

PEC 1

Análisis de datos ómicos

Eduardo A. Sánchez Torres

3 de mayo de 2020

Índice

1. Selección del estudio: Datos para el análisis	2
1.1. Estudio	2
1.2. Microarray	2
1.3. Organismo	2
1.4. Diseño experimental	3
2. Proceso de análisis	4
2.1. Preparación de los datos para el análisis	4
2.2. Control de calidad de datos sin procesar	4
2.3. Normalización	5
2.4. Control de calidad sobre los datos normalizados	6
2.5. Filtrado no específico	8
2.6. Selección de genes expresados diferencialmente	8
2.6.1. Matriz de diseño	8
2.6.2. Contrastes	9
2.6.3. Estimación del modelo y selección	10
2.7. Anotación de genes	10
2.8. Comparaciones múltiples	10
2.9. Análisis de significación biológica	11
3. Resumen de resultados	12
3.1. Etapa presintomática (60 días)	12
3.2. Etapa sintomática (90 días)	14
3.3. Etapa final (120 días)	14
3.4. Comparación del perfil de expresión génica entre las tres etapas	17
3.5. Significación biológica	21
4. Repositorio GitHub	23
Referencias	24

1. Selección del estudio: Datos para el análisis

1.1. Estudio

El conjunto de datos utilizado pertenece a un estudio publicado [1], que estudia el perfil de expresión de las neuronas motoras durante la progresión de la Esclerósis Lateral Amiotrófica (ELA) en el modelo de ratón SOD1 G93A. Los datos están disponibles desde marzo de 2008 en la base de datos Gene Expression Omnibus (GEO) bajo el código de identificación GSE10953.

La ELA es una enfermedad del sistema nervioso central, caracterizada por una degeneración progresiva de las neuronas motoras en la corteza cerebral, tronco del encéfalo y médula que implica debilidad y atrofia muscular. En el avance de la enfermedad se ve comprometida la autonomía motora, la respiración, la deglución y la comunicación oral. Por el contrario, se mantiene conservada la capacidad cognitiva, los sentidos, el intelecto y los músculos de los ojos y esfínteres [2]. La ELA se presenta tanto en formas esporádicas como familiares, clínicamente indistinguibles, con una incidencia de 1 a 2 casos por cada 100,000 individuos. Aproximadamente el 10 % de los casos son familiares, y el 20 % de éstos se han relacionado con una forma mutante dominante de la enzima Cu/Zn superóxido dismutasa 1 (SOD1) [3].

Los ratones transgénicos que portan formas mutantes de SOD1 desarrollan una enfermedad neuromuscular muy similar a la ELA humana tanto en el fenotipo como en las características histopatológicas, de modo que los ratones SOD1 mutantes se consideran un modelo confiable y se han utilizado ampliamente en la investigación de la ELA. Fuera del 2 % de los casos que portan una mutación SOD1, las causas de la ELA aún se desconocen principalmente, y la alta proporción de casos esporádicos sugiere que esta enfermedad es de origen multifactorial [4].

1.2. Microarray

El microarray utilizado para estudiar el perfil de expresión es concretamente el modelo GeneChip Mouse Genome 430A (Affy MOE430A) de la compañía Affymetrix.

1.3. Organismo

Los ratones transgénicos empleados portan la forma mutante G93A (glicina 93 sustituida por alanina) del gen humano SOD1. Concretamente poseen 20 copias de dicho gen y como se ha dicho, desarrollan una enfermedad neurodegenerativa que reproduce características fenotípicas e histopatológicas de la ELA humana. Al nacer, estos ratones no se distinguen de sus compañeros de camada, por lo que se identifican mediante detección por PCR para SOD1, utilizando ADN extraído de biopsias de cola.

La etapa presintomática dura aproximadamente 75 días desde el nacimiento, después de la cual los ratones comienzan a desarrollar temblores y pérdida del reflejo de aplastamiento de las extremidades posteriores. La enfermedad progresiona rápidamente, lo que lleva a una reducción de la marcha y al deterioro del rendimiento motor, con parálisis progresiva inicialmente de las extremidades posteriores, que culmina en parálisis total y muerte a los 140 ± 6 días. Este curso de la enfermedad determinó los tiempos elegidos en este estudio para la toma de

muestras: se fijaron 60 días como etapa presintomática, 90 días como etapa sintomática y 120 días como etapa final de la enfermedad.

1.4. Diseño experimental

Se aislaron neuronas motoras de la médula espinal mediante la tecnología láser LCM, de tres ratones machos transgénicos (G93A) y de 3 compañeros de camada machos no transgénicos (Control), para cada tiempo fijado: 60, 90 y 120 días desde el nacimiento.

De esta manera podemos considerar que el diseño exprimental se corresponde con un diseño factorial de 2 factores, *Ratón* con dos niveles (G93A y Control) y *Tiempo* con tres niveles (60, 90 y 120 días). Se dispone de 3 réplicas para cada combinación nivel-factor (grupos), constituyendo así un total de 18 muestras.

2. Proceso de análisis

El proceso de análisis de microarrays se ha llevado a cabo utilizando el lenguaje R versión 3.6.2 a través de los paquetes de Bioconductor versión 3.10. En este apartado se detallan cada uno de los pasos seguidos en dicho análisis.

2.1. Preparación de los datos para el análisis

A partir de los archivos CEL, propios de microarrays Affymetrix, descargados de la base de datos GEO, se ha generado el correspondiente **ExpressionSet** de datos crudos, con el que comenzar el análisis.

