Desregulación molecular inducida por el alcohol durante la diferenciación neural de las células madre embrionarias humanas

Marina Ballesteros

4/26/2020

Contents

1 Abstract

El estudio en el que se basa este análisis, ha realizado un experimento con microarray de expresión génica estudiando la diferenciación de células precursoras neurales desde células madre embrionarias (ESC) en presencia o ausencia del tratamiento con etanol (EtOH). Los perfiles transcriptómicos de todo el genoma identificaron las alteraciones moleculares inducidas por la exposición al etanol durante la diferenciación neuronal de las ESC en rosetas neuronales y poblaciones de células precursoras neuronales.

Los datos y el código del análisis se encuentran en el siguiente repositorio github [https://github.com/marinabf93/Effect-of-alcool-in-in-ESC-differentiation-]

2 Objetivos

En este estudio se pueden diferenciar dos claros objetivos:

a) Demostrar los posibles efectos teratogénicos del alcohol en el desarrollo fetal y, por tanto, los consecuentes defectos de desarrollo debidos al abuso del alcohol durante la gestación.

b) Determinar los mecanismos específicos por los que el alcohol media estas posibles lesiones y determinar si el alcohol tiene un efecto significativo en los mecanismos reguladores moleculares y celulares de la diferenciación de las células madre embrionarias (ESC), incluidos los genes que intervienen en el desarrollo neuronal.

3 Materiales

La razón de este trabajo es analizar bioinformáticamente los datos de un experimento con microarrays. El experimento en el que he basado mi análisis se recoge en dos artículos:

a) "Molecular effect of ethanol during neural differentiation of human embryonic stem cells in vitro" (1)

b) "Alcohol-Induced Molecular Dysregulation in Human Embryonic Stem Cell-Derived Neural Precursor Cells" (2)

3.1 Diseño experimental

El **tipo de experimento** corresponde al análisis de microarrays, donde a través del diseño de un experimento se intenta responder a las cuestiones biológicas planteadas en los objetivos. Con el uso de la estadística y las diferentes herramientas bioinformáticas, se pretende procesar, analizar, visualizar y analizar los datos con el fin de responder a las cuestiones biológicas de partida.

En el estudio se investiga el efecto del alcohol (EtOH) en el desarrollo de células madre neurales derivadas desde células madre de embriones humanos. La metodología del experimento es la siguiente:

Primero se cultivan células madre embrionarias (ESC) durante cinco días en un medio de inducción neural (NIM). A continuación, los agregados neuronales que se formaron fueron sembrados en placas recubiertas con poli-L-ornitina/laminina y cultivados con NIM durante siete días para desarrollar la estructura de roseta

neuronal. Al día siguiente de colocar los agregados neurales en dichas placas, se produjo el tratamiento con 20mM de EtOH. Las células fueron alimentadas con medio fresco todos los días alternando el tratamiento con 20 mM de etanol durante un día y dejando otro día de reposo sin tratamiento. Después de siete días, las rosetas neurales fueron desalojadas y luego replateadas en medio NIM para la expansión de las células precursoras neurales (NPC) durante 5 días, siguiendo el mismo procedimiento para el tratamiento con 20mM de EtOH.

3.2 Datos

El material en el que he basado mi análisis ha sido descargado desde la página del Gene Expression Omnibus (GEO) (3) . El GEO es un depósito de datos públicos de genómica funcional donde se aceptan datos basados en matrices y secuencias. Además, se proporcionan herramientas para ayudar a los usuarios a consultar y descargar experimentos y perfiles de expresión génica corregidos.

El conjunto de datos utilizados en este análisis se identifica con el número de adhesión: **GSE56906**:[https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE56906]. En este caso se han analizado 10 muestras distintas.

El **tipo de microarray** utilizado ha sido del tipo *Affymetrix Human Genome U133 Plus 2.0 Array*; cuyo fabricante es Affymetrix, uno de los principales vendedores de tecnología de microarray.

3.3 Software

Para comenzar el análisis se necesita instalar **R statistical software** el cual permite hacer análisis estadísticos, representaciones gráficas y lectura y creación de documentos en diferentes formatos. El software se puede descargar en la página web [https://cran.r-project.org/index.html] y solo deben seguirse las instrucciones indicadas en función del tipo de software del ordenador que se utilice para el análisis. El análisis de microarray que se presenta en este informe ha sido desarrollado con la versión 3.6.2 y todos los análisis se han llevado a cabo con la interfaz *RStudio*. Esta interfaz puede descargarse desde la página principal [https://www.rstudio.com/]

4 Métodos: Procedimiento general del análisis ("Workflow")

El flujo de trabajo se resumen en la siguiente imagen "FlujoTrabajo.jpg" dentro del directorio **figures**.

A continuación resumiré de forma muy general los métodos utilizados en cada paso del flujo de trabajo, así como los datos de entrada y los de salida. El desarrollo detallado de cada paso del análisis lo encontraréis en el archivo **Pipeline del análisis.Rmd** dentro del directorio principal del repositorio indicado al inicio de este informe.

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_EtOH_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.

4.1 Identificación de los grupos y clasificación de las muestras

En este caso se han introducido los 10 archivos **.CEL** correspondientes a las 10 muestras de partida del experimento. Estos archivos contienen los datos en crudo originados tras el escaneo y preprocesado de los microarrays. Además, se ha leído el archivo **targets.csv**, el cual ha sido creado manualmente y en el que se incluye la información de los diferentes grupos y variables.

Tras la lectura conjunta de ambos archivos, se crea un objeto **rawData** con el fin de resumir toda la información anterior. El resultado es la siguiente tabla:

Table 1: Contenido del objeto que contiene los datos crudos utilizados para el análisis actual

FileName	Group	CellType	Treatment	ShortName
GSM1371025	H1 EtOH 0 Rosett	Roseta diferenciada	Sin alcohol	Rosett.0.1
GSM1371026	H1 EtOH 0 Rosett	Roseta diferenciada	Sin alcohol	Rosett.0.2
GSM1371027	H1 EtOH 20 Rosett	Roseta diferenciada	Con alcohol	Rosett.20.1
GSM1371028	H1 EtOH 20 Rosett	Roseta diferenciada	Con alcohol	Rosett.20.2
GSM1371029	H1 p40 Und	ESC	Sin alcohol	P40.1
GSM1371030	H1 p40 Und	ESC	Sin alcohol	P40.2
GSM1371031	NPC H1 EtOH 0	NPC diferenciada	Sin alcohol	NPC.0.1
GSM1371032	NPC H1 EtOH 0	NPC diferenciada	Sin alcohol	NPC.0.2
GSM1371033	NPC H1 EtOH 20	NPC diferenciada	Con alcohol	NPC.20.1
GSM1371034	NPC H1 EtOH 20	NPC diferenciada	Con alcohol	NPC.20.2

4.2 Control de calidad de los datos crudos

El objetivo de esta etapa es saber si los datos en crudo tienen la calidad suficiente como para ser normalizados. El control de calidad se lleva a cabo con el paquete **arrayQualityMetrics** (4), este paquete efectúa distintos análisis con el fin de identificar los valores outliers a través de valores umbrales predefinidos.

