# PEC1 Informe

Beatriz Pardo Montenegro

03/05/2020

## Table of Contents



URL GitHub: https://github.com/bpardom/AO_PEC1.git

## 1.Abstract

La progresión metastásica en pacientes con cáncer de mama se asocia a la resistencia a terapias como la quimioterapia. Estudiamos las respuestas genéticas de células metastásicas de cáncer de mama tratadas con Paclitaxel y observamos una actividad

elevada de JNK así como de la expresión de SPP1 o TNC. Su inhibición puede ser una futura estrategia terapéutica en el cáncer de mama metastásico Insua-Rodríguez (2018).

# 2.Objetivos

Sabiendo que un gran problema en el tratamiento de las metástasis de cáncer de mama es la resistencia de estas a la quimioterapia, el objetivo de este estudio es estudiar qué genes y procesos intervienen en la resistencia a la quimioterapia de las células metastásicas para buscar así nuevas dianas terapéuticas.

# 3.Materiales y métodos

## 3.1 Naturaleza de los datos

Entro en la base de datos GEO, tal y como se suguiere en el enunciado de la PEC, y selecciono el estudio cuya referencia es GSE98238[1].

## 3.2 Tipo de experimento

El tipo de experimento que planteo es de Comparación de grupos o class comparison. El objetivo de los estudios comparativos es determinar si los perfiles de expresión génica difieren entre grupos previamente identificados, así como seleccionar genes diferencialmente expresados[2]. En mi experimento comparo 2 grupos, uno tratado con DMSO y otro con Paclitaxel.

## 3.3 Diseño general

Mi estudio lo componen 2 grupos, el primero son 3 muestras de la línea celular MDA231-LM2 tratadas con un vehículo (DMSO) y el segundo grupo son 3 muestras de la misma línea celular tratadas con Paclitaxel 5 nM durante 48 horas.

## 3.4 Materiales

Cultivo celular: línea celular MDA231-LM2, es una de las más utilizadas para el estudio experimental in vitro del cáncer de mama hormono-independiente. Estas células se aislaron en 1973 a partir de una muestra de derrame pleural de una paciente con cáncer de mama que falleció en Houston, estas células presentan un crecimiento extraordinariamente rápido en medios de cultivo poco enriquecidos.

---

[1] https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE98238

[2] http://materials.cv.uoc.edu/daisy/Materials/PID_00192730/pdf/PID_00192743.pdf

Array: Affymetrix GeneChip® Human Genome U133 Plus 2.0 microarray.

## 3.5 Ánalisis de datos

### 3.5.1 Datos y grupos

Tal y como se puede ver en el diseño del estudio, tengo 6 muestras divididas en 2 grupos. El primero de ellos son 3 muestras de células tratadas con DMSO y el segundo grupo son 3 muestras de células tratadas con Paclitaxel. En mi estudio se utiliza el array de Affymetrix que pertenece al grupo de los de un color. El resultado de escanear la imagen de este tipo de arrays es un archivo de extensión .CEL. Accedo a la web y me descargo los 6 archivos .CEL, uno por cada muestra[3].

### 3.5.2 Control de calidad de los datos crudos

A través de GEOquery, sin necesidad de descargarse ningún archivo, se pueden cargar los datos de cualquier estudio. Esto es lo primero que hago pero al hacer el Control de calidad de los datos crudos a partir del expressionset puedo comprobar que o bien los datos han sido tratados previamente o que los datos son comparables sin necesidad de normalizar.

[3] https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE98238

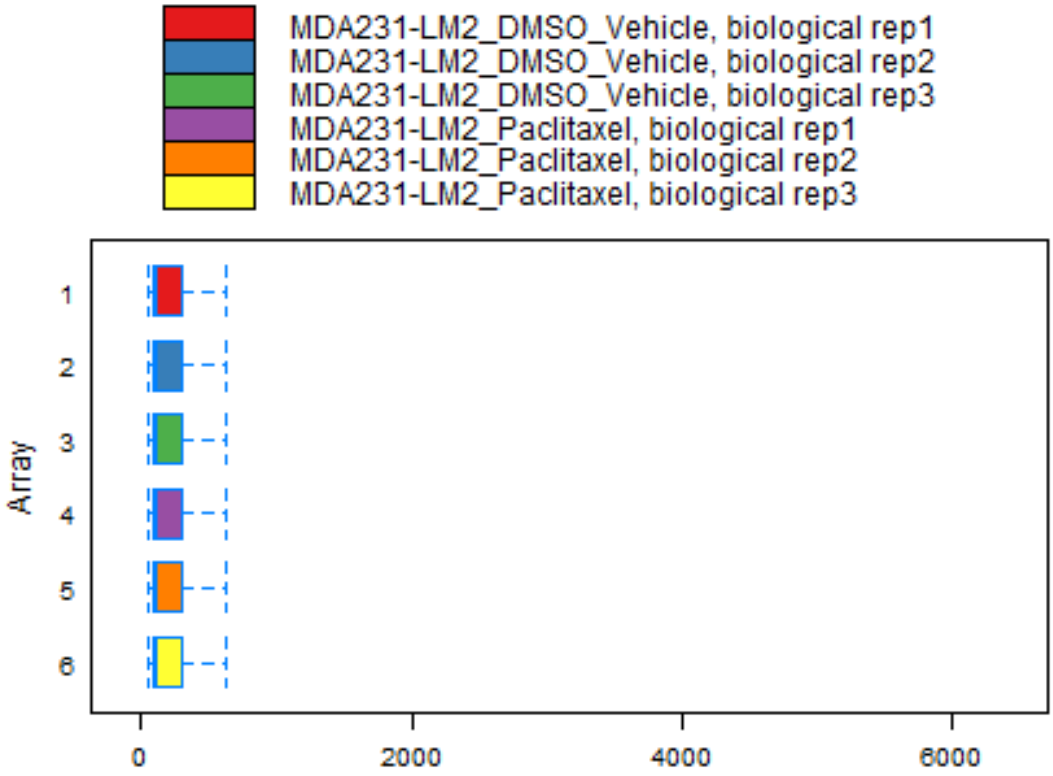



*Figura1: Boxplot aplicando control de calidad de datos cargados con GEOquery*

Cargo los datos de los archivos CEL descargados previamente. Instalo el paquete affy y con la función ReadAffy leo los archivos. Realizo sobre estos datos el control de calidad de los datos crudos con la función arrayQualityMetrics.Se genera un informe en el directorio elegido, en él puedo consultar el boxplot pero en este caso el boxplot lo genero yo con la función boxplot.