El conjunto de datos del estudio consistía en 18 archivos CEL, uno por cada muestra experimental, asociados según se muestra en la Tabla 1.

Tabla 1: Asociación de los 18 archivos CEL del estudio con cada condición experimental.

	Archivo	Grupo	Ratón	Tiempo	Identificador
1	GSM277558.CEL	G93A_60	G93A	60	G93A_60_1
2	GSM277559.CEL	G93A_60	G93A	60	G93A_60_2
3	GSM277560.CEL	G93A_60	G93A	60	G93A_60_3
4	GSM277561.CEL	Control_60	Control	60	Control_60_1
5	GSM277562.CEL	Control_60	Control	60	Control_60_2
6	GSM277563.CEL	Control_60	Control	60	Control_60_3
7	GSM277564.CEL	G93A_90	G93A	90	G93A_90_1
8	GSM277565.CEL	G93A_90	G93A	90	G93A_90_2
9	GSM277566.CEL	G93A_90	G93A	90	G93A_90_3
10	GSM277567.CEL	Control_90	Control	90	Control_90_1
11	GSM277568.CEL	Control_90	Control	90	Control_90_2
12	GSM277569.CEL	Control_90	Control	90	Control_90_3
13	GSM277570.CEL	G93A_120	G93A	120	G93A_120_1
14	GSM277571.CEL	G93A_120	G93A	120	G93A_120_2
15	GSM277572.CEL	G93A_120	G93A	120	G93A_120_3
16	GSM277573.CEL	Control_120	Control	120	Control_120_1
17	GSM277574.CEL	Control_120	Control	120	Control_120_2
18	GSM277575.CEL	Control_120	Control	120	Control_120_3

2.2. Control de calidad de datos sin procesar

Antes de realizar el proceso de normalización de los datos crudos del experimento, se ha realizado con control de calidad de los mismos, con la finalidad de verificar si presentan ciertas anomalías que introducirían ruido en el análisis.

Para ello se ha utilizado el paquete **ArrayQualityMetrics** que analiza los datos a partir de

diferentes enfoques de calidad. Los resultados obtenidos se pueden observar en la Figura 1.

array	sampleNames	*1	*2	*3	Group	Mouse	Days	ShortName
1	G93A_60_1	x			G93A_60	G93A	60	G93A_60_1
2	G93A_60_2				G93A_60	G93A	60	G93A_60_2
3	G93A_60_3		x		G93A_60	G93A	60	G93A_60_3
4	Control_60_1				Control_60	Control	60	Control_60_1
5	Control_60_2				Control_60	Control	60	Control_60_2
6	Control_60_3			x	Control_60	Control	60	Control_60_3
7	G93A_90_1		x		G93A_90	G93A	90	G93A_90_1
8	G93A_90_2				G93A_90	G93A	90	G93A_90_2
9	G93A_90_3		x		G93A_90	G93A	90	G93A_90_3
10	Control_90_1				Control_90	Control	90	Control_90_1
11	Control_90_2		x		Control_90	Control	90	Control_90_2
12	Control_90_3			x	Control_90	Control	90	Control_90_3
13	G93A_120_1		x		G93A_120	G93A	120	G93A_120_1
14	G93A_120_2			x	G93A_120	G93A	120	G93A_120_2
15	G93A_120_3				G93A_120	G93A	120	G93A_120_3
16	Control_120_1	x		x	Control_120	Control	120	Control_120_1
17	Control_120_2			x	Control_120	Control	120	Control_120_2
18	Control_120_3				Control_120	Control	120	Control_120_3

Figura 1: Resultados del análisis con el paquete `ArrayQualityMetrics` antes de la normalización.

Las columnas *1, *2 y *3 hacen referencia a los distintos métodos ejecutados por la función para la detección de outliers. Si el microarray de una muestra está por encima de un cierto umbral definido en la función según el método, se marca con un asterisco como un valor atípico (outlier). Cuando una determinada muestra se marca tres veces, tiene que revisarse cuidadosamente, y tal vez deba ser rechazada para mejorar la calidad general del experimento.

En nuestro caso ninguna muestra se ha marcado tres veces, por lo que todas se consideran en el análisis. Únicamente un microarray se ha marcado dos veces, para las columnas *1 y *3, y once microarrays se han marcado una vez para la columna *3.

Adicionalmente se ha construido un diagrama de cajas (Figura 2), donde podemos observar como se distribuye la intensidad de señal de cada microarray antes de la normalización. Se aprecia una ligera variación de intensidad entre las muestras, como es de esperar para datos crudos.

2.3. Normalización

Antes de comenzar con el análisis de expresión diferencial, es necesario normalizar la intensidad de cada microarray para que puedan ser comparables entre sí y tratar de reducir, y si es posible eliminar, toda la variabilidad en las muestras que no se deba a razones biológicas. El

