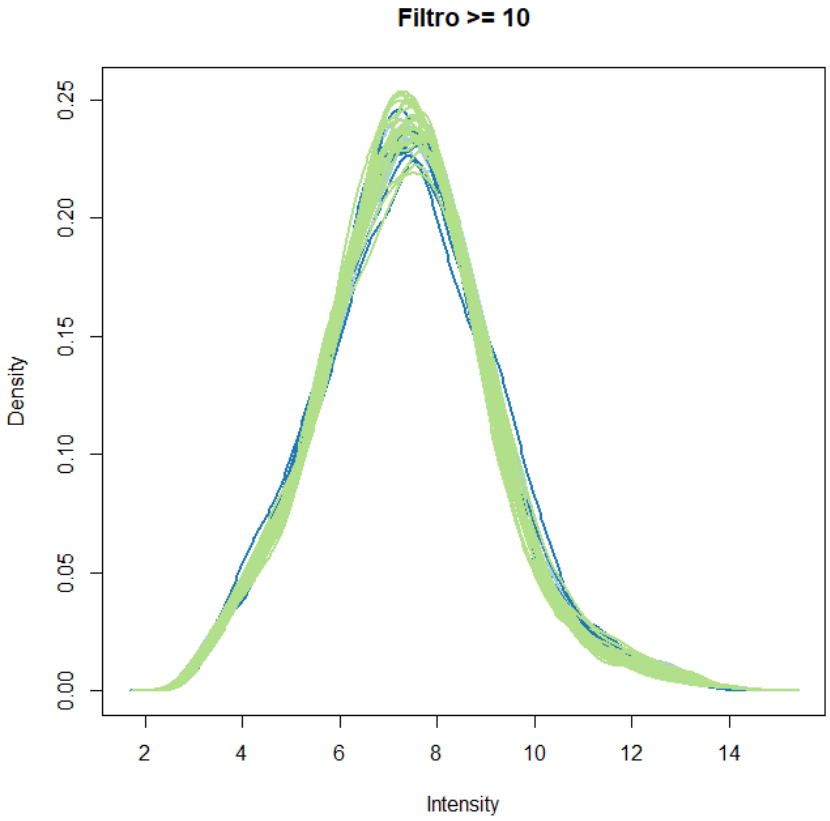

Figura 12 - Filtrado P >= 10.

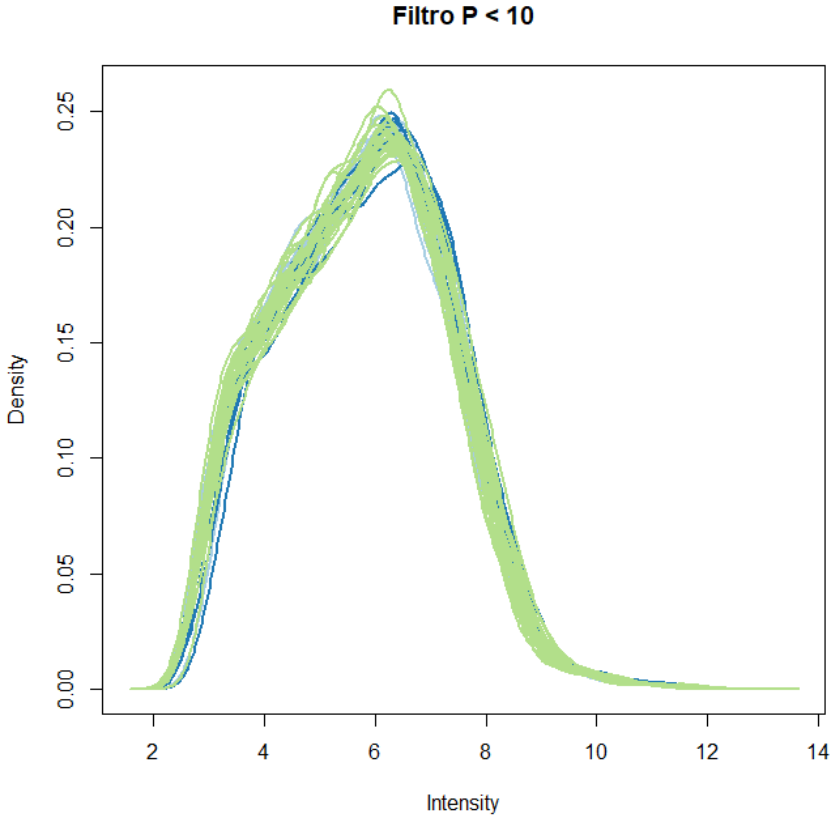

Figura 13 - Filtrado P < 10.

El filtrado también se puede realizar con la librería *genefilter*. En general, se debe esperar un número de genes distinto al obtenido con el método anterior.

Las opciones seleccionadas aquí para realizar el filtrado son las siguientes:

- require.entrez = TRUE,
- remove.dupEntrez = TRUE,
- var.func=IQR,
- var.filter=TRUE,
- var.cutoff=0.75,
- filterByQuantile=TRUE,
- feature.exclude = "^AFFX".

Se listan los objetos creados con `nsFilter`.

```
[1] "eset"       "filter.log"
```

Los resultados con la nueva opción de filtrado son los siguientes:

```
$numDupsRemoved
[1] 2856

$numLowVar
[1] 6438

$numRemoved.ENTREZID
[1] 1166

$feature.exclude
[1] 19

ExpressionSet (storageMode: lockedEnvironment)
assayData: 2146 features, 43 samples
  element names: exprs
protocolData
  sampleNames: GSM180626 GSM180627 ... GSM180668 (43 total)
  varLabels: ScanDate
  varMetadata: labelDescription
phenoData
  sampleNames: GSM180626 GSM180627 ... GSM180668 (43 total)
  varLabels: title geo_accession ... sample.levels (42 total)
  varMetadata: labelDescription
featureData: none
experimentData: use 'experimentData(object)'
Annotation: hgu95av2
```

Se observa que la nueva selección de genes tiene 2146; cantidad inferior a la obtenida con el procedimiento anterior.

Selección de genes expresados diferencialmente

Esta nueva selección se basará en una prueba estadística que realiza la comparación entre grupos de genes, no entre genes individuales. Aunque se han propuesto varias alternativas para realizar esta nueva prueba, existe un consenso general en que la aplicación de pruebas tipo t-test no es adecuada, y que las mejores opciones se deben basar en comparar la variación entre grupos. En general, se acepta que técnicas como SAM [15] o Linear Models for Microarrays [16] son adecuadas para esta tarea. En esta sección se presenta el método propuesto por Smyth e implantado en la librería **limma** de **Bioconductor** [17].

Antes de proceder con este paso, se realiza un control adicional de calidad con los genes filtrados para comprobar visualmente si existe cierta consistencia entre los datos seleccionados.

Control de calidad adicional

Las Figuras 14 y 15 presentan dos resultados gráficos, un gráfico de escala multidimensional (utilizando la función `plotMDS` disponible en **limma**) y otro con un dendograma obtenido utilizando distancias euclídeas (y la función `hclust` disponible en la librería *stats*).

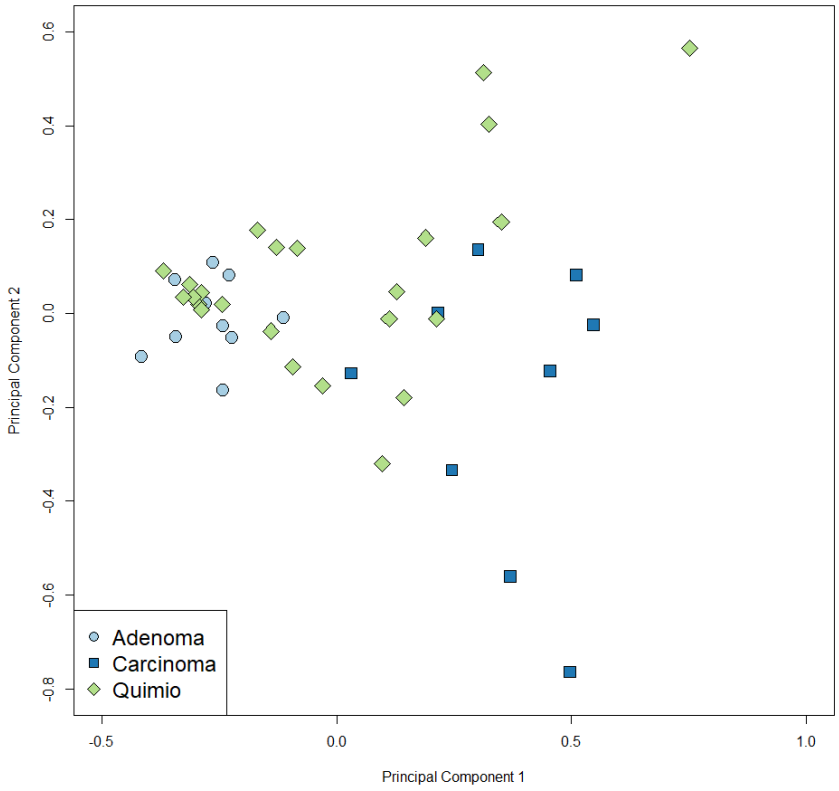

Figura 14. Gráfico de escala multidimensional con datos filtrados.

