Búsqueda de nuevas terapias que favorezcan la neuroregeneración en la esclerosis múltiple.

20 de enero de 2015

Miriam Mota Foix
NIU:1315280

Tutores:
Alex Sánchez
Juan Ramón González

Índice

Índice de figuras

Agradecimientos:

Me gustaría agradecer a mis tutores, Alex Sánchez y Juan Ramón González por haberme motivado y guiado en este trabajo. Al Dr. Manuel Comabella, Sra. Carme Riu y al Dr. Nicolás Miguel Fissolo del grupo de Neuroinmunología del VHIR por haberme proporcionado los datos y mostrarse dispuestos a ayudar en todo momento.

1. Introducción

Este trabajo lo he realizado en la Unidad d'Estadística y Bioinformática del VHIR (Vall Hebron Institut de Recerca). El grupo de Neuroinmunología del VHIR me proporcionó los datos y el problema práctico. Dado que se trata de un estudio real, el informe se ha realizado siguiendo la estructura habitual en este tipo de análisis:

1. **Introducción**: donde se describe el problema y los datos con los que trabajaremos
2. **Métodos**: descripción de los métodos estadísticos y bioinformáticos usados para los análisis.
3. **Resultados**: explicación de los resultados de los datos con los que estamos tratando.
4. **Discusión y conclusiones.**

1.1. Presentación

La esclerosis múltiple (EM) se considera una enfermedad autoinmune crónica del sistema nervioso central que afecta principalmente a adultos jóvenes entre 20-40 años de edad y conduce a una discapacidad neurológica significativa. La prevalencia de la enfermedad en España es aproximadamente de 70-90 casos / 100.000 habitantes, y la incidencia oscila entre 4-6 casos por 100.000 habitantes al año (Aladro [1]). La etiología de la esclerosis múltiple es desconocida, sin embargo, se supone que viene dado por un fondo genético complejo y desencadenantes ambientales tales como infecciones virales contribuyen a la manifestación de la enfermedad (Sospedra y Martin, [13]). La enfermedad generalmente comienza con un curso remitente-recidivante (RR) caracterizada por episodios agudos de disfunción neurológica (recaídas) seguidos de períodos de recuperación parcial o completa (remisiones). Con el tiempo, más de 50 % de los pacientes con EMRR entrará en una fase secundaria progresiva de la enfermedad que conducirá a déficits neurológicos permanentes (Debouverie [4]).

Las estrategias de tratamiento actuales son altamente efectivas en reducir o suprimir el componente inflamatorio de la esclerosis múltiple, que predomina en la mayoría de los pacientes en la fase RR y, clínicamente, está asociada con una reducción en el número de recaídas y la actividad radiológica del cerebro. Sin embargo, estas terapias no tienen ningún efecto sobre el componente neurodegenerativo de la esclerosis múltiple, que predomina en las fases progresivas crónicas de la enfermedad y conduce a déficit neurológicos permanentes. Desde un punto de vista teórico y considerando la relación existente entre el axón y los oligodendrocitos, los tratamientos que favorecen la oligodendrogenesis endógena y promueven la remielinización deben resultar en la preservación axonal y pueden representar estrategias neuroprotectoras prometedoras para las fases progresivas y discapacitantes de la enfermedad.

En conclusión, la esclerosis múltiple es una enfermedad autoinmune, lo que quiere decir que el sistema inmune se vuelve contra el propio cuerpo destruyendo los oligodendrocitos que

constituyen la mielina, la funda protectora que recubre las fibras nerviosas. La destrucción de la mielina detiene la transmisión de mensajes de los nervios y conduce a daos en las fibras nerviosas, lo que puede llevar a una prdida progresiva de movimiento, habla, visión,etc.

1.2. Objetivos

Los objetivos académicos de este estudio son :

- Aprender métodos estadísticos para la realización de estudios bioinformáticos.
- Aplicar estos métodos en datos reales.

Los objetivos concretos de este estudio son los que se enumeran a continuación:

- Encontrar genes diferencialmente expresados entre los diferentes tipos celulares.
- Determinar las vías relacionadas con los procesos de diferenciación hacia cada uno de los tres tipos celulares.
- Identificar compuestos relacionados con nuestros genes de interés.

Para alcanzar estos objetivos se propone investigar los cambios en la expresión de genes que tienen lugar durante la diferenciación de NSC ("*Neural Stem Cell*") a astrocito, neurona u oligodendrocito. Para ello se analizará la expresión génica que define la diferenciación a los diferentes tipos celulares y se definirá un perfil de expresión génica para cada uno. Está se utilizará para buscar compuestos químicos capaces de inducir la diferenciación de las NSC a astrocitos, neuronas y oligodedrocitos.

1.3. Los datos para el análisis

Para realizar el estudio se trabajará con 20 muestras procedentes de cultivos celulares obtenidos a partir de cerebro de ratón (4 ratones). Estas muestras han sido procesadas en microarrays (Mouse Gene array 1.0 ST).

Los datos con los que hemos trabajado son confidenciales dado que no han sido publicados los resultados.

Se dispone de diferentes tipos de células:

- **NSC**, células madre neurales. En nuestro caso se obtienen de la zona subventricular del cerebro de ratones adultos (3 meses de edad), y se mantienen en cultivo en suspensión donde se expanden y mantienen indiferenciadas.
- **Astrocitos**, son las principales y más numerosas células gliales, asumen un elevado número de funciones clave para la realización de la actividad nerviosa.
- **Neuronas**, son un tipo de células del sistema nervioso cuya principal función es la excitabilidad eléctrica de su membrana plasmática.

- **Oligodendrocito**, son un tipo de células gliales cuya función principal es la de generar la mielina que envuelve los axones de las neuronas.
- **Células progenitoras de oligodendrocitos**, es un subtipo de célula glial con capacidad para diferenciarse a oligodendrocitos o astrocitos tipo II.

Para obtener las muestras los investigadores diseccionaron la zona subventricular del cerebro del ratón. Las células obtenidas las mantuvieron en cultivo para aislar y expandir la población NSC. Una vez tuvieron suficientes NSC los separaron en 4 grupos diferentes y los diferenciaron a neuronas, otro a astrocitos y dos grupos a OPCs (células progenitoras de oligodendrocitos) (Samuel Weiss [2]). Uno de los grupos que se diferenció a OPC se siguió diferenciando hasta oligodendrocito.

Todo este proceso lo repitieron 4 veces, es decir, utilizaron 4 ratones diferentes, de los cuales obtuvieron 4 cultivos de NSC independientes, y los cuales fueron diferenciados hacia los diferentes tipos celulares.

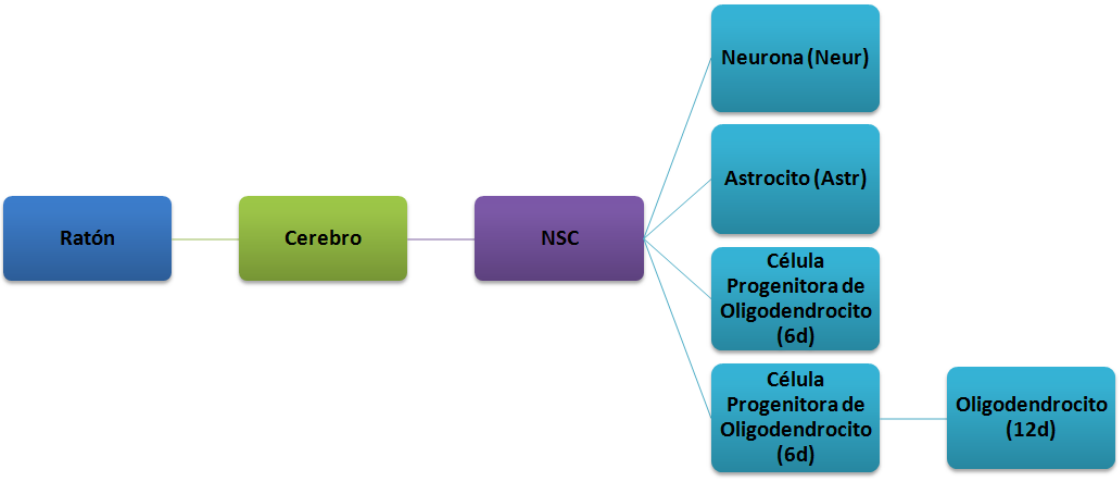


Figura 1: Proceso de diferenciación celular.

1.4. Diseño experimental

Este estudio consiste en un estudio comparativo cuyo objetivo es determinar si la expresión de los perfiles de expresión génica [1] difieren entre grupos. Se dispone de 20 muestras, y un factor (variable independiente) que es el tipo de célula. Los datos con los que tratamos son datos apareados, ya que de un mismo ratón obtenemos 5 muestras.

Las condiciones experimentales que se consideran en este estudio son los diferentes tipo de célula, se dispone de:

- 4 muestras de células madre neurales (*NSC*).
- 4 muestras de astrocitos (*Astr*).
- 4 muestras de neuronas (*Neur*).
- 4 muestras de células progenitoras de oligodendrocitos (*d6*).
- 4 muestras de oligodendrocitos (*d12*).

Ratón \Célula	**NSC**	**Astr**	**Neur**	**d6**	**d12**	**Total**
R7	1	1	1	1	1	**5**
R8	1	1	1	1	1	**5**
R9	1	1	1	1	1	**5**
R10	1	1	1	1	1	**5**
Total	**4**	**4**	**4**	**4**	**4**	**20**

En este estudio tenemos **replicación biológica** ya que se han analizado múltiples muestras de un mismo tipo de tejido bajo las mismas condiciones experimentales. Este tipo de replicación permite estimar la variabilidad a nivel de población.

Actualmente, es habitual la utilización de 3 a 5 replicas biólogicas por condición experimental, ya que la utilización de más es economicamente privativa y de que los análisis de potencia no soy excesivamente fiables.

Como indicaba Allison ([11]) aunque no hay consenso sobre qué procedimiento es mejor para determinar el tamaño de las muestra, si que lo hay sobre el hecho que un mayor número de réplicas generalmente proporcionan una mayor potencia.

[1] Un perfil de expresión génica de una célula es una medida de la actividad (expresión) de miles de genes simultáneamente, para crear una imagen global de la función celular.

1.4.1. Objetivos específicos

Uno de los objetivos de este estudio es encontrar diferencias de expresión entre los diferentes tipos de células, las comparaciones que se tendrán en cuenta son las que se enumeran a continuación:

1. Diferencias entre NSC con astrocitos y neuronas.
   a) Astr vs NSC = Astr - NSC
   b) Neur vs NSC = Neur - NSC
2. Diferencias entre NSC y los diferentes tipos de oligodendrocitos.
   a) d6 vs NSC = d6 - NSC
   b) d12 vs d6 = d12 - d6

1.5. Tecnologia: Microarrays

Que es un Microarray? Un microarray es un formato experimental basado en la síntesis o fijación de **sondas** que representan los genes sobre un sustrato sólido (cristal, plástico, sílice, etc.) y expuesto a las moléculas **diana** cuya expresión se desea analizar.

Lo que caracteriza ese método en comparación a los que había anteriormente para estudiar el transcriptoma no es lo que pueden medir, sino la cantidad de mediciones simultáneas que pueden realizar. Mientras que hasta hace apenas 15 años se estudiaban los genes uno a uno en profundidad, a partir del uso de estas nuevas tecnologías se pueden estudiar muchísimos genes a la vez, pero en contrapartida, con mucho menos detalle y más ruido.

Funcionamiento de un microarray. Su funcionamiento consiste, básicamente, en medir el nivel de hibridación entre la sonda específica (*probe*), y la molécula diana (*target*), y se indican generalmente mediante fluorescencia y a través de un análisis de imagen, lo cual indica el nivel de expresión del gen.

En los microarrays de un color, que son con el tipo que estamos trabajando, el proceso consiste en hibridar una muestra, donde el valor que se obtiene después de iluminar el array con el láser, es una medida numérica que se obtiene directamente del escáner, es decir, no está referida al valor de otra muestra, por lo que recibe el nombre de expresión absoluta.

La empresa Affymetrix (Santa Clara, California) es la casa comercial líder en la manufacturación y venta de este tipo de microarrays. Affymetrix sintetiza las sondas directamente sobre el chip mediante un proceso llamado fotolitografía. Este proceso consiste en la adición cíclica de los cuatro nucleótidos (adenina, timina, citosina y guanina) sobre la superficie rígida, donde existen ancladas unas especies químicas reactivas que se protegen y desprotegen para añadir el nucleótido deseado, mediante ciclos de luz y oscuridad. Así se consigue la síntesis de oligonucleótidos. En la Figura 2 se presenta el proceso de forma esquemática.

