Desregulación molecular inducida por tratamiento con alcohol en células precursoras neurales a partir de células embrionarias humanas pluripotentes

Marina Ballesteros

4/9/2020

Contents

Los datos del análisis se han cargado desde la base de datos Gene Expression Omnibus (GEO). El conjunto de datos seleccionado se identifica con el número de adhesión: **GSE56906**.

El estudio que generó dichos datos investiga el efecto del alcohol (EtOH) en el desarrollo de células madre neurales derivadas desde células madre de embriones humanos. Existen indicios de que el alcohol interviene negativamente en el desarrollo de las células madre neurales. Para corroborar estos indicios, se han cultivado células madre embrionarias durante cinco días en un medio de inducción neural (NIM). A continuación, los agregados neuronales fueron sembrados en placas recubiertas con poli-L-ornitina/laminina y cultivados con NIM durante siete días adiccionales para desarrollar la estructura de roseta neuronal. Después de siete días, las rosetas neurales fueron desalojadas y luego replateadas para la expansión de las células precursoras neurales (NPC) durante 3-5 días.

Los microarrays utilizados para este experimento fueron del siguiente tipo: Affymetrix Human Genome U133 Plus 2.0 Array.

##Preparación del ambiente de trabajo

Antes de empezar con el análisis y a manejar la enorme cantidad de datos y ficheros que ello conlleva, crearé tres carpetas para la organización del mismo:

- La carpeta principal del análisis será "Effect of alcool in ESC differentiation", la cual también será mi directorio de trabajo.
- Una carpeta llamada **data** para almecenar todo tipo de datos del experimento y en los cuales basaré mi análisis. En esta carpeta guardaré los archivos *.CEL* y el archivo *targets*, en el cual se decribirán los factores de estudio y sus niveles.
- En la carpeta **results** guardaré todos los resultados obtenidos en el análisis.
- La carpeta **figures** servirá para almacenar todo tipo de imágenes y figuras generadas durante el análisis.

Lectura de los datos

En el experimento se han analizado 10 muestras, las cuales se analizan en torno a dos factores: el tipo celular y el tipo de tratamiento durante la diferenciación celular. Los niveles de dichos factores son los siguientes:

Para el factor **tipo celular** existen tres niveles:

- Células madre embrionarias (ESC) indiferenciadas *H1 p40*
- Células neurales en forma de roseta *H1 Rosett*

- Células progenitoras neurales (NPC) *NPC H1*

Para el factor **tipo de tratamiento** existen dos niveles:

- Células diferenciadas sin tratamiento con alcohol *EtOH 0*
- Células diferenciadas bajo tratamiento con alcohol *EtOH 20*

Una tabla que recoge todas laas muestras sería la siguiente:

Referencia	Tipo de muestra	Características
GSM1371025	H1 EtOH0 Rosett1	Roseta sin alcohol
GSM1371026	H1 EtOH0 Rosett2	Roseta sin alcohol
GSM1371027	H1 EtOH20 Rosett1	Roseta con alcohol
GSM1371028	H1 EtOH20 Rosett2	Roseta con alcohol
GSM1371029	H1 p40 Und-1	ESC indiferenciada
GSM1371030	H1 p40 Und-2	ESC indiferenciada
GSM1371031	NPC H1 EtOH 0-1	NPC sin alcohol
GSM1371032	NPC H1 EtOH 0-2	NPC sin alcohol
GSM1371033	NPC H1 EtOH 20-1	NPC con alcohol
GSM1371034	NPC H1 EtOH 20-2	NPC con alcohol

Para la lectura de los datos, primero he descargado los 10 archivos *.CEL* desde la página de GEO y he creado un archivo llamado *targets.csv* donde recopilo las características de cada muestra así como su referencia en la base de datos GEO. Tras haber instalado Bioconductor, procedo a la lectura del archivo targets y de los datos.

```
targets <- read.csv2("./data/targets.csv", header = TRUE, sep = ";")
knitr::kable(
  targets, booktabs = TRUE,
  caption = 'Content of the targets file used for the current analysis')
```

```
library(oligo)
celFiles <- list.celfiles("./data", full.names = TRUE)
library(Biobase)
#Asociación de los archivos CEL con el archivo targets
my.targets <-read.AnnotatedDataFrame(file.path("./data","targets.csv"),
                                     header = TRUE, row.names = 1,
                                     sep=";")
rawData <- read.celfiles(celFiles, phenoData = my.targets)
print(pData(rawData))
```

En las líneas de código anterior, he asociado la información alamcenada en los archivos *.CEL* con la información del archivo *targets* en la variable llamada **rawData**.

La clase de archivo de la variable *rawData* se llama **ExpressionSet** y esta clase de archivos permite almacenar todo tipo de información del experimento. A continuación, cambio el largo nombre de cada variable por un nombre más corto, almacenado en la columna *ShortName* del archivo "targets".

```
my.targets@data$ShortName->rownames(pData(rawData))
colnames(rawData) <-rownames(pData(rawData))

head(rawData)
```

2. Contro de calidad de los datos crudos

El control de calidad de los datos en crudo nos permite conocer la calidad de los datos recogidos durante el experimento a través de un boxplot de intensidad o un estudio de los componentes principales (PCA). Este

paso es muy importante ya que es donde se evalua si los datos tienen suficiente calidad para la normalización de los mismos. Si alguno de los array no tiene la calidad suficiente, este será marcado y expuesto a revisión para decidir si mantenemos o no dicho array en el análisis.

El paquete que desarrolla el control de calidad se llama **ArrayQualityMetrics** y todos los resultados se recogen en el archivo **index.html** dentro de la carpeta *rawData_quality* del directorio *results*.

```
library(arrayQualityMetrics)
#Guardar los resultados del control calidad en el directorio results
arrayQualityMetrics(rawData, outdir = "./results/rawData_quality", force = TRUE)
```

La siguiente tabla resumen nos ofrece el resultado del control de calidad.

	array	sampleNames	*1	*2	*3	SampleTitle	Classification
☐	1	GSM1371025				H1 EtOH 0 Rosett-1	Roseta diferencida sin alcohol
☐	2	GSM1371026				H1 EtOH 0 Rosett-2	Roseta diferencida sin alcohol
☐	3	GSM1371027				H1 EtOH 20 Rosett-1	Roseta diferencida con alcohol
☐	4	GSM1371028				H1 EtOH 20 Rosett-2	Roseta diferencida con alcohol
☐	5	GSM1371029				H1 p40 Und-1	ESC indiferenciada
☐	6	GSM1371030				H1 p40 Und-2	ESC indiferenciada
☑	7	GSM1371031			x	NPC H1 EtOH 0-1	NPC diferenciada sin alcohol
☐	8	GSM1371032				NPC H1 EtOH 0-2	NPC diferenciada sin alcohol
☑	9	GSM1371033			x	NPC H1 EtOH 20-1	NPC diferenciada con alcohol
☐	10	GSM1371034				NPC H1 EtOH 20-2	NPC diferenciada con alcohol

Figure 1: Tabla resumen del archivo index.html, generdo por el paquete arrayQualityMetrics en los datos crudos

Los arrays 7 y 9 se consideran outliers a través de un plot MA, el cual visualiza las diferencias entre las mediciones tomadas en dos muestras. Sin embargo, ambos arrays los mantendré en el análisis puesto que solo una de las tres pruebas no es suficiente para considerar suprimir dichos arrays del análisis.