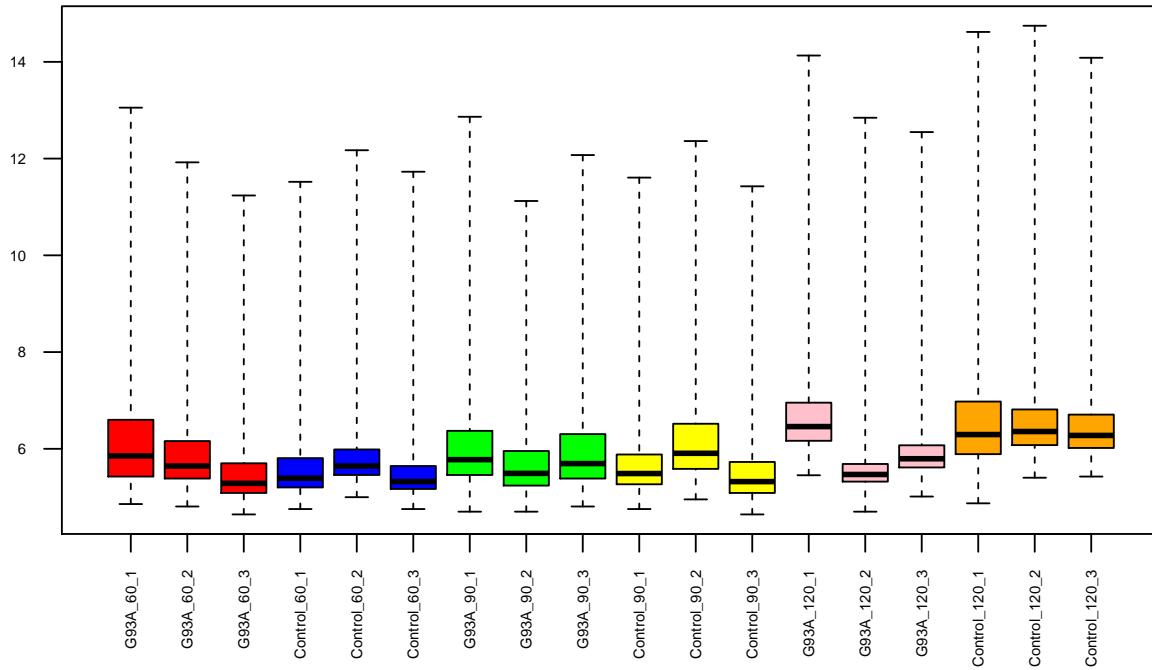


Figura 2: Distribución de intensidad de cada microarray antes de la normalización.

proceso de normalización intenta asegurar que las diferencias de intensidad presente entre muestras, refleje únicamente la expresión diferencial de los genes, en lugar de sesgos debidos a cuestiones técnicas.

El método de normalización que se ha utilizado en este análisis es el RMA (robust multi-array average) desarrollado por [5] en 2003. Para ello se ha utilizado la función `rma` del paquete `oligo`.

De esta manera, se ha generado en este punto el correspondiente `ExpressionSet` de los datos normalizados.

2.4. Control de calidad sobre los datos normalizados

Transcurrido el proceso de normalización, es conveniente realizar un control de calidad sobre los datos normalizados para explorar como ha afectado la normalización a los posibles valores atípicos detectados anteriormente.

Se ha procedido de la misma manera que en el control de calidad de los datos crudos. Los resultados del análisis con `ArrayQualityMetrics` se muestran en la Figura 3 y el diagrama de cajas de distribución de intensidades en la Figura 4.

Podemos destacar que no se han detectado valores atípicos para ninguna de las muestras y que la distribución de intensidades es más homogénea tras la normalización de los datos crudos.

array	sampleNames	*1	*2	*3	Group	Mouse	Days	ShortName
1	G93A_60_1				G93A_60	G93A	60	G93A_60_1
2	G93A_60_2				G93A_60	G93A	60	G93A_60_2
3	G93A_60_3				G93A_60	G93A	60	G93A_60_3
4	Control_60_1				Control_60	Control	60	Control_60_1
5	Control_60_2				Control_60	Control	60	Control_60_2
6	Control_60_3				Control_60	Control	60	Control_60_3
7	G93A_90_1				G93A_90	G93A	90	G93A_90_1
8	G93A_90_2				G93A_90	G93A	90	G93A_90_2
9	G93A_90_3				G93A_90	G93A	90	G93A_90_3
10	Control_90_1				Control_90	Control	90	Control_90_1
11	Control_90_2				Control_90	Control	90	Control_90_2
12	Control_90_3				Control_90	Control	90	Control_90_3
13	G93A_120_1				G93A_120	G93A	120	G93A_120_1
14	G93A_120_2				G93A_120	G93A	120	G93A_120_2
15	G93A_120_3				G93A_120	G93A	120	G93A_120_3
16	Control_120_1				Control_120	Control	120	Control_120_1
17	Control_120_2				Control_120	Control	120	Control_120_2
18	Control_120_3				Control_120	Control	120	Control_120_3

Figura 3: Resultados del análisis con el paquete `ArrayQualityMetrics` de los datos normalizados.