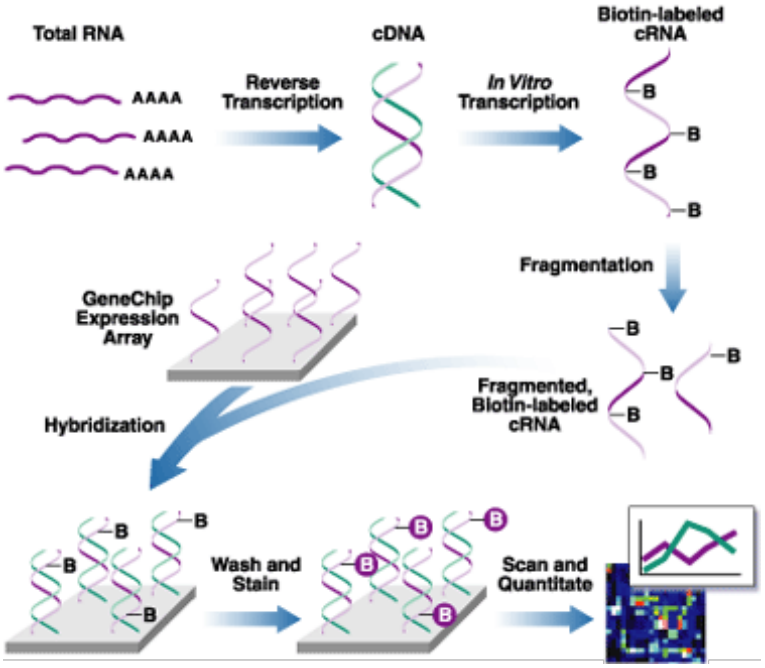


Figura 2: **Proceso microarrays.**Este diagrama representa el proceso de como la expresión de un organismo puede ser analizada utilizando microarrays de DNA (www.affymetrix.com)

El resultado de escanear la imagen de un array es un archivo de extensión .CEL, el cual esta en formato binario y solo puede ser leído con programas específicos para ello.

A partir de las intensidades de los archivos .CEL se genera la matriz de expresión que contiene una columna por chip con los valores de intensidad absoluta y una fila por grupo de sondas. En el caso de los arrays de affymetrix existe una gran variedad de algoritmos de sumarización y según cuál se utilice se obtendrá unos u otros valores de expresión pero estos serán siempre medidas absolutas, es decir, independientes del resto de muestras.

1.6. Análisis

El análisis se ha realizado siguiendo el esquema que se muestra en la Figura 3

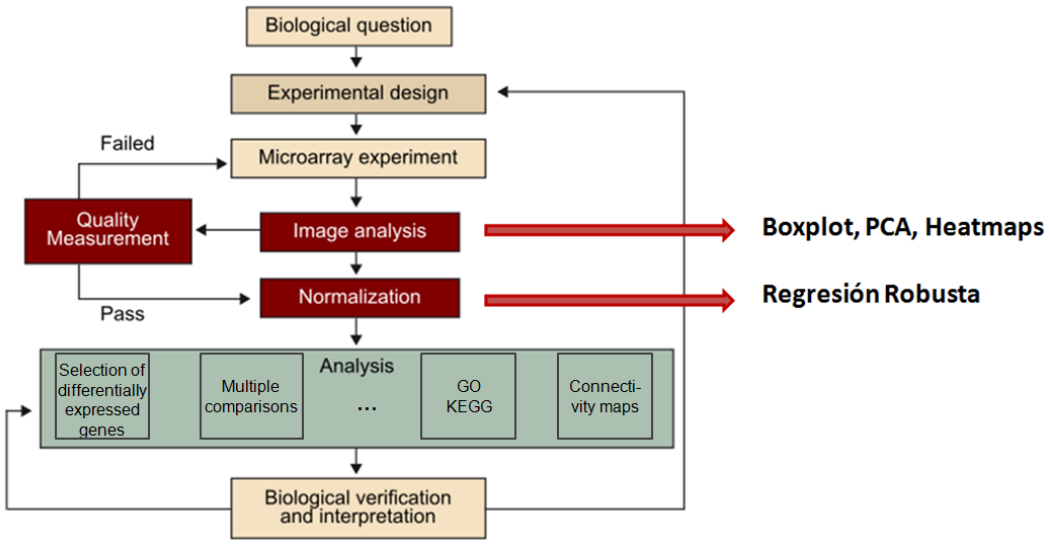


Figura 3: **Proceso de análisis de microarrays**. El análisis de microarrays puede ser fácilmente visualizado como un proceso que empieza por una pregunta biológica y concluye con una interpretación de los resultados de los análisis que, de alguna forma, confiamos que nos acerque un poco a la respuesta de la pregunta inicial .

Los análisis estadísticos se han realizado usando el lenguaje estadístico R y las librerias desarrolladas para el análisis de microarray en el proyecto de Bioconductor [2]. Para más detalle sobre los métodos ver sección 2.

[2] Bioconductor es un proyecto de código abierto para el análisis de datos en Genómica (http://www.bioconductor.org/)

2. Métodos

2.1. Control de calidad, Normalización y filtraje

2.1.1. Control de calidad y exploración

Como exploración inicial de los datos se realizan una serie de gráficos para comprobar la validez de los arrays. Estos gráficos incluyen diagramas de cajas de los datos brutos, histogramas de la distribución de la señal y controles de calidad internos de Affymetrix [3]. También se ha realizado un dendograma que representa una agrupación jerárquica de las muestras basado en los datos brutos. En este gráfico se pretende mostrar la presencia o ausencia de heterogeneidad entre las muestras. Idealmente, sería deseable que todas las réplicas de una condición específica se agruparan juntas. Por lo que si esto ocurre no se puede asumir heterogeneidad. En resumen, mediante gráficos se estudia la estructura de los datos con el fin de decidir si parecen correctos o presentan anomalías que deben ser corregidas.

2.1.2. Normalización

Una vez realizado el control de calidad, procedemos a normalizar los datos con el fin de eliminar los sesgos sistemáticos.

En el caso de los chips de Affymetrix, el método que se usa es el RMA (*Robust Multi-array Average*) (Irizarry *et al.*[7]) descrito en Gentleman (2005) ([6]). Este método realiza tres pasos de preprocesado:

1. Ajusta el ruido de fondo (background) basándose solo en los valores PM y utilizando un modelo estadístico complejo en el que combina la modelización de la señal mediante una distribución exponencial con la del ruido mediante una distribución normal.
2. Toma logaritmos base 2 de cada intensidad ajustada por el *background*.
3. Realiza una normalización por cuantiles de los valores del paso 2 consistente en sustituir cada valor individual por el que tendría la misma posición en la distribución empírica estimada sobre todas las muestras, es decir, los promedios de las distribuciones de los valores ordenados de cada array.
4. Estima las intensidades de cada gen separadamente para cada conjunto de sondas. Para ello realiza una técnica similar a una regresión robusta denominada median polish sobre una matriz de datos que tiene los arrays en filas y los grupos de sondas en columnas.

El método utiliza un modelo lineal de la forma:

$$T(PM_{ij}) = e_i + a_j + \epsilon_{ij}$$

[3] Affymetrix: compañía estadounidense especializada en el diseño de microarrays de ADN

donde T representa la transformación que corrige el ruido de fondo, normaliza y registra las intensidades PM , e_i representa el valor de la expresion en log_2 encontrado en arrays $i = 1, ..., I$, a_j representa la intesidad en escala logaritmica para los probes $j = 1, ..., J$, y ϵ_{ij} representa el error de la matriz de i y la sonda j. La regresión robusta es usada para estimar la expresión en escala logaritmica de los valores. A esto se le conoce como RMA.

Como resultado final de todos los pasos anteriores se obtiene la matriz con los datos sumarizados y normalizados.

Comprobación de la homogeneidad de la muestra con valores normalizados

Un gráfico de los dos primeros componentes principales de un análisis PCA, también se puede utilizar para ver la estructura general de los datos. En la mayoría de los casos se espera que las muestras de un mismo grupo tiendan a agruparse, lo que indica en general similitud en los patrones de expresión. Dada la naturaleza del método PCA la separación entre las muestras a lo largo del eje de coordenadas X suele ser más significativo que a lo largo del eje Y debido a que el primer componente está construido de tal manera que explica más que el segundo componente, que explica más que el tercero y así sucesivamente.

Se puede debatir acerca de si se deben aplicar dendrogramas o PCA antes o después de la normalización, aunque existen argumentos a favor de ambos enfoques. En principio, si existen similitudes entre las muestras que no se deben a problemas técnicos, estos deben realizarse después de la normalización.

2.1.3. Filtraje

El número de probesets disponibles en un microarray es muy grande (alrededor de 45.000 a 55.000). Incluir a todos los genes en el análisis implica tener que hacer grandes ajustes para comparaciones múltiples, por lo que se recomienda llevar a cabo algún tipo de filtraje no-especifico, con el fin de reducir el ruido.

1. Eliminación de los spots marcados como erróneos mediante flags y que son debidos a problemas en la hibridación o en el escaneo.

2. Eliminación de spots con señales muy bajas debido a problemas en el *spotting* o a que no ha habido hibridación en ese spot (filtraje por señal).

3. Eliminación de los genes que no presenten una variación significativa en su señal, entre distintas condiciones experimentales (filtraje por variabilidad). Esto se hace manteniendo sólo aquellos genes cuya desviación estándar es mayor de un umbral. Un umbral conservador razonable puede ser el tercer cuartil de toda desviación estándar. Es razonable ya que, las desviaciones estándar se pueden trazar de manera ordenada y aumentan moderadamente desde el valor más pequeño hasta el percentil 90-95. El último 5 %, aumenta a

un ritmo mucho mayor que sugiere que la variación importante se encuentra en este 5 %, por lo que tomar la parte superior del 25 % es claramente un enfoque seguro.

2.2. Selección de genes diferencialmente expresados

El problema de seleccionar genes diferencialmente expresados se traduce de manera casi inmediata al problema estadístico de comparar variables y, en años recientes, se han desarrollado un gran número de métodos estadísticos para resolverlo. La mayoría son extensiones de los métodos estadísticos clásicos -pruebas t o análisis de la varianza- adaptados en uno u otro sentido para tener en cuenta las peculiaridades de los microarrays.

La selección de genes diferencialmente expresados se realiza usando un modelo lineal con una componente bayesiana, ya que se supone que los mismos parámetros a estimar son variables (no constantes) con distribuciones que se estimarán a partir de la información de los genes (descrito por G. Smyth en [6] e implementado en el paquete "*limma*" de Bioconductor).

A continuación se obtienen las estimaciones de los parámetros del modelo. La aproximación utilizada garantiza que estos estimadores tienen un comportamiento robusto incluso para un pequeño número de arrays.

El análisis utiliza varios estadísticos del test que ayudan a decidir qué genes aparecen expresados diferencialmente.

- *logFC* es la estimación del cambio entre condiciones: $\overline{R} = \frac{1}{n}\sum_{i=1} R_i$, donde se realiza de forma indirecta la aproximación siguiente:

$$\overline{X}_1 - \overline{X}_2 = \overline{\log Y}_1 - \overline{\log Y}_2 \simeq \log(\overline{Y}_1) - \log(\overline{Y}_2) = \log(\frac{Y_1}{Y_2})$$

- *t–test moderada* : En el númerado se encuentra el fold-change y en el denominador una expresión en la que se combina la estimación del error estándar de cada gen SE_g con una estimación de dicho error obtenida a partir de todos los genes analizados.

$$t = \frac{R_g}{\sqrt{\frac{d_0 \cdot SE_0^2 + d \cdot SE_g^2}{d_0 + d}}}$$

- p es el p-valor correspondiente a la t-moderada.

- El *estadístico B*, introducido por G.Smyth [12] representa el logaritmo de la probabilidad posterior (en el sentido Bayesiano) de que un gen este diferencialmente expresado frente a la probabilidad de que no lo esté, por lo que su interpretación es similar a la de un log-odds-ratio.

$$B = \log \frac{P[Afectado|M_{ij}]}{P[NoAfectado|M_{ij}]},$$

gen=i ($i = 1...N$), réplica=j ($j = 1, ..., n$).

El hecho de trabajar con logaritmos permite poner el punto de corte en el cero: a mayor valor positivo más probable es que el gen esté diferencialmente expresado. A mayor valor negativo, más probable es que no lo esté.

En el modelo lineal, las comparaciones de interés se definen como "contrastes", que pueden ser descritos como operaciones (combinaciones lineales) entre columnas de la matriz de diseño.

Un conjunto de uno o más contrastes se almacena en la "*matriz de contrastes*".

Usando la matriz de diseño y, si está disponible, la matriz de contrastes se ajusta un modelo lineal para cada gen

El resultado de estos análisis consiste en una lista de los genes seleccionados ordenados del más al menos diferencialmente expresado.

Con el fin de hacer frente a los problemas de comparaciones múltiples derivados de realizar muchas pruebas (una por gen), los p-valores se ajustan para obtener un fuerte control sobre la tasa de falsos positivos (FDR) utilizando los métodos Benjamini y Hochberg ([3]) que se describe a continuación.