A continuación desarrollo el gráfico resultado del **Análisis de Componentes Principales (PCA)**, donde se recogen los dos primeros componentes principales.Dicho gráfico se recoge en el directorio *figures*.

```
library(ggplot2)
library(ggrepel)
plotPCA3 <- function (datos, labels, factor, title, scale,colores, size = 1.5, glineas = 0.25) {
  data <- prcomp(t(datos),scale=scale)
  # plot adjustments
  dataDf <- data.frame(data$x)
  Group <- factor
  loads <- round(data$sdev^2/sum(data$sdev^2)*100,1)
  # main plot
  p1 <- ggplot(dataDf,aes(x=PC1, y=PC2)) +
    theme_classic() +
    geom_hline(yintercept = 0, color = "gray70") +
    geom_vline(xintercept = 0, color = "gray70") +
    geom_point(aes(color = Group), alpha = 0.55, size = 3) +
```

```
    coord_cartesian(xlim = c(min(data$x[,1])-5,max(data$x[,1])+5)) +
    scale_fill_discrete(name = "Group")
  # avoiding labels superposition
  p1 + geom_text_repel(aes(y = PC2 + 0.25, label = labels),segment.size = 0.25, size = size) +
    labs(x = c(paste("PC1",loads[1],"%")),y=c(paste("PC2",loads[2],"%"))) +
    ggtitle(paste("Análisis de componentes principales para: ",title,sep=" "))+
    theme(plot.title = element_text(hjust = 0.5)) +
    scale_color_manual(values=colores)
  }
```

```
require(ggplot2)
plotPCA3(exprs(rawData), labels = targets$ShortName, factor = targets$Group,
         title="Datos crudos", scale = FALSE, size = 4,
         colores = c("red", "blue", "green", "yellow", "pink"))
```

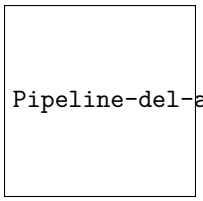

Figure 2: Visualization of the two first Principal Components for raw data

El análisis de componentes principales indica que el 57,4% de la variabilidad total de las muestras se explica con el primer componente. En el gráfico se observa cómo el factor *tipo celular* es la principal fuente de variabilidad, puesto que las células indiferenciadas (p40) se encuentran a la derecha del gráfico mientras que las células diferenciadas (NPC) se sitúan a la izquierda del mismo. Del mismo modo, el tipo celular *Rosetta* se encuentra en mitad del gráfico ya que es el grupo intermedio entre las células madre embrionarias (p40) y las células progenitoras neurales (NPC).

Existen otros gráficos que nos deja el control de calidad de los datos en crudo, entre ellos destaca el boxplot múltiple con la distribución de las intensidades a lo largo de todas las muestras.

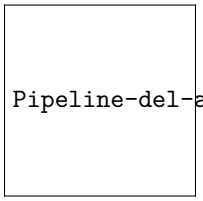

Figure 3: Boxplot para las intensidades de los arrays (Raw Data)

3. Normalización de los datos crudos

El objetivo de la normalización es hacer comparables los arrays entre sí además de eliminar cualquier variabilidad en las muestras no debida a razones biológicas. Es decir, la normalización de los datos asegura que las diferencias de intensidades en las muestras se deban a diferencias en la expresión de los genes y no a sesgos debidos a cuestiones técnicas del experimento.

El proceso consta de tres etapas: eliminación del ruido de fondo, normalización y sumarización de los datos. Los tres procesos se llevan a cabo gracias al método **Robust Multichip Analysis** a través de la función *rma*.

```
eset_rma <- rma(rawData)
```

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

Una vez tenemos los datos normalizados, repetimos el proceso de control de calidad de los datos pero con los datos normalizados.

4. Control de calidad de los datos normalizados

El resultado del control de calidad de los datos normalizados los podemos ver nuevamente en el archivo **index.html** de la carpeta *normalized_quality* en el directorio *results*. El resumen del resultado se muestra en la siguiente tabla, donde se observan los arrays 5 y 6 como outliers a través de la medida "distancias entre arrays". Estos arrays se mantendrán en el análisis por la misma justificación dada anteriormente, una sola prueba de las tres pruebas realizadas para detectar outliers no es suficiente como para eliminar un array del análisis.

	array	sampleNames	*1	*2	*3	Group	CellType	Treatment
☐	1	Rosett.0.1				H1 EtOH 0 Rosett-1	Roseta diferenciada	Sin alcohol
☐	2	Rosett.0.2				H1 EtOH 0 Rosett-2	Roseta diferenciada	Sin alcohol
☐	3	Rosett.20.1				H1 EtOH 20 Rosett-1	Roseta diferenciada	Con alcohol
☐	4	Rosett.20.2				H1 EtOH 20 Rosett-2	Roseta diferenciada	Con alcohol
☑	5	P40.1	x			H1 p40 Und-1	ESC	Sin alcohol
☑	6	P40.2	x			H1 p40 Und-2	ESC	Sin alcohol
☐	7	NPC.0.1				NPC H1 EtOH 0-1	NPC diferenciada	Sin alcohol
☐	8	NPC.0.2				NPC H1 EtOH 0-2	NPC diferenciada	Sin alcohol
☐	9	NPC.20.1				NPC H1 EtOH 20-1	NPC diferenciada	Con alcohol
☐	10	NPC.20.2				NPC H1 EtOH 20-2	NPC diferenciada	Con alcohol

Figure 4: Tabla resumen del archivo index.html, generdo por el paquete arrayQualityMetrics en los datos normalizados

Se puede observar en el gráfico de **Componentes Principales** de los datos normalizados que las muestras se siguen separando en función al *tipo celular*, indicando que este sigue siendo el factor principal de variación entre las muestras y no el tratamiento con alcohol. En este caso, dicho factor explica el 69,5% de la variabilidad total de las muestras.

```
plotPCA3(exprs(eset_rma), labels = targets$ShortName, factor = targets$Group,
         title="Datos normalizados", scale = FALSE, size = 4,
         colores = c("red", "blue", "green", "yellow", "pink"))
```