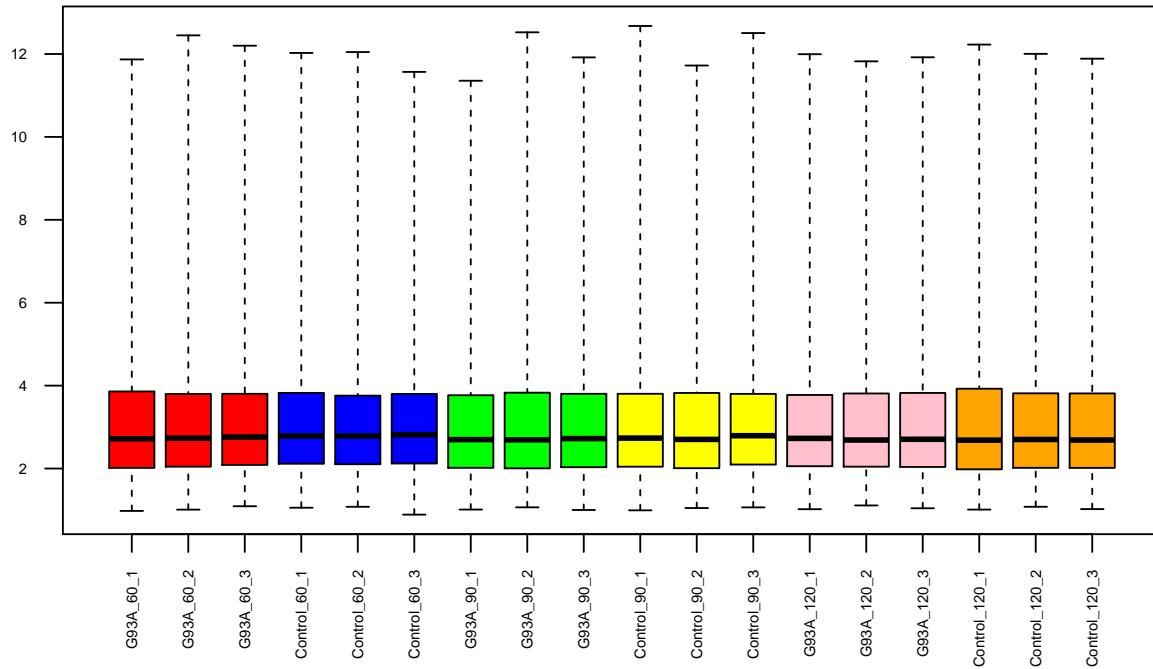


Figura 4: Distribución de intensidad de cada microarray después de la normalización.

2.5. Filtrado no específico

El filtrado no específico se utiliza para eliminar los genes que constituyen básicamente ruido, bien porque sus señales son muy bajas o bien porque apenas varían entre condiciones, por lo que no aportan nada a la selección de genes diferencialmente expresados. Además, también puede usarse para eliminar conjuntos de sondas que no tienen un identificador de gen asociado.

Para ello, se ha utilizado la función `nsFilter` del paquete `genefilter`. El umbral de variabilidad (`var.cutoff`) aplicado ha sido de 0.5. Este valor se puede interpretar como un cuantil, por lo que aproximadamente el 50 % de los genes han sido filtrados. Además se ha empleado el paquete de anotación `moe430a.db`, correspondiente al tipo de microarray del estudio, para realizar la correcta anotación del `ExpressionSet` de trabajo, antes del filtrado.

Así, consultando las filas de la matriz de expresiones (`assayData`) del `ExpressionSet` antes y después del filtrado, se ha observado que de 22,690 genes iniciales quedan 11,313.

2.6. Selección de genes expresados diferencialmente

Para la selección de genes expresados diferencialmente se ha realizado un análisis basado en modelos lineales [6] utilizando el paquete `limma`.

De esta manera ajustamos nuestros datos a un modelo lineal como el mostrado en la ecuación 1:

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\alpha} + \boldsymbol{\epsilon} \quad (1)$$

Donde

- \mathbf{Y} es el vector de valores de expresión observados de cada gen para cada una de las muestras.
- \mathbf{X} es la matriz de diseño del modelo lineal.
- $\boldsymbol{\alpha}$ es el vector de parámetros (efectos).
- $\boldsymbol{\epsilon}$ es el vector de errores.

2.6.1. Matriz de diseño

Se ha creado la matriz de diseño del modelo reformulando el diseño experimental hasta ahora descrito, pasando a considerar a *Grupo* como único factor con 6 niveles, derivados de todas las posibles combinaciones de los factores *Ratón* y *Tiempo* (Tabla 1).

De esta manera nuestra matriz de diseño \mathbf{X} tiene una dimensión de 18x6, tantas filas como muestras y tantas columnas como niveles del factor *Grupo*. Cada fila contiene un 1 en la columna del grupo al que pertenece la muestra y un 0 en las demás.

Particularizando la ecuación 1 en función del modelo planteado obtenemos la ecuación 2.

$$\begin{pmatrix} y_1 \\ y_2 \\ y_3 \\ \vdots \\ y_{16} \\ y_{17} \\ y_{18} \end{pmatrix} = \begin{pmatrix} 1 & \cdots & 0 \\ 1 & \cdots & 0 \\ 1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & 1 \\ 0 & \cdots & 1 \\ 0 & \cdots & 1 \end{pmatrix} \begin{pmatrix} \alpha_1 \\ \alpha_2 \\ \alpha_3 \\ \alpha_4 \\ \alpha_5 \\ \alpha_6 \end{pmatrix} + \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \epsilon_3 \\ \vdots \\ \epsilon_{16} \\ \epsilon_{17} \\ \epsilon_{18} \end{pmatrix} \quad (2)$$

Los efectos de cada grupo, vector $\boldsymbol{\alpha}$ en la ecuación 2, se asocian como se puede observar en la Tabla 2.

Tabla 2: Correspondencia de los efectos del modelo con cada nivel del factor *Grupo*.