La figura muestra una separación entre los tres grupos: todas las muestras de adenoma están a la izquierda, la mayoría de muestras de carcionoma están a la derecha, y (con excepción de una muestra) todas las de quimio en el centro-izquierda.

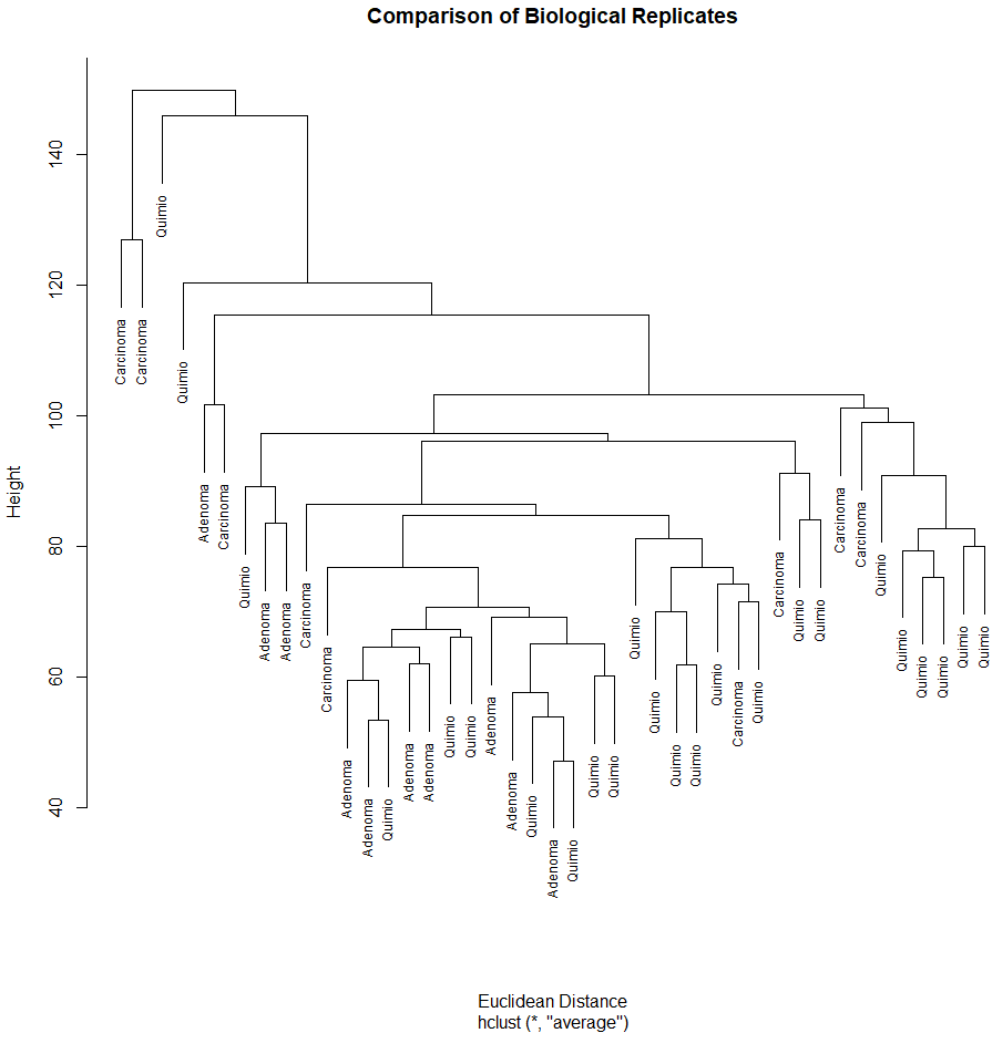

Figura 15. Dendograma con distancias euclídeas.

En ambas figuras se observa una cierta separación entre el grupo *carcinoma* y los otros dos, pero también algún dato atípico, como alguna muestra del grupo *adenoma* cercana a muestras de tipo *carcinoma*.

Definición del experimento - Matriz de diseño

El primer paso de un análisis basado en modelos lineales consiste en la creación de la conocida como matriz de diseño. Se trata de una tabla que presenta la separación de muestras (archivo *GSM*) a los distintos grupos en los que se ha dividido el ensayo. La matriz o tabla tiene tantas filas como muestras y tantas columnas como grupos. Cada fila de la tabla tiene un solo valor "1" en la columna del grupo al que pertenece la muestra. Se recuerda que los resultados del ensayo están divididos en 43 muestras y que se han seleccionado tres grupos o factores: Adenoma, Carcinoma, y Quimio.

Puesto que en este trabajo no se editado el archivo *targets*, la matriz de diseño se tiene que definir ahora manualmente. Las muestras han sido clasificadas en tres grupos que servirán para editar la matriz.

Si las filas se nombran con los códigos originales de las muestras, la decoración de la matriz será la siguiente:

	GAdenoma	GCarcinoma	GQuimio
GSM180626	1	0	0
GSM180627	1	0	0
GSM180628	1	0	0
GSM180629	1	0	0
GSM180630	1	0	0
GSM180631	1	0	0
GSM180632	1	0	0
GSM180633	1	0	0
GSM180634	1	0	0
GSM180635	1	0	0
GSM180636	0	1	0
GSM180637	0	1	0
GSM180638	0	1	0
GSM180639	0	1	0
GSM180640	0	1	0
GSM180641	0	1	0
GSM180642	0	1	0
GSM180643	0	1	0
GSM180644	0	1	0
GSM180645	0	0	1
GSM180646	0	0	1
GSM180647	0	0	1
GSM180648	0	0	1
GSM180649	0	0	1
GSM180650	0	0	1
GSM180651	0	0	1
GSM180652	0	0	1
GSM180653	0	0	1
GSM180654	0	0	1
GSM180655	0	0	1
GSM180656	0	0	1
GSM180657	0	0	1
GSM180658	0	0	1
GSM180659	0	0	1
GSM180660	0	0	1
GSM180661	0	0	1
GSM180662	0	0	1
GSM180663	0	0	1

GSM180664	0	0	1
GSM180665	0	0	1
GSM180666	0	0	1
GSM180667	0	0	1
GSM180668	0	0	1

Comparación entre muestras - Matriz de contrastes

Una segunda matriz, conocida como matriz de contrastes, es necesaria para describir las comparaciones entre grupos. La matriz tiene tantas filas como grupos y tantas columnas como comparaciones. Una comparación (o contraste) se representa con "1" y "-1" en las filas de los grupos que se han de comparar, y ceros en las restantes filas. La columna tendrá tantos elementos no nulos como grupos intervienen en la comparación; la única restricción es que la suma de elementos de una columna debe ser cero.

En este estudio se compara la expresión de genes en muestras con adenoma, carcinoma y tratamiento con quimioterapia. Esto se puede realizar mediante las tres siguientes comparaciones:

- AC = GAdenoma - GCarcinoma
- AQ = GAdenoma - GQuimio
- CQ = GCarcinoma - GQuimio.

La matriz de contrastes tiene la siguiente decoración:

	AC	AQ	CQ
GAdenoma	1	1	0
GCarcinoma	-1	0	1
GQuimio	0	-1	-1

Selección de genes

Una vez definidas las matrices de diseño y de contrastes se continúa con una estimación del modelo y de los contrastes, y la realización de pruebas que permitan seleccionar los genes expresados diferencialmente. El método implantado utiliza modelos de Bayes empíricos para combinar la información de toda la matriz de datos y de cada gen individual y obtener estimaciones de error mejoradas. El análisis proporciona los estadísticos de test habituales (por ejemplo, valores p ajustados) que se utilizan para ordenar los genes diferencialmente expresados. A fin de controlar el porcentaje de falsos positivos que puedan resultar de un alto número de contrastes realizados simultáneamente, los valores p se ajustan de forma que haya control sobre la tasa de falsos positivos utilizando el método de Benjamini y Hochberg [18].

El proceso completo consiste en tres pasos en los que se aplican sucesivamente las funciones `lmFit`, `contrasts.fit`, y `eBayes`.

La aplicación de la función `lmFit` genera un archivo que aquí se nombra como *my.fit*. La información resultante de aplicar sucesivamente `contrasts.fit` y `eBayes` se almacena en

un archivo con el nombre *fit.main*. Ambos archivos son de clase `MArrayLM`, definida en la librería **limma**.