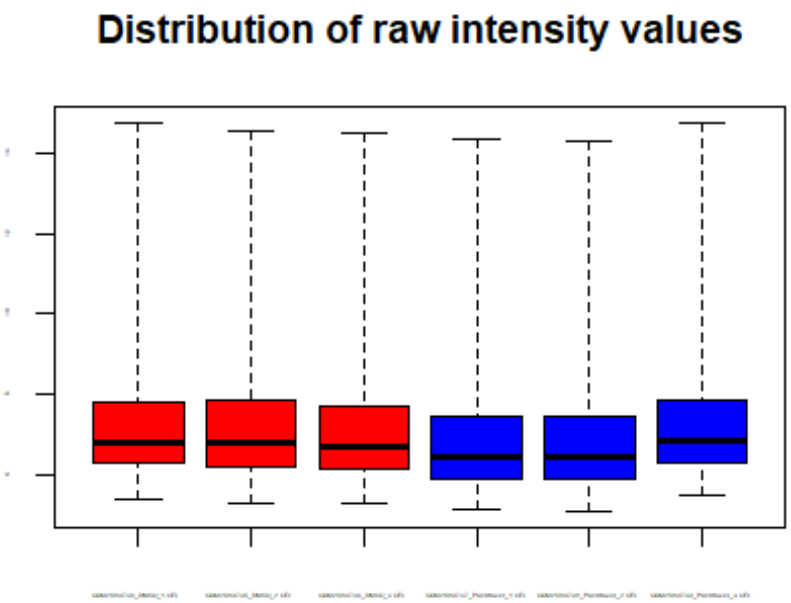



*Figura2: Boxplot de datos cargados desde archivos CEL*

Se observa un mínimo desplazamiento de unos arrays respecto a otros pero el resultado es realmente bueno sin necesidad de rechazar ninguna muestra. Los 3 primeros, en rojo, son los tratados con DMSO y en azul, los tratados con Paclitaxel.

En el archivo index.html que se genera tras realizar el control de calidad, en el resumen inicial aparece el análisis de los datos evaluando varios criterios de calidad. En este primero se evalún 6 criterios y solamente en el array 2 y en el 5 aparece una cruz en alguno de los criterios. Con sólo una cruz no es necesario descartar dichos arrays.

| | array | sampleNames | *1 | *2 | *3 | *4 | *5 | *6 | sample | ScanDate |
|---|---|---|---|---|---|---|---|---|---|---|
| ☐ | 1 | GSM2589734_DMSO_1.CEL | | | | | | | 1 | 2016-04-06T14:12:58Z |
| ☑ | 2 | GSM2589735_DMSO_2.CEL | | | x | | | | 2 | 2016-04-06T11:31:14Z |
| ☐ | 3 | GSM2589736_DMSO_3.CEL | | | | | | | 3 | 2016-04-06T11:42:27Z |
| ☐ | 4 | GSM2589737_Paclitaxel_1.CEL | | | | | | | 4 | 2016-04-06T13:40:49Z |
| ☑ | 5 | GSM2589738_Paclitaxel_2.CEL | | | | | x | | 5 | 2016-04-06T12:27:24Z |
| ☐ | 6 | GSM2589739_Paclitaxel_3.CEL | | | | | | | 6 | 2016-04-06T13:12:36Z |

*Figura3: Tabla resumen de index*

En el gráfico de análisis de componentes principales generado por arrayQualityMetrics con estos datos sin normalizar, detecto que los arrays 2 y 3, ambos de las líneas celulares tratadas con DMSO y los arrays 4 y 5, de las tratadas con

Paclitaxel, se agrupan de forma natural. Sin embargo en el array 1 y el 6 no se ve proximidad en este gráfico.

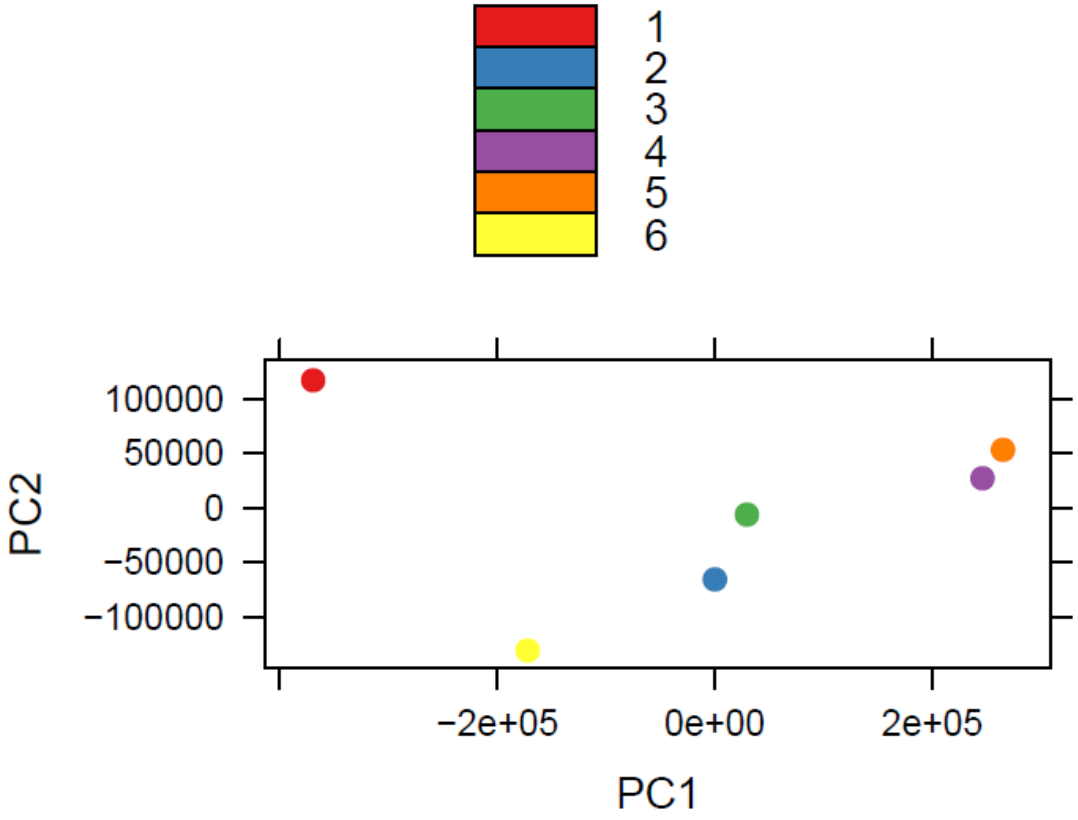



*Figura4:PCA de index*

### 3.5.3 Normalización

Una vez que acepto que la calidad de los datos es aceptable, paso a la normalización. Muchas son las ventajas de este paso del análisis de datos, las dos principales: nos permite eliminar el ruido de fondo y nos permite hacer comparables todos los valores del estudio. Uno de los problemas de la normalización es que podemos detectar sesgos, ya que todos los datos se han igualado y no podemos eliminarlos, por lo que el análisis de calidad previo es importante.

Hay distintas formas de normalizar, al tratarse de un array de Affymetrix voy a utilzar el método RMA. Es un método basado en la modelización de las intensidades de las sondas que, en vez de basarse en las distintas sondas de un gen dentro de un mismo array, se basa en los distintos valores de la misma sonda entre todos los arrays disponibles[4].