α_i	Grupo
1	G93A_60
2	Control_60
3	G93A_90
4	Control_90
5	G93A_120
6	Control_120

2.6.2. Contrastes

Se han realizado 6 comparaciones entre los 6 niveles del factor *Grupos*, con las que se cubren la totalidad de posibles expresiones diferenciales de interés, cumpliendo así con el objetivo del estudio. De esta manera, cada ratón G93A se ha comparado con su Control para cada tiempo (60, 90 y 120 días). Y por otro lado, se han comparado entre sí los ratones G93A para cada tiempo.

Para realizar las distintas comparaciones se ha definido la matriz de contrastes \mathbf{C} mostrada en las ecuaciones 3 y 4, que tiene una dimensión de 6x6, tantas columnas como niveles de *Grupos* y tantas filas como comparaciones.

$$\boldsymbol{\beta} = \mathbf{C}\boldsymbol{\alpha} \quad (3)$$

$$\begin{pmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \\ \beta_4 \\ \beta_5 \\ \beta_6 \end{pmatrix} = \begin{pmatrix} 1 & -1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & -1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & -1 \\ 1 & 0 & -1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & -1 & 0 \\ 0 & 0 & 1 & 0 & -1 & 0 \end{pmatrix} \begin{pmatrix} \alpha_1 \\ \alpha_2 \\ \alpha_3 \\ \alpha_4 \\ \alpha_5 \\ \alpha_6 \end{pmatrix} \quad (4)$$

Así, se han generado a partir de la ecuación 4, los siguientes parámetros β de contraste:

$$\begin{aligned}\beta_1 &= \alpha_1 - \alpha_2 \\ \beta_2 &= \alpha_3 - \alpha_4 \\ \beta_3 &= \alpha_5 - \alpha_6 \\ \beta_4 &= \alpha_1 - \alpha_3 \\ \beta_5 &= \alpha_1 - \alpha_5 \\ \beta_6 &= \alpha_3 - \alpha_5\end{aligned}$$

2.6.3. Estimación del modelo y selección

Definida la matriz de diseño y de contrastes se ha procedido con la estimación del modelo con las funciones `lmFit`, `contrasts.fit` y `eBayes`, generando el correspondiente `MArrayLM`.

Para la obtención de listados de genes expresados diferencialmente en cada contraste se ha utilizado la función `topTable`. Para controlar el porcentaje de falsos positivos que pueden resultar del alto número de contrastes realizados simultáneamente entre genes, los p-valores se han ajustado por el método de Benjamini y Hochberg (BH) [7].

La distinción entre genes *up* o *down* regulados se ha hecho con la función `decideTests` bajo `method = "separate"`, por lo que únicamente se ha realizado un ajuste BH de p-valores entre genes, descartando el ajuste entre comparaciones.

Por último, la visualización global de la expresión diferencial en cada contraste se ha llevado a cabo mediante gráficos tipo `volcano plot`.

Los genes que se han considerado significativamente diferencialmente expresados en cada contraste son aquellos que presentan un fold change (FC) ≥ 2 (`lfc = 1`) y un p-valor < 0.05 (criterios utilizados en el estudio que se analiza [1]). Para transformar el ratio `logFC` (`log2-fold-change`) a valores de fold change, se ha empleado la función `logratio2foldchange` del paquete `gtools`.

2.7. Anotación de genes

Para establecer la correspondencia entre las sondas del microarray y el gen al que pertenecen se ha utilizado el paquete de anotaciones `moe430a.db`, específico para la plataforma utilizada modelo GeneChip Mouse Genome 430A. Así se han asociado con los siguientes identificadores:

- *Gene Symbol*: Símbolo del gen.
- *Gene name*: Nombre del gen.
- *Entrez gene Id*: Identificador del gen en la base de datos Entrez.

Nótese que se han detectado algunas sondas asociadas con más de un gen, por lo que en esos casos no es posible determinar cuál es el gen que está en realidad diferencialmente expresado.

2.8. Comparaciones múltiples

Con el fin de conocer qué genes cambian simultáneamente entre comparaciones se han utilizado diagramas de Venn (`vennDiagram`), sin diferenciar entre genes *up* o *down* regulados. Para la

construcción de los mismos ha sido necesario emplear la función `decideTests` bajo `method = "separate"` con ajuste BH de p-valores.

2.9. Análisis de significación biológica

Por último, para el análisis de significación biológica se ha llevado a cabo un análisis de enriquecimiento bajo las categorías de Gene Ontology (funciones moleculares, procesos biológicos o componentes celulares) con el paquete `clusterProfiler` utilizando la función `enrichGO`. Como objeto `OrgDb` se ha utilizado `org.Mm.eg.db`. De esta manera, se ha podido determinar si una determinada categoría GO aparece significativamente más a menudo entre los genes diferencialmente expresados en cada contraste, con un p-valor < 0.15 y sin fold change (FC) mínimo.

3. Resumen de resultados

3.1. Etapa presintomática (60 días)

El análisis del ratón SOD1 G93A a los 60 días muestra un significativo cambio en el perfil de expresión correspondiente a 55 genes, cuando se compara con su compañero de camada no transgénico (Control). Todos ellos se encuentran *up* regulados (Tabla 3). El perfil de expresión diferencial se puede observar de manera global en la Figura 5.