Los datos de entrada son los datos en crudo (*rawData*) y se devuelve un informe del control de calidad llamado *index.html* guardado en el directorio **results**. En este informe se encontrarán un boxplot de intensidad, análisis de componentes principales y MA plots entre otros. El criterio para eliminar un array del experimento es que este debe ser marcado tres veces como outlier. En este caso sólo se han marcado los arrays 7 y 9 como outliers bajo la detección del MA plot; sin embargo, ambos arrays se conservarán pues sólo han sido marcados por uno de los tres criterios.

A continuación se expone el **análisis de componentes principales** para los datos en crudo. Este gráfico nos permite ver los dos componentes principales, capaces de explicar el mayor porcentaje de variabilidad de las muestras.

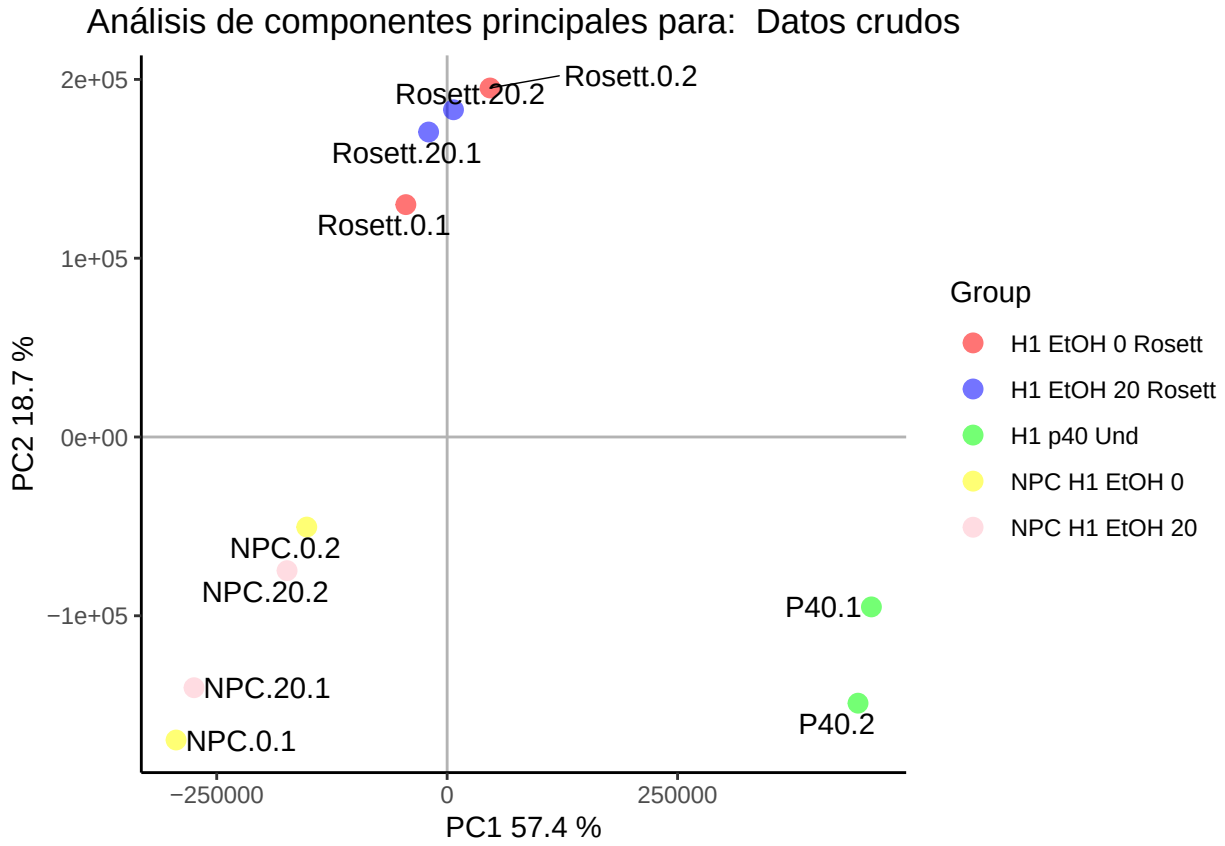

Figure 1: Visualización de los dos componentes principales para los datos en crudo.

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 muestras están separadas por su tipo celular y no por el tratamiento con EtOH. Las células indiferenciadas (p40) se encuentran a la derecha del gráfico mientras que las células más 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).

4.3 Normalización

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*.

Los datos de entrada son nuevamente los datos crudos *rawData* y como output se obtiene un ExpressionSet llamado **eset_rma**, que contiene los datos normalizados.

4.4 Control de calidad de los datos normalizados

Se realiza el mismo procedimiento que para el control de calidad de los datos en crudo; sin embargo, esta vez el input es el vector *eset_rma* y como output se obtiene nuevamente una carpeta donde se encuentra informe del control de calidad *index.html*. Para los datos normalizados se detectaron dos outliers, los arrays 5 y 6, que también se conservan en el estudio puesto que tan sólo fueron detectados por el método de distancias entre arrays.

Tras la normalización de los datos, se observa cómo el análisis de componentes principales ha variado ligeramente. El componente principal de variabilidad de los datos normalizados sigue siendo el *tipo celular*. Sin embargo, en este caso, dicho factor explica el 69,5% de la variabilidad total de las muestras, un porcentaje mayor de variabilidad que para los datos en crudo. Es decir, el factor *tipo celular* en los datos normalizados es responsable de la mayor variabilidad de las muestras un 12.1% más que en los datos en crudo.