Se analizan ambos archivos. Primero, se presenta la clase y se listan los objetos creados en *my.fit*.

```
[1] "MArrayLM"
attr(,"package")
[1] "limma"
```

Los objetos generados con la función `lmFit` son los siguientes:

```
 [1] "coefficients"     "rank"             "assign"           "qr"
 [5] "df.residual"      "sigma"            "cov.coefficients" "stdev.unscaled"
 [9] "pivot"            "Amean"            "method"           "design"
```

Se presenta una selección de los datos incluidos en el objeto *coefficients*.

	GAdenoma	GCarcinoma	GQuimio
1001_at	6.46	6.17	6.52
1002_f_at	4.91	4.92	4.95
1003_s_at	6.93	6.99	6.94
1004_at	6.70	6.73	6.65
1006_at	4.78	5.19	4.79
1010_at	6.32	6.34	6.28
1012_at	4.41	4.15	4.35
1015_s_at	6.60	6.66	6.55
1016_s_at	4.07	4.16	4.12
1018_at	6.87	6.78	6.89

Y ahora para *fit.main*, que incluye los resultados de los contrastes.

```
[1] "MArrayLM"
attr(,"package")
[1] "limma"
```

La lista de los nombres de objetos generados al crear *fit.main* es la siguiente:

```
 [1] "coefficients"     "rank"             "assign"           "qr"
 [5] "df.residual"      "sigma"            "cov.coefficients" "stdev.unsca
led"
 [9] "Amean"            "method"           "design"           "contrasts"
[13] "df.prior"         "s2.prior"         "var.prior"        "proportion"
[17] "s2.post"          "t"                "df.total"         "p.value"
[21] "lods"             "F"                "F.p.value"
```

Se presenta una selección de los datos incluidos en el objeto *coefficients*.

	AC	AQ	CQ
1001_at	0.2927468	-0.0558547	-0.3486016
1002_f_at	-0.0097472	-0.0377162	-0.0279690
1003_s_at	-0.0600305	-0.0133144	0.0467162
1004_at	-0.0273705	0.0513806	0.0787511
1006_at	-0.4184498	-0.0195870	0.3988628
1010_at	-0.0236116	0.0415675	0.0651792
1012_at	0.2591659	0.0584240	-0.2007419
1015_s_at	-0.0602258	0.0479843	0.1082101
1016_s_at	-0.0925666	-0.0505557	0.0420109
1018_at	0.0936885	-0.0227364	-0.1164250

Finalmente, se crea un archivo en formato texto con los resultados obtenidos con **limma** y se guarda en el subdirectorio de resultados.

Listas de genes diferencialmente expresados

La librería **limma** incluye la opción `topTable` que genera, para cada contraste una lista de genes ordenados de menor a mayor según el valor *p* [19]. Para cada gen se proporcionan los siguientes estadísticos:

- logFC: Mean difference between groups
- AveExpr: Average expression of all genes in the comparison
- t : Moderated t-statistic (t-test-like statistic for the comparison)
- P.Value: Test p–value
- adj.P.Val: Adjusted p–value following Benjamini and Hochberg (1995)
- B-statistic: Posterior log odds of the gene of being vs non being differential expressed.

Los primeros resultados de cada tabla son los siguientes:

```
Contraste Adenoma-Carcinoma
```

	logFC	AveExpr	t	P.Value	adj.P.Val	B
36562_at	0.933	8.048	7.538	0	0	11.457
38348_at	0.725	5.140	7.016	0	0	9.809
419_at	-1.084	5.944	-6.680	0	0	8.741
153_f_at	-1.168	4.945	-6.617	0	0	8.541
41400_at	-0.971	7.321	-6.463	0	0	8.049
511_s_at	-0.567	3.609	-6.456	0	0	8.026

```
Contraste Adenoma-Quimioterapia
```

	logFC	AveExpr	t	P.Value	adj.P.Val	B
33576_at	1.039	5.239	6.862	0	0.000	8.643
41607_at	2.702	6.787	5.739	0	0.002	5.394
36562_at	0.576	8.048	5.682	0	0.002	5.229
34010_at	0.475	6.814	5.353	0	0.004	4.274
36842_at	0.722	4.471	5.221	0	0.005	3.894
34823_at	1.216	5.027	5.061	0	0.007	3.432

```
Contraste Carcinoma-Quimioterapia
```

	logFC	AveExpr	t	P.Value	adj.P.Val	B
38536_at	0.4267706	3.572383	6.973503	0e+00	0.0000365	9.628744
153_f_at	1.0436498	4.944693	6.950724	0e+00	0.0000365	9.557146
38167_at	0.6301467	6.143752	6.793508	0e+00	0.0000414	9.062256
419_at	0.8773142	5.944305	6.355088	1e-07	0.0001201	7.677312
37305_at	1.2920511	6.266616	6.308434	1e-07	0.0001201	7.529700
38920_at	0.5977512	3.713327	6.276175	1e-07	0.0001201	7.427624

Se analiza lo obtenido en las tres tablas.

- Dimensiones de las tablas

```
Contraste Adenoma-Carcinoma

[1] 5575    6

Contraste Adenoma-Quimioterapia

[1] 5575    6

Contraste Carcinoma-Quimioterapia

[1] 5575    6
```

- Número de $p - valores$ por debajo de 0.05

```
Contraste Adenoma-Carcinoma

[1] 1325

Contraste Adenoma-Quimioterapia

[1] 350

Contraste Carcinoma-Quimioterapia

[1] 1329
```

Anotación de genes

Obtener información adicional o anotaciones sobre los genes seleccionados y disponibles en bases de datos públicas puede ser muy útil. Este proceso busca información funcional

(Gene Symbol, Entrez Gene identifier, Gene description) utilizando los identificadores de los genes seleccionados.

Se define una función para facilitar el trabajo de anotación y se guardan las tres tablas de anotaciones.

Se analiza lo que se ha obtenido con la primera tabla (contraste Adenoma-Carcinoma).

Primero, la lista de objetos.

```
Resumen de resultados

 [1] "PROBEID"   "SYMBOL"    "ENTREZID"  "GENENAME"  "logFC"     "AveExpr"
 [7] "t"         "P.Value"   "adj.P.Val" "B"
```

Se observa que el resultado incluye tanto anotaciones (objeto SYMBOL) como resultados de pruebas estadísticas (objeto B).

A continuación, los nombres de los genes según los distintos identificadores.

```
Genes e identificadores - Constraste Adenoma-Carcinoma
```

PROBEID	SYMBOL	ENTREZID	GENENAME
1001_at	TIE1	7075	tyrosine kinase with immunoglobulin like and EGF like domains 1
1002_f_at	CYP2C19	1557	cytochrome P450 family 2 subfamily C member 19
1003_s_at	CXCR5	643	C-X-C motif chemokine receptor 5
1004_at	CXCR5	643	C-X-C motif chemokine receptor 5
1006_at	MMP10	4319	matrix metallopeptidase 10
1010_at	MAPK11	5600	mitogen-activated protein kinase 11

Finalmente, se presentan algunos valores. Por ejemplo, los 10 primeros valores del objeto t.

```
Valores t - Contraste Adenoma-Carcinoma

 [1]  2.3271110 -0.1314626 -0.6405913 -0.3313000 -1.9869267 -0.3087171
 [7]  1.9611516 -0.8312651 -0.9806068  0.8567301
```

Visualizando las diferencias de expresiones

Una visualización general de la expresión diferencial se puede obtener con un *volcano plot*. Este tipo de gráficos presenta una relación entre los cambios de expresión en escala logarítmica ("efecto biologico") y el negativo del logaritmo del p valor ("efecto estadístico"), y permite detectar la existencia de genes diferencialmente expresados. La Figura 16 muestra un *volcano plot* a partir de la comparación Adenoma-Quimioterapia con los nombres de los 4 (número a seleccionar) primeros genes de la tabla correspondiente.