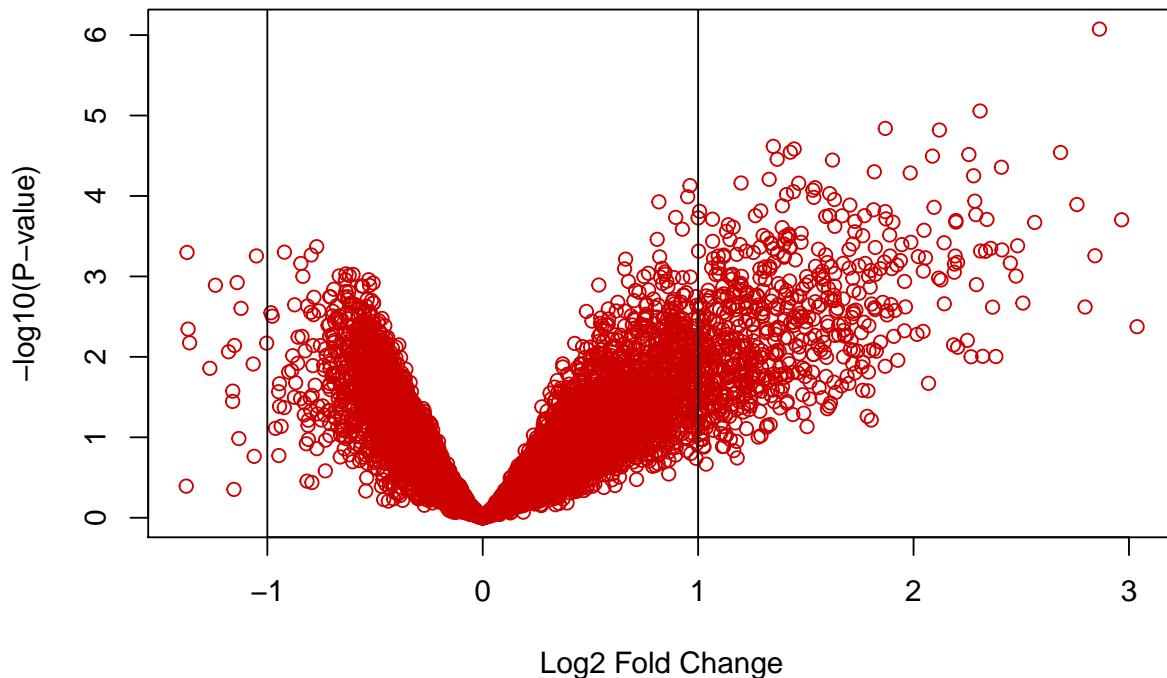


Figura 5: Volcano plot con los genes diferencialmente expresados en la etapa presintomática (60 días).

Tabla 3: Genes diferencialmente expresados en la etapa presintomática (60 días).

Identificador de sonda	Gene Symbol	Entrez Gene Id.	log2(Fold-Change)	Fold Change
1415796_at	Dazap2	23994	1.07	2.09
1415820_x_at	Nono	53610	1.33	2.51
1416300_a_at	Slc25a3	18674	1.81	3.52
1417133_at	Pmp22	18858	2.20	4.58
1418273_a_at	Rpl30	19946	2.20	4.59
1418273_a_at	Gm5481	433003	2.20	4.59
1418273_a_at	Gm6109	619883	2.20	4.59
1418273_a_at	Gm6570	625281	2.20	4.59
1418273_a_at	Gm7429	664969	2.20	4.59

Identificador de sonda	Gene Symbol	Entrez Gene Id.	log2(Fold-Change)	Fold Change
1418273_a_at	Gm12191	666899	2.20	4.59
1419246_s_at	Rab14	68365	2.05	4.14
1419663_at	Ogn	18295	2.76	6.77
1420196_s_at	Tbc1d14	100855	1.67	3.18
1421772_a_at	Cox7a2l	20463	1.82	3.53
1422241_a_at	Ndufa1	54405	1.54	2.90
1423095_s_at	Crbn	58799	2.41	5.31
1423798_a_at	Eif1	20918	1.87	3.66
1423883_at	Acsl1	14081	1.59	3.02
1424071_s_at	Ice1	218333	1.29	2.45
1424251_a_at	Hnrnpdl	50926	2.09	4.25
1426371_at	Far1	67420	1.27	2.40
1432263_a_at	Cox7a2l	20463	1.70	3.26
1433603_at	Ndufs6	407785	1.35	2.55
1433654_at	Oga	76055	1.98	3.96
1433924_at	Peg3	18616	2.29	4.89
1434435_s_at	Cox17	12856	1.62	3.08
1434853_x_at	Mkrn1	54484	1.90	3.74
1434875_a_at	Hmgm3	94353	1.44	2.72
1434888_a_at	Matr3	17184	2.97	7.81
1434935_at	Aak1	269774	1.87	3.65
1435112_a_at	Atp5h	71679	1.45	2.72
1435362_at	Foxj3	230700	1.63	3.10
1435534_a_at	Tomm20	67952	2.10	4.27
1435791_x_at	Rpl17	319195	2.26	4.78
1435791_x_at	Gm10362	100040929	2.26	4.78
1435791_x_at	Rpl17-ps10	100042880	2.26	4.78
1435791_x_at	Rpl17-ps9	100043872	2.26	4.78
1436451_a_at	Tmed2	56334	2.31	4.95
1436451_a_at	Gm10698	100862175	2.31	4.95
1436452_x_at	Tmed2	56334	1.20	2.30
1436452_x_at	Gm10698	100862175	1.20	2.30
1437327_x_at	Enoph1	67870	1.53	2.89
1437690_x_at	Csnk1d	104318	2.12	4.35
1438171_x_at	Mettl9	59052	1.73	3.31
1438318_x_at	Ngdn	68966	1.39	2.62
1439257_x_at	Rpn1	103963	2.68	6.42
1439271_x_at	Ik	24010	1.63	3.10
1448283_a_at	Uba2	50995	2.56	5.91
1448358_s_at	Snrpg	68011	1.39	2.62
1448391_at	Rab9	56382	1.47	2.76
1448504_a_at	Cbx3	12417	1.54	2.91
1448610_a_at	Sod2	20656	1.43	2.69