En el boxplot de intensidades de los datos normalizados se espera que todos los boxplots tengan el mismo aspecto; es decir, las mismas intensidades. Esto es debido a que en el proceso de normaliación se incluye la normalización de los cuantiles, en el que la distribución empírica de todas las muestras se establece con los mismos valores. Aquí se muestra el boxplot de **Distribución de los valores de intensidad de los datos normalizados**.

```
boxplot(eset_rma, cex.axis=0.5, las=2,  which="all",
         col = c(rep("red", 2), rep("blue", 2), rep("green", 2), rep("yellow", 2), rep("pink",2)),
```

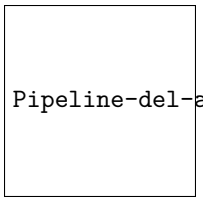

Figure 5: Visualization of first two principal components for normalized data

```
        main="Distribución de los valores de intensidad \n de los datos normalizados")
```

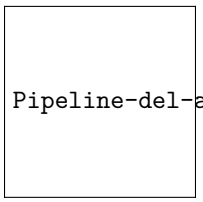

Figure 6: Boxplot para las intensidades de los arrays (Datos Normalizados)

Un control que acompaña al análisis de componentes principales, es el llamado **Batch Detection**. Este control consiste en conocer la procedencia de la mayor fuente de variabilidad introducido por las variaciones experimentales dependientes del tiempo y lugar del experimento a la hora de recolectar y analizar las muestras.

Existen distintos métodos de llevar a cabo este control, uno de los más conocidos es el llamado **Combat and Principal variation component analysis (PVCA)**. Esta técnica estima la fuente y la proporción de la variación en dos pasos, el análisis de componentes principales y el análisis de componentes de variación.

En el gráfico de estimación de PVCA se muestra una barra por cada fuente de variación incluída en el análisis. El gráfico indica que el factor *Group* es el de mayor variabilidad, al cual se le atribuye aproximadamente el 79% de la variación; esto coincide con lo observado en los gráficos de componentes principales tanto de los datos en crudo como de los datos normalizados. Además, se indica que el factor *Tratamiento con EtOH* tan sólo supone un 2,7% de la variabilidad de las muestras.

```
#plot the results
bp <- barplot(pvcaObj$dat, xlab = "Factores",
  ylab = "Variación de la proporción media ponderada",
  ylim= c(0,1.1),col = c("turquoise"), las=2,
  main="Estimación PVCA")
axis(1, at = bp, labels = pvcaObj$label, cex.axis = 0.55, las=2)
values = pvcaObj$dat
new_values = round(values , 3)
text(bp,pvcaObj$dat,labels = new_values, pos=3, cex = 0.5)
```

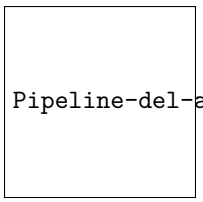

Figure 7: Relative importance of the different factors -genotype, temperature and interaction- affecting gene expression

5. Filtraje no específico

El filtraje no específico es el proceso de identificación y filtración de los genes que no se espera que se expresen diferencialmente, sino que su variación pueda deberse a la variación aleatoria.

La función encargada de hacer esta tarea es *nsFilter* del paquete **genefilter**. El criterio para identificar los genes que no se expresan diferencialmente de los que si lo hacen es un umbral dado por la persona encargada de hacer el análisis. Además, la función *nsFilter* tiene una segunda función, la de eliminar las muestras que no tienen un identificador de genes asociado.

Antes de realizar el filtraje debemos conocer el tipo de microarray utilizado y, posteriormente, descargar la librería de anotaciones asociada a dicho tipo de microarray. En este análisis, el microarray utilizado corresponde con el modelo *Affymetrix Human Genome U133 Plus 2.0 Array*.

```
#Descarga del paquete genefilter para la función nsfilter
library(genefilter)
#Descarga de la libreria de anotaciones del microarray utilizado
library(hgu133plus2.db)
annotation(eset_rma) <- "hgu133plus2.db"
filtered <- nsFilter(eset_rma,
                     require.entrez = TRUE, remove.dupEntrez = TRUE,
                     var.filter=TRUE, var.func=IQR, var.cutoff=0.75,
                     filterByQuantile=TRUE, feature.exclude = "^AFFX")
```

El filtraje nos devuelve los genes filtrados en un objeto llamado en este caso **eset_filtered**. La clase de este objeto sigue siendo *ExpressionSet*; ya que este objeto se ha obtenido a través de los datos normalizados *eset_rma*, que a su vez se obtuvieron de los datos en bruto *rawData*, todos de la misma clase.

```
print(filtered$filter.log)
```

```
## $numDupsRemoved
## [1] 21738
##
## $numLowVar
## [1] 15130
##
## $numRemoved.ENTREZID
## [1] 12753
##
## $feature.exclude
## [1] 10
```

```
eset_filtered <-filtered$eset
```

6. Identificación de genes diferencialmente expresados

Para identificar los genes diferencialmente expresados existen varios métodos. Hasta el momento, el método aue mejores resultados ofrece es el de **Modelos lineales para microarrays**. Dicho método está implementado en el paquete **limma**.

El **primer paso** para el análisis basado en modelos lineales es crear la *matriz de diseño*, que es una tabla que describe la asignación de cada muestra a un grupo o condición experimental. Tiene tantas filas como muestras y tantas columnas como grupos, en este caso hay cinco grupos si juntamos los dos factores: tipo celular y tratamiento. Cada fila contiene un uno en la columna del grupo al que pertenece la muestra y un cero en las demás.

La matriz de diseño se elabora a partir de los datos filtrados *eset_filtered*. El resultado es el siguiente:

```
##    Rosett.0 Rosett.20 p40 NPC.0 NPC.20
## 1         1         0   0     0      0
```

```
## 2         1         0   0     0      0
## 3         0         1   0     0      0
## 4         0         1   0     0      0
## 5         0         0   1     0      0
## 6         0         0   1     0      0
## 7         0         0   0     1      0
## 8         0         0   0     1      0
## 9         0         0   0     0      1
## 10        0         0   0     0      1
## attr(,"assign")
## [1] 1 1 1 1 1
## attr(,"contrasts")
## attr(,"contrasts")$Group
## [1] "contr.treatment"
```

Una vez hecha la matriz de diseño, el **segundo paso** es realizar comparaciones entre los grupos de genes a través de la *matriz de contraste*. La matriz de contraste tendrá tantas columnas como comparaciones se hagan y tantas filas como grupos existentes. Esta matriz estará compuesta de 1 y -1 en las filas de grupos a comparar y ceros en el resto.