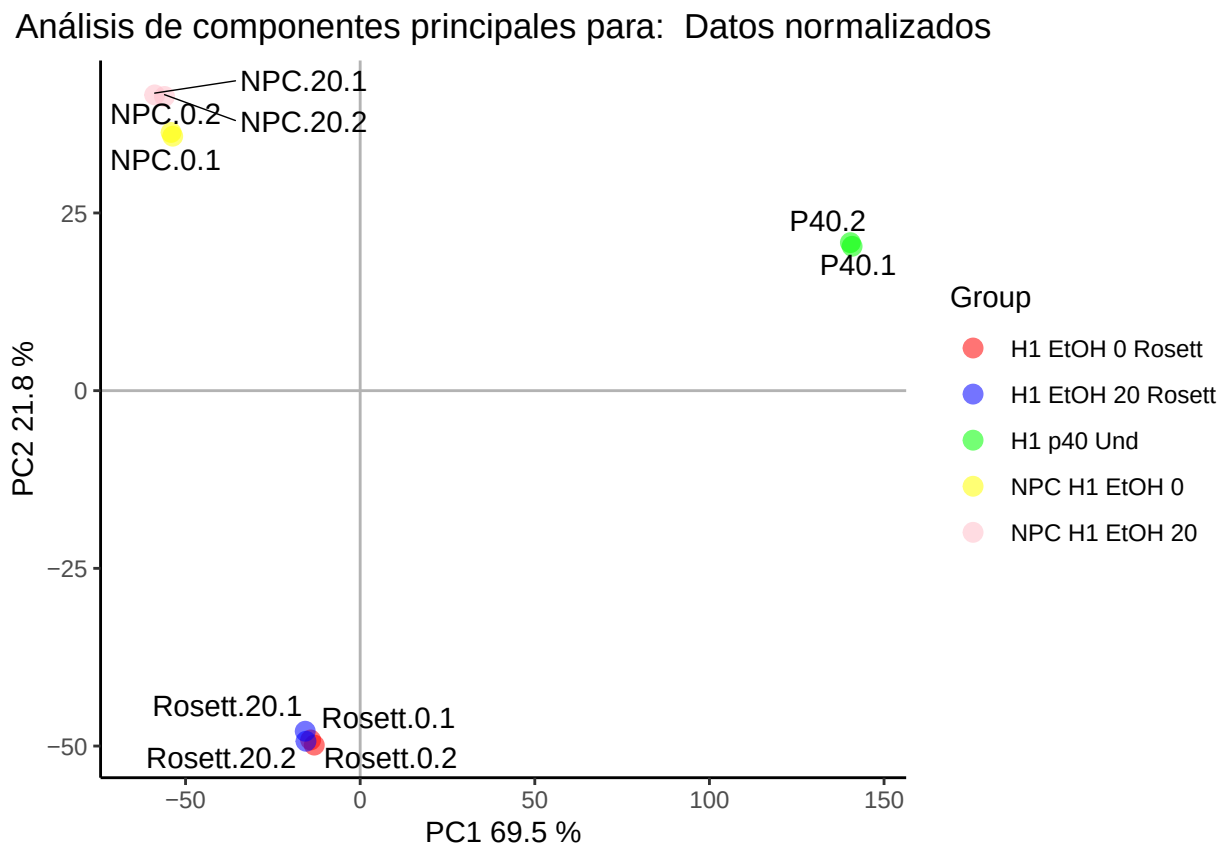

Figure 2: Visualización de los dos componentes principales para los datos normalizados.

Finalmente, se puede comprobar en la *Figura 3* como las intensidades de todas las muestras estan alineadas 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.

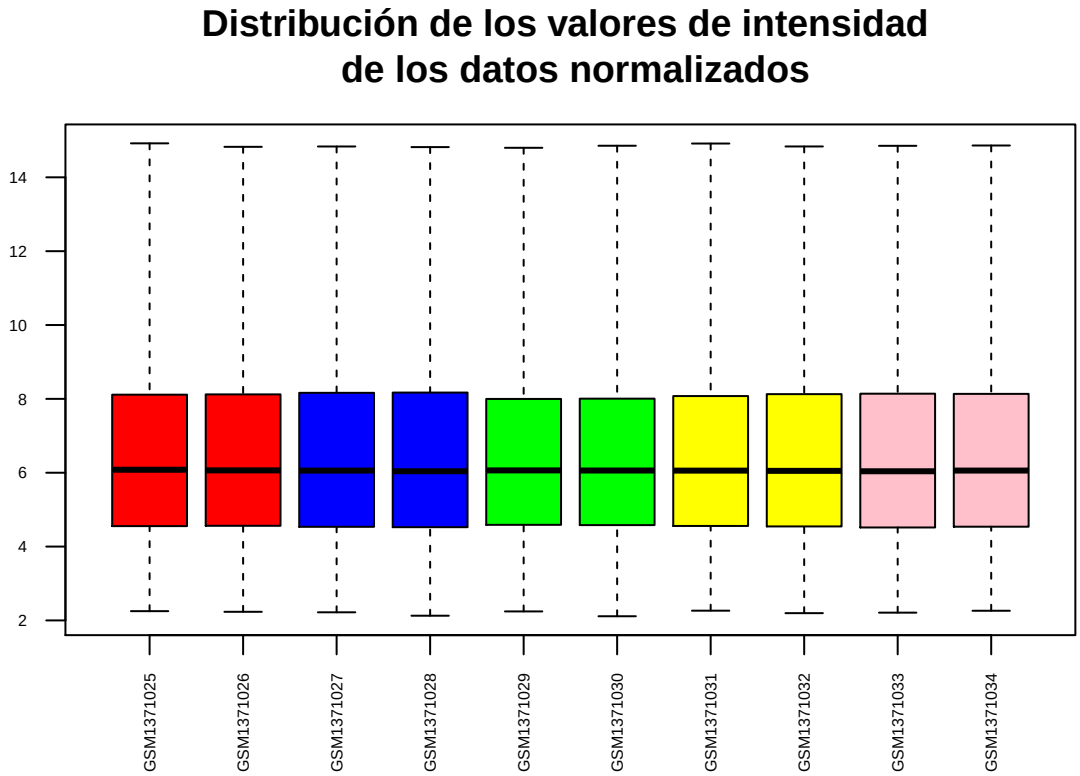

Figure 3: Boxplot para las intensidades de los arrays (Datos Normalizados)

4.5 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 porque varían muy poco entre las distintas condiciones; por lo que su variación podría deberse únicamente a la variación aleatoria. La función encargada de hacer esta tarea es *nsFilter* del paquete **genefilter** (5). 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* y su librería de anotaciones es **hgu133plus2.db**.

En esta etapa el input son los datos normalizados y el output son los genes filtrados en un objeto llamado **eset_filtered** cuya clase sigue siendo del tipo *ExpressionSet*.

4.6 Identificación de genes diferencialmente expresados

Para identificar los genes diferencialmente expresados existen varios métodos. Hasta el momento, el método que mejores resultados ofrece es el de **Modelos lineales para microarrays**. Dicho método está implementado en el paquete **limma** (6).

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* para devolvernos la siguiente matriz de diseño.

```
##            Rosett.0 Rosett.20 p40 NPC.0 NPC.20
## GSM1371025        1         0   0     0      0
## GSM1371026        1         0   0     0      0
## GSM1371027        0         1   0     0      0
## GSM1371028        0         1   0     0      0
## GSM1371029        0         0   1     0      0
## GSM1371030        0         0   1     0      0
## GSM1371031        0         0   0     1      0
## GSM1371032        0         0   0     1      0
## GSM1371033        0         0   0     0      1
## GSM1371034        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
```

Una vez que las matrices de diseño y contraste están creadas, el último paso sería 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.*(7)

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*. Por tanto, el intput de esta etapa son los datos filtrados mientras que el output será un objeto que se llamará **fit.main**.

Para terminar con la identificación de los genes diferencialmente expresados, debe obtenerse 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.

En este caso los datos de entrada será el objeto *fit.main* obtenido anteriormente y los datos de salida una tabla para cada uno de los cuatro contrastes llevados a cabo.