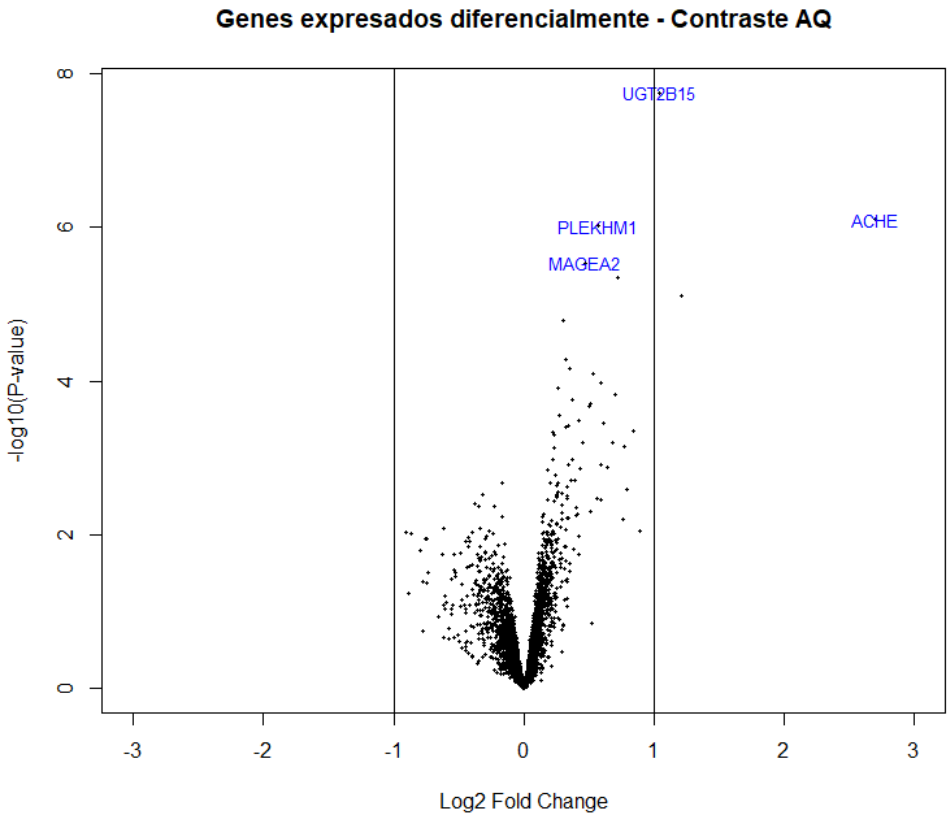

Figura 16. Volcano plot.

Múltiples comparaciones

Después de realizar varias comparaciones, puede ser interesante conocer los genes comunes dos o más comparación. A veces los genes más relevantes son los que se encuentran en un grupo; otras veces, pueden ser los genes comunes a todos los grupos. Las funciones `decideTests` y `VennDiagram` disponibles en la librería **limma** se pueden utilizar para anotar y contar el número de genes seleccionados en cada comparación, o los genes comunes a dos o más grupos.

Se presenta una tabla con el resumen de resultados después de aplicar `decideTests`.

	AC	AQ	CQ
Down	222	0	15
NotSig	5284	5566	5435
Up	69	9	125

La suma en todas las columnas es 5575. Se comprueba que tiene tantas columnas como grupos de comparación y tres filas que clasifican los genes de acuerdo con los siguientes criterios:

- Up: t-test values > 0, FDR < selected cutoff,
- NotSig: FDR > selected cutoff,
- Down: t-test values < 0, FDR < selected cutoff.

NOTA: FDR = False Discovery Rate - Tasa de falsos negativos

Estos resultados se pueden resumir con un diagrama de Venn. El diagrama de la Figura 17 muestra los genes diferencialmente expresados en cada comparación según los criterios seleccionados (aquí son "FDR < 0.1" and "logFC > 1""), y los genes que son comunes a todas las combinaciones de los grupos obtenidos en cada comparación.

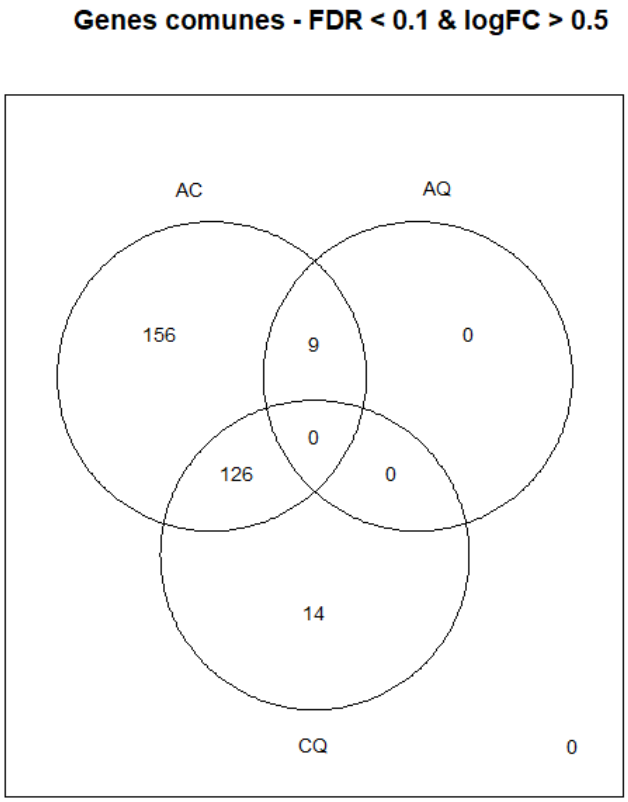

Figura 17. Diagrama de Venn.

Visualización de los perfiles de expresión

Las expresiones de cada gen se agrupan para destacar los genes regulados simultáneamente constituyendo perfiles de expresión. Una de los métodos más utilizados para visualizar estos perfiles se basa en el uso de mapas de color o *heatmaps*. Un *heatmap* es una técnica de visualización de datos que muestra la expresión de los genes como un fenómeno de color en dos dimensiones: los valores numéricos de la matriz de datos se convierten en tonos de color, de forma que se pueda analizar como varían los valores entre muestras o entre genes.

Las Figuras 18 y 19 muestran los *heatmaps* que corresponden a las matrices de datos sin agrupar y después de agrupados, respectivamente.

De la comparación de estos resultados, se observa como varía la intensidad de color (comparar, por ejemplo, las franjas izquierdas de las dos figuras) como consecuencia del reagrupamiento de los genes (filas) y las muestras (columnas), que se puede deducir de las dos barras de colores que hay en la parte superior de cada heatmap.

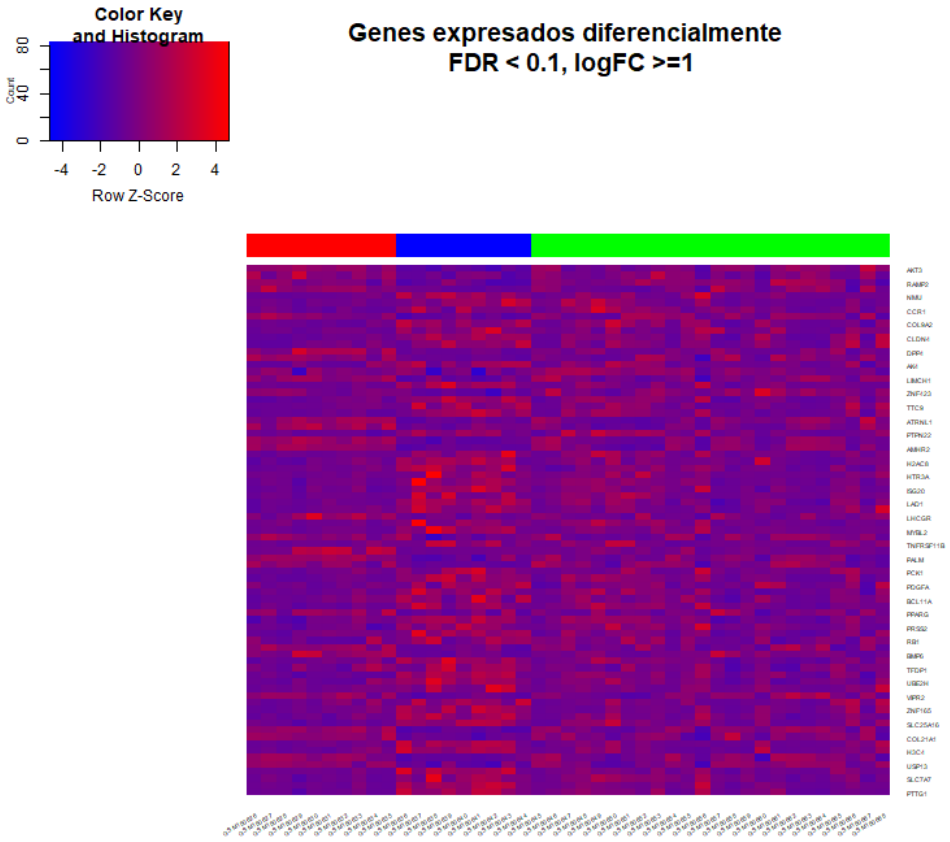

Figura 18. Heatmap sin agrupamiento.