Según el artículo del experimento,la correlación de la expresión génica con el tratamiento con EtOH no fue lo suficientemente fuerte sobre el efecto de la diferenciación. Por lo tanto, se decide analizar el conjunto de datos para el efecto del EtOH en las células de la roseta y NPC por separado.

La matriz de contraste la he definido para realizar cuatro comparaciones, dos comparaciones para cada tipo celular (Roseta y NPC), con el fin de responder a las siguientes preguntas:

Contrastes para las rosetas neurales:

- Efecto de la diferenciación celular hacia rosetas neurales
- Efecto del tratamineto con EtOH en rosetas neurales

Contrastes para las células progenitoras neurales (NPC):

- Efecto de la diferenciación celular hacia células progenitoras neurales
- Efecto del tratamineto con EtOH en células progenitoras neurales

```
##            Contrasts
## Levels      p40vsRosett0 Rosett20vsRosett0 p40vsNPC0 NPC20vsNPC0
##   Rosett.0            -1                -1         0           0
##   Rosett.20            0                 1         0           0
##   p40                  1                 0         1           0
##   NPC.0                0                 0        -1          -1
##   NPC.20               0                 0         0           1
```

Ahora que las matrices de diseño y contraste están creadas, puedo pasar a estimar el modelo y hacer las pruebas de significación para decidir qué genes se expresan diferencialmente en cada condición.

El método del paquete *limma* para la selección de genes utiliza modelos empíricos de Bayes para combinar una estimación de la variabilidad basada en toda la matriz con estimaciones individuales basadas en cada uno de los valores individuales. Los estadísticos de prueba se utilizan para ordenar los genes de más a menos expresados.

Este método, además, controla el número de falsos positivos a través de un ajsute de los p-valores por el método de *Benjamini and Hochberg*.

Finalmente, la información relevante para la posterior exploración de los resultados se almacena en un objeto R de la clase **MArrayLM** definido en el paquete *limma*. El objeto se llamará **fit.main**.

```
## [1] "MArrayLM"
```

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

Para terminar con la identificación de los genes diferencialmente expresados, debo obtener la lista de dichos genes. Esto se hace a través de la función **topTable** del paquete *limma*. La función *topTable* nos devolverá una lista de genes ordenados de menor a mayor p-valor, lo que se traduce en genes de más a menos expresados diferencialmente. Junto al estadístico p-valor aparecen otros estadísticos, destacando el p-valor ajustado o el estadístico B, que es el posterior logaritmo de probabilidades del gen de ser contra no ser expresado diferencialmente.

Obtendré una tabla para cada uno de los cuatro contrastes llevados a cabo.

Comparación 1 (p40vsRosett0): Genes que cambian su expresión entre células indiferenciadas y rosetas neurales.

```
topTab_p40vsRosett0 <- topTable (fit.main, number=nrow(fit.main), coef="p40vsRosett0", adjust="fdr")
head(topTab_p40vsRosett0)
```

```
##                 logFC   AveExpr         t      P.Value    adj.P.Val        B
## 206373_at   -7.247784 11.916089 -145.9622 5.563830e-15 2.806396e-11 23.80169
## 221086_s_at -7.053506  8.486889 -118.9733 2.849593e-14 4.155992e-11 22.79328
## 204069_at   -5.951557  9.496864 -115.2565 3.671996e-14 4.155992e-11 22.62029
## 206018_at   -8.175788 11.804071 -113.8389 4.053567e-14 4.155992e-11 22.55168
## 201012_at   -6.151703  9.144047 -113.6084 4.119738e-14 4.155992e-11 22.54038
## 212764_at   -6.997682 10.028004 -110.2744 5.226443e-14 4.393696e-11 22.37234
```

Comparación 2 (Rosett20vsRosett0): Genes que cambian su expresión bajo el tratamiento con 20mM de EtOH en rosetas neurales.

```
topTab_Rosett20vsRosett0 <- topTable (fit.main, number=nrow(fit.main), coef="Rosett20vsRosett0", adjust=
head(topTab_Rosett20vsRosett0)
```

```
##                   logFC  AveExpr         t      P.Value   adj.P.Val        B
## 221019_s_at  -0.8904331 7.331574 -16.18214 2.157364e-07 0.001088174 7.255788
## 209875_s_at  -0.7385711 8.636895 -12.47501 1.606207e-06 0.004023319 5.686700
## 206552_s_at  -1.4997448 8.686413 -11.25229 3.519545e-06 0.004023319 5.014387
## 222549_at    -0.6611985 6.824317 -11.21858 3.600320e-06 0.004023319 4.994504
## 206924_at    -0.7477088 7.215421 -11.06771 3.988222e-06 0.004023319 4.904552
## 1555450_a_at -0.7991543 5.825407 -10.28585 6.919160e-06 0.005252623 4.412336
```

Comparación 3 (p40vsNPC0): Genes que cambian su expresión entre células indiferenciadas y células progenitoras neurales.

```
topTab_p40vsNPC0 <- topTable (fit.main, number=nrow(fit.main), coef="p40vsNPC0", adjust="fdr")
head(topTab_p40vsNPC0)
```

```
##                 logFC   AveExpr         t      P.Value    adj.P.Val        B
## 206373_at   -7.793669 11.916089 -156.9557 3.114442e-15 1.570925e-11 24.14319
## 209988_s_at -8.534566  9.158885 -132.8936 1.177220e-14 1.988252e-11 23.38483
## 238878_at   -7.762829  9.563890 -127.9598 1.592695e-14 1.988252e-11 23.19516
## 207443_at   -7.559835  9.644671 -124.1422 2.028729e-14 1.988252e-11 23.03877
## 204069_at   -6.349737  9.496864 -122.9675 2.188834e-14 1.988252e-11 22.98885
## 221086_s_at -7.219977  8.486889 -121.7812 2.365089e-14 1.988252e-11 22.93755
```

Comparación 4 (NPC20vsNPC0): Genes que cambian su expresión bajo el tratamiento con 20mM de EtOH en células progenitoras neurales.

```
topTab_NPC20vsNPC0 <- topTable (fit.main, number=nrow(fit.main), coef="NPC20vsNPC0", adjust="fdr")
head(topTab_NPC20vsNPC0)
```

```
##                   logFC   AveExpr         t      P.Value    adj.P.Val        B
## 206349_at    -1.3770890  8.688765 -20.54262 3.336970e-08 0.0001308226 9.509704
## 200650_s_at   0.9974532 12.828520  19.28520 5.479087e-08 0.0001308226 9.086425
## 213880_at    -1.1437077  7.626006 -18.44092 7.780882e-08 0.0001308226 8.779723
## 1556057_s_at -1.0640304  7.933443 -16.92764 1.519035e-07 0.0001547235 8.179092
## 201848_s_at   1.1429809  8.833883  16.72138 1.671371e-07 0.0001547235 8.091710
## 225342_at     1.1032601  9.176274  16.34496 1.995667e-07 0.0001547235 7.928574
```

La primera columna de cada una de las tablas obtenidas contiene la identificación del fabricante (Affymetrix) de cada conjunto de sondas; mientras que el resto de columnas son variables numéricas y estadísticas para indicar el cambio de pliegue del gen. Con estas tablas finalizaría el proceso de identificación de los genes diferencialmente expresados.