Tasa de falsos positivos (FDR). Los genes están ordenados por los p-valores ascendentemente. Sin embargo, el umbral para el p-valor de orden i será: $p_i < \frac{i}{n}\frac{\alpha_e}{p_0}$ donde α_e es el nivel de significación y p_0 es la proporción de hipotesis nulas H_i que realmente son ciertas. Dado que esta proporción no se sabe, los p_0 se pueden estimar conservadoramente como 1. Esto supone que todas las hipotesis nulas son realmente ciertas y no hay genes diferencialmente expresados. El aspecto conservador significa que si la hipótesis nula puede ser rechazada por un gen en particular, en estas circunstancias, entonces la hipótesis nula será todavía rechazada si algunas son en realidad falsas $p_0 < 1$

1. Elegir nivel de significación del experimento α_e.
2. Ordenar los p-valores de los genes ascendentemente.

Genes	g_{i1}	g_{i2}	...	g_{ik}	...	g_{iR}
P.valores (ascendentes)	p_1	p_2	...	p_k	...	p_R

3. Comparar los valores de p de cada gen con un umbral que depende de la posición del gen en la lista de p-valores ordenados. Los umbrales son los siguientes : $\frac{1}{R}\alpha_e$ para el primer gen, $\frac{2}{R}\alpha_e$ para el segundo gen, etc.

Genes	g_{i1}	g_{i2}	...	g_{ik}	...	g_{iR}
P.valores (ascendentes)	p_1	p_2	...	p_k	...	p_R
Test	$p_1 < \frac{1}{R}\alpha_e$	$p_2 < \frac{2}{R}\alpha_e$	...	$p_k < \frac{k}{R}\alpha_e$	...	$p_R < \alpha_e$

4. Sea k el mayor y para los que $p_i < \frac{i}{R}\alpha_e$. Rechazamos la hipotesis nula H_i para $i = 1, 2, .., k$. Estos genes son de hecho diferentes entre los dos grupos para el nivel de significación, α_e elegido.

Un enfoque estándar es llamar genes diferencialmente expresados aquellos que tienen una B mayor que 0 en valor absoluto o un p-valor ajustado más pequeño que un umbral como por ejemplo 0,05 o 0,01.

Cuando no hay genes que superen este umbral se puede confiar en los p-valores ajustados o simplemente en el rango de los genes en la tabla: los genes en la parte superior de la tabla son los que están diferencialmente más expresados.

Los genes diferencialmente expresados se pueden mostrar gráficamente utilizando "Volcano plots", que organizan los genes a lo largo de las dimensiones de importancia biológica(*fold change*) y estadística (*p-valor*). La primera dimensión (horizontal) es el cambio entre los dos grupos (en un escala logarítmica, de modo que la regulación positiva o negativa se representa de forma simétrica), y la segunda (vertical) representa el p-valor del test moderado en una escala logarítmica negativa, por lo que los p-valores más pequeños aparecen más arriba. El primer eje indica impacto biológico del cambio; el segundo indica la evidencia estadística, o la fiabilidad del cambio.

2.3. Comparaciones múltiples a partir de listas de genes

Cuando se realizan varias comparaciones a la vez puede resultar importante ver qué genes cambian simultaneámente en más de una comparación. Si el número de comparaciones es alto, también puede ser necesario realizar un ajuste de p-valores entre las comparaciones, distinto del realizado entre genes.

Dado un número pequeño de repeticiones es común el uso de este enfoque simplemente para resumir qué genes cambian en cada condición sin hacer ninguna corrección adicional. En ese caso debe verse más como un análisis descriptivo que un resumen inferencial. Un diagrama de Venn permite visualizar los genes diferencialmente expresados para las diferentes comparaciones.

2.4. Visualización de perfiles de expresión

Tras seleccionar los genes diferencialmente expresados, podemos visualizar las expresiones de cada gen agrupándolas para destacar los genes que se encuentran *up* o *down* regulados simultáneamente, constituyendo perfiles de expresión.

Para cada agrupación de un conjunto de genes seleccionados de las muestras seleccionadas, se realiza un heatmap como se describe en [6], (capítulo *Visualization*). Este gráfico representa los valores de expresión de los genes seleccionados (con las muestras en las columnas y los

genes en las filas) donde las intensidades están codificadas por colores con el fin de resaltar posibles patrones de expresión entre las diferentes condiciones. El código de colores es: "azul" para los valores bajos, "rojo" para los valores de expresión elevados y "blanco" para los que no se expresan. Se suele realizar una agrupación jerárquica de las columnas (muestras), las filas(genes), o ambos a la vez, lo que ayudára a visualizar la variación conjunta.

2.5. Análisis de la significación biológica

Los resultados del análisis de los microarrays consiste en una o más listas de genes, es decir, un archivo o una tabla con los identificadores de los genes considerados, diferencialmente expresados en una o más comparaciones.

Una vez que el investigador tiene la lista de genes disponible por lo general va a investigar la mayoría de los "genes relevantes" por separado. Sin embargo, parece evidente que, dado que algunos genes interactúan juntos en procesos biológicos también puede ser informativo para tratar de entender la lista como un todo. Esto es a menudo descrito como la búsqueda de la importancia biológica.

El objetivo general de analizar el significado biológico es ayudar a responder preguntas tales como si los genes que aparecen en la lista tienen funcionalidades similares o están involucrados en los mismos procesos, y también, por supuesto, para saber cuáles son estos procesos y cómo se relacionan con el problema biológico de interés.

Una forma común para investigar la importancia biológica es proyectar los genes seleccionados en la *"Gene Ontology"* (GO). La GO es una base de datos que contiene las anotaciones genéricas que describen las funciones moleculares, los procesos o componentes celulares biológicos asociados con cada gen. Está organizado en una jerarquía que relaciona todos los términos en sucesivos refinamientos.

Hay muchos métodos y aún más herramientas para realizar análisis GO aunque pueden reducirse a unas pocas categorías (ver [9]). Aquí consideramos sólo el *Gene Enrichment Analysis*.

La misma base y metodología se pueden aplicar adecuadamente dentro del marco de otras bases de datos biológicos integrales, como la Enciclopedia de Kyoto de genes y genomas (KEGG). Por lo tanto, además, se aplica el siguiente enfoque también para realizar una génica basada en el análisis de enriquecimiento KEGG a lo largo de las vías biológicas conocidas en la actualidad.

Análisis de enriquecimiento genético

Este es probablemente el método más común para hacer un análisis basado en GO y más de una docena de herramientas diferentes implementan versiones ligeramente diferentes.

El objetivo de este análisis es realizar una de las pruebas estadísticas disponibles para determinar si un conjunto de genes determinado, por lo general de una categoría particular de la GO, está más o menos representado en la lista de genes seleccionados (la muestra) con respeto (es decir, comparado) a un conjunto de referencia (la población) de donde ha sido seleccionada. El conjunto de referencia son por lo general todos los genes del microarrays.

Por ejemplo, consideremos un experimento que da una lista de genes y el 10 % de los genes más diferencialmente expresados están asociados con el término apoptosis en la GO (GO:0006915). Esto puede parecer una proporción inusualmente grande de la lista de genes, dado que la apoptosis es un proceso biológico muy específico. Para determinar cuánto más grande de lo normal es esta proporción, debe compararse con la proporción de genes relacionados con apoptosis en la lista de genes de referencia, por ejemplo el conjunto de todos los genes del microarray, más a menudo, los genes que se han incluido en el análisis tras descartar los que pueden considerarse ruido de fondo.

El análisis estadístico realizado para comparar proporciones es el test exacto de Fisher que se utiliza para probar la hipótesis:

$$H_0 : p^A_{sel} = p^A_{all} \qquad vs \qquad H_1 : p^A_{sel} \neq p^A_{all}$$

donde A representa el conjunto de genes cuya mas/menos reresentación esta siendo considerada , p^A_{sel} es la proporción de genes seleccionados, que están incluidos en este conjunto de genes y p^A_{all} es la proporción de la lista de referencia.

Todo esto es equivalente al análisis "típico" para una tabla de contingencia de dos vías que por lo general se lleva a cabo mediante una prueba de chi-cuadrado o una prueba exacta de Fisher.

Hay muchas herramientas que pueden ayudar a realizar este análisis. aquí vamos a utilizar los contenidos en *Bioconductor*, principalmente en el paquete *GOstats*.

El análisis ha sido realizado mediante este paquete tal y como se explica a continuación:

- Toma como entrada el *Entrez* o los identificadores *Affy* de la lista de los genes seleccionados, así como el nombre del paquete de la anotación correspondiente al array que ha sido usado para el análisis.
- La salida del análisis es la lista de categorías que aparece más/menos representado en cada conjunto seleccionado.

2.6. Mapas de conectividad

El desafío fundamental que se plantea a lo largo de la biomecdicina es la necesidad de establecer la relación entre las enfermedades, proceso fisiológicos y la acción de terapias de pequeñas moléculas. Los mapas de conectividad son un recurso que puede ser utilizado para encontrar conexiones entre las pequeñas moléculas que comparten un mecanismo de acción, los productos químicos, los procesos fisiológicos y enfermedades. Para ello usaremos el paquete CTDq proporcionado por el Dr. Juan Ramón González (CREAL).

El objetivo de los mapas de conectividad es ayudar a los investigadores a analizar y establecer una relación entre los genes, los productos químicos y las enfermedades mediante la realización de un análisis de enriquecimiento de genes o de enriquecimiento de análisis químico. Para ello, el paquete CTDq utiliza la información de la base de datos El Toxicogenomics comparativo [4] y desde DisGeNET [5]

[4] CTD: The Comparative Toxicogenomics Database. http://ctdbase.org
[5] DisGeNET - una base de datos de la asociación entre genes y enfermedades. http://www.disgenet.org

3. Resultados

3.1. Preprocesado: Control de Calidad, Normalización y Filtraje

Los datos procedentes de la lectura de los microarrays se denominan datos "crudos" y deben ser preprocesados de diversas formas antes de analizarlos. El término "preprocesado" engloba varios procesos descritos a continuación.

3.1.1. Control de calidad

En primer lugar, se ha realizado una exploración de los datos, basándose en técnicas univariantes como los histogramas o diagramas de caja o técnicas multivariantes como los análisis de conglomerados ("clústers"), de distancias o de análisis de componentes.

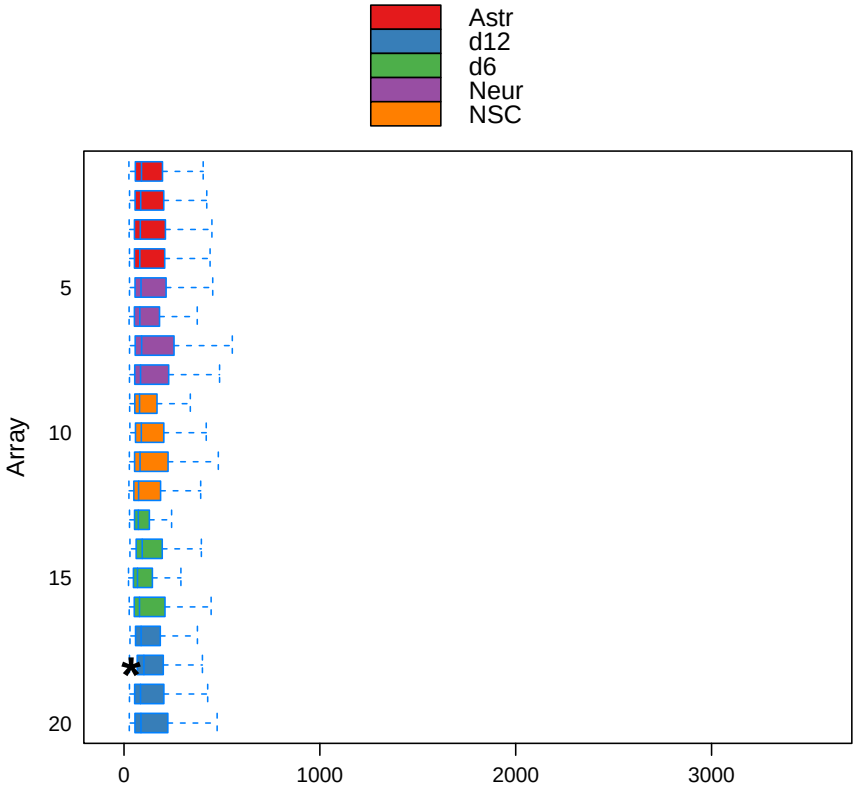


Figura 4: **Diagrama de cajas de datos brutos**. Muestra, cómo las distribuciones de los datos son relativamente similares. Hay heterogeneidad pero no se muestra ninguna diferencia sistemática, lo que sugiere que es conveniente normalizar. Vemos que parece haber algun problema en una de las muestras de "d12".