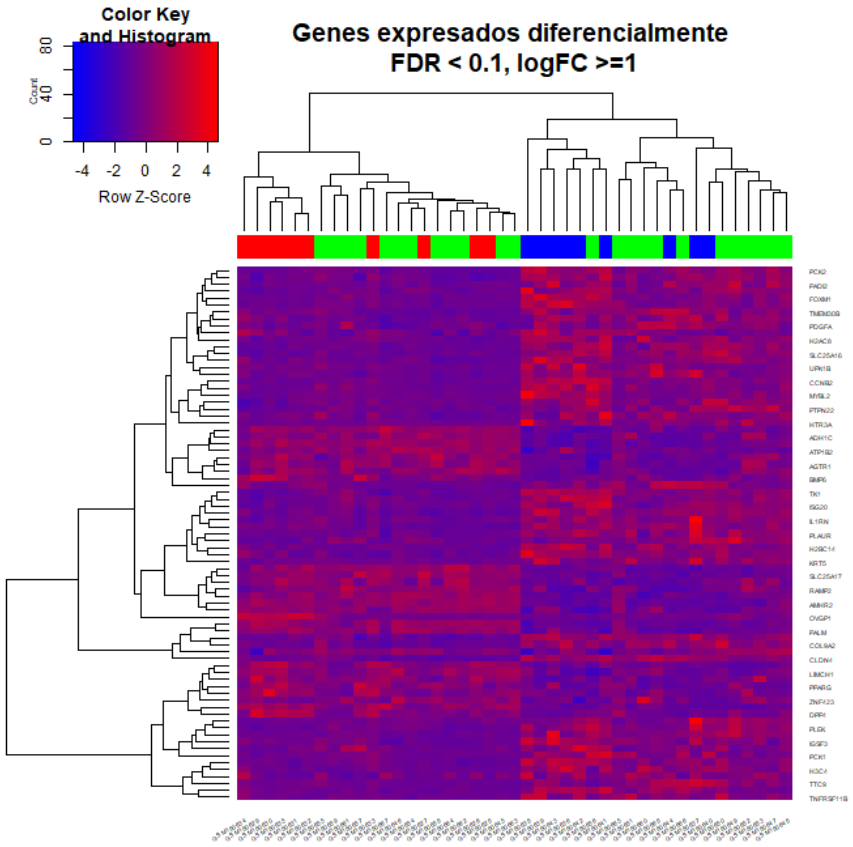

Figura 19. Heatmap con agrupamiento.

Análisis de significación biológica de los resultados

Una vez obtenida la lista de genes, se realiza su interpretación biológica, o lo que es igual una aproximación estadística conocida como *Análisis de Conjuntos de Genes* ("Gene Set Analysis"). Dada una lista de genes expresada diferencialmente entre dos condiciones (ver matriz de contrastes), el objetivo es determinar si hay funciones, procesos biológicos o sendas moleculares que caractericen estos genes más frecuentemente que al resto de genes analizado. De las varias alternativas que se pueden utilizar [20], aquí se ha escogido la opción implantada en la librería **ReactomePA** de **Bioconductor** [21].

Esta librería utiliza la base de anotación de datos ReactomePA. Para llevar a cabo la tarea es recomendable realizar una selección de genes poco restrictiva en el momento de filtrar genes.

En el primer paso se prepara una lista de todos genes analizados, y que se puede obtener a partir de las tres tablas obtenidas anteriormente:

```
  AC   AQ   CQ
1166   32 1144
```

Este análisis también necesita los identificadores **Entrez** para todos los genes analizados. Aquí se entiende que los genes son todos los genes disponibles, y se define el universo como todos los genes que tienen al menos una anotación en Gene Ontology.

El número de genes que forman el "universo" en este estudio es:

```
Número total de genes (universo) = 20525
```

El análisis de significación biológica será aplicado a la primera y tercera listas. Usando la librería **ReactomePA** y la opción `enrichPathway` se realiza un estudio ORA.

```
####################################
Comparison:  AC
                           ID
R-HSA-3214815 R-HSA-3214815
R-HSA-2299718 R-HSA-2299718
R-HSA-912446   R-HSA-912446
R-HSA-5625886 R-HSA-5625886
R-HSA-9616222 R-HSA-9616222
R-HSA-3214858 R-HSA-3214858

Description
R-HSA-3214815
HDACs deacetylate histones
R-HSA-2299718
Condensation of Prophase Chromosomes
R-HSA-912446
Meiotic recombination
R-HSA-5625886 Activated PKN1 stimulates transcription of AR (androgen rece
ptor) regulated genes KLK2 and KLK3
R-HSA-9616222                                                    Transcrip
tional regulation of granulopoiesis
R-HSA-3214858
RMTs methylate histone arginines
```

```
              GeneRatio  BgRatio       pvalue     p.adjust       qvalue
R-HSA-3214815    38/732 94/10616 2.121547e-20 2.622232e-17 2.161745e-17
R-HSA-2299718    32/732 73/10616 1.202836e-18 7.433525e-16 6.128132e-16
R-HSA-912446     34/732 85/10616 3.286109e-18 1.279492e-15 1.054802e-15
R-HSA-5625886    30/732 66/10616 4.140750e-18 1.279492e-15 1.054802e-15
R-HSA-9616222    34/732 89/10616 1.804318e-17 4.307867e-15 3.551367e-15
R-HSA-3214858    32/732 79/10616 2.091197e-17 4.307867e-15 3.551367e-15

geneID
R-HSA-3214815 H2BC11/H2BC14/H2AC16/H2AC8/H3C4/H3C1/H3C3/H3C6/H3C11/H3C8/H3
C12/H3C10/H3C2/H3C7/H2AC11/H2AC13/H2AC15/H2AC17/H2AC12/ARID4B/H4C9/H4C1/H4
C4/H4C6/H4C12/H4C11/H4C3/H4C8/H4C2/H4C5/H4C13/H4C14/H4-16/H4C15/BRMS1/TBL1
X/H2AC7/REST
R-HSA-2299718                                                  H2BC11/H2BC14
/CDK1/H2AC8/H3C4/SMC2/H3C1/H3C3/H3C6/H3C11/H3C8/H3C12/H3C10/H3C2/H3C7/PLK1
/H4C9/H4C1/H4C4/H4C6/H4C12/H4C11/H4C3/H4C8/H4C2/H4C5/H4C13/H4C14/H4-16/H4C
15/RB1/H2AC7
R-HSA-912446                                        H2BC11/H2BC14/BLM/H2AC8/H3C4
/RPA3/H3C1/H3C3/H3C6/H3C11/H3C8/H3C12/H3C10/H3C2/H3C7/BRCA2/RAD51/H4C9/H4C
1/H4C4/H4C6/H4C12/H4C11/H4C3/H4C8/H4C2/H4C5/H4C13/H4C14/H4-16/H4C15/RAD51C
/H2AC7/BRCA1
R-HSA-5625886                                                          H2BC
11/H2BC14/H2AC8/H3C4/H3C1/H3C3/H3C6/H3C11/H3C8/H3C12/H3C10/H3C2/H3C7/H4C9/
H4C1/H4C4/H4C6/H4C12/H4C11/H4C3/H4C8/H4C2/H4C5/H4C13/H4C14/H4-16/H4C15/H2A
C7/KLK3/KLK2
R-HSA-9616222                                     H2BC11/H2BC14/H2AC8/H3C4
/H3C1/H3C3/H3C6/H3C11/H3C8/H3C12/H3C10/H3C2/H3C7/CSF3R/TFDP2/H4C9/H4C1/H4C
4/H4C6/H4C12/H4C11/H4C3/H4C8/H4C2/H4C5/H4C13/H4C14/H4-16/H4C15/SPI1/MYB/H2
AC7/TAL1/PML
R-HSA-3214858                                       H2AC16/H2AC8/H3C4/H3C1
/H3C3/H3C6/H3C11/H3C8/H3C12/H3C10/H3C2/H3C7/H2AC11/H2AC13/H2AC15/H2AC17/H2
AC12/H4C9/H4C1/H4C4/H4C6/H4C12/H4C11/H4C3/H4C8/H4C2/H4C5/H4C13/H4C14/H4-16
/H4C15/H2AC7
              Count
R-HSA-3214815    38
R-HSA-2299718    32
R-HSA-912446     34
R-HSA-5625886    30
R-HSA-9616222    34
R-HSA-3214858    32

###################################
Comparison:  AQ
                        ID
R-HSA-400511 R-HSA-400511
R-HSA-400508 R-HSA-400508

Description
R-HSA-400511 Synthesis, secretion, and inactivation of Glucose-dependent I
nsulinotropic Polypeptide (GIP)
R-HSA-400508                                                 Incretin synthes
is, secretion, and inactivation
             GeneRatio  BgRatio       pvalue   p.adjust     qvalue     gen
eID
```

```
R-HSA-400511      2/19 13/10616 0.0002339579 0.01871663 0.01305239 DPP4/GA
TA4
R-HSA-400508      2/19 23/10616 0.0007508072 0.03003229 0.02094357 DPP4/GA
TA4
             Count
R-HSA-400511     2
R-HSA-400508     2

###################################
Comparison:  CQ
                          ID                                  Description
R-HSA-3214815 R-HSA-3214815                 HDACs deacetylate histones
R-HSA-2299718 R-HSA-2299718       Condensation of Prophase Chromosomes
R-HSA-912446   R-HSA-912446                      Meiotic recombination
R-HSA-73728     R-HSA-73728          RNA Polymerase I Promoter Opening
R-HSA-427359   R-HSA-427359 SIRT1 negatively regulates rRNA expression
R-HSA-3214858 R-HSA-3214858           RMTs methylate histone arginines
              GeneRatio  BgRatio       pvalue     p.adjust       qvalue
R-HSA-3214815    35/684 94/10616 1.824816e-18 1.119036e-15 9.177813e-16
R-HSA-2299718    31/684 73/10616 1.837497e-18 1.119036e-15 9.177813e-16
R-HSA-912446     33/684 85/10616 3.927225e-18 1.594453e-15 1.307697e-15
R-HSA-73728      28/684 62/10616 1.146650e-17 3.051054e-15 2.502333e-15
R-HSA-427359     29/684 67/10616 1.252485e-17 3.051054e-15 2.502333e-15
R-HSA-3214858    31/684 79/10616 2.874516e-17 5.643304e-15 4.628378e-15

geneID
R-HSA-3214815 H2BC11/H2BC14/H3C4/H2AC16/H2AC8/H3C11/REST/H2AC11/H2AC13/H2A
C15/H2AC17/H2AC12/H4C9/H4C1/H4C4/H4C6/H4C12/H4C11/H4C3/H4C8/H4C2/H4C5/H4C1
3/H4C14/H4-16/H4C15/H3C1/H3C3/H3C6/H3C8/H3C12/H3C10/H3C2/H3C7/H2BC15
R-HSA-2299718                                      H2BC11/H2BC14/SMC2/H3C4/H2AC
8/CDK1/H3C11/H4C9/H4C1/H4C4/H4C6/H4C12/H4C11/H4C3/H4C8/H4C2/H4C5/H4C13/H4C
14/H4-16/H4C15/H3C1/H3C3/H3C6/H3C8/H3C12/H3C10/H3C2/H3C7/H2BC15/PLK1
R-HSA-912446                         H2BC11/H2BC14/H3C4/BLM/H2AC8/RPA3/H3C11/
H4C9/H4C1/H4C4/H4C6/H4C12/H4C11/H4C3/H4C8/H4C2/H4C5/H4C13/H4C14/H4-16/H4C1
5/RAD51/H3C1/H3C3/H3C6/H3C8/H3C12/H3C10/H3C2/H3C7/BRCA2/H2BC15/BRCA1
R-HSA-73728                                                  H2BC11/H2BC14
/H3C4/H2AC8/H3C11/H4C9/H4C1/H4C4/H4C6/H4C12/H4C11/H4C3/H4C8/H4C2/H4C5/H4C1
3/H4C14/H4-16/H4C15/H3C1/H3C3/H3C6/H3C8/H3C12/H3C10/H3C2/H3C7/H2BC15
R-HSA-427359                                               H2BC11/H2BC14/H3C4/
H2AC8/H3C11/H4C9/H4C1/H4C4/H4C6/H4C12/H4C11/H4C3/H4C8/H4C2/H4C5/H4C13/H4C1
4/H4-16/H4C15/H3C1/H3C3/H3C6/H3C8/H3C12/H3C10/H3C2/H3C7/H2BC15/TAF1B
R-HSA-3214858                            H3C4/H2AC16/H2AC8/H3C11/H2AC11/H2A
C13/H2AC15/H2AC17/H2AC12/H4C9/H4C1/H4C4/H4C6/H4C12/H4C11/H4C3/H4C8/H4C2/H4
C5/H4C13/H4C14/H4-16/H4C15/H3C1/H3C3/H3C6/H3C8/H3C12/H3C10/H3C2/H3C7
              Count
R-HSA-3214815    35
R-HSA-2299718    31
R-HSA-912446     33
R-HSA-73728      28
R-HSA-427359     29
R-HSA-3214858    31
```