Identificador de sonda	Gene Symbol	Entrez Gene Id.	log2(Fold-Change)	Fold Change
1450084_s_at	Ivns1abp	117198	2.28	4.87
1452774_at	Hnrnpa3	229279	1.41	2.66
1452774_at	Gm6793	627828	1.41	2.66
1454641_at	Cggbp1	106143	1.16	2.24
1454697_at	Sec62	69276	1.61	3.05
1454793_x_at	Ddx5	13207	1.77	3.42
1455222_a_at	Ubp1	22221	1.14	2.20
1455462_at	Adcy2	210044	1.01	2.01
1455875_x_at	Tm9sf2	68059	1.87	3.65
1455897_x_at	Hmgm1	15312	2.28	4.85
1455968_x_at	Tmed2	56334	1.61	3.05
1455968_x_at	Gm10698	100862175	1.61	3.05
1456174_x_at	Ndrg1	17988	2.86	7.27
1456438_x_at	Rpn1	103963	2.34	5.07
1460656_a_at	Sft2d1	106489	1.37	2.58

3.2. Etapa sintomática (90 días)

En la etapa sintomática de la enfermedad, la comparación entre el ratón mutante (SOD1 G93A) y su compañero de camada no transgénico (Control), no presenta ningún cambio significativo en el perfil de expresión, es decir, ningún gen se ha visto significativamente alterado. En la Figura 6 se puede comprobar que los p-valores obtenidos en el análisis son $>> 0.05$.

3.3. Etapa final (120 días)

A los 120 días, el perfil de expresión del ratón SOD1 G93A en comparación con su control, muestra un significativo cambio en la expresión de 104 genes, entre los cuales 64 son *down* y 40 son *up* regulados (Tabla 4). El perfil de expresión diferencial se puede observar de manera global en la Figura 7.

Tabla 4: Genes diferencialmente expresados en la etapa final (120 días).

Identificador de sonda	Gene Symbol	Entrez Gene Id.	log2(Fold-Change)	Fold Change
1416351_at	Map2k1	26395	-1.14	-2.20
1416431_at	Tubb6	67951	1.08	2.11
1416485_at	Timm23	53600	-1.25	-2.38
1416547_at	Ndufb3	66495	-1.23	-2.34
1416908_s_at	Tsn	22099	-1.41	-2.65
1416972_at	Snu13	20826	-1.03	-2.04
1416972_at	LOC100862468	100862468	-1.03	-2.04

Identificador de sonda	Gene Symbol	Entrez Gene Id.	log2(Fold-Change)	Fold Change
1417005_at	Klc1	16593	-1.81	-3.50
1417126_a_at	Rpl22l1	68028	-1.85	-3.60
1417154_at	Slc25a14	20523	-1.13	-2.19
1417346_at	Pycard	66824	1.20	2.29
1417381_at	C1qa	12259	1.21	2.31
1417516_at	Ddit3	13198	1.15	2.22
1417533_a_at	Itgb5	16419	1.05	2.07
1417869_s_at	Ctsz	64138	1.58	3.00
1418021_at	C4b	12268	1.64	3.12
1418021_at	C4a	625018	1.64	3.12
1418031_at	Myo9b	17925	1.30	2.47
1418133_at	Bcl3	12051	1.50	2.82
1418201_at	Plekhg2	101497	1.05	2.07
1418364_a_at	Ftl1	14325	1.04	2.05
1418364_a_at	LOC100862446	100862446	1.04	2.05
1418457_at	Cxcl14	57266	1.52	2.87
1418744_s_at	Tesc	57816	-1.70	-3.26
1418892_at	Rhoj	80837	1.88	3.69
1419100_at	Serpina3n	20716	1.59	3.01
1419666_x_at	Nupr1	56312	1.21	2.32
1419873_s_at	Csf1r	12978	1.53	2.88
1420361_at	Slc11a1	18173	1.43	2.69
1420394_s_at	Lilrb4a	14728	1.07	2.10
1420394_s_at	Lilr4b	14727	1.07	2.10
1420402_at	Atp2b2	11941	1.24	2.37
1420489_at	Mrps14	64659	-1.78	-3.43
1420649_at	Zfhx3	11906	1.11	2.16
1421374_a_at	Fxyd1	56188	1.28	2.44
1421904_at	Tgs1	116940	-1.39	-2.61
1422557_s_at	Mt1	17748	1.39	2.62
1422831_at	Fbn2	14119	1.14	2.21
1422861_s_at	Pdlim5	56376	1.02	2.02
1422903_at	Ly86	17084	1.52	2.87
1423486_at	Cript	56724	-1.84	-3.57
1423547_at	Lyz2	17105	2.19	4.55
1423608_at	Itm2a	16431	-1.48	-2.79
1423625_a_at	Dnajc19	67713	-1.56	-2.95
1423625_a_at	Dnajc19-ps	100503724	-1.56	-2.95
1423804_a_at	Idi1	319554	-2.41	-5.30
1423954_at	C3	12266	1.22	2.33
1423995_at	Kif1b	16561	1.55	2.94
1424470_a_at	Rapgef3	223864	1.43	2.69
1424714_at	Aldoc	11676	1.38	2.60