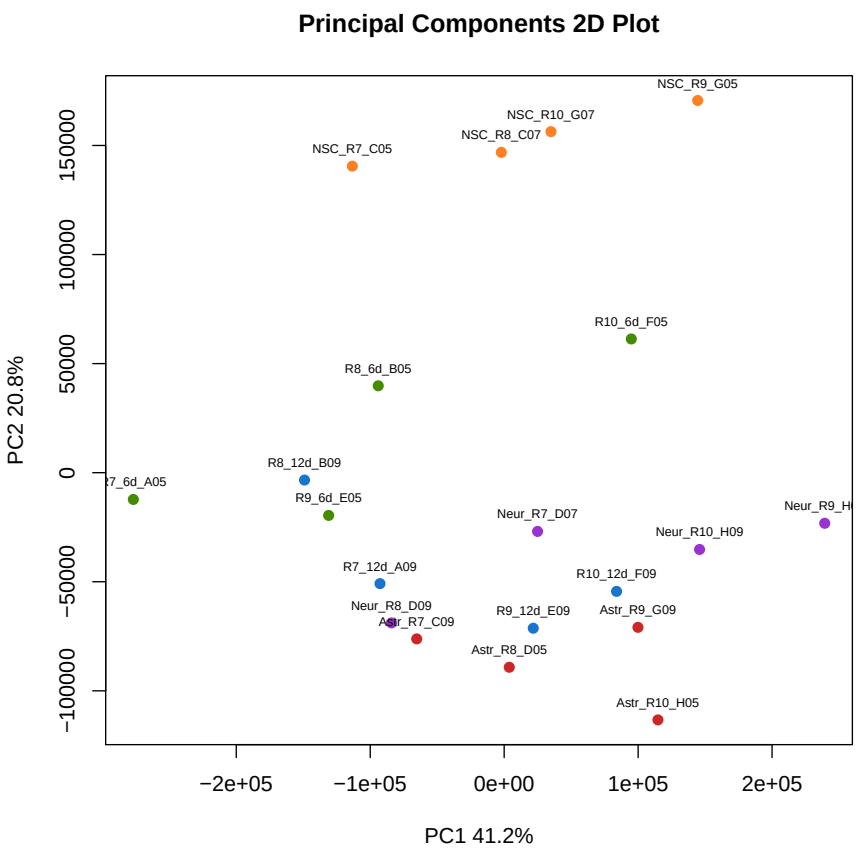


Figura 5: **PCA de los datos brutos**. La primera componente explica un 41.2 % de la varianza y la segunda un 20.8 %. Esta segunda componente vemos que nos diferencia los tipos de células, donde se puede observar claramente separada del resto las NSC (células no diferenciadas). Lo que significa que sucede un cambio en la diferenciación de las células, mientras que entre las neuronas, astrocitos y oligodendrocitos, las diferencias no son tan obvias. También podemos ver que los ratones (R7, R8, R9, R10) no se agrupan por individuo, lo que es buena señal, ya que significa que no hay factor individuo, es decir , que las diferencias que encontremos serán debidas al tipo de célula.

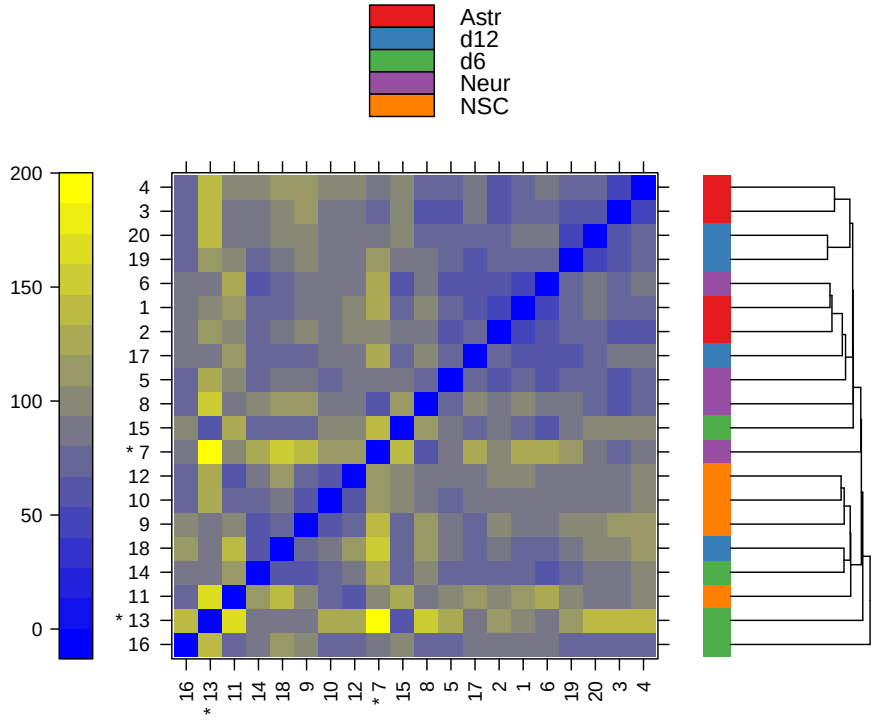


Figura 6: **Heatmap de los datos brutos**. tenemos a la derecha una escala de colores, donde para los valores elevados tenemos colores claros (amarillo) y cuanto más pequeña sea la expresión más oscuro (azul). También tenemos a la derecha una agrupación jerárquica de las muestras donde vemos bastante heterogeneidad. Por último, vemos que para los arrays 13 y 7 podría haber algun tipo de problema.

Las exploraciones anteriores nos proporcionan una idea general acerca de la distribución de los datos y la posible presencia de grupos naturales (p.e. tipos de células) o artificiales (días de procesado). En este caso, no se han mostrado claras agrupaciones entre condiciones.

Por lo general, sólo los array que muestran posibles valores atípicos en más de un criterio son descartados. En este caso, hemos visto 3 arrays que mostraban posibles valores atípicos en algunas de las pruebas realizadas (sólo en una prueba por cada array), por lo que, no se ha descartado ninguno y se usarán todas las muestras en el análisis.

3.1.2. Normalización

Con el fin de poder comparar los datos, así como para eliminar sesgos técnicos los arrays se ha procedido a normalizar con el método de RMA ([7]) que se describe en la sección 2.1.2.

El proceso de control de calidad realizado ha demostrado que (I) de forma individual todos los arrays son buenos y (II) no existe efecto de lotes asociados con la amplificación o el proceso de hibridación. De acuerdo con eso, hemos aceptado todos los arrays para el análisis.

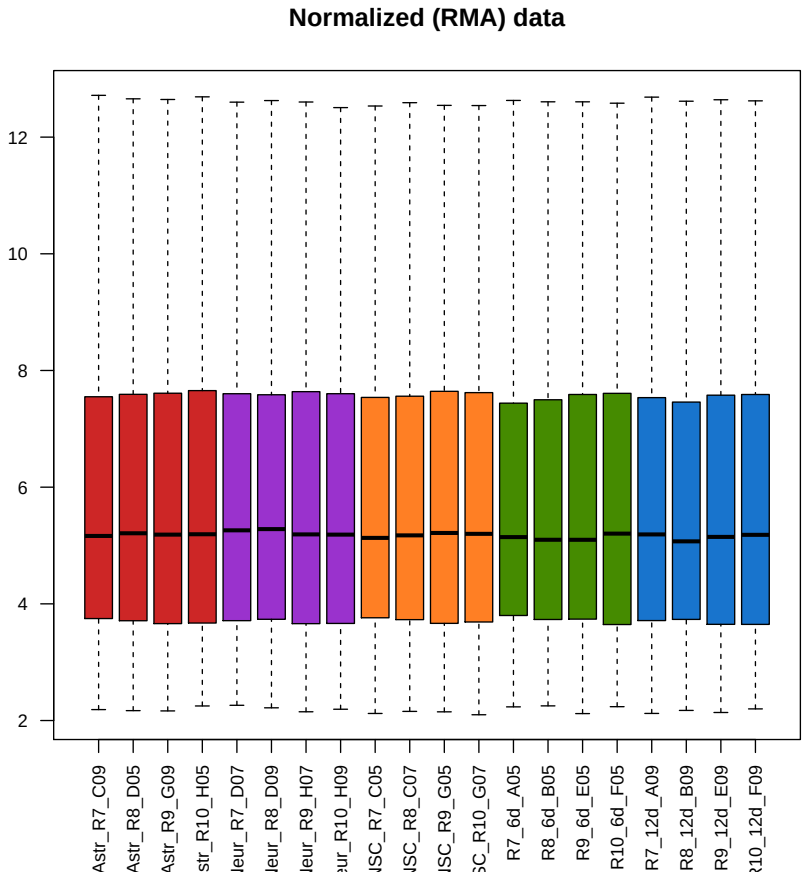


Figura 7: **Diagrama de cajas de los datos normalizados**.Los valores normalizados han quedado claramente en una escala común donde se pueden comparar. Esto no significa que si hubieran errores los habría arreglado por lo que dicho proceso debe hacerse tras descartar la ausencia de arrays problemáticos.

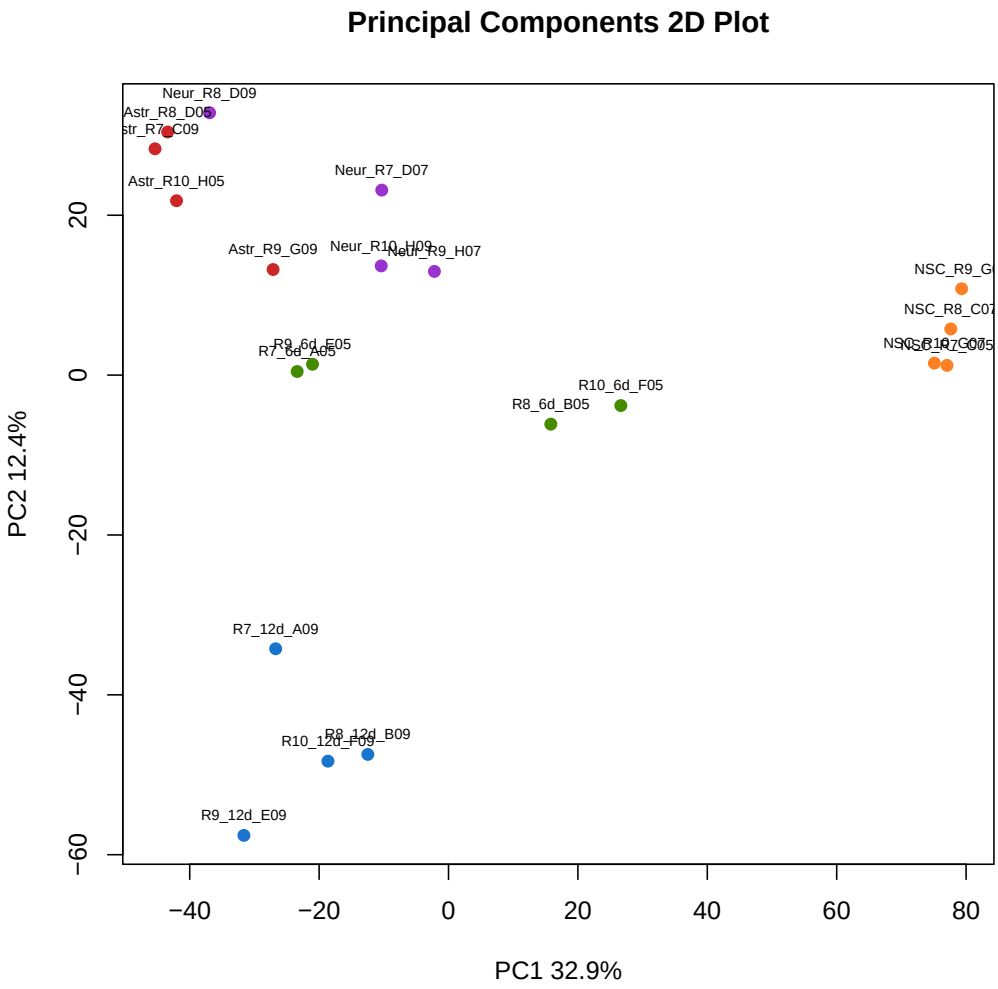


Figura 8: **PCA de los datos normalizados.** Una vez normalizados los datos, los arrays se agrupan por condiciones. Cabe destacar, el porcentaje de variabilidad explicada por cada una de las dos componentes. La primera explica un 32.9 % y la segunda un 12.4 %. Podemos ver claramente que la primera componente explica la diferenciación de las células, donde NSC (no diferenciadas) está totalmente separada del resto. Un poco más a la izquierda vemos los oligodendrocitos (d6), los cuales eran una diferenciación intermedia y por último el resto de tipos de células. La segunda explica el tipo de célula, donde quedan separadas entre ellas a excepción de las neuronas y astrócitos que no se ve una clara agrupación.