[4] http://materials.cv.uoc.edu/daisy/Materials/PID_00192730/pdf/PID_00192743.pdf

```
affyceles <- rma(affycel)

## Background correcting
## Normalizing
## Calculating Expression
```

### 3.5.4 Control de calidad de los datos normalizados

Este es un paso opcional en el pipeline del análisis del enunciado de la PEC1. Lo realizo ya que se pueden observar muy fácilmente los efectos de la normalización de los datos. En el resumen del archivo index.html puedo comprobar que ya no aparece ninguna cruz en ninguno de los 3 criterios de calidad que aplica, esto quiere decir que los datos son comparables.

| | array | sampleNames | *1 | *2 | *3 | sample | ScanDate |
|---|---|---|---|---|---|---|---|
| ☐ | 1 | GSM2589734_DMSO_1.CEL | | | | 1 | 2016-04-06T14:12:58Z |
| ☐ | 2 | GSM2589735_DMSO_2.CEL | | | | 2 | 2016-04-06T11:31:14Z |
| ☐ | 3 | GSM2589736_DMSO_3.CEL | | | | 3 | 2016-04-06T11:42:27Z |
| ☐ | 4 | GSM2589737_Paclitaxel_1.CEL | | | | 4 | 2016-04-06T13:40:49Z |
| ☐ | 5 | GSM2589738_Paclitaxel_2.CEL | | | | 5 | 2016-04-06T12:27:24Z |
| ☐ | 6 | GSM2589739_Paclitaxel_3.CEL | | | | 6 | 2016-04-06T13:12:36Z |

*Figura5: Tabla resumen de index datos normalizados*

En el análisis de componentes principales podemos observar como se agrupan los arrays del grupo tratado con DMSO y los arrays del grupo tratado con Paclitaxel.

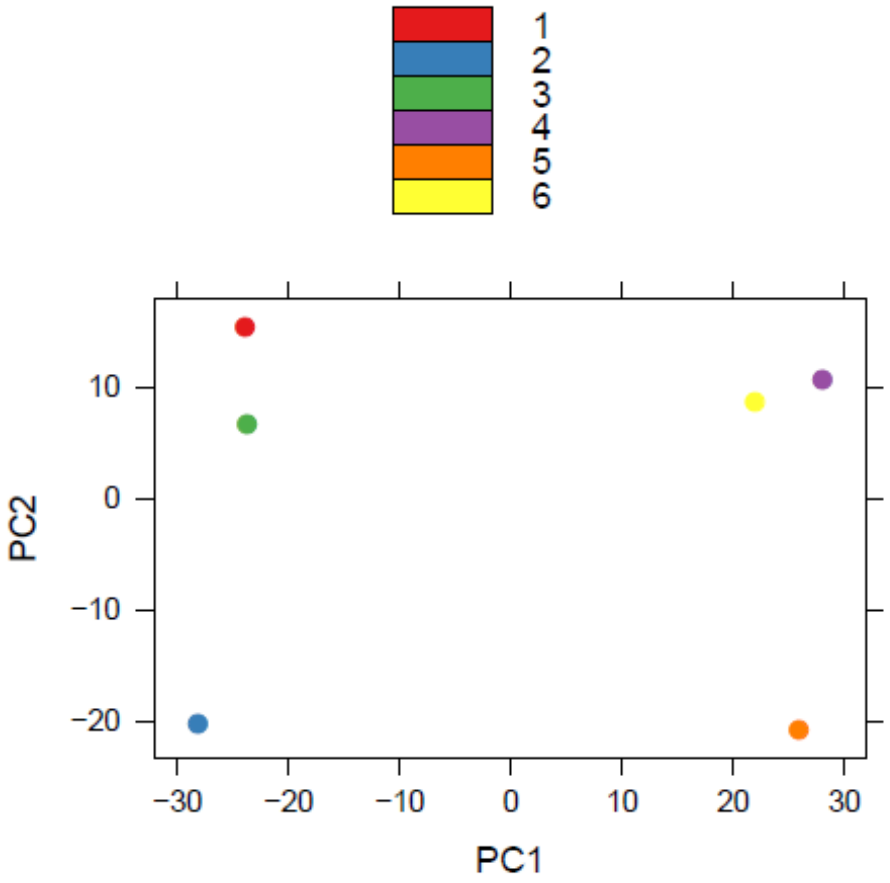



*Figura6: PCA datos normalizados*

Realizo también el boxplot donde se puede ver claramente que los valores están en una escala en donde se pueden comparar.

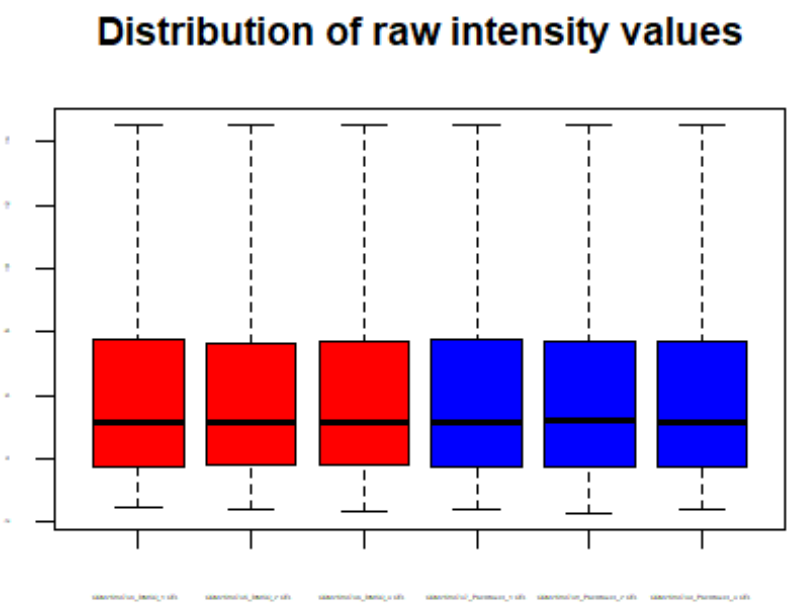



*Figura7: Boxplot de datos normalizados*

Los arrays de Affymetrix contienen millones de sondas por lo que no pueden examinarse a simple vista. A pesar de ello se puede obtener por ejemplo un MAplot de un canal. La única forma de definir M (el log ratio) es comparar entre cada array o bien con otros arrays o bien con un array de referencia creado, por ejemplo tomando gen a gen la mediana de todas las expresiones[5].

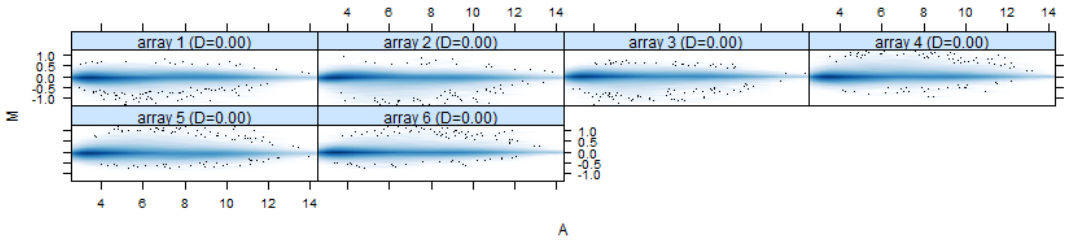



*Figura8: MAplot de un canal*

[5] http://materials.cv.uoc.edu/daisy/Materials/PID_00192730/pdf/PID_00192743.pdf

### 3.5.5 Filtrado de genes

Utilizo la función nsFilter del paquete genefilter para quitar los genes que no superan un umbral de expresión, y que por lo tanto no se espera que estén diferencialmente expresados.