6. Anotación de los resultados

El proceso de anotación consiste en relacionar los identificadores de la primera columna de las tablas, los cuales corresponden a conjuntos de sondas o tránscritos, con información más fácil de manejar como el *Gene Symbol*, *Entrez Gene* o *Gene description*. Es decir, el proceso de anotación es la identificación de cada gen con cada ID de Affymetrix, además de complementar dichos genes con la máxima información posible encontrada en diferentes bases de datos.

Para realizar la anotación de los resultados, sigo una función de base cuyas funciones serán:

-Identificar a la primera columna de las tablas obtenidas anteriormente como *PROBEID*, esta columna contiene las identificaciones de las sondas de Affymetrix. -Asociar a cada sonda de Affymetrix los identificadores *SYMBOL*, *ENTREZID* y *GENENAME* disponibles en el archivo de anotaciones del microarray utilizado.

El paquete de anotaciones para el microarray utilizado es **hgu133plus2.db**. Una vez que creemos las nuevas tablas, derivadas de las tablas obtenidas con *topTab* y a las que se les han añadido las anotaciones pertinentes, se guardarán en diferentes archivos *.csv* dentro del directorio **resultados**.

```
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)
}
```

```
require(hgu133plus2.db)
topAnnotated_p40vsRosett0 <- annotatedTopTable(topTab_p40vsRosett0,
anotPackage="hgu133plus2.db")
topAnnotated_Rosett20vsRosett0 <- annotatedTopTable(topTab_Rosett20vsRosett0,
anotPackage="hgu133plus2.db")
topAnnotated_p40vsNPC0 <- annotatedTopTable(topTab_p40vsNPC0,
anotPackage="hgu133plus2.db")
topAnnotated_NPC20vsNPC0 <- annotatedTopTable(topTab_NPC20vsNPC0,
anotPackage="hgu133plus2.db")
write.csv(topAnnotated_p40vsRosett0, file="./results/topAnnotated_p40vsRosett0.csv")
write.csv(topAnnotated_Rosett20vsRosett0, file="./results/topAnnotated_Rosett20vsRosett0.csv")
write.csv(topAnnotated_p40vsNPC0, file="./results/topAnnotated_p40vsNPC0.csv")
write.csv(topAnnotated_NPC20vsNPC0, file="./results/topAnnotated_NPC20vsNPC0.csv")
```

Los resultados de las tablas de selección se pueden visualizar a través de un gráfico llamado **volcano plot**. En este gráfico se representa en el eje de abscisas los cambios de expresión en escala logarítmica; mientras que en el eje de ordenadas se representa el estadístico B o probabilidad del gen de estar contra la probabilidad de

no estar expresado diferencialmente (en escala logarítmica). En este caso se representan los cuatro primeros genes más diferencialmente expresados de cada tabla.

Los resultados se recogen en el archivo *Volvanos.pdf* del directorio **figures**.

8. Comparación entre distintas comparaciones

En este análisis se han realizado cuatro contrastes, dos contrastes para las rosetas neurales y dos para las células progenitoras neurales. El objetivo es extraer los genes que cambian simultáneamente entre las distintas comparaciones para cada tipo celular por separado.

Para anotar y contar los genes que cambian en una o más condiciones se utiliza la función **decideTest** del paquete *limma*. Esta función nos devolverá una tabla llamada **res** cuya interpretación es la siguiente:

- 1: El gen esta sobreexpresado (Up)
- 0: No hay cambio significativo en la expresión del gen (NotSig)
- -1: El gen ha bajado su expresión (Down)

A continuación se muestra un resumen de la tabla *res* obtenida, donde ya puede observarse que el efecto de la diferenciación celular (contrastes 1 y 3) es mucho más fuerte que el efecto del tratamiento con EtOH (contrastes 2 y 4).

```
##        p40vsRosett0 Rosett20vsRosett0 p40vsNPC0 NPC20vsNPC0
## Down           2049                5      2512          10
## NotSig         2590             5038      1899        5012
## Up              405                1       633          22
```

La representación visual de la tabla *res* es un **Diagrama de Venn** a través de la función *VenDiagram*.

Como en este análisis se ha estudiado cada tipo celular por separado, haré un Diagrama de Venn para las rosetas neurales y otro para las células progenitoras neurales. Con este diagrama se encontrarán los genes que han cambiado su expresión (tanto Upregulated como Downregulated) en cada comparación y cuántos genes han cambiado su expresión simultáneamente debido a los dos efectos: diferenciación celular y tratamiento con EtOH.

Diagrama de Venn para las rosetas neurales

```
vennDiagram (res.selected[,c(1,2)], cex=0.9, circle.col = "lightpink")
title("Genes que cambian su expresión en Rosetas \n Genes seleccionados con FDR < 0.1 y logFC > 1")
```

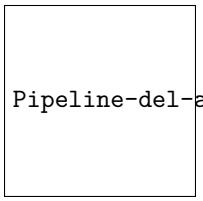

Figure 8: Venn diagram showing the genes in common between the three performed for neural rosette

Del diagrama de Venn para las *rosetas neurales* se obtienen las siguientes conclusiones:

- Existen $2449 + 5 = 2454$ genes que cambian su expresión por el efecto de la diferenciación celular desde H1 hESC hasta rosetas neurales.
- Tan sólo $5 + 1 = 6$ genes cambian su expresión por el efecto del tratamiento con 20mM de EtOH durante la diferenciación hacia rosetas neurales.
- Un total de 5 genes ven alterada su expresión debido a la interación de ambas condiciones: diferenciación celular y tratamiento con EtOH.
  - 1178 genes no cambian sus perfiles de expresión ni por el efecto de la diferenciación celular ni por efecto del tratamiento con EtOH.