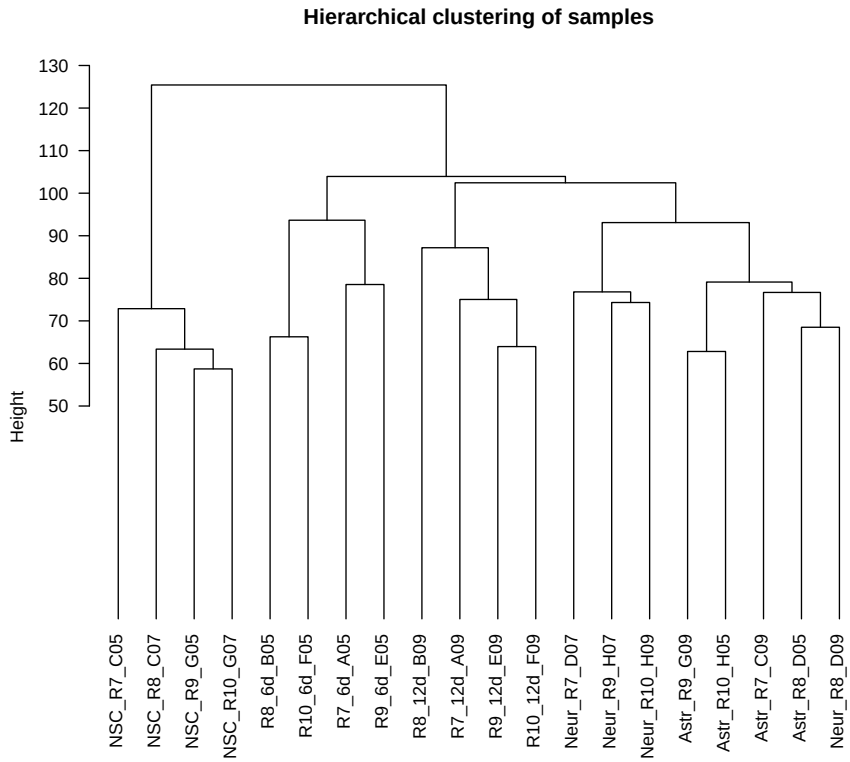


Figura 9: **Dendrograma de los datos normalizados**, resultante de un agrupamiento jerárquico entre las muestras. Como en otras representaciones, como por ejemplo el gráfico de PCA se observa una cierta agrupación de muestras por tipos similares. Vemos como la mayoria de las muestras están agrupadas por tipo de célula, a excepción de "`Neur_R8_D09`".

3.1.3. Filtraje

Por lo general, con el fin de aumentar el poder estadístico y reducir el ruido innecesario, se realiza un filtraje para aquellos genes lo cuales han estado marcados como erroneos. También se filtra por señal y variabilidad descritos en la Sección 2.1.3.

Antes de filtrar disponemos de 35017 genes, una vez se han aplicado los diferentes tipos de filtraje a todos los grupos, obtenemos una lista de 6772 genes que se incluirán en el análisis (genes con los que trabajaremos en adelante).

3.2. Selección de genes diferencialmente expresados (DEG)

El análisis para seleccionar los genes diferencialmente expresados se ha basado en el ajuste de un modelo lineal descrito en la sección 2.2 ([12]). Al tratarse de datos apareados (los datos de distintas células se obtenían de un mismo individuo) se requerirá algún tipo de análisis de medidas repetidas. Aunque éste análisis no se encuentra disponible en la metodología utilizada (modelos lineales aplicados a microarrays) se ha incluido el individuo en el análisis considerándolo como un bloque (en el sentido del diseño de experimentos) lo que proporciona una forma de tener en cuenta la correlación entre las observaciones de un mismo individuo.

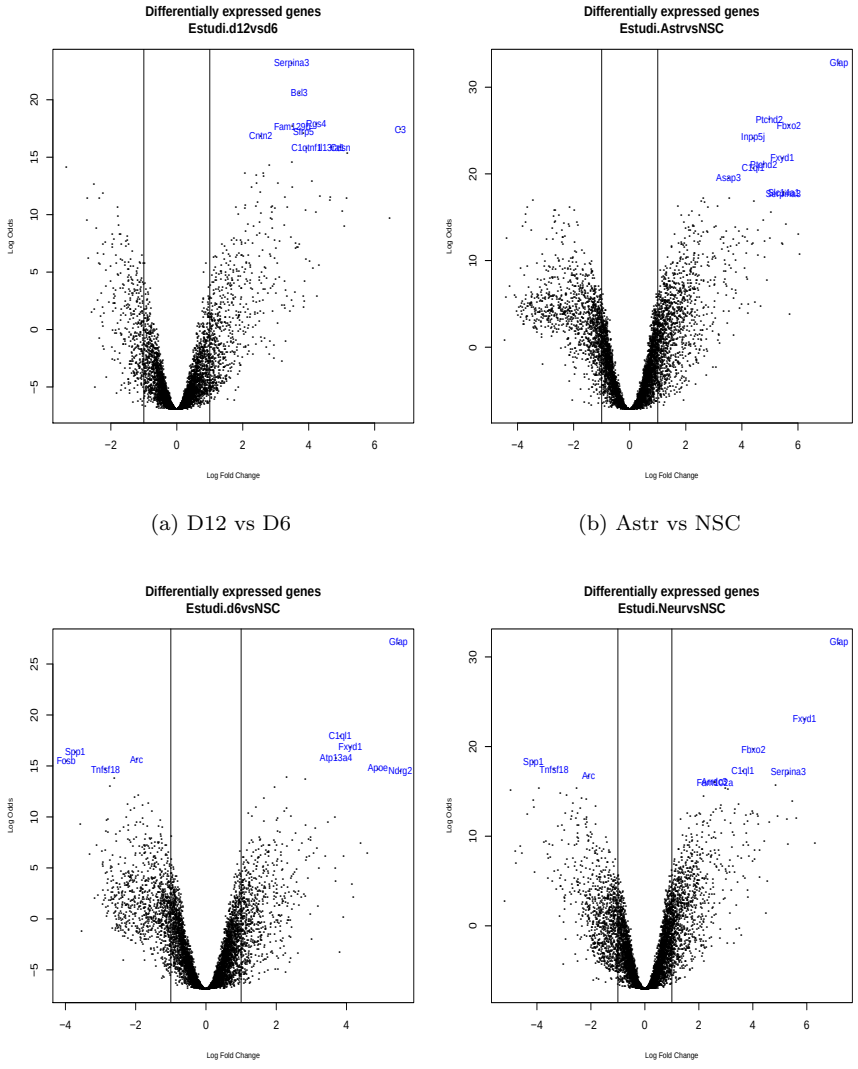
Para cada comparación se obtiene una lista de genes ordenados de mayor a menor diferencialmente expresados. Esto se llama genéricamente "top table" descrita en detalle en la sección 2.2. Un ejemplo de esta en la Figura 10 para la comparación d12 vs d6.

EntrezID	affyIDs	GeneSymbols	AveExpr	t	P.Value	adj.P.Val	B
546546	10398052	Serpina3h Serpina3g Serpina3i	4.9429	21.2080	0.0000	0.0000	23.1882
12051	10560685	Bcl3	5.0160	18.0369	0.0000	0.0000	20.6093
19736	10359908	Rgs4	5.7974	15.2491	0.0000	0.0000	17.8625
227737	10471535	Fam129b	6.9258	15.0838	0.0000	0.0000	17.6829
12266	10452316	C3	8.3858	14.8440	0.0000	0.0000	17.4185
54612	10467744	Sfrp5	4.9618	14.5859	0.0000	0.0000	17.1290
21367	10357705	Cntn2	5.0232	14.3576	0.0000	0.0000	16.8684
386463	10444883	Cdsn	6.6987	13.5304	0.0000	0.0000	15.8882
16164	10599174	Il13ra1	5.7529	13.5167	0.0000	0.0000	15.8715
56745	10383025	C1qtnf1	5.3115	13.4421	0.0000	0.0000	15.7801
14955	10569335	H19	5.0672	13.0853	0.0000	0.0000	15.3361
15561	10593233	Htr3a	6.2859	12.4707	0.0000	0.0000	14.5438
75750	10531887	Slc10a6	4.0750	12.2955	0.0000	0.0000	14.3112
224079	10438854	Atp13a4	5.4547	-12.1981	0.0000	0.0000	14.1807
320265	10540233	Fam19a1	5.1793	11.8154	0.0000	0.0000	13.6582
58220	10478907	Pardób	5.6701	11.7800	0.0000	0.0000	13.6091
74563	10375650	Rasgef1c	4.4505	11.6437	0.0000	0.0000	13.4188
13830	10482030	Stom	9.7121	11.5989	0.0000	0.0000	13.3559
108052	10459866	Slc14a1	7.8035	11.2792	0.0000	0.0000	12.9004
226421	10349661	5430435G22Rik	6.2449	11.1852	0.0000	0.0000	12.7643
12363	10582997	Casp4	4.2020	11.1730	0.0000	0.0000	12.7466
50500	10503502	Ttpa	5.7465	-11.0943	0.0000	0.0000	12.6319
667742	10459335	Fam38b	6.0403	10.9607	0.0000	0.0000	12.4355
382639	10398881	Zbtb42	4.8896	10.9439	0.0000	0.0000	12.4107
15460	10416302	Hr	6.2939	10.6690	0.0000	0.0000	11.9996
245578	10601569	Pcdh11x	8.6636	-10.5900	0.0000	0.0000	11.8799
19293	10430297	Pvalb	6.0501	10.4444	0.0000	0.0000	11.6574
12609	10433885	Cebpd	6.0192	10.3851	0.0000	0.0000	11.5661
22403	10478415	Wisp2	6.2356	10.3445	0.0000	0.0000	11.5032
16803	10478048	Lbp	6.7937	10.3306	0.0000	0.0000	11.4817
50500	10503520	Ttpa	5.3716	-10.3130	0.0000	0.0000	11.4544
108069	10528145	Grm3	7.6911	-10.2781	0.0000	0.0000	11.4001
226115	10467529	Opalin	5.0145	10.1963	0.0000	0.0000	11.2723
22066	10492006	Trpc4	4.1761	10.0818	0.0000	0.0000	11.0921
52840	10478495	Dbndd2	6.2556	9.9075	0.0000	0.0000	10.8147
54132	10467420	Pdlim1	6.8142	9.8775	0.0000	0.0000	10.7667

Figura 10: TopTable para la comparación d12 vs d6.

Los archivos resultantes son `Estudi.XXXvsYYY.html` (donde XXXvsYYY se refiere a la comparación).

Los Volcano plot permiten visualizar si hay muchos o pocos genes con un gran fold-change y significativamente expresados. Estos gráficos representa en las abscisas los cambios de expresión en escala logarítimica y en ordenadas el "logarítmo negativo" del p-valor.



(a) D12 vs D6

(b) Astr vs NSC

(c) D6 vs NSC

(d) Neur vs NSC

Figura 11: **Volcano plot** que muestra los genes candidatos a considerarse como diferencialmente expresados en las diferentes comparaciones. Cuanto más "hacia arriba" y "hacia afuera" se encuentra un gen, más fuerte es la evidencia a favor de que esté diferencialmente expresado.

Resultados de la selección de DEG La tabla 1 muestra el número de genes que han sido seleccionados en cada comparación teniendo en cuenta los diferentes criterios descritos en las filas y con un fold-change superior a 2.

	d12vsd6	AstrvsNSC	d6vsNSC	NeurvsNSC
upReg-B>0	282	180	152	101
downReg-B>0	377	166	28	197
upReg-Adjusted-p-val < 0.01	306	208	164	119
downReg-Adjusted-p-val < 0.01	387	176	29	238
upReg-Adjusted-p-val < 0.05	312	212	174	128
downReg-Adjusted-p-val < 0.05	388	181	31	240
upReg-Adjusted-p-val < 0.25	312	212	174	131
downReg-Adjusted-p-val < 0.25	389	181	33	240
upReg-P value < 0.01	308	211	174	123
downReg-P value < 0.01	387	178	31	239
upReg-P value < 0.05	312	212	174	129
downReg-P value < 0.05	388	181	32	240

Cuadro 1: Numero de DEG bajo diferentes criterios, comparacion simple. FC superior a 2.

Ajuste p-valor Si se desea tener un criterio estadísticamente potente, la selección de los genes diferencialmente expresados debe estar basado en los p-valores ajustados (menores a 0,05) o la B (mayor que 0). Las corrección del p-valor que se aplica en este paso se describe en la sección 2.2.

En este estudio la selección de genes diferencialmente expresados se ha basado en los p-valores ajustados menores de 0,01.