### 3.5.6 Identificación de genes diferencialmente expresados

Realizo el análisis basado en modelos lineales. Para la aplicación de dicho modelo lineal, lo primero que hago es definir la matriz de diseño.

```
##                              DMSO Paclitaxel
## GSM2589734_DMSO_1.CEL           1          0
## GSM2589735_DMSO_2.CEL           1          0
## GSM2589736_DMSO_3.CEL           1          0
## GSM2589737_Paclitaxel_1.CEL     0          1
## GSM2589738_Paclitaxel_2.CEL     0          1
## GSM2589739_Paclitaxel_3.CEL     0          1
```

Con el modelo lineal definido a través de una matriz de diseño, pueden formularse las preguntas de interés como contrastes, es decir, comparaciones entre los parámetros del modelo. Para ello creo la matriz de contrastes donde en mi estudio se compara que la línea celular haya sido tratada con DMSO o con Paclitaxel.

```
##             Contrasts
## Levels       Paclitaxel - DMSO
##   DMSO                      -1
##   Paclitaxel                 1
```

Una vez creadas ambas matrices paso a estimar el modelo donde puedo comparar la expresión de los genes según las muestras hayan sido tratadas con DMSO o Paclitaxel. Podré observar si hay genes diferencialmente expresados.

El análisis proporciona los estadísticos de test habituales como Fold-change o p-valores ajustados, que se utilizan para ordenar los genes de más a menos diferencialmente expresados.

En los arrays de Affymetrix sabemos que cada valor no corresponde a la expresión de un gen sino de una sonda y hay varios valores (sondas) por cada gen.

Fijando el lfc en 1 salen 172 valores sobreexpresados en los tratados con Paclitaxel respecto a los tratados con DMSO.

```
##        Paclitaxel - DMSO
## Down                   0
## NotSig              4872
## Up                   172
```

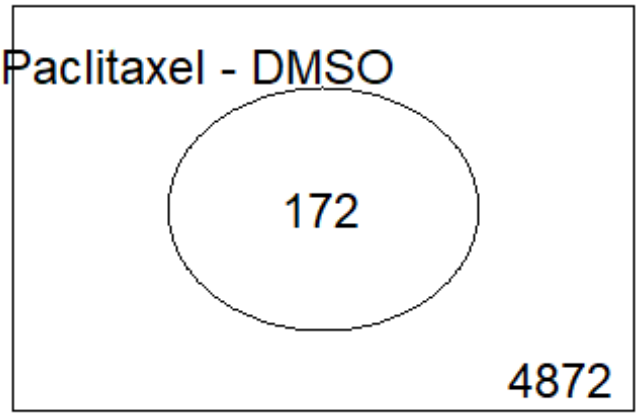



*Figura9: Diagrama genes diferencialmente expresados*

Pruebo a bajar a 0,5 el valor del lfc, así observo 186 genes down regulados y 894 genes up regulados.

```
##        Paclitaxel - DMSO
## Down                186
## NotSig             3964
## Up                  894
```

### 3.5.7 Anotación de los resultados

La identificación de los genes seleccionados será más sencilla si le asigno el nombre y el símbolo del gen. A este proceso se le llama anotación, utilizo la tabla generada por la función toptable y me descargo la libreria específica del array utilizado en mi estudio (Affymetrix GeneChip® Human Genome U133 Plus 2.0 microarray). Haciendo esto se asocian los identificadores que aparecen en la tabla, con carácterísticas como el Símbolo del gen, el identificador del gen (EntrezID) o la descripción del gen.

```
connombres <- annotatedTopTable(toptarmacont,
anotPackage="hgu133plus2.db")

## 'select()' returned 1:1 mapping between keys and columns
```

Ordeno de menor a mayor p valor ajustado y muestro los 4 con menor p valor ajustado.

```
##         PROBEID   SYMBOL ENTREZID
## 117  1553736_at   ZFC3H1   196441
## 1124  204614_at SERPINB2     5055
## 1712  209189_at      FOS     2353
## 4681  239336_at    THBS1     7057
##                                                    GENENAME    logFC
AveExpr
## 117                       zinc finger C3H1-type containing 2.074206
5.248166
## 1124                               serpin family B member 2 3.273535
8.711623
## 1712 Fos proto-oncogene, AP-1 transcription factor subunit 2.424858
8.289016
## 4681                                        thrombospondin 1 2.441313
5.823216
##             t      P.Value    adj.P.Val        B
## 117  19.29552 9.379043e-10 1.460107e-06 12.89797
## 1124 18.91492 1.157896e-09 1.460107e-06 12.70666
## 1712 19.71424 7.472029e-10 1.460107e-06 13.10296
## 4681 19.42717 8.727761e-10 1.460107e-06 12.96303
```

Puede obtenerse una visualización de la expresión diferencial global utilizando gráficos volcano plot. Estas gráficas muestran si hay muchos o pocos genes diferenciados, en mi caso como son muchos, en lugar de ponerles el nombre los resalto en otro color.