Diagrama de Venn para células progenitoras neurales (NPC)

```
vennDiagram (res.selected[,c(3,4)], cex=0.9, circle.col = "lightblue")
title("Genes que cambian su expresión en NPC \n Genes seleccionados con FDR < 0.1 y logFC > 1")
```

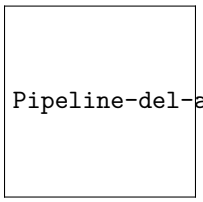

Figure 9: Venn diagram showing the genes in common between the comparisons 3 and 4 performed

El diagrama de Venn para las *NPC* nos deja los siguientes datos:

- Existen $3122 + 23 = 3145$ genes que cambian su expresión por el efecto de la diferenciación celular desde H1 hESC hasta NPC.
- Tan sólo $23 + 9 = 32$ genes cambian su expresión por el efecto del tratamiento con 20mM de EtOH durante la diferenciación hacia NPC.
- Un total de 23 genes ven alterada su expresión debido a la interación de ambas condiciones: diferenciación celular y tratamiento con EtOH en NPC.
  - 479 genes no cambian sus perfiles de expresión ni por el efecto de la diferenciación celular ni por efecto del tratamiento con EtOH en NPC.

Las **conclusiones** que se obtiene de ambos diagramas es que el efecto de la diferenciación celular es mucho más pronunciado por sí solo que el hecho de que dicha diferenciación celular se produzca bajo el tratamiento con 20mM de EtOH. Una segunda conclusión es que las NPC se ven más afectadas que las rosetas neurales bajo el tratamiento con EtOH, puesto que 32 genes cambian su expresión en NPC mientras que tan sólo 6 genes lo hacen en las rosetas neurales por efecto exclusivo de dicho tratamiento.

Otra manera de visualizar los perfiles de expresión de los genes seleccionados a través de la tabla *res* es creando un **Heatmap**. Dichos gráficos nos permiten visualizar las expresiones de cada gen, pero esta vez si vamos a distinguir entre genes *up* o *down* regulados. Además, tanto los genes seleccionados como las 10 muestras se agruparán con el fin de encontrar grupos con patrones de expresión similares.

El *Heatmap* final se recoge en el directorio **figures** como *Heatmap.tiff*.

```
##
## Attaching package: 'gplots'

## The following object is masked from 'package:IRanges':
##
##     space

## The following object is masked from 'package:S4Vectors':
##
##     space

## The following object is masked from 'package:stats':
##
##     lowess
```

#Tabla de anotaciones sencillas

Es posible hacer una tabla de anotaciones con hiperenlaces a las bases de datos para cada anotación de los genes seleccionados anteriormente. Esta tabla es posible gracias al paquete **annaffy** y el resultado se recoge en un archivo llamado *anotaciones.html* dentro del directorio **resultados**.

9. Análisis de significación biológica ("Gene Enrichment Analysis")

Una vez he seleccionado los genes que cambian su expresión entre las diferentes comparaciones, debo darle un sentido biológico a dicha selección. Además de los conocimientos biológicos necesarios para resolver esta cuestión, existe una aproximación estadística llamada **Gene Set Enrichment Analysis** que ayuda a interpretar la significación biológica de estos genes.

El modo de operación del *Gene Set Enrichment Analysis* es identificar las funciones, rutas moleculares y procesos biológicos que aparecen con mayor frecuencia entre las listas de los genes seleccionados.

Existen distintos paquetes para llevar a cabo este análisis de significación biológica, uno de ellos es el paquete de Bioconductor **ReactomePA**. El paquete se basa en un modelo hipergeométrico para evaluar si el número de genes seleccionados asociados a las vías es mayor de lo esperado.

Los pasos a seguir serán preparar la lista de genes que van a ser analizados (estos serán los diferentes contrastes realizados) y asegurarnos de que en todas las listas exista el identificador *Entrez*.

#Preparación de la lista de genes a analizar

Las listas de genes corresponden a los cuatro contrastes realizados. En cada una de las cuatro listas se seleccionarán los genes cuyo p-valor ajustado sea menor a 0.15. De estos genes seleccionados, se convertirá su identificador en un identificador *Entrez*, seleccionado a partir de las anotaciones *hgu133plus2* correspondientes al microarray utilizado.

```
listOfTables <- list(p40vsRosett0 = topTab_p40vsRosett0,
                     Rosett20vsRosett0 = topTab_Rosett20vsRosett0,
                     p40vsNPC0 = topTab_p40vsNPC0,
                     NPC20vsNPC0 = topTab_NPC20vsNPC0)
listOfSelected <- list()
for (i in 1:length(listOfTables)){
  # select the toptable
  topTab <- listOfTables[[i]]
  # select the genes to be included in the analysis
  whichGenes<-topTab["adj.P.Val"]<0.15
  selectedIDs <- rownames(topTab)[whichGenes]
  # convert the ID to Entrez
  EntrezIDs<- select(hgu133plus2.db, selectedIDs, c("ENTREZID"))
  EntrezIDs <- EntrezIDs$ENTREZID
  listOfSelected[[i]] <- EntrezIDs
  names(listOfSelected)[i] <- names(listOfTables)[i]
}
```

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

```
sapply(listOfSelected, length)
```

```
##      p40vsRosett0 Rosett20vsRosett0         p40vsNPC0       NPC20vsNPC0
##              4747               542              4874              1403
```

#Preparación de los genes con anotaciones para el *Homo sapiens*

En este estudio el organismo utilizado es el *Homo sapiens*, por lo que debemos indicar que los genes que han de mapearse tanto para las anotaciones en GO como en KEEG corresponden a dicho organismo.

```
mapped_genes2GO <- mappedkeys(org.Hs.egGO)
mapped_genes2KEGG <- mappedkeys(org.Hs.egPATH)
mapped_genes <- union(mapped_genes2GO , mapped_genes2KEGG)
```