4.7 Anotación de los resultados

El proceso de anotación es la asociación de cada identificador de Affymetrix con cada gen o conjuntos de sondas de cada array a través de identificadores más sencillos de manejar como el *Gene Symbol*, *Entrez Gene* o *Gene description*, además de complementar los genes seleccionados con la máxima información posible.

Los datos de entrada del proceso de anotación serán el paquete de anotaciones asociado al tipo de microarray utilizado, en este caso este paquete es **hgu133plus2.db**, y las tablas de genes seleccionados creadas anteriormente. El output de esta etapa serán nuevamente tablas de selección de genes a las que se les han añadido las anotaciones e informaciones pertinentes.

Además del listado de genes junto a sus anotaciones, se obtiene también una salida visual a través de un **VolcannoPlot**. A continuación, se muestra el ejemplo del resultado de la selección de genes para la primera comparación *p40vsRosett0*.

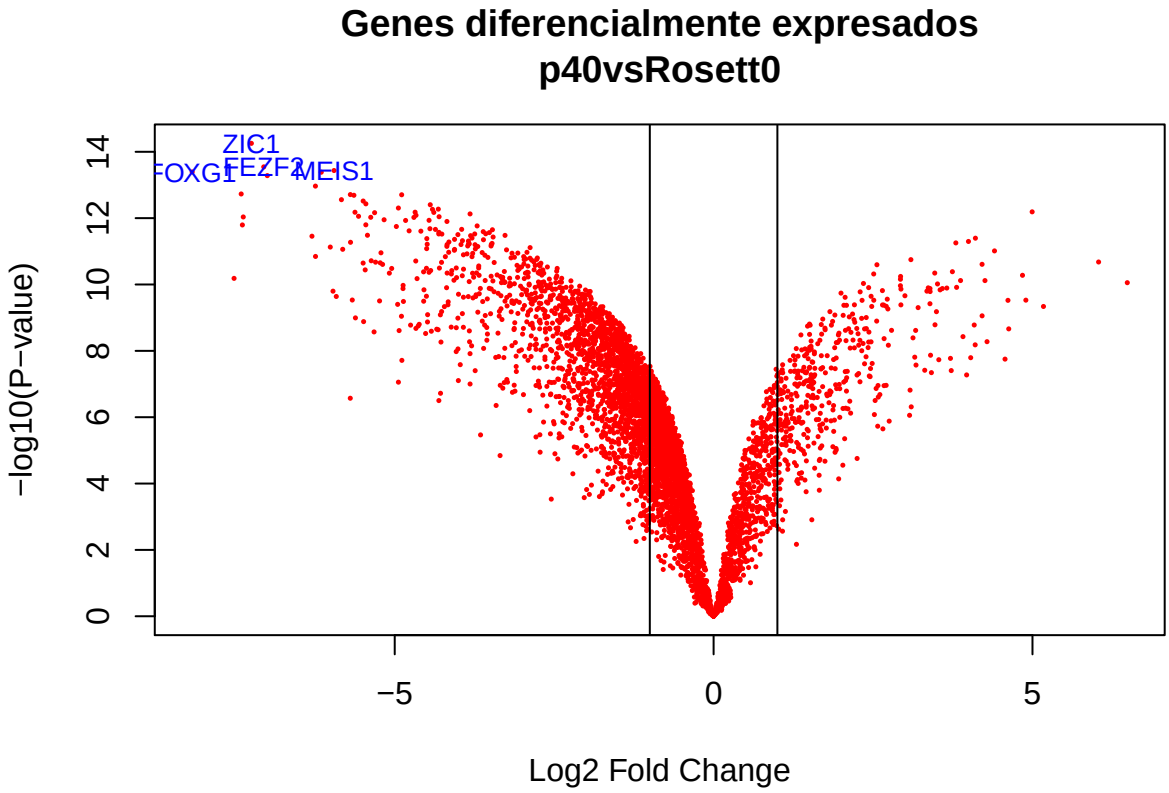

Figure 4: Selección de los cuatro primeros genes en la topTab para el contraste p40vsRosett0

4.8 Comparación entre distintas comparaciones

El objetivo es extraer los genes que cambian simultáneamente entre las distintas condiciones experimentales.

Para lograr dicho objetivo, se utilizará una función llamda **decideTest** del paquete *lima* donde los datos de entrada serán los genes seleccionados en *fit.main*. Como salida obtendremos una tabla donde se contabilizan los genes que estan sobreexpresados (Up), los genes que han reducido su expresión (Down) y los genes que no han cambiado su nivel de expresión (NotSig).

Visualmente esta tabla se representa a través de un **Diagrama de Venn** que se expone en la sección de resultados.

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

4.9 Análisis de significación biológica ("Gene Enrichment Analysis")

En esta etapa se intenta dar un sentido biológico a la lista de genes obtenida por expresarse diferencialmente en distintas condiciones. El objetivo es identificar las funciones, rutas metabólicas o procesos biológicos donde interviene los genes seleccionados con el fin de saber si estas funciones, vías o procesos aparecen con mayor frecuencia en la lista y, por tanto, son susceptibles de alterarse.

Los datos de entrada serán las listas de genes obtenidas con la función *topTab* y los datos de salida serán un archivo excel por cada comparación realizada en el estudio así como diferentes gráficos donde se representan las vías enriquecidas en cada proceso y los genes que intervienen en ellas.

Esta etapa se lleva a cabo gracias al paquete **ReactomePA** (8) de Bioconductor.

5 Resultados

Este análisis de microarrays nos deja varios resultados. Para reflejar los resultados de forma más ordenada, he decidido agrupar los resultados en función de las distintas conclusiones que nos dejan.

5.1 Principal componente de variación

Uno de los resultados obtenidos al analizar el microarray es que el principal factor de variación en las 10 muestras estudiadas es el factor **tipo celular** y no el tratamiento con alcohol. Por lo tanto, la primera conclusión es que la diferenciación celular por si misma afecta en mayor medida al buen desarrollo de las células neurales más que el tratamiento con 20mM de EtOH.

El gráfico "Estimación PVCA" representa el análisis de variación de componentes principales. En él se muestra tanto el porcentaje de variación que representa cada factor de estudio por separado como el porcentaje de variación debida a la interación de ambos factores en conjunto sobre las muestras estudiadas.