3.3. Comparaciones múltiples entre listas de genes

Con el fin de encontrar los genes que están diferencialmente expresados en múltiples contrastes entre grupos de las comparaciones (AstrvsNSC - NeurvsNSC y d6vsNSC - d12vsd6), hemos realizado un análisis de comparaciones múltiples, como se describe en la sección 2.3.

Asociado a cada grupo de comparaciones múltiples, tenemos los diagramas de Venn donde se muestra la superposición entre contrastes (Figura 12 y 13).

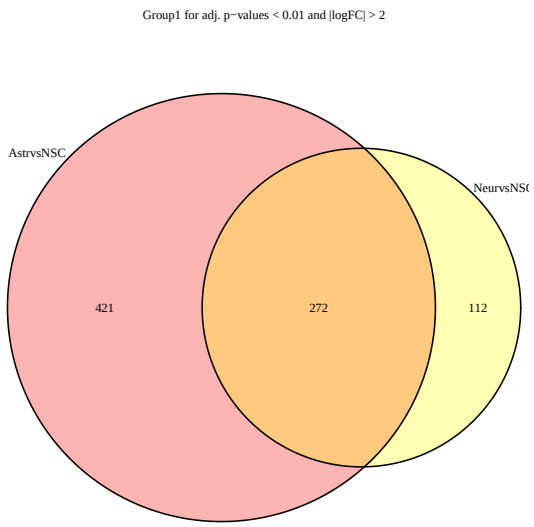


Figura 12: **Diagrama de Venn AstrvsNSC - NeurvsNSC.**En este grupo de comparaciones tenemos en común 272 genes diferencialmente expresados ($p.valor < 0{,}01$ y $|logFC| < 2$). Además se muestran 421 para la comparación Astr vs NSC y 112 para Neur vs NSC

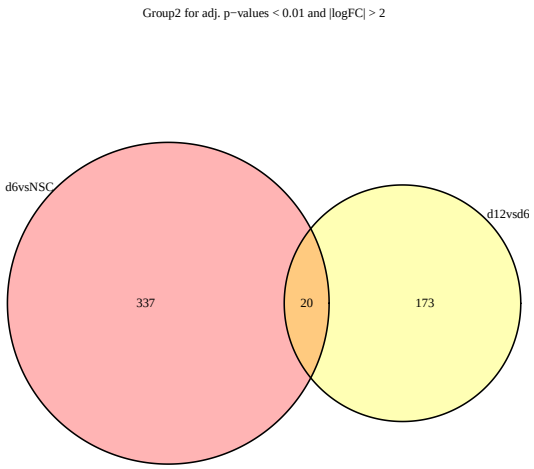


Figura 13: **Diagrama de Venn d6vsNSC - d12vsd6**. Esta figura muestra 20 genes en común diferencialmente expresados ($p.valor < 0{,}01$ y $|logFC| < 2$). Además se muestran 337 únicos para la comparación d6 vs NSC y 173 para d12 vs d6

3.4. Visualización de los perfiles de expresión

Tras seleccionar los genes diferencialmente expresados podemos visualizar las expresiones de cada gen agrupándolas para destacar los genes que se encuentran up o down regulados simultáneamente constituyendo *perfiles de expresión*.

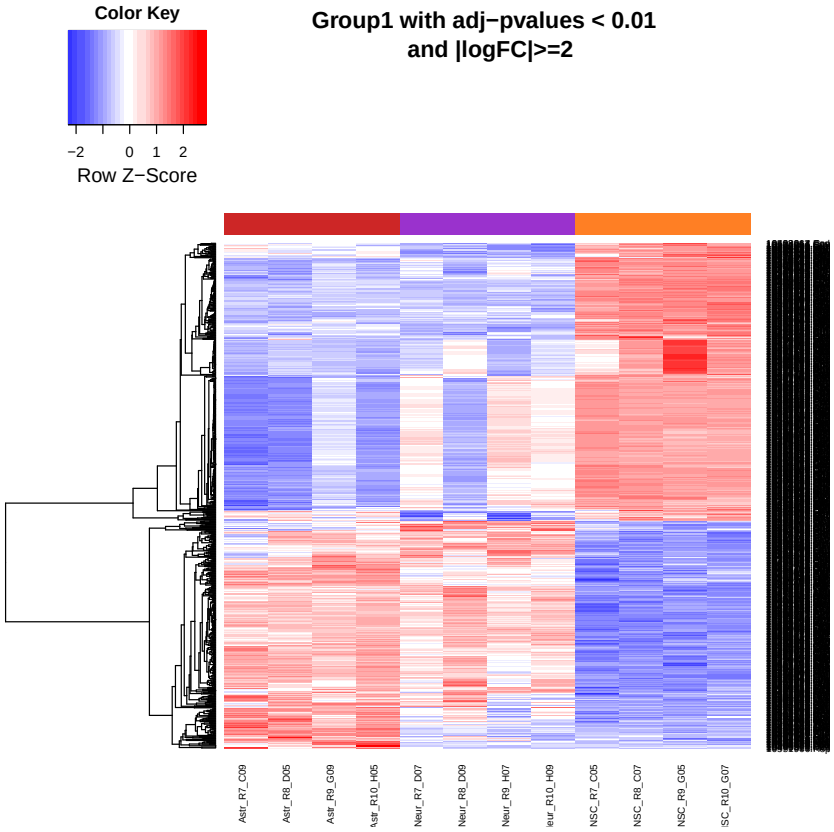


Figura 14: **Heatmap AstrvsNSC - NeurvsNSC.** Vemos dos patrones claramente diferenciados, por un lado tenemos las neuronas y astrocitos, mientras que por otro lado tenemos las células no diferenciadas (NSC). Como ya habíamos visto en el apartado 2.1.2 la muestra `Neur_R8_D09` tiende a agruparse con el patrón de los astrocitos

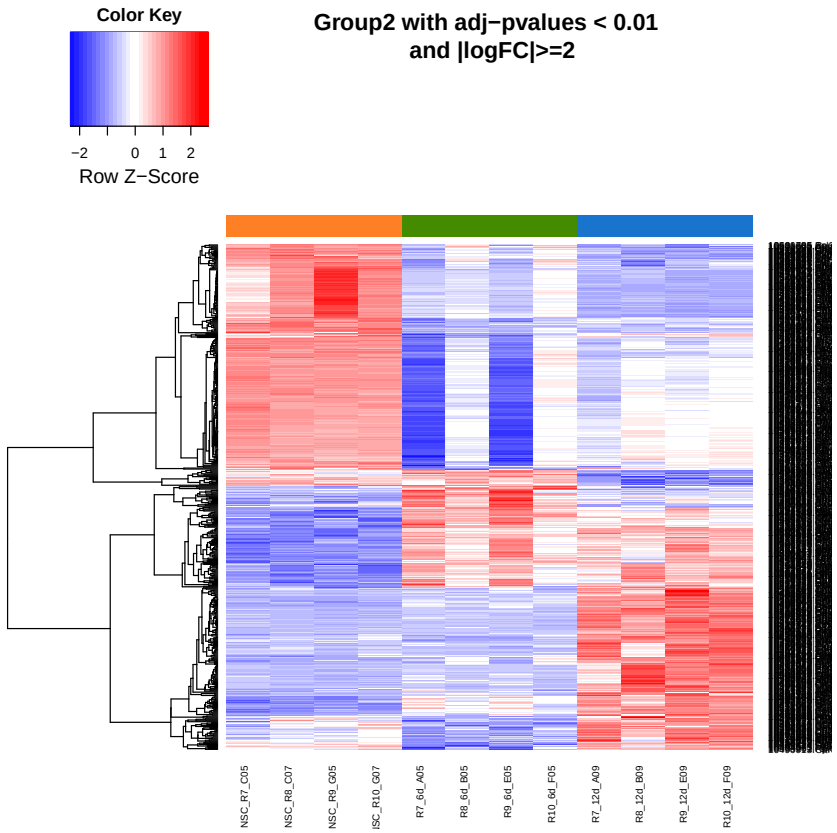


Figura 15: **Heatmap d6vsNSC - d12vsd6**. Este heatmap muestra la evolución de diferenciación de las células (recordemos que el proceso de diferenciación es NSC - d6 - d12). En el heatmap se muestran dos claros patrones (NSC y d12) mientras que d6 es un intermedio, es decir, el cambio de uno a otro.

3.5. Análisis de la significación biológica

El análisis de significación biológica busca establecer si, dada una lista de genes seleccionados por estar diferencialemente expresados entre dos condiciones, las funciones y procesos biólogicos que los caracterizan aparecen en la lista con mayor frecuencia que entre el resto de genes analizados.

Se han desarrollado multitud de variantes de estos tipos de análisis ([8]) pero aquí utilizaremos el análisis básico de enriquecimiento tal como describe Gentleman [5] implementado en el paquete `GOstats` de Bioconductor.

El análisis se realiza sobre dos bases de datos de anotaciones, la "Gene Ontology" y la "Kyoto Encyclopedia of genes and genomes".

Análisis GO El análisis de la "Gene Ontology" (GO) tiene que basarse en un determinado conjunto de genes. Por esa razón, en este estudio hemos decidido seleccionar los genes con un p-valor ajustado inferior a 0.01 para todas las comparaciones.

El análisis llevado a cabo con la GO ha detectado un cierto número de GO categorías enriquecidas. Los genes con "up" o "down" regulados han sido considerados en diferentes listas. Los archivos resultantes son `SignificantGO.Estudi.XXXvsYYY.Up.html y SignificantGO.Estudi.XXXvsYYY.Down.html` (donde XXXvsYYY se refiere a la comparación), el aspecto de estos se muestra en la Figura 16.

Donde los términos de GO han sido separados por tipo ontología (MF = Función Molecular, BP = Proceso biológico, y CC = Componente Celular), ordenados por p-valor y vinculados a la entrada correspondiente GO en AmiGO.