#Generación de los resultados

Gracias a la función **enrichPathway** del paquete *ReactomePA* se consigue obtener el análisis de significación biológica. Esta función selecciona los genes cuyo p-valor es inferior a 0.05 dentro del universo de genes, que son todos los genes disponibles en la anotación **org.Hs.eg** para el *Homo sapines*.

```
library(ReactomePA)
```

```
##
```

```
## Registered S3 method overwritten by 'enrichplot':
##   method               from
##   fortify.enrichResult DOSE
```

```
## ReactomePA v1.30.0  For help: https://guangchuangyu.github.io/ReactomePA
##
## If you use ReactomePA in published research, please cite:
## Guangchuang Yu, Qing-Yu He. ReactomePA: an R/Bioconductor package for reactome pathway analysis and
```

```
listOfData <- listOfSelected[1:4]
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("./results/","ReactomePA.Results.",comparison,".csv"),
            row.names = FALSE)

  pdf(file=paste0("./results/","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("./results/","ReactomePAcnetplot.",comparison,".pdf"))
    print(cnetplot(enrich.result, categorySize = "geneNum", schowCategory = 15,
         vertex.label.cex = 0.75))
  dev.off()

  pdf(file = paste0("./results/","ReactomePAemapplot.",comparison,".pdf"))
    print(emapplot(enrich.result, categorySize = "geneNum", schowCategory = 15,
         vertex.label.cex = 0.75))
  dev.off()
  }
}
```

```
## ##################################
## Comparison:  p40vsRosett0
##                           ID                                    Description
## R-HSA-1474244 R-HSA-1474244           Extracellular matrix organization
## R-HSA-9006934 R-HSA-9006934      Signaling by Receptor Tyrosine Kinases
## R-HSA-3000171 R-HSA-3000171      Non-integrin membrane-ECM interactions
## R-HSA-216083   R-HSA-216083          Integrin cell surface interactions
## R-HSA-1474290 R-HSA-1474290                          Collagen formation
## R-HSA-1650814 R-HSA-1650814 Collagen biosynthesis and modifying enzymes
##               GeneRatio   BgRatio       pvalue     p.adjust       qvalue
## R-HSA-1474244  132/2496 301/10616 2.483537e-15 3.439698e-12 2.990701e-12
## R-HSA-9006934  177/2496 473/10616 2.768214e-12 1.916988e-09 1.666756e-09
## R-HSA-3000171   35/2496  59/10616 3.838557e-09 1.772134e-06 1.540810e-06
## R-HSA-216083    44/2496  85/10616 1.378495e-08 4.773040e-06 4.149997e-06
## R-HSA-1474290   45/2496  90/10616 3.746741e-08 1.037847e-05 9.023731e-06
## R-HSA-1650814   36/2496  67/10616 8.214807e-08 1.896251e-05 1.648726e-05
##
## R-HSA-1474244
## R-HSA-9006934 COL2A1/COL11A1/PDGFC/COL5A2/FLRT3/DUSP4/FGF8/FGFBP3/CXCL12/COL5A1/GABRB3/FN1/ESRP1/COL4
## R-HSA-3000171
## R-HSA-216083
## R-HSA-1474290
## R-HSA-1650814
##               Count
## R-HSA-1474244   132
## R-HSA-9006934   177
## R-HSA-3000171    35
## R-HSA-216083     44
## R-HSA-1474290    45
## R-HSA-1650814    36
## ##################################
## Comparison:  Rosett20vsRosett0
##                           ID
## R-HSA-1474244 R-HSA-1474244
## R-HSA-3000178 R-HSA-3000178
## R-HSA-3000171 R-HSA-3000171
## R-HSA-216083   R-HSA-216083
## R-HSA-1474228 R-HSA-1474228
## R-HSA-2022090 R-HSA-2022090
##                                                          Description
## R-HSA-1474244                          Extracellular matrix organization
## R-HSA-3000178                                      ECM proteoglycans
## R-HSA-3000171                      Non-integrin membrane-ECM interactions
## R-HSA-216083                           Integrin cell surface interactions
## R-HSA-1474228                     Degradation of the extracellular matrix
## R-HSA-2022090 Assembly of collagen fibrils and other multimeric structures
##               GeneRatio   BgRatio       pvalue     p.adjust       qvalue
## R-HSA-1474244    41/290 301/10616 6.950093e-18 5.462773e-15 5.238175e-15
## R-HSA-3000178    18/290  76/10616 1.275950e-12 5.014484e-10 4.808318e-10
## R-HSA-3000171    15/290  59/10616 3.322878e-11 8.705941e-09 8.348003e-09
## R-HSA-216083     17/290  85/10616 9.866529e-11 1.938773e-08 1.859062e-08
## R-HSA-1474228    21/290 140/10616 1.801590e-10 2.832099e-08 2.715660e-08
## R-HSA-2022090    13/290  61/10616 7.316348e-09 9.584416e-07 9.190360e-07
```

```
##
## R-HSA-1474244 SPP1/LUM/LTBP2/COL8A1/COL12A1/COL4A1/COL5A1/TNC/MMP2/LOXL2/CD44/HAPLN1/TGFB2/ADAM12/FN
## R-HSA-3000178
## R-HSA-3000171
## R-HSA-216083
## R-HSA-1474228
## R-HSA-2022090
##               Count
## R-HSA-1474244    41
## R-HSA-3000178    18
## R-HSA-3000171    15
## R-HSA-216083     17
## R-HSA-1474228    21
## R-HSA-2022090    13
## ####################################
## Comparison:  p40vsNPC0
##                            ID                                 Description GeneRatio
## R-HSA-9006934 R-HSA-9006934 Signaling by Receptor Tyrosine Kinases  184/2548
## R-HSA-1474244 R-HSA-1474244      Extracellular matrix organization  126/2548
## R-HSA-112316   R-HSA-112316                         Neuronal System  158/2548
## R-HSA-3000171 R-HSA-3000171 Non-integrin membrane-ECM interactions   38/2548
## R-HSA-399956   R-HSA-399956              CRMPs in Sema3A signaling   15/2548
## R-HSA-2022928 R-HSA-2022928                    HS-GAG biosynthesis   22/2548
##                 BgRatio       pvalue     p.adjust       qvalue
## R-HSA-9006934 473/10616 1.199528e-13 1.657748e-10 1.416706e-10
## R-HSA-1474244 301/10616 3.155830e-12 2.180679e-09 1.863601e-09
## R-HSA-112316  413/10616 3.274638e-11 1.508517e-08 1.289173e-08
## R-HSA-3000171  59/10616 4.765929e-11 1.646628e-08 1.407203e-08
## R-HSA-399956   16/10616 6.080268e-09 1.680586e-06 1.436223e-06
## R-HSA-2022928  31/10616 4.268041e-08 9.830722e-06 8.401302e-06
##
## R-HSA-9006934 COL2A1/DUSP4/ESRP1/COL11A1/APOE/KITLG/FLRT3/PDGFC/COL5A2/FGFBP3/CXCL12/RPS6KA2/AKT3/ERI
## R-HSA-1474244
## R-HSA-112316
## R-HSA-3000171
## R-HSA-399956
## R-HSA-2022928
##               Count
## R-HSA-9006934   184
## R-HSA-1474244   126
## R-HSA-112316    158
## R-HSA-3000171    38
## R-HSA-399956     15
## R-HSA-2022928    22
## ####################################
## Comparison:  NPC20vsNPC0
##                            ID                                 Description GeneRatio
## R-HSA-3000171 R-HSA-3000171   Non-integrin membrane-ECM interactions    23/743
## R-HSA-1474244 R-HSA-1474244        Extracellular matrix organization    57/743
## R-HSA-9006934 R-HSA-9006934   Signaling by Receptor Tyrosine Kinases    71/743
## R-HSA-3000178 R-HSA-3000178                          ECM proteoglycans    20/743
## R-HSA-1474228 R-HSA-1474228  Degradation of the extracellular matrix    28/743
## R-HSA-6794362 R-HSA-6794362 Protein-protein interactions at synapses    21/743
```

```
##                  BgRatio       pvalue     p.adjust       qvalue
## R-HSA-3000171  59/10616 2.530799e-12 1.509679e-09 1.375924e-09
## R-HSA-1474244 301/10616 2.700677e-12 1.509679e-09 1.375924e-09
## R-HSA-9006934 473/10616 4.644724e-10 1.730934e-07 1.577576e-07
## R-HSA-3000178  76/10616 1.598340e-07 4.467359e-05 4.071560e-05
## R-HSA-1474228 140/10616 3.343902e-07 7.196692e-05 6.559079e-05
## R-HSA-6794362  87/10616 3.862268e-07 7.196692e-05 6.559079e-05
##
## R-HSA-3000171
## R-HSA-1474244                                                                                                                                LUM
## R-HSA-9006934 FGFR2/ANOS1/DUSP6/SPRY2/SPRY1/VEGFA/FGF8/LYN/LAMC1/KITLG/PRKCB/SPP1/IRS4/SPRED2/VEGFC/
## R-HSA-3000178
## R-HSA-1474228
## R-HSA-6794362
##                Count
## R-HSA-3000171     23
## R-HSA-1474244     57
## R-HSA-9006934     71
## R-HSA-3000178     20
## R-HSA-1474228     28
## R-HSA-6794362     21
```