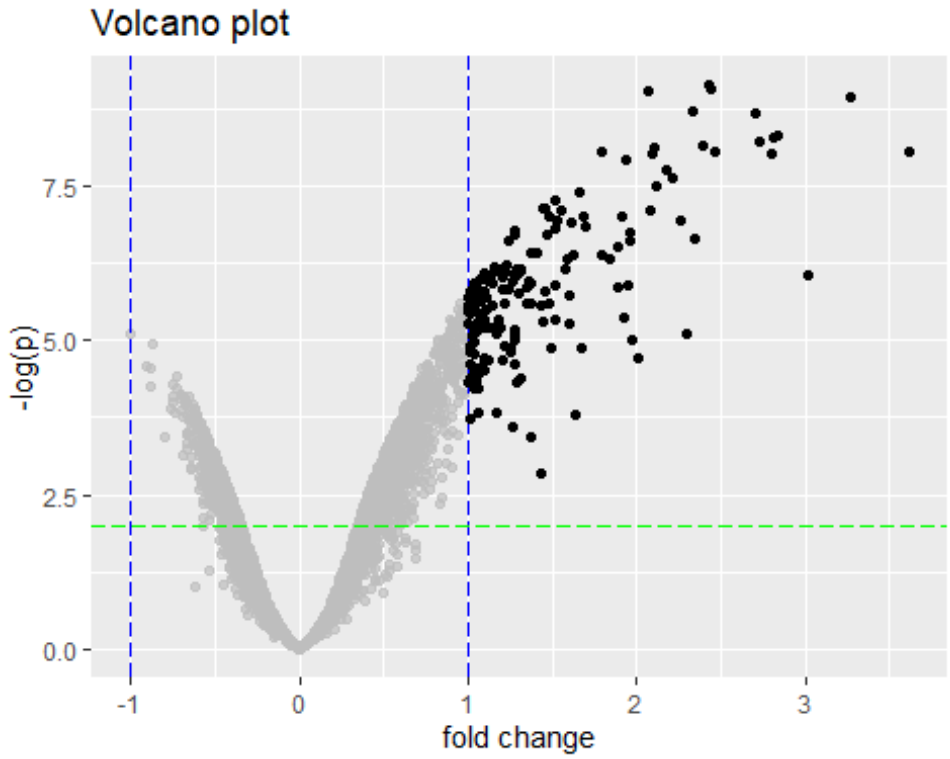



*Figura10: Volcano plot genes diferencialmente expresados*

Selecciono aquellos con logFC>2 y con el menos logaritmo del p valor mayor a 1. Los ordenos de menor a mayor p valor ajustado y saco el resultado de los 4 primeros.

```
##          PROBEID SYMBOL ENTREZID                          GENENAME
logFC
## 1102   204470_at  CXCL1     2919 C-X-C motif chemokine ligand 1
2.008540
## 1790 209774_x_at  CXCL2     2920 C-X-C motif chemokine ligand 2
2.302282
## 1968 211506_s_at  CXCL8     3576 C-X-C motif chemokine ligand 8
3.022097
## 886  202708_s_at H2BC21     8349       H2B clustered histone 21
2.345237
##       AveExpr         t      P.Value    adj.P.Val        B
## 1102 8.489757  7.159065 1.972579e-05 4.737947e-04 3.010665
## 1790 9.168898  7.901608 7.919792e-06 2.610943e-04 3.958776
## 1968 8.414236  9.882097 9.182325e-07 7.236820e-05 6.180741
## 886  6.896901 11.327447 2.352318e-07 3.042332e-05 7.566102
```

### 3.5.8 Análisis de significación biológica

Con este gráfico podemos ver el resultado del análisis de enriquecimiento cuyo objetivo es establecer si un determinado proceso biológico o una vía metabólica aparece con mayor o menor frecuencia en la lista de genes seleccionados que en la población de genes.

```
## 'select()' returned 1:1 mapping between keys and columns

## DMSOvsPaclitaxel
##             3315

## ####################################
## Comparison:  DMSOvsPaclitaxel
##                            ID                                   Description
## R-HSA-9614085 R-HSA-9614085                FOXO-mediated transcription
## R-HSA-400253   R-HSA-400253                            Circadian Clock
## R-HSA-2559583 R-HSA-2559583                          Cellular Senescence
## R-HSA-3000171 R-HSA-3000171     Non-integrin membrane-ECM interactions
## R-HSA-6785807 R-HSA-6785807 Interleukin-4 and Interleukin-13 signaling
## R-HSA-3000170 R-HSA-3000170                       Syndecan interactions
##               GeneRatio   BgRatio       pvalue    p.adjust
qvalue
## R-HSA-9614085   28/1861  65/10616 1.313053e-06 0.001089990
0.0009404849
## R-HSA-400253    29/1861  69/10616 1.578552e-06 0.001089990
0.0009404849
## R-HSA-2559583   60/1861 193/10616 2.525018e-06 0.001162350
0.0010029195
## R-HSA-3000171   25/1861  59/10616 6.884506e-06 0.002376876
0.0020508582
## R-HSA-6785807   36/1861 108/10616 4.978346e-05 0.011574782
```

```
0.0099871594
## R-HSA-3000170   14/1861  27/10616 5.028870e-05 0.011574782
0.0099871594
##
geneID
## R-HSA-9614085
FBXO32/DDIT3/GADD45A/BCL6/CDKN1A/CCNG2/CITED2/ATXN3/TXNIP/BCL2L11/AKT3/SM
AD2/PPARGC1A/SREBF1/SIRT1/BTG1/KLF4/AKT1/INS/EP300/NR3C1/AKT2/NFYA/FOXO3/
STK11/FOXO1/RBL2/IGFBP1
## R-HSA-400253
NOCT/FBXL3/CRTC3/TBL1XR1/RBM4/NAMPT/MEF2C/NFIL3/RXRA/PPARGC1A/NPAS2/BHLHE
40/SREBF1/SIRT1/ARNTL2/TBL1X/CUL1/ATF2/NRIP1/BTRC/ARNTL/CREM/EP300/CPT1A/
NCOA6/NR3C1/RORA/NCOA1/MEF2D
## R-HSA-2559583
FOS/H4C8/IL1A/H2BC21/CXCL8/H2BC12/H2BC6/H2BC5/H2BC9/H2BC11/H2AC6/H1-
2/H2BC4/H2AJ/EED/ETS1/PHC3/CDKN1A/H2AC18/HMGA2/TFDP2/MAPK9/KDM6B/CBX4/CDC
23/CBX2/UBE2D1/LMNB1/AGO3/IL6/STAT3/MAPK11/H2AZ2/TNRC6A/ETS2/TERF2IP/TERF
2/TFDP1/MAP3K5/SCMH1/H2BC3/MAPKAPK3/CCNE2/CDKN2C/MINK1/TNIK/RPS6KA2/FZR1/
MAPK1/MDM2/ANAPC4/H4C5/CDK2/SP1/HMGA1/EP400/MAPK8/RBBP4/CDK6/H2BC17
## R-HSA-3000171
THBS1/SDC1/COL5A2/LAMA5/COL1A1/ITGB5/LAMA3/LAMA1/SDC3/SDC2/LAMA2/CASK/LAM
C2/FN1/ITGA6/LAMB1/DAG1/ITGB4/PRKCA/SDC4/DDR2/NTN4/TGFB1/LAMA4/PDGFB
## R-HSA-6785807
FOS/MMP3/HMOX1/IL1A/CXCL8/MMP1/BCL6/JUNB/PTGS2/IL12A/CDKN1A/ICAM1/IL1B/MA
OA/IL6R/IL6/LAMA5/STAT3/BCL2/BIRC5/CEBPD/PIK3R1/AKT1/FN1/RORC/SOX2/BCL2L1
/FOXO3/JAK2/RORA/IL18/FOXO1/TIMP1/TGFB1/S1PR1/ITGAM
## R-HSA-3000170
THBS1/SDC1/COL5A2/COL1A1/ITGB5/SDC3/SDC2/CASK/FN1/ITGA6/ITGB4/PRKCA/SDC4/
TGFB1
##               Count
## R-HSA-9614085    28
## R-HSA-400253     29
## R-HSA-2559583    60
## R-HSA-3000171    25
## R-HSA-6785807    36
## R-HSA-3000170    14
```


*Figura11: Análisis de significación biológica*

## 4.Resultados

Se observan un número bastante elevado de genes que incrementan su expresión entre las células tratadas con Paclitaxel. Estos genes pueden estar implicados en los mecanísmos de resistencia a quimioterapia que desarrollan las células metastásicas de cáncer de mama. Su inhibición por lo tanto podría ser una nueva estrategia terapéutica. En el artículo publicado con los datos de mi experimento se concluye que la quimioterapia produce una actividad elevada de JNK que incrementa la actividad en la matriz extracelular (ECM), cicatrización de heridas y una red de células madre en las células cancerosas lo que conduce a una menor eficacia terapéutica. El tratamiento con fármacos quimioterapéuticos induce la actividad de JNK en las células de cáncer de mama, lo que refuerza la producción de SPP1 y TNC que promueven las metástasis pulmonares M (2020). Con lo que la inhibición tanto de JNK como de la expresión de SPP1 o TNC hace que los tumores mamarios sean más sensibles a la quimioterapia, pudiendo ser esta una futura estrategia en el tratamiento del cáncer de mama metastásico Insua-Rodríguez (2018).