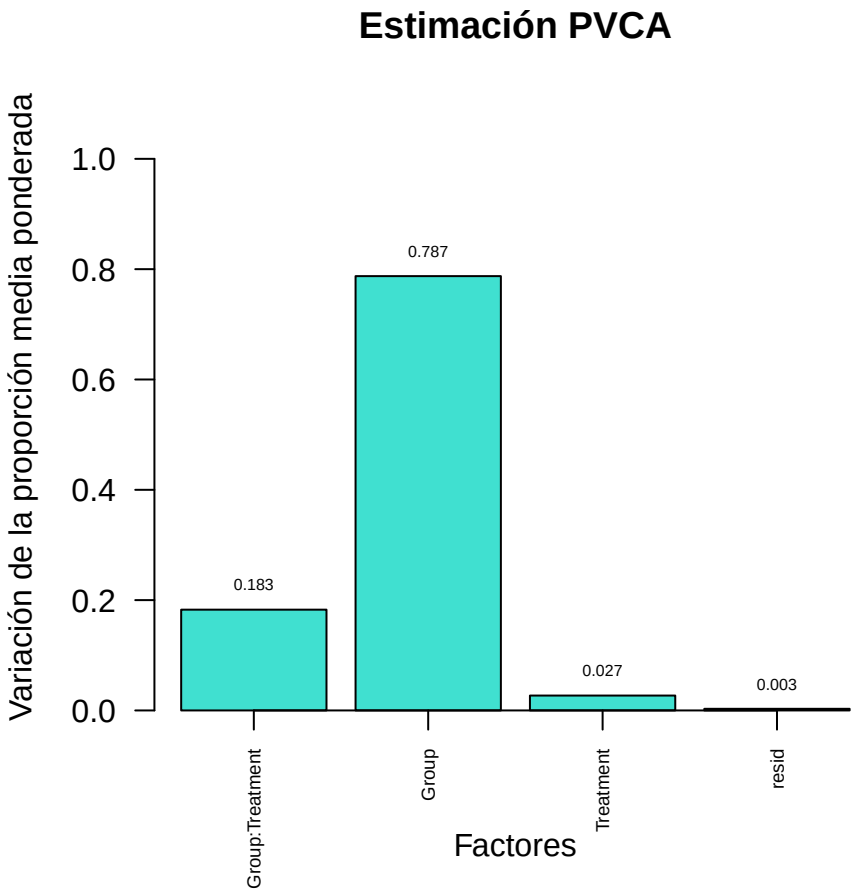

Figure 5: Análisis de variación de componentes principales. Importancia relativa de los dos factores: tipo celular y tratamiento con EtOH sobre la variabilidad de las muestras estudiadas.

5.2 Tratamiento con EtOH

Tras haber comparado el tratamiento con 20mM de etanol durane 7 días alternos en rosetas neurales y durante 5 días alternos en células precursoras neurales, se ha demostrado que el tratamiento con alcohol afecta de manera más agresiva al desarrollo de las células precursoras neurales; mientras que la diferenciación desde agregados neurales hasta rosetas se ve menos afectado.

Esta conclusión ha sido posible gracias al **Diagrama de Venn**. Al ser dos tipos celulares (rosetas y NPC) bastantes diferentes, en el experimento se decidió estudiar el desarrollo celular de ambos tipos por separado pero siguiendo su evolución bajo el tratamiento con etanol. A continuación se muestran los Diagramas de Venn obtenidos para cada tipo celular, donde se representan los genes que han modificado su expresión por efecto de la diferenciación celular, los genes que han modificado su expresión por efecto del tratamiento con 20mM de EtOH y los genes que han cambiado significativamente su expresión por efecto de la interacción de ambos tratamientos.

Diagrama de Venn para las rosetas neurales

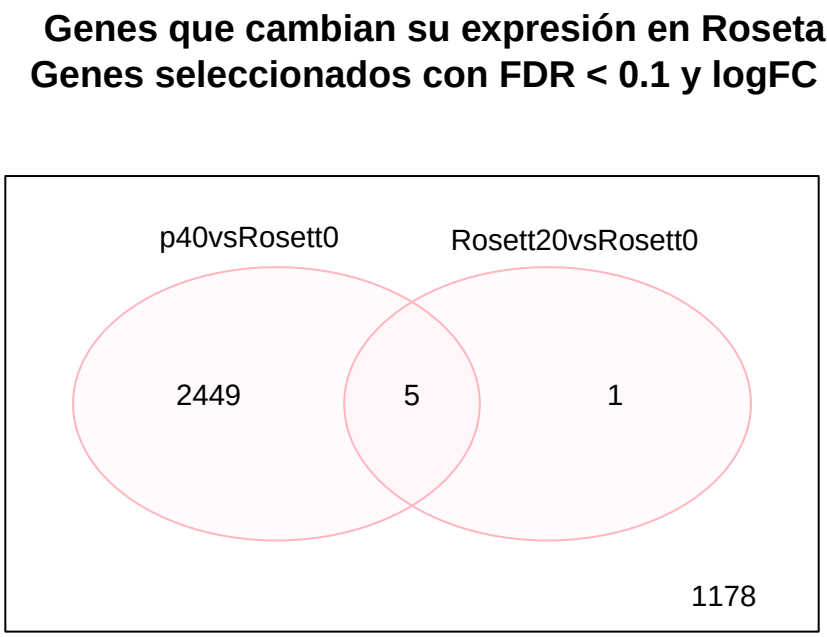

Figure 6: Diagrama de Venn para mostrar los genes en común entre el efecto de la diferenciación celular y del tratamiento con EtOH en rosetas neurales.

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)

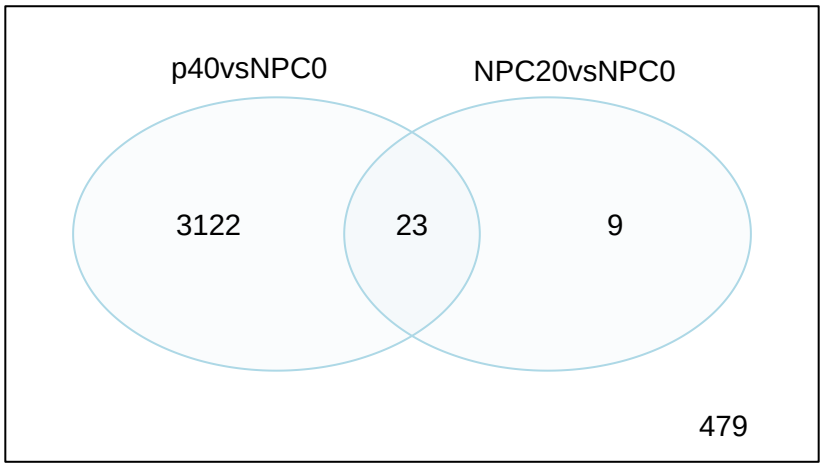

Figure 7: Diagrama de Venn para mostrar los genes en común entre el efecto de la diferenciación celular y del tratamiento con EtOH en NPC.

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.

5.3 Visualización de los perfiles de expresión génica

La manera de visualizar los perfiles de expresión de los genes seleccionados es creando un **Heatmap** o mapa de color. Dichos gráficos nos permiten visualizar las expresiones de cada gen permitiendo la distincción entre genes regulados positivamente (upregulated) o genes regulados negativamente (downregulated).

En este análisis, se han representado sólamente aquellos genes que se han considerado diferencialmente expresados en alguna de las cuatro comparaciones realizadas. Tanto los genes seleccionados como las 10 muestras estudiadas se agrupan con el fin de encontrar grupos con patrones de expresión similares.