Una vez el análisis de significación ha finalizado, obtenemos cuatro tipos de documentos para cada una de las cuatro comparaciones:

- Un **archivo excel** de extensión *.csv* donde se resumen de todas las vías enriquecidas en cada proceso y sus correspondientes estadísticos.
- Un **barplot** que recoge las mejores vías enriquecidas en cada proceso. La altura de cada barra corresponde al número de genes que están relacionados con dicha ruta metabólica enriquecida. Además, las vías están ordenadas en función de su significación estadística.
- Un **cnetplot** que representa la red de vías enriquecidas en cada proceso, donde se incluyen las relaciones entre los genes incluídos en dichas vías. Este plot permite extraer la compleja relación entre los distintos genes y las enfermedades asociadas.
- Un **emapplot** donde nuevamente se visualizan los genes en forma de red, en este caso los conjuntos de genes que se superponen mutuamente tienden a agruparse para facilitar su interpretación.

Para este estudio se han analizado cuatro comparaciones o conjuntos de genes, los cuales, además, incluyen una enorme cantidad de genes. Para tener una visión más clara de los resultados del análisis de significación biológica, he construído una tabla para cada comparación donde se resumen las cinco vías más enriquecidas y sus estadísticos. Sin embargo, ya podemos adelantar que en las cuatro situaciones analizadas la ruta de organización de la matriz extracelular se encuentra muy enriquecida.

Tabla resumen para p40vsRosett0

Table 2: First rows and columns for Reactome results on p40vsRosett0.csv comparison

	Description	GeneRatio	BgRatio	pvalue	p.adjust
R-HSA-1474244	Extracellular matrix organization	132/2496	301/10616	2.48353664965071e-15	3.4396982
R-HSA-9006934	Signaling by Receptor Tyrosine Kinases	177/2496	473/10616	2.76821416020616e-12	1.9169883
R-HSA-3000171	Non-integrin membrane-ECM interactions	35/2496	59/10616	3.83855674324901e-09	1.7721336
R-HSA-216083	Integrin cell surface interactions	44/2496	85/10616	1.37849541626087e-08	4.7730403
R-HSA-1474290	Collagen formation	45/2496	90/10616	3.74674126835196e-08	1.0378473

Tabla resumen para Rosett20vsRosett0

Table 3: First rows and columns for Reactome results on Rosett20vsRosett0.csv comparison

	Description	GeneRatio	BgRatio	pvalue	p.adjust
R-HSA-1474244	Extracellular matrix organization	41/290	301/10616	6.95009314084962e-18	5.4627732
R-HSA-3000178	ECM proteoglycans	18/290	76/10616	1.27595018881832e-12	5.0144842
R-HSA-3000171	Non-integrin membrane-ECM interactions	15/290	59/10616	3.32287817292416e-11	8.7059408
R-HSA-216083	Integrin cell surface interactions	17/290	85/10616	9.86652937566947e-11	1.9387730
R-HSA-1474228	Degradation of the extracellular matrix	21/290	140/10616	1.80158983714342e-10	2.8320992

Tabla resumen para p40vsNPC0

Table 4: First rows and columns for Reactome results on p40vsNPC0.csv comparison

	Description	GeneRatio	BgRatio	pvalue	p.adjust
R-HSA-9006934	Signaling by Receptor Tyrosine Kinases	184/2548	473/10616	1.199527923865e-13	1.6577475
R-HSA-1474244	Extracellular matrix organization	126/2548	301/10616	3.15583025702968e-12	2.1806787
R-HSA-112316	Neuronal System	158/2548	413/10616	3.27463790914985e-11	1.5085165
R-HSA-3000171	Non-integrin membrane-ECM interactions	38/2548	59/10616	4.7659285783892e-11	1.6466283
R-HSA-399956	CRMPs in Sema3A signaling	15/2548	16/10616	6.08026782429621e-09	1.6805860

Tabla resumen para NPC20vsNPC0

Table 5: First rows and columns for Reactome results on NPC20vsNPC0.csv comparison

	Description	GeneRatio	BgRatio	pvalue	p.adjust
R-HSA-3000171	Non-integrin membrane-ECM interactions	23/743	59/10616	2.53079932659888e-12	1.5096785
R-HSA-1474244	Extracellular matrix organization	57/743	301/10616	2.70067714678017e-12	1.5096785
R-HSA-9006934	Signaling by Receptor Tyrosine Kinases	71/743	473/10616	4.64472384774152e-10	1.7309337
R-HSA-3000178	ECM proteoglycans	20/743	76/10616	1.59833958740309e-07	4.4673591
R-HSA-1474228	Degradation of the extracellular matrix	28/743	140/10616	3.3439024171866e-07	7.1966923