GO Enrichment Analysis for down-regulated genes in comparison: d12vsd6

Ontology	GOID	Term	GeneNames	Size	Count	ExpCount	OddsRatio	Pvalue	OverUnder
MF	GO:0003680	AT DNA binding	Mef2c, Pax6, Pou3f4	6	3	0.0924	64.6	6.99e-05	over
MF	GO:0005215	transporter activity	Cplx1, Slc29a2, Gabra2, Grid1, Hrh1, Kcnd2, Kcnj16, Rlbp1, Rph3a, Scn8a, Slc7a5, Slc27a1, Abcd2, Ttpa, Slc4a4, Syt7, Calcrl, Kcnip3, Slc15a2, Kcnmb4, Aqp11, Kcnv1, Kcnip1, Pitpnc1, Ttyh3, Slc24a3, Kcnh8, Atp13a4, Slc30a10, Slc6a1, Slc9a7, Grin3a, Slc6a11	1043	33	16.1	2.22	7.32e-05	over
MF	GO:0005539	glycosaminoglycan binding	Bcan, Bmp7, Ncan, Fgf1, Fgf12, Fgf14, Ptn, Vegfb, Nell2, Nav2	160	10	2.46	4.37	0.000193	over
MF	GO:0015079	potassium ion transmembrane transporter activity	Kcnd2, Kcnj16, Kcnip3, Kcnmb4, Kcnv1, Kcnip1, Slc24a3, Kcnh8, Slc9a7	136	9	2.09	4.64	0.000262	over
MF	GO:0022892	substrate-specific transporter activity	Slc29a2, Gabra2, Grid1, Hrh1, Kcnd2, Kcnj16, Scn8a, Slc7a5, Slc27a1, Slc4a4, Calcrl, Kcnip3, Slc15a2, Kcnmb4, Aqp11, Kcnv1, Kcnip1, Pitpnc1, Ttyh3, Slc24a3, Kcnh8, Atp13a4, Slc30a10, Slc6a1, Slc9a7, Grin3a, Slc6a11	862	27	13.3	2.17	0.000396	over
MF	GO:0005515	protein binding	Adcy8, Alpl, Bckdhb, Bmp7, Chka, Cplx1, Cpt1a, Dab1, Dll3, Egfr, Emp2, Eps8, F2r, Fgd1, Fgf1, Fgf12, Fgf14, Fxn, Fzd9, Gjd2, Gpr56, Grid1, Magi1, Gyk, Hes5, Il18, Jag1, Kcnd2, Klf12, Lhx2, Lyn, Ascl1, Mef2c, Myo6, Ndp, Nid1, Nrxn1, Nsg2, Ntrk3, Pax6, Pcp4, Pcsk2, Phka1, Prkg2, Pstpip2, Ptn, Rph3a, Scn8a, Sfrp2, Sema3a, Sh3bp1, Sh3gl2, Sntb1, Sox5, Suclg2, Sdc3, Tead2, Nr2e1, Vegfb, Wnt7a, Spry2, Map2k6, Slc27a1, Abcd2, Npas3, Amot, Cttnbp2, Lmcd1, Rgs7bp, Nell2, Gabbr1, Slc4a4, Syt7, Hes6, Smad9, Kcnip3, Vav3, Ppp1r1a, Dact1, Chchd3, Fam69a, Rab3c, Sncaip, Cmtm3, Spire1, Fam195a, Nadk2, Tyw5, Reep6, Kcnip1, Nos1ap, Arhgap26, Pygo1, Ppp2r2b, Pcdh18, Rhobtb3, Agl, Glrx, Trim9, Abtb2, Lphn2, Tdrd7, Cd276, Gldc, Rhoq, Aif1l, Nudt19, Impa2, Rims2, Acot2, Bai3, Spsb4, Satb2, Sema6d, Appl2, Trib2, Rorb, Dner, Kctd12, Pik3c2b, Lrrc7, Grin3a, Raver2, Actr3b, Mtss1l, Pcdh11x, Caskin1, Rasl10b, Mical2, AI464131, Txlnb, Pfn4, Znrf3, Dok6	6851	134	105	1.49	0.000405	over
MF	GO:0043167	ion binding	Acadl, Adcy8, Alpl, Aox1, Bmp7, Car2, Cdh6, Chka, Ncan, Dab1, Egfr, Fgd6, Fgd1, Fgf1, Fgf12, Fgf14, Fxn, Magi1, Gyk, Hrh1, Jag1, Kcnd2, Klf12, Lhx2, Lyn, Mertk, Myo6, Nid1, Nptx1, Nrxn1, Ntrk3, Pip4k2a, Pld1, Prkg2, Ptn, Rph3a, Scn8a, Soat1, Sall3, Suclg2, Nr2e1, Vegfb, Cdh20, Pde10a, Map2k6, Abcd2, Pde7b, Lmcd1, Ttpa, Sall2, Sept9, Nell2, Syt7, Smad9, Kcnip3, Sept6, Pnkd, Vav3, Rab3c, Akr1b10, Mto1, Nadk2, Tyw5, Dpf3, Kcnip1, Arhgap26, Cyp2u1, Pygo1, Mblac2, Pcdh18, Rhobtb3, Lonrf3, Cyp2j9, Ppa2, Pank1, Nav2, Trim9, Efhd1, Gldc, Rhoq, Adamts6, Aif1l, Lmo3, Nudt19, Impa2, Rims2, Zkscan2, Trib2, Dgkb, Atp13a4, Rorb, Dner, Adamts12, Pik3c2b, Zfp697, Grin3a, Actr3b, Nim1k, Fat3, Rasl10b, Pcdh19, B3gat2, Agmo, Dctd, Cdh10, Mical2, Lonrf2, Zfp69, Znrf3, Zfp827	5406	110	83.2	1.51	0.000451	over
MF	GO:0046873	metal ion transmembrane transporter activity	Hrh1, Kcnd2, Kcnj16, Scn8a, Slc4a4, Kcnip3, Kcnmb4, Kcnv1, Kcnip1, Slc24a3, Kcnh8, Slc6a1, Slc9a7, Grin3a, Slc6a11	361	15	5.56	2.86	0.000511	over
MF	GO:0008201	heparin binding	Bmp7, Fgf1, Fgf12, Fgf14, Ptn, Vegfb, Nell2, Nav2	122	8	1.88	4.58	0.000609	over
MF	GO:0015077	monovalent inorganic cation transmembrane transporter activity	Kcnd2, Kcnj16, Scn8a, Slc4a4, Kcnip3, Kcnmb4, Kcnv1, Kcnip1, Slc24a3, Kcnh8, Slc6a1, Slc9a7, Slc6a11	295	13	4.54	3.03	0.000696	over

Figura 16: Resultados GO

Análisis KEGG El análisis de *La Enciclopedia de Kyoto de genes y genomas* (KEGG) está basado también en un conjunto fijo de genes. En este estudio, los genes seleccionados han sido solo aquellos cuyos p.valores ajustados son inferiores a 0.01. El análisis ha detectado un cierto número de vías KEGG enriquecidas, con la correspondiente lista en el fichero `SignificantKEGG.Estudi.XXXvsYYY.html` , ordenado por el p-valor. La Figura 17 muestra un ejemplo del contenido del fichero.

KEGG Enrichment Analysis for regulated genes in comparison: d12vsd6

KEGGID	Pvalue	OddsRatio	ExpCount	Count	Size	Term	GeneNames
04010	8.95e-06	2.79	11.5	28	263	MAPK signaling pathway	Gadd45a, Egfr, Fas, Fgf1, Fgf12, Fgf14, Fgf2, Fos, Hspb1, Il1r1, Stmn1, Mef2c, Gadd45b, Nfkb1, Pdgfa, Pdgfb, Ppp3ca, Dusp1, Relb, Rras, Tnfrsf1a, Gadd45g, Map2k6, Map3k8, Rras2, Arrb1, Hspa1a, Flnb
04380	0.000479	3.12	5.06	14	116	Osteoclast differentiation	Socs3, Fcgr2b, Fos, Fosl2, Ifnar2, Ifngr1, Il1r1, Junb, Nfkb1, Nfkbia, Ppp3ca, Relb, Tnfrsf1a, Map2k6
00564	0.000638	3.6	3.49	11	80	Glycerophospholipid metabolism	Chka, Gpd1, Gpam, Pld1, Agpat4, Pgs1, Dgkg, Dgkb, Mboat1, Lpgat1, Pcyt1b
05218	0.000925	3.69	3.1	10	71	Melanoma	Egfr, Fgf1, Fgf12, Fgf14, Fgf2, Mdm2, Met, Pdgfa, Pdgfb, Rb1
04810	0.00093	2.37	9.29	20	213	Regulation of actin cytoskeleton	Arpc1b, Egfr, F2r, Fgd1, Fgf1, Fgf12, Fgf14, Fgf2, Itga2b, Pdgfa, Pdgfb, Pip4k2a, Rras, Ezr, Fgd3, Vav3, Rras2, Actn1, Gsn, Pfn4
04540	0.00118	3.31	3.75	11	86	Gap junction	Adcy8, Egfr, Gjd2, Lpar1, Pdgfa, Pdgfb, Prkg2, Tjp1, Tubb3, Tubb4a, Tubb4b
04514	0.00144	2.64	6.28	15	144	Cell adhesion molecules (CAMs)	Cd80, H2-D1, H2-K1, H2-Q7, Mag, Nrxn1, Cldn11, Sdc1, Sdc3, Cntn2, Icosl, Cldn10, Cd276, Nfasc, Nrcam
04144	0.00266	2.2	9.42	19	216	Endocytosis	Egfr, Ehd1, F2r, H2-D1, H2-K1, H2-Q7, Mdm2, Met, Pld1, Sh3gl2, Ehd3, Pard6b, Cltb, Pard6g, Ldlrap1, Arrb1, Hspa1a, Asap3, Ehd2
04620	0.00399	2.78	4.36	11	100	Toll-like receptor signaling pathway	Cd80, Fos, Cxcl10, Ifnar2, Lbp, Nfkb1, Nfkbia, Map2k6, Map3k8, Ikbke, Irak4
00561	0.00621	3.55	2.22	7	51	Glycerolipid metabolism	Gpam, Gyk, Dgat2, Agpat4, Dgkg, Dgkb, Mboat1
04115	0.00826	3.03	2.92	8	67	p53 signaling pathway	Cd82, Gadd45a, Fas, Mdm2, Gadd45b, Gadd45g, Shisa5, Steap3
04060	0.00868	1.94	10.5	19	241	Cytokine-cytokine receptor interaction	Bmp7, Egfr, Fas, Cxcl10, Ifnar2, Ifngr1, Il13ra1, Il18, Il1r1, Il6ra, Il6st, Met, Osmr, Pdgfa, Pdgfb, Tnfrsf1a, Vegfb, Tnfrsf21, Tnfsf18
04210	0.0113	2.65	3.71	9	85	Apoptosis	Fas, Il1r1, Nfkb1, Nfkbia, Ppp3ca, Prkar1b, Tnfrsf1a, Endod1, Irak4
05142	0.0115	2.49	4.36	10	100	Chagas disease (American trypanosomiasis)	Ace, C3, Fas, Fos, Ifngr1, Nfkb1, Nfkbia, Tnfrsf1a, Ppp2r2b, Irak4
05200	0.012	1.76	14	23	320	Pathways in cancer	Runx1, Col4a5, Egfr, Fas, Fgf1, Fgf12, Fgf14, Fgf2, Fos, Fzd9, Itga2b, Mdm2, Met, Nfkb1, Nfkbia, Pdgfa, Pdgfb, Rb1, Stat3, Vegfb, Wnt5b, Wnt7a, Foxo1
05160	0.0144	2.19	5.89	12	135	Hepatitis C	Socs3, Egfr, Ifnar2, Irf1, Nfkb1, Nfkbia, Cldn11, Stat3, Tnfrsf1a, Ikbke, Cldn10, Ppp2r2b
00534	0.0248	4.03	1.13	4	26	Glycosaminoglycan biosynthesis - heparan sulfate	Ndst1, Hs3st3b1, Xylt1, B3gat2
00601	0.0248	4.03	1.13	4	26	Glycosphingolipid biosynthesis - lacto and neolacto series	Fut9, Gcnt2, Ggta1, B3galt1
04512	0.0293	2.35	3.66	8	84	ECM-receptor interaction	Col11a1, Col4a5, Col6a1, Col6a2, Itga2b, Sdc1, Sdc3, Vtn
05215	0.0395	2.2	3.88	8	89	Prostate cancer	Egfr, Mdm2, Nfkb1, Nfkbia, Pdgfa, Pdgfb, Rb1, Foxo1
04610	0.0442	2.29	3.27	7	75	Complement and coagulation cascades	Serping1, C3, C4b, F2r, Pros1, C1s, A2m
04662	0.0442	2.29	3.27	7	75	B cell receptor signaling pathway	Fcgr2b, Fos, Lyn, Nfkb1, Nfkbia, Ppp3ca, Vav3
04070	0.0498	2.22	3.36	7	77	Phosphatidylinositol signaling system	Pip4k2a, Plce1, Dgkg, Impa2, Dgkb, Inpp4b, Pik3c2b

Figura 17: Resultados KEGG

3.6. Mapas de conectividad

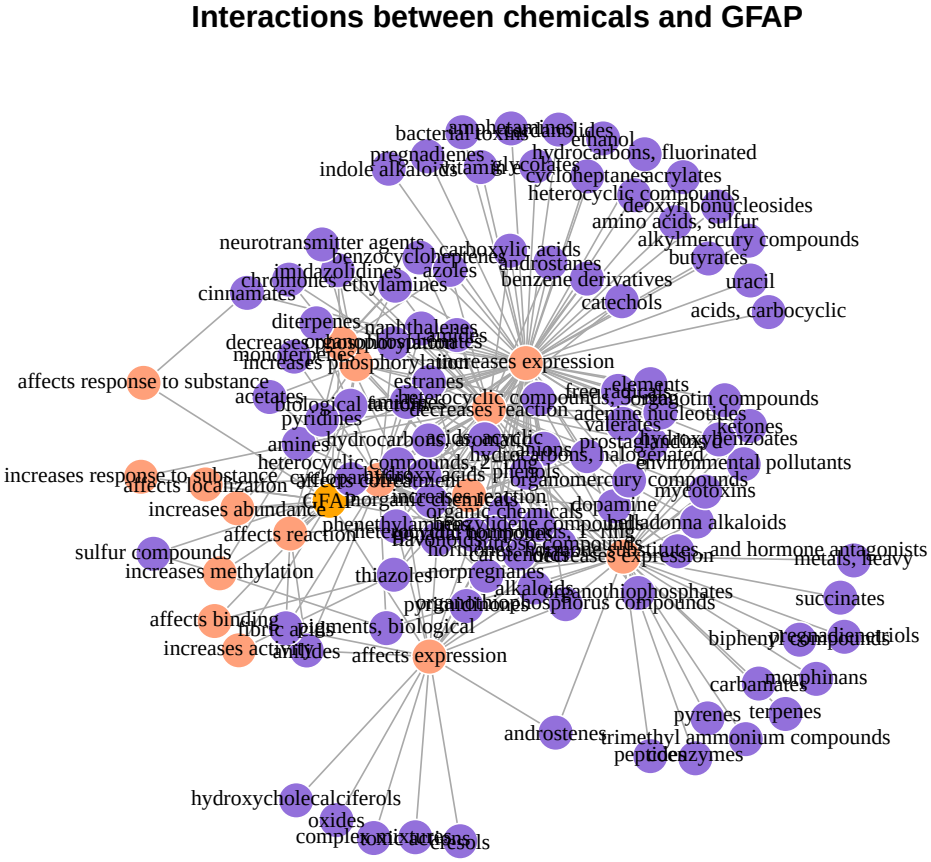


Figura 18: **Mapa de conectividad para GFAP vs químicos**. Muestra los productos químicos que interactúan con el gen `GFAP` en forma de red.