El archivo generado es de clase *enrichResult*.

```
[1] "enrichResult"
attr(,"package")
[1] "DOSE"
```

Para poder analizar el contenido se convierte en data frame. Se presenta la lista de objetos:

```
[1] "ID"          "Description" "GeneRatio"   "BgRatio"     "pvalue"
[6] "p.adjust"    "qvalue"      "geneID"      "Count"
```

Las dimensiones del archivo son:

```
[1] 134   9
```

Se muestra una selección de la información contenida en el data frame (5 primeras filas, columnas 1, 4, 6 y7; ver la lista con los nombres de los objetos).

	ID	BgRatio	p.adjust	qvalue
R-HSA-3214815	R-HSA-3214815	94/10616	0	0
R-HSA-2299718	R-HSA-2299718	73/10616	0	0
R-HSA-912446	R-HSA-912446	85/10616	0	0
R-HSA-73728	R-HSA-73728	62/10616	0	0
R-HSA-427359	R-HSA-427359	67/10616	0	0

Se averigua la información disponible en el archivo siguiendo otra opción:

- $geneID

```
[1] "H2BC11/H2BC14/H3C4/H2AC16/H2AC8/H3C11/REST/H2AC11/H2AC13/H2AC15/H2AC1
7/H2AC12/H4C9/H4C1/H4C4/H4C6/H4C12/H4C11/H4C3/H4C8/H4C2/H4C5/H4C13/H4C14/H
4-16/H4C15/H3C1/H3C3/H3C6/H3C8/H3C12/H3C10/H3C2/H3C7/H2BC15"
[2] "H2BC11/H2BC14/SMC2/H3C4/H2AC8/CDK1/H3C11/H4C9/H4C1/H4C4/H4C6/H4C12/H4
C11/H4C3/H4C8/H4C2/H4C5/H4C13/H4C14/H4-16/H4C15/H3C1/H3C3/H3C6/H3C8/H3C12/
H3C10/H3C2/H3C7/H2BC15/PLK1"
[3] "H2BC11/H2BC14/H3C4/BLM/H2AC8/RPA3/H3C11/H4C9/H4C1/H4C4/H4C6/H4C12/H4C
11/H4C3/H4C8/H4C2/H4C5/H4C13/H4C14/H4-16/H4C15/RAD51/H3C1/H3C3/H3C6/H3C8/H
3C12/H3C10/H3C2/H3C7/BRCA2/H2BC15/BRCA1"
[4] "H2BC11/H2BC14/H3C4/H2AC8/H3C11/H4C9/H4C1/H4C4/H4C6/H4C12/H4C11/H4C3/H
4C8/H4C2/H4C5/H4C13/H4C14/H4-16/H4C15/H3C1/H3C3/H3C6/H3C8/H3C12/H3C10/H3C2
/H3C7/H2BC15"
[5] "H2BC11/H2BC14/H3C4/H2AC8/H3C11/H4C9/H4C1/H4C4/H4C6/H4C12/H4C11/H4C3/H
4C8/H4C2/H4C5/H4C13/H4C14/H4-16/H4C15/H3C1/H3C3/H3C6/H3C8/H3C12/H3C10/H3C2
/H3C7/H2BC15/TAF1B"
[6] "H3C4/H2AC16/H2AC8/H3C11/H2AC11/H2AC13/H2AC15/H2AC17/H2AC12/H4C9/H4C1/
H4C4/H4C6/H4C12/H4C11/H4C3/H4C8/H4C2/H4C5/H4C13/H4C14/H4-16/H4C15/H3C1/H3C
3/H3C6/H3C8/H3C12/H3C10/H3C2/H3C7"
```

- $geneID

enrich.result$ID
R-HSA-3214815
R-HSA-2299718
R-HSA-912446

R-HSA-73728
R-HSA-427359
R-HSA-3214858

La información generada con este análisis de significación biológica es la siguiente:

(i) Un archivo .csv con un resumen de los "pathways" enriquecidos y los estadísticos asociados (p-valor, q-valor, p-valor ajustado, etc);

(ii) Un diagrama de barras con las mejores enriched pathways. En este último tipo de resultado, la altura/longitud de una barra indica el número de genes relacionado con ese pathway. Por otra parte, los pathways están ordenados por significado estadístico (valor p ajustado).

(iii) Un gráfico con una red de las enriched pathways y una relación entre los genes incluidos.

La Figura 20 muestra la red creada con los genes seleccionados.

Resumen de resultados

Se crea una tabla con una lista de los archivos creados durante el estudio.