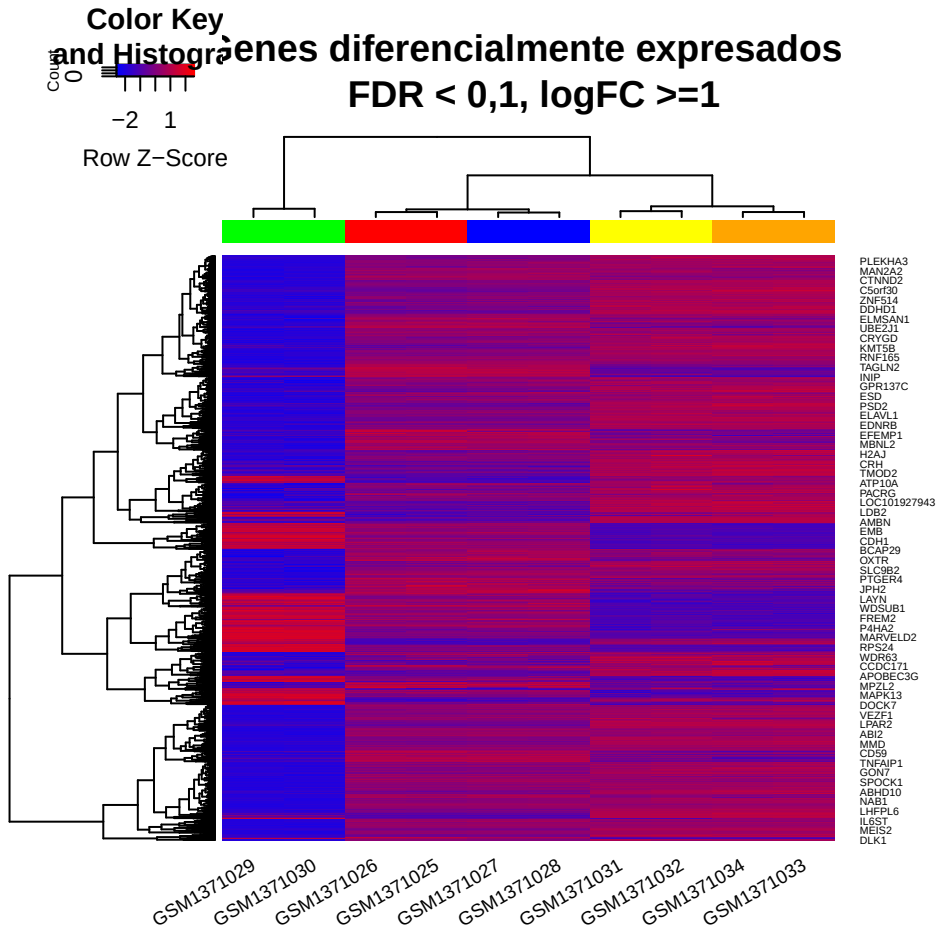

Figure 8: Heatmap de los perfiles de expresión de los genes seleccionados en las muestras estudiadas.

La escala de colores del Heatmap representa en azul los genes que reducen su expresión (downregulated) y en rojo los genes que aumentan su expresión normal (upregulated). En el gráfico se aprecian claramente grupos de genes con patrones de variación comunes, que coinciden con los tres tipos celulares:

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

5.4 Análisis de significación biológica

Gene Set Enrichment Analysis es la aproximación estadística que ayuda a interpretar la significación biológica de los genes seleccionados que cambian su expresión. Su objetivo es identificar las funciones, rutas moleculares y procesos biológicos que aparecen con mayor frecuencia entre las listas de los genes seleccionados para deducir cuáles son los procesos y funciones alterados por la diferente expresión de esos genes.

En este análisis se ha utilizado 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. Gracias a la función **enrichPathway** de este mismo paquete, se seleccionan 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.db** para el *Homo sapines*.

Los resultados completos están disponibles en el directorio **results** del repertorio Github proporcionado al inicio de este informe.

Para visualizar en este informe un ejemplo concreto de este análisis de significación biológica, he incluído los resultados de las vías enriquecidas debido al tratamiento con 20mM de EtOH durante la diferenciación de las ESC hasta NPC en comparación con el proceso de diferenciación sin el tratamiento.

El primer gráfico es un **barplot** donde se recogen las vías más enriquecidas y se representan en forma de barras. La longitud de cada barra corresponde al número de genes que están implicados en las vías más enriquecidas. Además, las vías están ordenadas en función de su significación estadística.

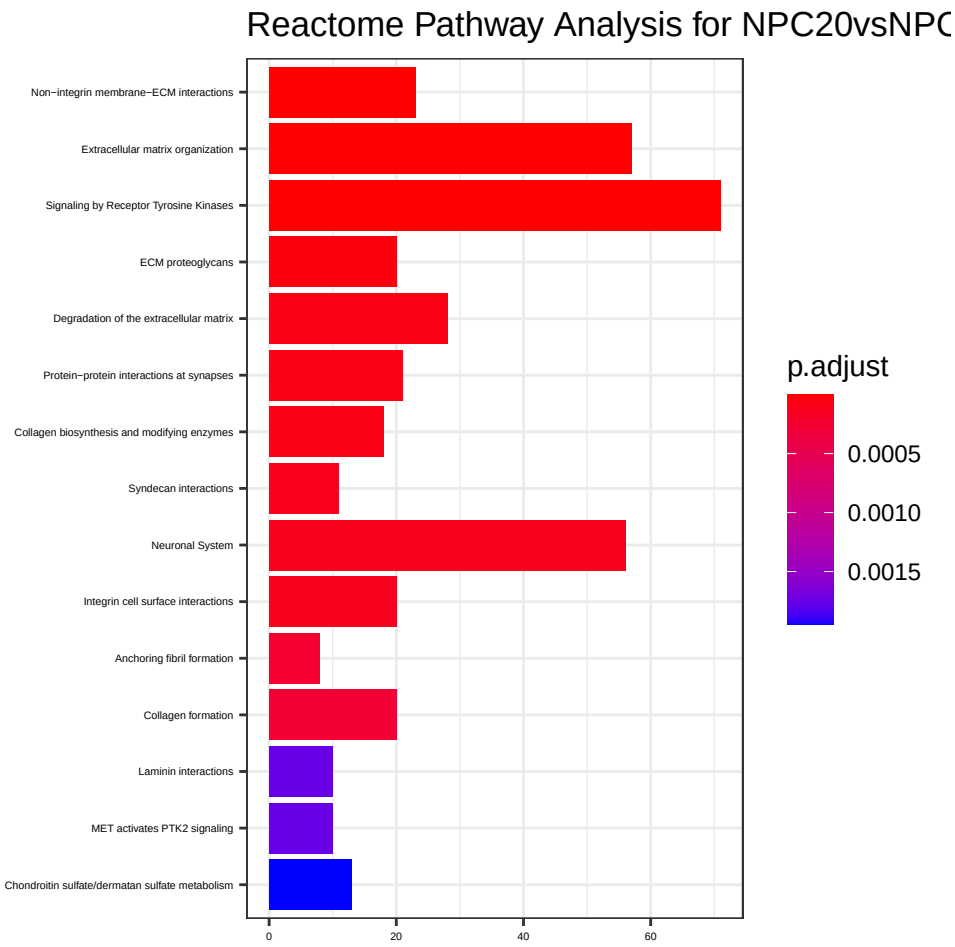

Figure 9: Barplot para el análisis de enriquecimiento de las NPC bajo el tratamiento con alcohol.