## 5.Discusión

El análisis de microarrays presentado en este informe es un análisis sencillo cuyo fin principal era aprender las herramientas disponibles para el análisis de datos. Un

inconveniente es el número de muestras, solamente 6 con lo que el estudio tendrá muy poca potencia estadística. Es muy importante la elección de un correcto tamaño muestral. Si queremos ganar precisión podremos recurrir a la replicación o repetición de un experimento de forma idéntica en un número determinado de unidades.Para futuros experimentos sería interesante la aleatorización, es decir, la asignación de todos los factores al azar a las unidades experimentales. Con ello se consigue disminuir el efecto de los factores no controlados por el experimentador en el diseño experimental y que podrían influir en los resultados. Otra herramienta con la que podríamos mejorar los resultados en futuros estudios es el bloqueo o control local que consiste en agrupar las unidades experimentales de forma que la variabilidad dentro de los grupos sea inferior a la variabilidad de todas las unidades antes de agrupar.

## 6.Apéndice

Tal y cómo se ha hablado en repetidas ocasiones, la reproducibilidad del estudio es fundamental a la hora de trabajar como bioinformáticos, con lo que creo un repositorio en Github con todo lo relativo al proyecto de forma que se pueda clonar en otro ordenador y reproducir mi trabajo[6].

URL(puesta también al inicio del informe): https://github.com/bpardom/AO_PEC1.git

Pongo a continuación el código de R utilizado para la realización del análisis. También está disponible en el documento RMD disponible en repositorio GitHub:

```
#Cargo las librerías que utilizo para la realización del análisis

library(BiocManager)
library(GEOquery)
library(affy)
library(arrayQualityMetrics)
library(limma)
library(hgu133plus2.db)
library(ggplot2)
library(gghighlight)
library(org.Hs.eg.db)
library(ReactomePA)

#Cargo los datos utilizando GEOquery

library(BiocManager)
library(GEOquery)
Mi_gse <- getGEO("GSE98238", destdir = "C:/Bea/Master/Datos omicos
Bea/PEC1",
              GSEMatrix = TRUE)
```

[6] https://cfss.uchicago.edu/setup/git-with-rstudio/

```
## Found 1 file(s)

## GSE98238_series_matrix.txt.gz

## Using locally cached version: C:/Bea/Master/Datos omicos
Bea/PEC1/GSE98238_series_matrix.txt.gz

## Parsed with column specification:
## cols(
##   ID_REF = col_character(),
##   GSM2589734 = col_double(),
##   GSM2589735 = col_double(),
##   GSM2589736 = col_double(),
##   GSM2589737 = col_double(),
##   GSM2589738 = col_double(),
##   GSM2589739 = col_double()
## )

## Using locally cached version of GPL570 found here:
## C:/Bea/Master/Datos omicos Bea/PEC1/GPL570.soft

#Control de calidad de datos cargados con GEOquery

arrayQualityMetrics(Mi_gse$GSE98238_series_matrix.txt.gz,
                    outdir = "C:/Bea/Master/Datos omicos
Bea/PEC1/arrayQM",
                    intgroup =
colnames(Mi_gse$GSE98238_series_matrix.txt.gz@phenoData),
                    reporttitle = "arrayQualityMetrics",
                    do.logtransform = FALSE, force = TRUE)

## The report will be written into directory 'C:/Bea/Master/Datos omicos
Bea/PEC1/arrayQM'.

## (loaded the KernSmooth namespace)

#Cargo los archivos CEL

archivos_cel <- list.celfiles("C:/Bea/Master/Datos omicos
Bea/PEC1/Data/CEL")
affycel <- ReadAffy(celfile.path = "C:/Bea/Master/Datos omicos
Bea/PEC1/Data")

#Control de calidad de datos contenidos en archivos CEL

arrayQualityMetrics(affycel,
                    outdir = "C:/Bea/Master/Datos omicos
Bea/PEC1/arrayQMcel",
                    intgroup = colnames(affycel@phenoData@data),
                    reporttitle = "affycelquality",
                    do.logtransform = FALSE, force = TRUE)
```

```
## The report will be written into directory 'C:/Bea/Master/Datos omicos
Bea/PEC1/arrayQMcel'.

#Normalización RMA

affyceles <- rma(affycel)

## Background correcting
## Normalizing
## Calculating Expression

#Control de calidad de datos normalizados

arrayQualityMetrics(affyceles,
                    outdir = "C:/Bea/Master/Datos omicos
Bea/PEC1/arrayQMceles",
                    intgroup = colnames(affyceles@phenoData),
                    reporttitle = "affycelesquality",
                    do.logtransform = FALSE, force = TRUE)

## The report will be written into directory 'C:/Bea/Master/Datos omicos
Bea/PEC1/arrayQMceles'.

library(genefilter)
annotation(affyceles) <- "hgu133plus2.db"
fil <-nsFilter(affyceles, require.entrez =T, remove.dupEntrez =T,
var.filter =T, var.func =IQR, var.cutoff =0.75, filterByQuantile =T,
feature.exclude ="^AFFX")
affycelesfil<-fil$eset

#Creación de matriz de diseño

madisf <- cbind(DMSO = c(1,1,1,0,0,0),
                Paclitaxel = c(0,0,0,1,1,1))
rownames(madisf) <- rownames(affycelesfil@phenoData)
madisf

##                             DMSO Paclitaxel
## GSM2589734_DMSO_1.CEL          1          0
## GSM2589735_DMSO_2.CEL          1          0
## GSM2589736_DMSO_3.CEL          1          0
## GSM2589737_Paclitaxel_1.CEL    0          1
## GSM2589738_Paclitaxel_2.CEL    0          1
## GSM2589739_Paclitaxel_3.CEL    0          1

#Creación de matriz de contraste

macont <- makeContrasts(Paclitaxel - DMSO, levels= madisf)

#Estimación del modelo lineal

lineartar <- lmFit(affycelesfil, madisf)
```