Lista de archivos generados en el estudio

Lista
data4Heatmap.csv
GSE7463_Limma_Coeff.txt
GSE7463_RMA_Norm.txt
ReactomePA.Results.AC.csv
ReactomePA.Results.AQ.csv
ReactomePA.Results.CQ.csv
ReactomePABarplot.AC.pdf
ReactomePABarplot.AQ.pdf
ReactomePABarplot.CQ.pdf
ReactomePAcnetplot.AC.pdf
ReactomePAcnetplot.AQ.pdf
ReactomePAcnetplot.CQ.pdf
topAnnotated_AC.csv
topAnnotated_AQ.csv
topAnnotated_CQ.csv

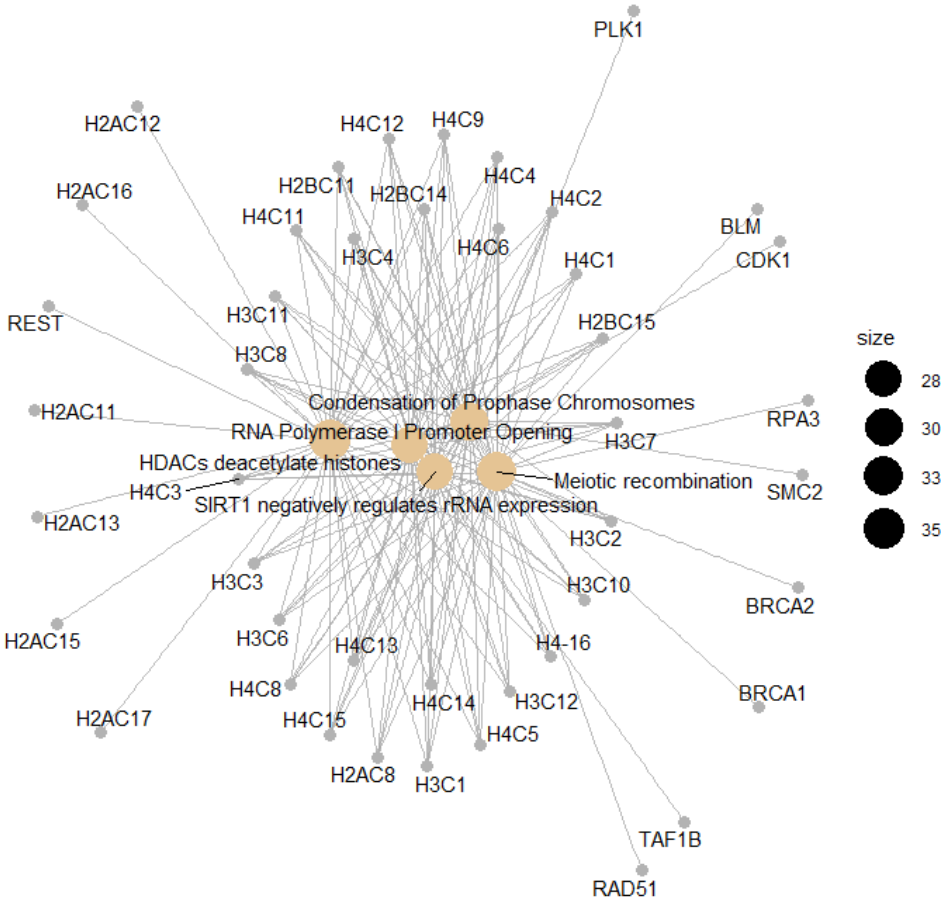

Figura 20. Términos GO y genes asociados.

Notas

A - El proceso de normalización ha producido un resultado que se puede contrastar con la figura disponible en GEO. Se puede comprobar que este es el resultado presentado en la Figura 9.

B - Es discutible que el resultado obtenido en la etapa de filtrado sea el más adecuado. Después de varios intentos, se llegó a la conclusión de que era mejor analizar un número de genes relativamente elevado.

C - En el inicio del análisis de significación biológica se realiza un proceso de anotación en el que se origina varias entradas "NA" en alguna tabla. La prueba realizada proporciona las siguientes respuestas:

```
NAs en Tabla AC?

[1] TRUE

NAs en Tabla AQ?

[1] FALSE

NAs en Tabla CQ?

[1] TRUE
```

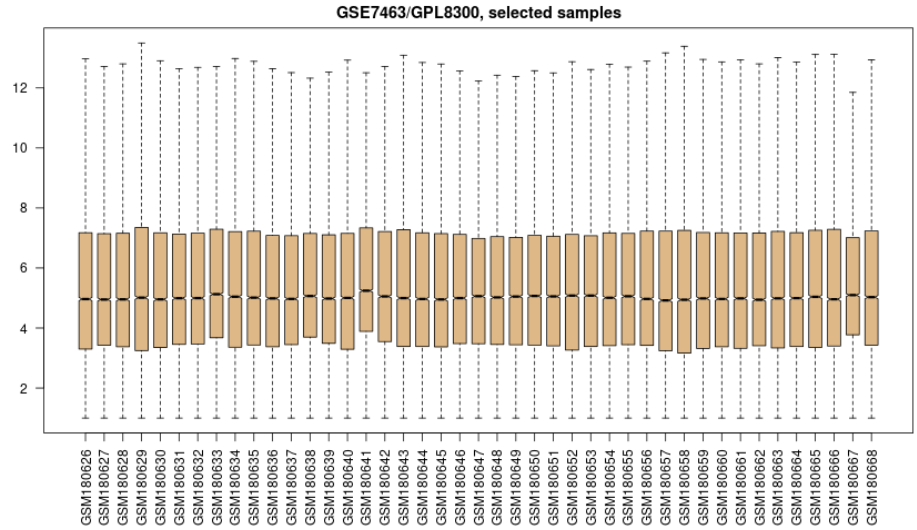

Figura 21. Boxplot - GEO-GSE7463.

Esto no sido un inconveniente a la hora de usar la librería **ReactomePA** para realizar un ORA. Sin embargo, dificulta la realización de un GSEA ya que esta librería requiere que la lista de genes esté ordenada.

D - El número de genes que hay que analizar con las tres comparaciones realizadas (AC0, AC1, C0C1) es muy pequeño en la segunda comparación. Esto puede ser un problema con la librería **ReactomePA** Para aumentar el número de genes se ha elevado el umbral en la etapa de filtrado hasta 15. En tal caso el número de genes que hay que analizar con **ReactomePA** asciende a 1325, 350 y 1329 en las tres comparaciones, y no presenta ningún problema para llevar a cabo el análisis con todas ellas. Ver diagrama de Venn.

E - La Figura 22 resume el proceso de análisis de datos.

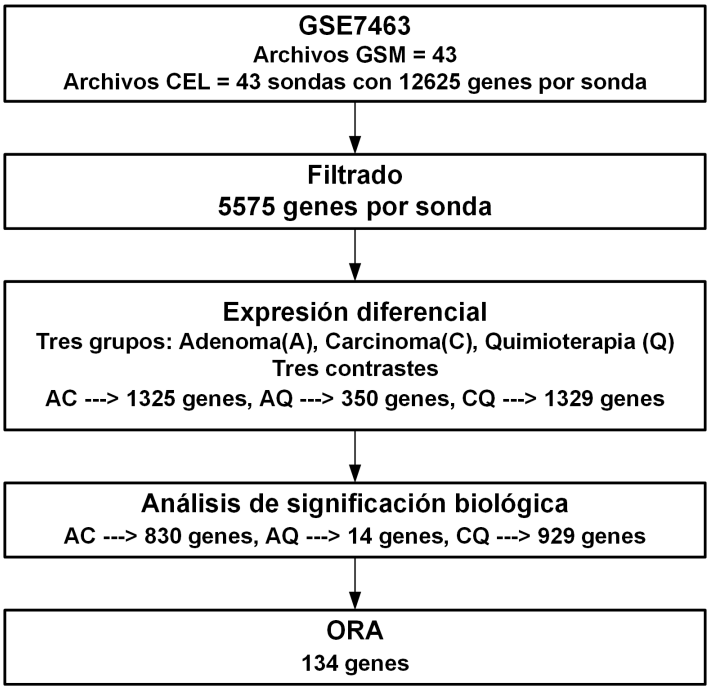

Figura 22. Resumen.