En el gráfico se confirma que las vías más enriquecidas por el tratamiento con 20mM de EtOH durante la diferenciación celular desde ESC hasta NPC son:

- La vía de señalización del receptor tirosina quinasa.
- El proceso de organización de la matriz extracelular.
- El sistema neuronal.

El segundo gráfico es un **cnetplot** que representa los vínculos entre los genes y los conceptos biológicos implicados como una red; es decir, se dibuja la red de vías enriquecidas en este 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 a ellos y a los procesos donde intervienen.

Para el caso de la comparación entre las células NPC tratadas con alcohol y las NPC que no han sido tratadas con EtOH se ven afectados los siguientes procesos: en mayor medida la organización de la matriz extracelular y la vía de señalización del receptor tirosina quinasa. En menor medida se alteran la degradación de la matriz extracelular, los proteoglicanos de la matriz y el proceso de interacción entre las proteínas no integradas de la membrana con las proteínas de la matriz extracelular.

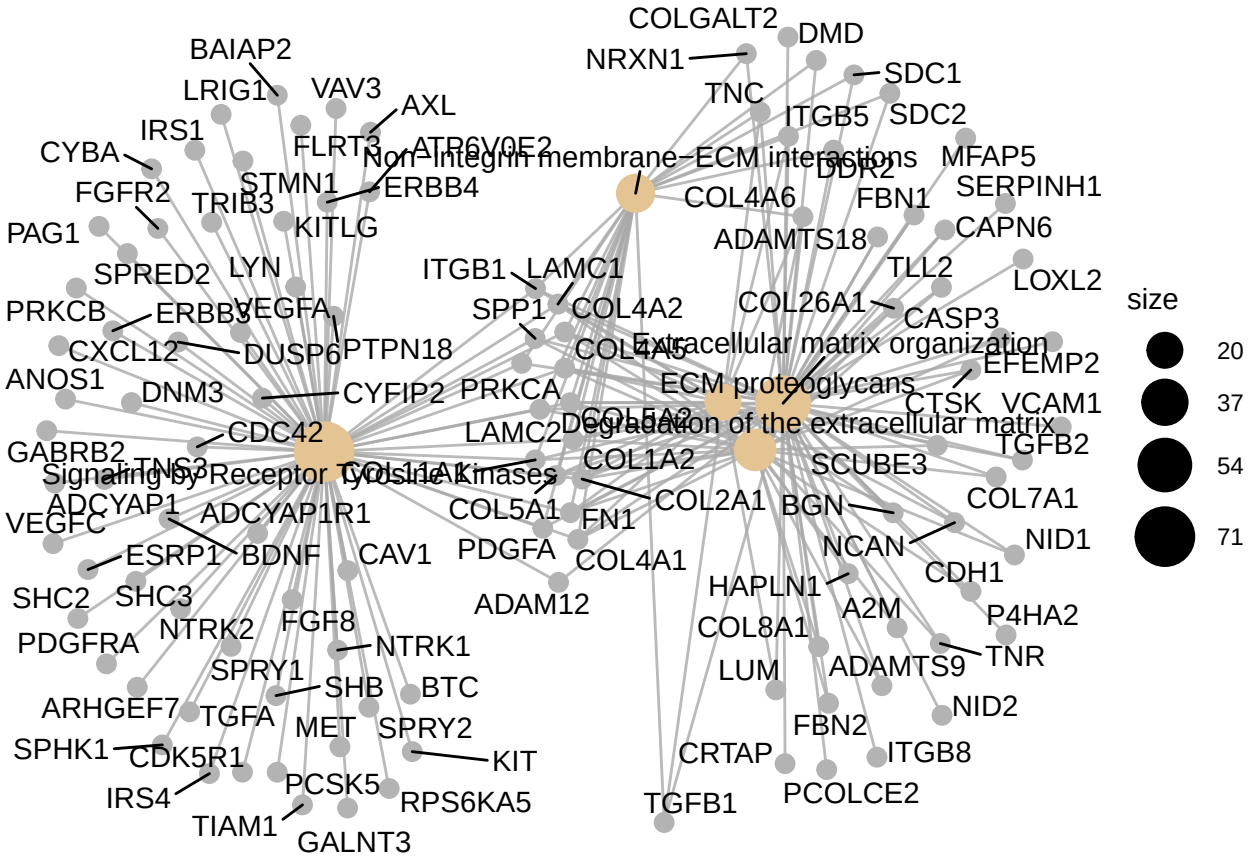

Figure 10: Representación de las vías más enriquecidas en las NPC bajo el tratamiento con alcohol. De cada vía salen líneas que nos conducen a los genes implicados en dichos procesos enriquecidos.

Por último, se incluye un **emapplot** donde nuevamente se visualizan las vías más enriquecidas en el proceso. En este caso los conjuntos de genes que se superponen mutuamente tienden a agruparse para facilitar la interpretación del módulo funcional.

Nuevamente, entre las vías más enriquecidas destacan la organización de la matriz extracelular, el sistema neuronal y la señalización del receptor tirosina quinasa. En el *emapplot* se indican, además, las posibles enfermedades asociadas a la alteración de estas vías. El enriquecimiento de estas vías podría alterar los siguientes procesos: el desarrollo de los huesos intramembranosos y endocondrales, aparición de enfermedades asociadas al metabolismo de los glicosaminoglicanos o aparición de exotosis, que es un trastorno caracterizado por el desarrollo de múltiples masas osteocartilaginosas benignas en los extremos de los huesos largos de los miembros inferiores.