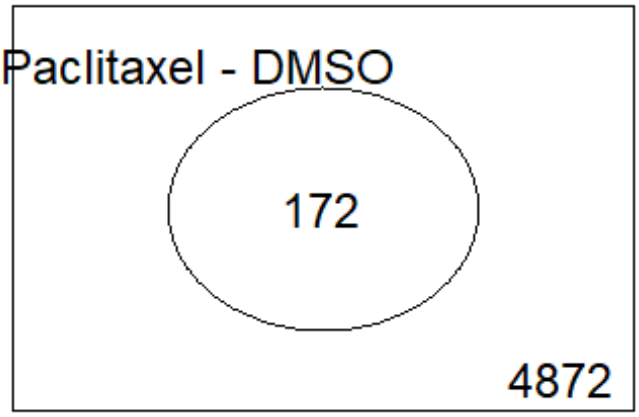
```
tarmacont <- contrasts.fit(lineartar, macont)
tarmacont <- eBayes(tarmacont)
toptarmacont <- topTable(tarmacont, number=nrow(tarmacont),
         coef = "Paclitaxel - DMSO" , adjust="fdr")
head(toptarmacont, n = 2)
```

```
##              logFC  AveExpr        t      P.Value    adj.P.Val
B
## 209189_at 2.424858 8.289016 19.71424 7.472029e-10 1.460107e-06
13.10296
## 239336_at 2.441313 5.823216 19.42717 8.727761e-10 1.460107e-06
12.96303
```

```
resultados<-decideTests(tarmacont, method="separate",
adjust.method="fdr", p.value=0.1, lfc=1)
sumabs<-apply(abs(resultados),1,sum)
rescero<-resultados[sumabs!=0,]
print(summary(resultados))
```

```
##        Paclitaxel - DMSO
## Down                   0
## NotSig              4872
## Up                   172
```

```
vennDiagram(resultados)
```




```
resultados2<-decideTests(tarmacont, method="separate",
adjust.method="fdr", p.value=0.1, lfc=0.5)
```

```r
sumabs<-apply(abs(resultados2),1,sum)
rescero<-resultados2[sumabs!=0,]
print(summary(resultados2))
```

```
##        Paclitaxel - DMSO
## Down                186
## NotSig             3964
## Up                  894
```

```r
#Anotación de genes

annotatedTopTable <- function(topTab, anotPackage)
{
  topTab <- cbind(PROBEID=rownames(topTab), topTab)
  myProbes <- rownames(topTab)
  thePackage <- eval(parse(text = anotPackage))
  geneAnots <- select(thePackage, myProbes, c("SYMBOL", "ENTREZID",
"GENENAME"))
  annotatedTopTab<- merge(x=geneAnots, y=topTab, by.x="PROBEID",
by.y="PROBEID")
  return(annotatedTopTab)
}

connombres <- annotatedTopTable(toptarmacont,
anotPackage="hgu133plus2.db")
```

```
## 'select()' returned 1:1 mapping between keys and columns
```