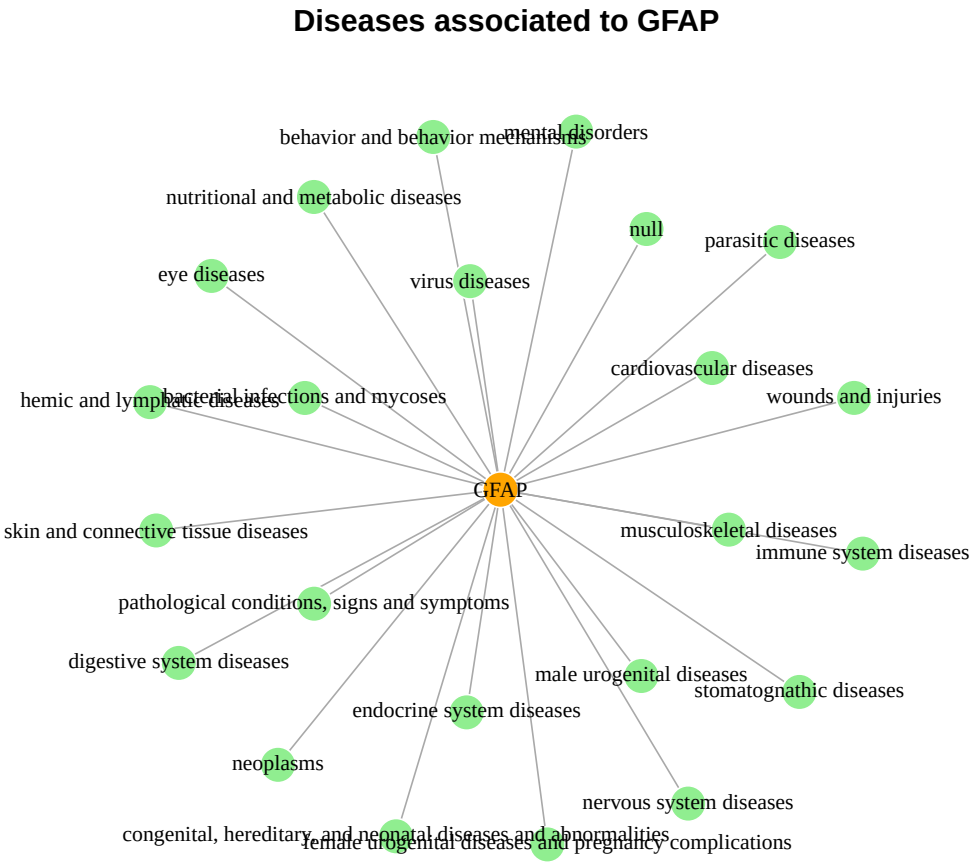


Figura 19: **Mapa de conectividad para GFAP vs enfermedades**. De la misma manera que podemos ver los productos químicos que interactuan con el gen `GFAP`, podemos trazar uan red con las enfermedades asociadas a este determinado gen.

Y por último la Figura 20 nos muestra la relación del gen de interés (`GFAP`) con otros genes.

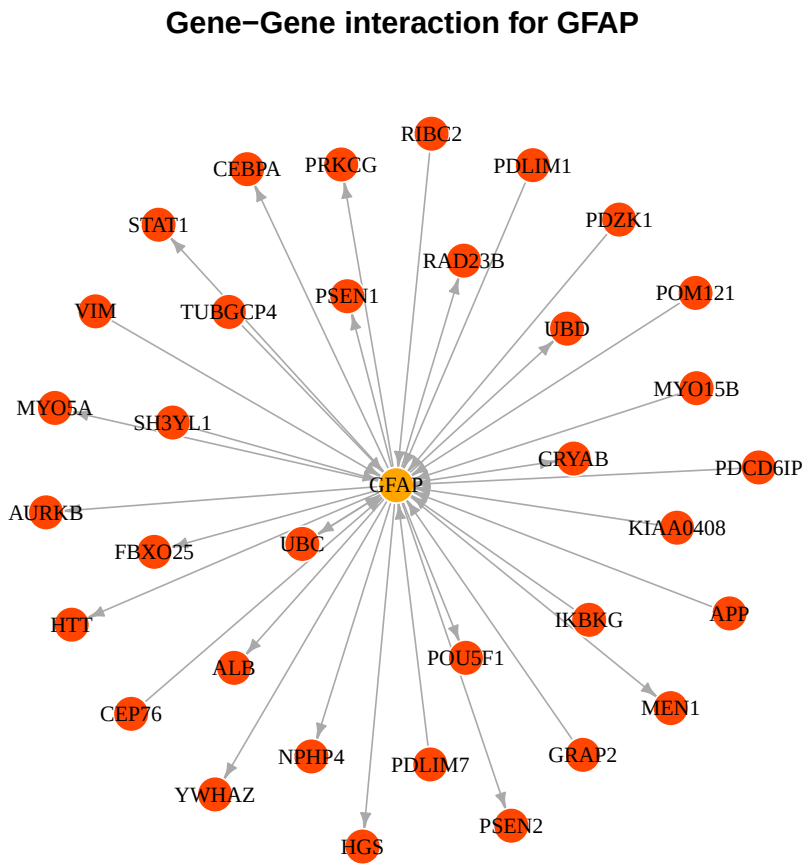


Figura 20: Mapa de conectividad GFAP vs Genes

4. Discusión y conclusiones

4.1. Discusión y conclusiones

Los controles de calidad llevados a cabo en la primera parte de este estudio han permitido establecer que los datos con los que se ha trabajado eran de buena calidad, a pesar de presentar cierta heterogeneidad que queda ilustrada en la incompleta agrupación de las muestras según tipos de célula.

Los análisis llevados a cabo sobre los datos normalizados y filtrados han permitido detectar un cierto número de genes diferencialmente expresados (Tabla 1) que se prestan a estudios posteriores acerca de su significación biológica.

El análisis de significación biológica ha permitido detectar una serie de funciones y procesos biológicos que caracterizan las listas de genes seleccionadas en las bases de datos de anotaciones más populares, la Gene Ontology (GO) y la "Kyoto Encyclopedia of Genes and Genomes".

Los mapas de conectividad han permitido ver con que compuestos químicos, enfermedades o genes están relacionados los genes seleccionados por los investigadores, y dan lugar a un posterior análisis de estos por parte de los investigadores.

Como limitaciones más importantes de ese estudio podemos destacar:

El tamaño de las muestras utilizadas es bastante limitado lo que determina que el estudio tenga poca potencia por lo que probablemente habrá menos reproducibilidad y más falsos negativos de lo que sería deseables si se utilizara un mayor número de muestras.

En cada paso del proceso se han tomado decisiones relativamente arbitrarias acerca de los métodos a seguir para la normalización, filtración, selección de los genes, etc. La decisión de si estos métodos son los más adecuados o no es probablemente subjetiva (véase por ejemplo [[14],[10]]) por lo que sería interesante saber como cambian los resultados si se tomaran otras decisiones.

Los problemas anteriores no son, sin embargo, problemas de este estudio concreto, sino en general de los estudios basados en microarrays por lo que, limitaciones aparte, el estudio aportará probablemente información valiosa que permitirá un seguimiento posterior del problema.

Asmismo, cabe destacar la importancia del uso correcto de las técnicas estadísticas en análisis de microarray y estudios científicos.

5. Bibliografía

Referencias

[1] M.J.;Prez-Vieitez M.C. Aladro, Y.; Alemany. Prevalencia e incidencia de la esclerosis mltiple en la ciudad de las palama, islas canarias,espaa. *Neuroepidemiologa*, 24:70–75, 2005.

[2] Samuel Weiss Andrew Chojnacki. Production of neurons, astrocytes and oligodendrocytes from mammalian cns stem cells. *Nature Protocols*, 3:935–940, 2008.

[3] Y. Benjamini and Y. Hochberg. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society B*, 57:289–300, 1995.

[4] Louis S. Guillemin F.; LORSEP Group. Debouverie M., Pittion-Vouyovitch S. Natural history of multiple sclerosis in a population-based cohort. *Eur J Neurol.*, 15:916–21, 2008.

[5] Robert Gentleman. Using go for statistical analysis. *Bioconductor's compendiums*, 2004.

[6] Robert Gentleman, Vince Carey, Wolfgang Huber, Rafael Irizarry, and Sandrine Dudoit. *Bioinformatics and Computational Biology Solutions using R and Bioconductor*. Springer, New York, 2005.

[7] R. A. Irizarry, B. Hobbs, F. Collin, Y. D. Beazer-Barclay, K. J. Antonellis, U. Scherf, and T. P. Speed. Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics*, 4:249–264, 2003.

[8] Drghici S. Khatri P. Ontological analysis of gene expression data: current tools, limitations, and open problems. *Bioinformatics.*, 18:3587–95, 2005.

[9] J L Mosquera and A Sánchez-Pla. Serbgo: searching for the best go tool. *Nucleic Acids Res*, 36(Web Server issue):W368–71, July 2008.

[10] Jeffrey C Miecznikowski Qianqian Zhu and Marc S Halfon. Preferred analysis methods for affymetrix genechips. ii. an expanded, balanced, wholly-defined spike-in dataset. *BMC Bioinformatics*, 11:285, 2005.

[11] D. B. Allison; X. Cui; G. P. Page; M. Sabripour. Microarray data analysis:from disarray to consolidation and consensus. *Nat. Rev. Genet.*, 1:55–65, 2006.

[12] Gordon K. Smyth. Linear models and empirical bayes methods for assessing differential expression in microarray experiments. *Statistical applications in genetics and molecular biology*, 3(1), 2004.

[13] Martin R. Sospedra M. Immunology of multiple sclerosis. *PAnnu Rev Immunol*, 23:683–747, 2005.

[14] Alan M Michelson George M Church Sung E Choe, Michael Boutros and Marc S Halfon. Preferred analysis methods for affymetrix genechips revealed by a wholly defined control dataset. *Genome Biol.*, 6, 2005.

6. Anexo

6.1. Metadata

- **Título:** Búsqueda de nuevas terapias que favorezcan la neuroregeneración en la esclerosis múltiple.
- **Autor:** Miriam Mota
- **Tutores:** Alex Sánchez y Juan Ramón González.

Abstract. The aim of this study is to determine the genetic profiling that defines the differentiation from NSC into astrocyte, neuron and oligodendrocyte . The objective is to specify the channels related to the differentiation processes toward each of the three cell types. And finally, use the defined genetic profiling in order to identify compounds that would promote the differentiation of the NSC towards each cellular types in a independent way .

Resumen. Los objetivos de este estudio son determinar la huella genética que define la diferenciación de NSC a astrocito, neurona y oligodendrocito. Determinar las vías relacionadas con los procesos de diferenciación hacia cada uno de los tres tipos celulares. Y finalmente, utilizar la huella genética definida para identificar compuestos que favorezcan la diferenciación de las NSC hacia cada uno de los tipos celulares independientemente.

Resum. Els objectius d'aquest estudi són determinar la petjada genética que defineix la diferenciació de NSC a astròcit, neurona o oligodendrocit. Determinar les vies relacionades amb els processos de diferenciació cap a cadascun dels tres tipus cel.lulars. I finalment, utilitzar la petjada genética definida per identificar compostos que afavoreixin la diferenciació de la NSC cap a cadascun dels tipus cel.lulars independentment.

6.2. Glosario de algunos términos utilizados

- **PCA:** es una técnica utilizada para reducir la dimensionalidad de un conjunto de datos. Intuitivamente la tcnica sirve para hallar las causas de la variabilidad de un conjunto de datos y ordenarlas por importancia.
- **Heatmap:** Representación gráfica de los datos donde los valores se representan en una escala de colores.
- **Dendograma:** diagrama de datos en forma de árbol que organiza los datos en subcategorías que se van dividiendo en otras hasta llegar al nivel de detalle deseado.
- **Median polish:** es un procedimiento de exploración propuesto por John Tukey. Donde se ajusta un modelo aditivo para dlos datos, cuya forma seria efecto filas + efecto columnas + mediana global.
- **RMA:** Robust Multichip Average, método de normalización en el cual se usan todos los arrays simultaneamente. Este procedimiento solamente utiliza los valores PM.
- **Odds Ratio:** es una medida de asociación entre una exposición y un resultado. El odds representa la probabilidad de que un resultado se produzca dada una exposición particular, en comparación con las odds del resultado que ocurre en ausencia de esa exposición.
- **Volcano plot:** es un tipo de *scatterplot* que se utiliza para identificar rápidamente los cambios en grandes conjuntos de datos compuestos por datos replicados.
- **Diagrama de Venn:** son esquemas usados en la teoría de conjuntos. Estos diagramas muestran colecciones (conjuntos) de cosas (elementos) por medio de líneas cerradas. La línea cerrada exterior abarca a todos los elementos bajo consideración, el conjunto universal U.