Información de la sesión

Se solicita un resumen con la información de la sesión.

```
R version 3.6.3 (2020-02-29)
Platform: x86_64-w64-mingw32/x64 (64-bit)
Running under: Windows 10 x64 (build 18362)

Matrix products: default

locale:
[1] LC_COLLATE=Spanish_Spain.1252  LC_CTYPE=Spanish_Spain.1252
[3] LC_MONETARY=Spanish_Spain.1252 LC_NUMERIC=C
[5] LC_TIME=C

attached base packages:
 [1] stats4    tools     parallel  stats     graphics  grDevices utils
 [8] datasets  methods   base

other attached packages:
 [1] ReactomePA_1.30.0          gplots_3.0.3
 [3] hgu95av2.db_3.2.3          org.Hs.eg.db_3.10.0
 [5] AnnotationDbi_1.48.0       IRanges_2.20.2
 [7] S4Vectors_0.24.3           genefilter_1.68.0
 [9] RColorBrewer_1.1-2         limma_3.42.2
[11] hgu95av2cdf_2.18.0         affy_1.64.0
[13] MVA_1.0-6                  HSAUR2_1.1-17
[15] arrayQualityMetrics_3.42.0 GEOquery_2.54.1
[17] Biobase_2.46.0             BiocGenerics_0.32.0
[19] printr_0.1                 knitr_1.28

loaded via a namespace (and not attached):
  [1] backports_1.1.5        Hmisc_4.4-0            fastmatch_1.1-0
  [4] systemfonts_0.1.1      igraph_1.2.5           plyr_1.8.6
  [7] splines_3.6.3          BiocParallel_1.20.1    GenomeInfoDb_1.22.1
 [10] ggplot2_3.3.0          urltools_1.7.3         digest_0.6.25
 [13] htmltools_0.4.0        GOSemSim_2.12.1        viridis_0.5.1
 [16] GO.db_3.10.0           gdata_2.18.0           magrittr_1.5
 [19] checkmate_2.0.0        memoise_1.1.0          affyPLM_1.62.0
 [22] cluster_2.1.0          gcrma_2.58.0           graphlayouts_0.6.0
 [25] Biostrings_2.54.0      readr_1.3.1            annotate_1.64.0
 [28] beadarray_2.36.0       svglite_1.2.3          askpass_1.1
 [31] prettyunits_1.1.1      enrichplot_1.6.1       jpeg_0.1-8.1
 [34] colorspace_1.4-1       rappdirs_0.3.1         ggrepel_0.8.2
 [37] blob_1.2.1             xfun_0.13              dplyr_0.8.5
 [40] crayon_1.3.4           RCurl_1.98-1.1         jsonlite_1.6.1
 [43] hexbin_1.28.1          graph_1.64.0           survival_3.1-11
 [46] glue_1.3.2             polyclip_1.10-0        gtable_0.3.0
 [49] zlibbioc_1.32.0        XVector_0.26.0         BeadDataPackR_1.38.0
 [52] graphite_1.32.0        scales_1.1.0           DOSE_3.12.0
 [55] setRNG_2013.9-1        vsn_3.54.0             DBI_1.1.0
 [58] Rcpp_1.0.3             progress_1.2.2         viridisLite_0.3.0
 [61] xtable_1.8-4           htmlTable_1.13.3       gridGraphics_0.5-0
```

```
 [64] reactome.db_1.70.0     europepmc_0.3          foreign_0.8-75
 [67] bit_1.1-15.2           preprocessCore_1.48.0  Formula_1.2-3
 [70] httr_1.4.1             htmlwidgets_1.5.1      fgsea_1.12.0
 [73] acepack_1.4.1          ellipsis_0.3.0         pkgconfig_2.0.3
 [76] XML_3.99-0.3           farver_2.0.3           nnet_7.3-12
 [79] ggplotify_0.0.5        tidyselect_1.0.0       labeling_0.3
 [82] rlang_0.4.5            reshape2_1.4.3         munsell_0.5.0
 [85] RSQLite_2.2.0          ggridges_0.5.2         evaluate_0.14
 [88] stringr_1.4.0          yaml_2.2.1             bit64_0.9-7
 [91] tidygraph_1.1.2        caTools_1.18.0         purrr_0.3.3
 [94] ggraph_2.0.2           DO.db_2.9              xml2_1.2.5
 [97] compiler_3.6.3         rstudioapi_0.11        curl_4.3
[100] png_0.1-7              affyio_1.56.0          tweenr_1.0.1
[103] tibble_2.1.3           stringi_1.4.6          highr_0.8
[106] gdtools_0.2.1          lattice_0.20-38        Matrix_1.2-18
[109] vctrs_0.2.4            pillar_1.4.3           lifecycle_0.2.0
[112] BiocManager_1.30.10    triebeard_0.3.0        cowplot_1.0.0
[115] data.table_1.12.8      bitops_1.0-6           GenomicRanges_1.38.0
[118] qvalue_2.18.0          R6_2.4.1               latticeExtra_0.6-29
[121] hwriter_1.3.2          KernSmooth_2.23-16     gridSVG_1.7-1
[124] gridExtra_2.3          codetools_0.2-16       MASS_7.3-51.5
[127] gtools_3.8.2           assertthat_0.2.1       openssl_1.4.1
[130] GenomeInfoDbData_1.2.2 hms_0.5.3              grid_3.6.3
[133] rpart_4.1-15           tidyr_1.0.2            base64_2.0
[136] rvcheck_0.1.8          rmarkdown_2.1          illuminaio_0.28.0
[139] ggforce_0.3.1          base64enc_0.1-3
```

Referencias

[1] C.S. Moreno et al. (2007) "Evidence that p53-mediated cell-cycle-arrest inhibits chemotherapeutic treatment of ovarian carcinomas," PLoS ONE, vol. 2, no. 5, e441, May 2007. https://doi.org/10.1371/journal.pone.0000441.

[2] C.D. Scharer et al. "Aurora kinase inhibitors synergize with paclitaxel to induce apoptosis in ovarian cancer cells," J Transl Med, vol. 6, December 2008. https://doi.org/10.1186/1479-5876-6-79.

[3] S. Imbeaud and C. Auffray, "The 39 steps'in gene expression profiling: critical issues and proposed best practices for microarray experiments," Drug Discovery Today, vol. 10, no. 17, pp.1175-1182, September 2005. https://doi.org/10.1016/S1359-6446(05)03565-8.

[4] R. Gonzalo and A. Sanchez-Pla, "Statistical Analysis of Microarray data," March 2020. Disponible en https://github.com/ASPteaching/Omics_Data_Analysis-Case_Study_1-Microarrays.

[5] Ver https://www.stat.purdue.edu/bigtap/online/docs/Introduction_to_Microarray_Analysis_GSE15947.html.

[6] Ver https://rdrr.io/bioc/Biobase/man/class.ExpressionSet.html.

[7] Ver https://www.rdocumentation.org/packages/Biobase/versions/2.32.0/topics/eSet.

[8] MIAME (Minimum information about a microarray experiment). Ver http://fged.org/projects/miame/

[9] C. Genolini, "A (not so) short introduction to S4-Object Oriented Programming in R V0.5.1," August 2008. Ver https://cran.r-project.org/doc/contrib/Genolini-S4tutorialV0-5en.pdf.

[10] CEL format https://www.affymetrix.com/support/developer/powertools/changelog/gcos-agcc/cel.html.

[11] P. Stafford (Editor), Methods in Microarray Normalization, CRC Press, 2008. ISBN 9781420052787.

[12] R.A. Irizarry et al., "Exploration, normalization, and summaries of high density oligonucleotide array probe level data," Biostatistics, vol. 4, no. 2, pp. 249–264, 2003. Ver https://doi.org/10.1093/biostatistics/4.2.249.

[13] L. Gautier, R. Irizarry, L. Cope, and B. Bolstad, "Description of affy," October 2019. Ver https://www.bioconductor.org/packages/release/bioc/vignettes/affy/inst/doc/affy.pdf.

[14] Package affy, April 2020. Ver en https://www.bioconductor.org/packages/release/bioc/manuals/affy/man/affy.pdf.

[15] V.G. Tusher, R. Tibshirani, and G. Chu, "Significance analysis of microarrays applied to the ionizing radiation response," Proc. of the National Academy of Sciences of the United States of America (PNAS), vol. 98, no. 9, pp. 5116–5121, April 2001. Vere n https://doi.org/10.1073/pnas.091062498.

[16] G.K. Smyth, "Linear Models and Empirical Bayes Methods for Assessing Differential Expression in Microarray Experiments," Statistical Applications in Genetics and Molecular Biology, vol. 3, no. 1, pp. 1-25, 2004. https://doi.org/10.2202/1544-6115.1027.

[17] G.K. Smyth, "limma: Linear Models for Microarray Data," In Bioinformatics and Computational Biology Solutions Using R and Bioconductor, pp. 397–420, R. Gentleman et al. (Editors), New York: Springer-Verlag, 2005.

[18] Y. Benjamini and Y. Hochberg, "Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing," Journal of the Royal Statistical Society, vol. 57, no. 1, pp. 289-300, 1995, https://doi.org/10.1111/j.2517-6161.1995.tb02031.x

[19] Package limma. Ver https://bioconductor.org/packages/release/bioc/manuals/limma/man/limma.pdf.

[20] P. Khatri, M. Sirota, and A. J. Butte, "Ten years of pathway analysis: current approaches and outstanding challenges," PLOS Computational Biology, vol. 8, no. 2, e1002375, 2012, https://doi.org/10.1371/journal.pcbi.1002375.

[21] Y. Guangchuang and Q.Y. He, "ReactomePA: an R/Bioconductor package for reactome pathway analysis and visualization," Molecular BioSystems, vol. 12, no. 2, pp. 477-479, 2016. Ver https://doi.org/10.1039/C5MB00663E.