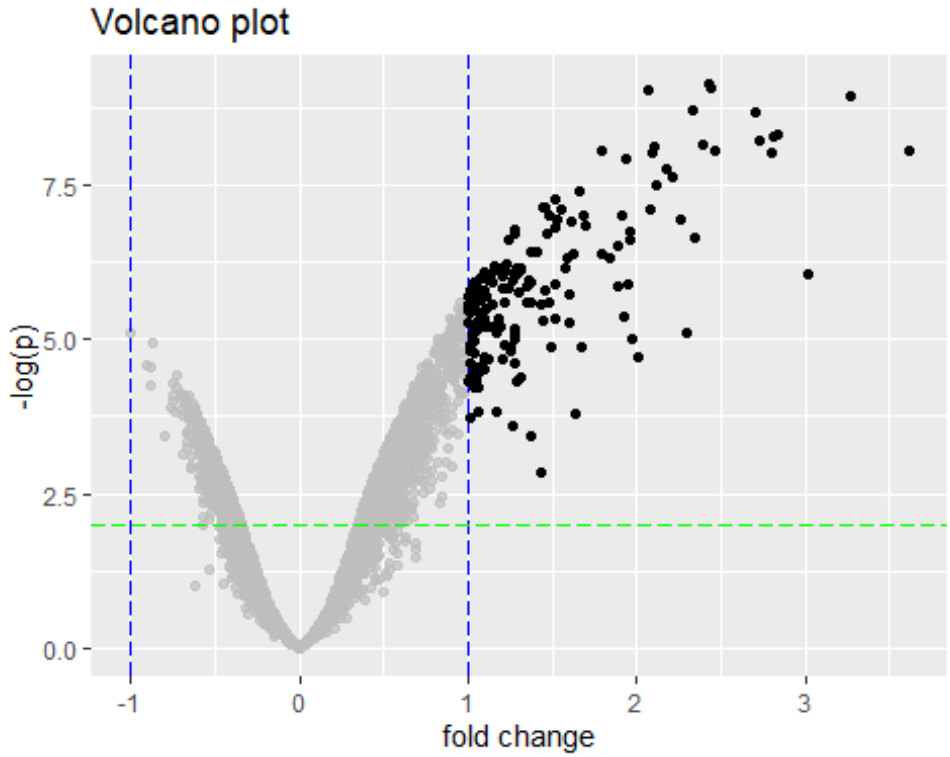
```r
Topgenedif = connombres [order(connombres$adj.P.Val,decreasing = FALSE),
]
head(Topgenedif, n = 4)
```

```
##         PROBEID   SYMBOL ENTREZID
## 117  1553736_at   ZFC3H1   196441
## 1124  204614_at SERPINB2     5055
## 1712  209189_at      FOS     2353
## 4681  239336_at    THBS1     7057
##                                                GENENAME    logFC
AveExpr
## 117                      zinc finger C3H1-type containing 2.074206
5.248166
## 1124                               serpin family B member 2 3.273535
8.711623
## 1712 Fos proto-oncogene, AP-1 transcription factor subunit 2.424858
8.289016
## 4681                                       thrombospondin 1 2.441313
5.823216
##             t      P.Value    adj.P.Val        B
## 117  19.29552 9.379043e-10 1.460107e-06 12.89797
## 1124 18.91492 1.157896e-09 1.460107e-06 12.70666
```

```
## 1712 19.71424 7.472029e-10 1.460107e-06 13.10296
## 4681 19.42717 8.727761e-10 1.460107e-06 12.96303

#Elaboración Volcano plot

attach(connombres)
ggplot(data = connombres,  aes(x = logFC,  y = -log10(P.Value))) +
  geom_point()+
  geom_vline(xintercept = -1,colour="blue", linetype = "longdash")+
  geom_vline(xintercept = 1,colour="blue", linetype = "longdash")+
  geom_hline(yintercept = -log10(0.01),colour="green", linetype =
"longdash")+
  xlab("fold change")+
  ylab("-log(p)")+
  labs(title = "Volcano plot") +
  gghighlight(-log10(P.Value) > 2 & abs(logFC) > 1)
```




```
detach(connombres)

attach(connombres)
genesdif<-subset(connombres, logFC > 2 & abs(-log10(P.Value)) > 1)
detach(connombres)
genes = genesdif [order(genesdif$adj.P.Val,decreasing = TRUE), ]
head(genes,4)

##          PROBEID SYMBOL ENTREZID                          GENENAME
logFC
```

```
## 1102   204470_at  CXCL1     2919 C-X-C motif chemokine ligand 1
2.008540
## 1790 209774_x_at  CXCL2     2920 C-X-C motif chemokine ligand 2
2.302282
## 1968 211506_s_at  CXCL8     3576 C-X-C motif chemokine ligand 8
3.022097
## 886  202708_s_at H2BC21     8349       H2B clustered histone 21
2.345237
##       AveExpr         t      P.Value    adj.P.Val        B
## 1102 8.489757  7.159065 1.972579e-05 4.737947e-04 3.010665
## 1790 9.168898  7.901608 7.919792e-06 2.610943e-04 3.958776
## 1968 8.414236  9.882097 9.182325e-07 7.236820e-05 6.180741
## 886  6.896901 11.327447 2.352318e-07 3.042332e-05 7.566102

signitables <- list(DMSOvsPaclitaxel = toptarmacont)
signiselect <- list()
for (i in 1:length(signitables)){
  topTab <- signitables[[i]]
  whichGenes<-topTab["adj.P.Val"]<0.15
  selectedIDs <- rownames(topTab)[whichGenes]
  EntrezIDs<- AnnotationDbi::select(hgu133plus2.db, selectedIDs,
c("ENTREZID"))
  EntrezIDs <- EntrezIDs$ENTREZID
  signiselect[[i]] <- EntrezIDs
  names(signiselect)[i] <- names(signitables)[i]
}

## 'select()' returned 1:1 mapping between keys and columns

sapply(signiselect, length)

## DMSOvsPaclitaxel
##             3315

mapped_genes2GO <- mappedkeys(org.Hs.egGO)
mapped_genes2KEGG <- mappedkeys(org.Hs.egPATH)
mapped_genes <- union(mapped_genes2GO , mapped_genes2KEGG)
listOfData <- signiselect[1]
comparisonsNames <- names(listOfData)
universe <- mapped_genes

for (i in 1:length(listOfData)){
  genesIn <- listOfData[[i]]
  comparison <- comparisonsNames[i]
  enrich.result <- enrichPathway(gene = genesIn, pvalueCutoff = 0.05,
                                 readable = T,
                                 pAdjustMethod = "BH",
                                 organism = "human",
                                 universe = universe)

  cat("####################################")
```

```
  cat("\nComparison: ", comparison,"\n")
  print(head(enrich.result))

  if (length(rownames(enrich.result@result)) != 0) {
  write.csv(as.data.frame(enrich.result),
            file=paste0("C:/Bea/Master/Datos omicos
Bea/PEC1/Results","ReactomePA.Results.",comparison,".csv"), row.names =
FALSE)

  pdf(file=paste0("C:/Bea/Master/Datos omicos
Bea/PEC1/Result","ReactomePABarplot.",comparison,".pdf"))
  print(barplot(enrich.result, showCategory = 15, font.size = 4, title =
paste0("Reactome Pathway Analysis for ", comparison, ". Barplot")))
  dev.off()

  pdf(file = paste0("C:/Bea/Master/Datos omicos
Bea/PEC1/Result","ReactomePAcnetplot.",comparison,".pdf"))
    print(cnetplot(enrich.result, categorySize = "geneNum", schowCategory
= 15, vertex.label.cex = 0.75))
  dev.off()
  }
}

## ####################################
## Comparison:  DMSOvsPaclitaxel
##                            ID                                  Description
## R-HSA-9614085 R-HSA-9614085                FOXO-mediated transcription
## R-HSA-400253   R-HSA-400253                             Circadian Clock
## R-HSA-2559583 R-HSA-2559583                          Cellular Senescence
## R-HSA-3000171 R-HSA-3000171     Non-integrin membrane-ECM interactions
## R-HSA-6785807 R-HSA-6785807 Interleukin-4 and Interleukin-13 signaling
## R-HSA-3000170 R-HSA-3000170                       Syndecan interactions
##               GeneRatio   BgRatio       pvalue    p.adjust
qvalue
## R-HSA-9614085   28/1861  65/10616 1.313053e-06 0.001089990
0.0009404849
## R-HSA-400253    29/1861  69/10616 1.578552e-06 0.001089990
0.0009404849
## R-HSA-2559583   60/1861 193/10616 2.525018e-06 0.001162350
0.0010029195
## R-HSA-3000171   25/1861  59/10616 6.884506e-06 0.002376876
0.0020508582
## R-HSA-6785807   36/1861 108/10616 4.978346e-05 0.011574782
0.0099871594
## R-HSA-3000170   14/1861  27/10616 5.028870e-05 0.011574782
0.0099871594
##
geneID
## R-HSA-9614085
FBXO32/DDIT3/GADD45A/BCL6/CDKN1A/CCNG2/CITED2/ATXN3/TXNIP/BCL2L11/AKT3/SM
```

```
AD2/PPARGC1A/SREBF1/SIRT1/BTG1/KLF4/AKT1/INS/EP300/NR3C1/AKT2/NFYA/FOXO3/
STK11/FOXO1/RBL2/IGFBP1
## R-HSA-400253
NOCT/FBXL3/CRTC3/TBL1XR1/RBM4/NAMPT/MEF2C/NFIL3/RXRA/PPARGC1A/NPAS2/BHLHE
40/SREBF1/SIRT1/ARNTL2/TBL1X/CUL1/ATF2/NRIP1/BTRC/ARNTL/CREM/EP300/CPT1A/
NCOA6/NR3C1/RORA/NCOA1/MEF2D
## R-HSA-2559583
FOS/H4C8/IL1A/H2BC21/CXCL8/H2BC12/H2BC6/H2BC5/H2BC9/H2BC11/H2AC6/H1-
2/H2BC4/H2AJ/EED/ETS1/PHC3/CDKN1A/H2AC18/HMGA2/TFDP2/MAPK9/KDM6B/CBX4/CDC
23/CBX2/UBE2D1/LMNB1/AGO3/IL6/STAT3/MAPK11/H2AZ2/TNRC6A/ETS2/TERF2IP/TERF
2/TFDP1/MAP3K5/SCMH1/H2BC3/MAPKAPK3/CCNE2/CDKN2C/MINK1/TNIK/RPS6KA2/FZR1/
MAPK1/MDM2/ANAPC4/H4C5/CDK2/SP1/HMGA1/EP400/MAPK8/RBBP4/CDK6/H2BC17
## R-HSA-3000171
THBS1/SDC1/COL5A2/LAMA5/COL1A1/ITGB5/LAMA3/LAMA1/SDC3/SDC2/LAMA2/CASK/LAM
C2/FN1/ITGA6/LAMB1/DAG1/ITGB4/PRKCA/SDC4/DDR2/NTN4/TGFB1/LAMA4/PDGFB
## R-HSA-6785807
FOS/MMP3/HMOX1/IL1A/CXCL8/MMP1/BCL6/JUNB/PTGS2/IL12A/CDKN1A/ICAM1/IL1B/MA
OA/IL6R/IL6/LAMA5/STAT3/BCL2/BIRC5/CEBPD/PIK3R1/AKT1/FN1/RORC/SOX2/BCL2L1
/FOXO3/JAK2/RORA/IL18/FOXO1/TIMP1/TGFB1/S1PR1/ITGAM
## R-HSA-3000170
THBS1/SDC1/COL5A2/COL1A1/ITGB5/SDC3/SDC2/CASK/FN1/ITGA6/ITGB4/PRKCA/SDC4/
TGFB1
##               Count
## R-HSA-9614085    28
## R-HSA-400253     29
## R-HSA-2559583    60
## R-HSA-3000171    25
## R-HSA-6785807    36
## R-HSA-3000170    14

cnetplot(enrich.result, categorySize = "geneNum", schowCategory = 15,
vertex.label.cex = 0.75)
```


## 7.Referencias bibliográficas


Insua-Rodríguez, Jacob. 2018. “Stress Signaling in Breast Cancer Cells Induces Matrix Components That Promote Chemoresistant Metastasis.” *EMBO Mol Med* online.

M, Pein. 2020. “Metastasis-Initiating Cells Induce and Exploit a Fibroblast Niche to Fuel Malignant Colonization of the Lungs.” *Nat Commun* online.