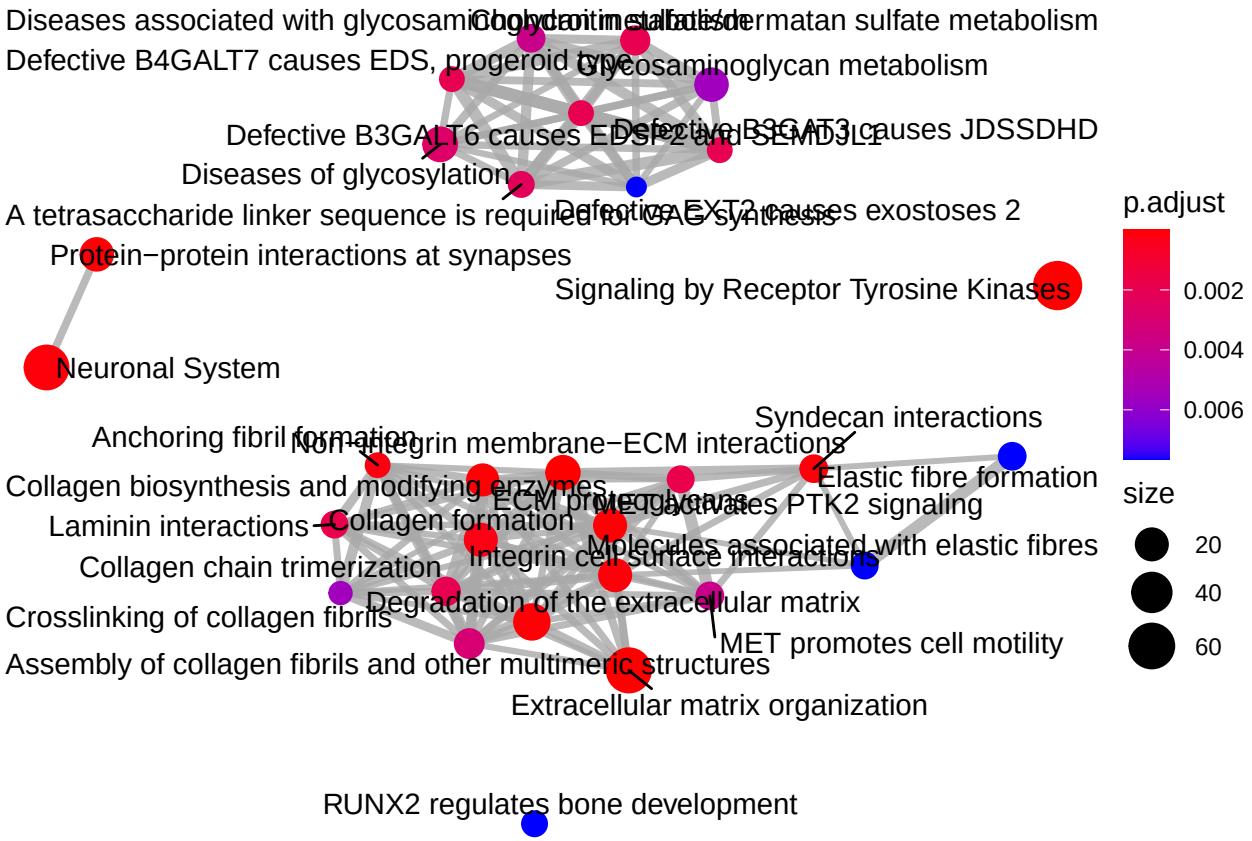

Figure 11: Red de los procesos biológicos y de las vías más enriquecidas en las NPC bajo el tratamiento con alcohol.

Para acabar los resultados que nos deja el análisis de enriquecimiento biológico, dejo cuatro listas donde se agrupan los cinco procesos más enriquecidos en cada comparación llevada a cabo. En las cuatro situaciones analizadas la ruta de organización de la matriz extracelular se encuentra muy enriquecida y en tres de ellas la vía de señalización del receptor tiosina quinasa también se ve alterado, confirmando así los resultados obtenidos en el barplot mostrado anteriormente.

Tabla resumen para p40vsRosett0

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

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

Tabla resumen para Rosett20vsRosett0

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

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

Tabla resumen para p40vsNPC0

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

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

Tabla resumen para NPC20vsNPC0

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

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

6 Discusión

La principal limitación que se encuentra en el estudio es la gran diferencia entre las rosetas neurales y las células progenitoras celulares. Esta diferencia es tan grande que mientras se realizaba el estudio se decidió bifurcar el estudio en dos: uno para las rosetas neurales y otro para las células progenitoras.

En este sentido, la limitación es que el efecto que tiene el tratamiento con 20mM de EtOH debe hacerse de manera individualizada para cada grupo celular sin poder hacer una extrapolación o comparación directa entre grupos.

Otra posible limitación es que el efecto del tratamiento con alcohol podría quedar enmascarado parcialmente por los enormes cambios que el proceso de diferenciación celular produce por sí mismo.

7 Conclusión

Si volvemos a los objetivos de este estudio para ver si se han cumplido, podemos decir que ambos objetivos han quedado demostrados pues se ha visto que el tratamiento con alcohol hace que determinados genes cambien su expresión y, además, se han determinado específicamente las vías enriquecidas por este efecto.

Queda demostrado también que los efectos que el EtOH causa en las células neurales es siempre dañino y nunca tiene un sentido positivo. Como el alcohol inhibe la diferenciación de las células madre neurales (NSC), el crecimiento celular, la migración y la viabilidad celular se ven afectados negativamente. En concreto, los NSC tratados con EtOH muestran retrasos en el ciclo celular, reducción de la proliferación de NSC y aumento de la fragmentación del ADN.

En este estudio, se han generado células madre neurales a partir de hESCs pluripotentes y se han examinado los cambios globales de la firma transcriptómica afectados por el tratamiento con etanol, identificanso así los posibles efectos moleculares de la exposición fetal al alcohol en la diferenciación neural del desarrollo embrionario temprano. En particular, se han identificado y verificado varios genes candidatos que el alcohol podría interferir en la regulación de las células madre neurales (ver el directorio **resultados**). Se ha demostrado que las principales vías moleculares de las NPC afectadas por el EtOH están asociadas con la exposición al alcohol y no por el proceso normal de diferenciación celular. Prueba de ello es que en los Diagramas de Venn el número de genes que cambian su expresión por el tratamiento con el alcohol es mucho mayor que en las rosetas neurales.

Bibliografía

1. Kim JJ, Duan L, Tu TG, Elie O, Kim Y, Mathiyakom N, et al. Molecular effect of ethanol during neural differentiation of human embryonic stem cells in vitro. Genomics data. 2014;2:139–43.

2. Kim YY, Roubal I, Lee YS, Kim JS, Hoang M, Mathiyakom N, et al. Alcohol-induced molecular dysregulation in human embryonic stem cell-derived neural precursor cells. PloS one. 2016;11(9).

3. Edgar R, Domrachev M, Lash AE. Gene expression omnibus: NCBI gene expression and hybridization array data repository. Nucleic acids research. 2002;30(1):207–10.

4. Kauffmann A, Gentleman R, Huber W. ArrayQualityMetrics—a bioconductor package for quality assessment of microarray data. Bioinformatics. 2009;25(3):415–6.

5. Gentleman R, Carey V, Huber W, Hahne F. Genefilter: Methods for filtering genes from microarray experiments. R package version. 2011;1(0).

6. Smyth GK. Limma: Linear models for microarray data. In: Bioinformatics and computational biology solutions using r and bioconductor. Springer; 2005. pp. 397–420.

7. Thissen D, Steinberg L, Kuang D. Quick and easy implementation of the benjamini-hochberg procedure for controlling the false positive rate in multiple comparisons. Journal of educational and behavioral statistics. 2002;27(1):77–83.

8. Yu G, He Q-Y. ReactomePA: An r/bioconductor package for reactome pathway analysis and visualization. Molecular BioSystems. 2016;12(2):477–9.