Identificador de sonda	Gene Symbol	Entrez Gene Id.	log2(Fold-Change)	Fold Change
1425052_at	Isoc1	66307	-1.12	-2.17
1426432_a_at	Slc4a4	54403	1.80	3.49
1426509_s_at	Gfap	14580	2.38	5.21
1426574_a_at	Add3	27360	1.06	2.08
1426808_at	Lgals3	16854	4.12	17.40
1426917_s_at	Scrn3	74616	-1.42	-2.67
1427044_a_at	Amph	218038	-1.30	-2.46
1427076_at	Mpeg1	17476	1.02	2.03
1427262_at	Xist	213742	1.69	3.23
1427351_s_at	Ighm	16019	-1.54	-2.91
1427699_a_at	Ptpn11	19247	1.25	2.37
1427718_a_at	Mdm2	17246	1.11	2.16
1427914_a_at	Eloc	67923	-1.09	-2.12
1427988_s_at	Safb2	224902	1.04	2.06
1428502_at	Actr6	67019	-1.28	-2.43
1430979_a_at	Prdx2	21672	-1.06	-2.08
1433474_at	Edil3	13612	-1.43	-2.69
1434366_x_at	C1qb	12260	2.50	5.66
1434976_x_at	Eif4ebp1	13685	1.62	3.08
1435137_s_at	1200015M12Rik	71739	1.30	2.47
1435137_s_at	A130040M12Rik	319269	1.30	2.47
1435454_a_at	Helz2	229003	1.28	2.43
1435906_x_at	Gbp2	14469	1.15	2.22
1436451_a_at	Tmed2	56334	-1.52	-2.86
1436451_a_at	Gm10698	100862175	-1.52	-2.86
1436506_a_at	Snhg6	73824	-1.47	-2.77
1436905_x_at	Laptm5	16792	1.72	3.30
1437012_x_at	Rapgef3	223864	1.15	2.22
1437945_x_at	Nap1l1	53605	-1.24	-2.37
1438250_s_at	Taf9	108143	-1.35	-2.55
1438250_s_at	Taf9-ps	545600	-1.35	-2.55
1439079_a_at	Erbin	59079	1.70	3.25
1439426_x_at	Lyz2	17105	2.64	6.23
1439426_x_at	Lyz1	17110	2.64	6.23
1448179_at	Atp5md	66477	-1.24	-2.36
1448682_at	Dynll1	56455	-1.42	-2.67
1448793_a_at	Sdc4	20971	1.68	3.20
1449401_at	C1qc	12262	1.45	2.73
1449648_s_at	Polr1c	20016	-1.34	-2.53
1450688_at	Rgl2	19732	1.17	2.26
1450792_at	Tyrobp	22177	2.09	4.25
1450871_a_at	Bcat1	12035	-1.32	-2.50
1451071_a_at	Atp1a1	11928	-2.58	-5.97

Identificador de sonda	Gene Symbol	Entrez Gene Id.	log2(Fold-Change)	Fold Change
1451122_at	Idi1	319554	-1.14	-2.21
1451132_at	Pbxip1	229534	1.25	2.37
1451205_at	Psmb4	19172	-1.72	-3.30
1451294_s_at	Snrpe	20643	-1.09	-2.13
1451500_at	Ushbp1	234395	1.01	2.01
1451640_a_at	Rsrc2	208606	1.38	2.60
1452005_at	Dlat	235339	-1.36	-2.57
1452159_at	Apmap	71881	-1.10	-2.14
1452882_at	Pgrmc2	70804	-1.13	-2.19
1454690_at	Ikbkg	16151	-1.03	-2.04
1455899_x_at	Socs3	12702	2.96	7.79
1455956_x_at	Ccnd2	12444	1.99	3.97
1456014_s_at	Trpt1	107328	1.60	3.02
1456014_s_at	Fermt3	108101	1.60	3.02
1456055_x_at	Pold1	18971	1.14	2.20
1456133_x_at	Itgb5	16419	1.38	2.60
1456244_x_at	Glrx3	30926	-1.66	-3.17
1456244_x_at	Gm12669	620016	-1.66	-3.17
1456319_at	NA	NA	2.70	6.49
1460218_at	Cd52	23833	1.04	2.06
1460220_a_at	Csf1	12977	1.16	2.23
1460349_at	Prr14	233895	1.06	2.09

3.4. Comparación del perfil de expresión génica entre las tres etapas

En la comparación del perfil de expresión del ratón transgénico SOD1 G93A entre los 60 y 120 días, se ha observado un cambio en la expresión de 623 genes (Figura 8), de los cuales 419 son *down* y 204 son *up* regulados.

Entre la etapa sintomática (90 días) y final (120 días), cambia el perfil de expresión en 1382 genes en el ratón SOD1 G93A. De éstos, 765 son *down* y 617 son *up* regulados (Figura 9).

Sin embargo, la comparación de SOD1 G93A entre los 60 y 90 días, no arroja ningún cambio significativo en la expresión de los genes (Figura 10).

Por otro lado, se han comparado mediante diagrama de Venn las tres etapas de la enfermedad en su comparación con el ratón control (Figura 11). Además, se han comparado a su vez, la expresión diferencial de SOD1 G93A entre las tres etapas de la enfermedad (Figura 12).

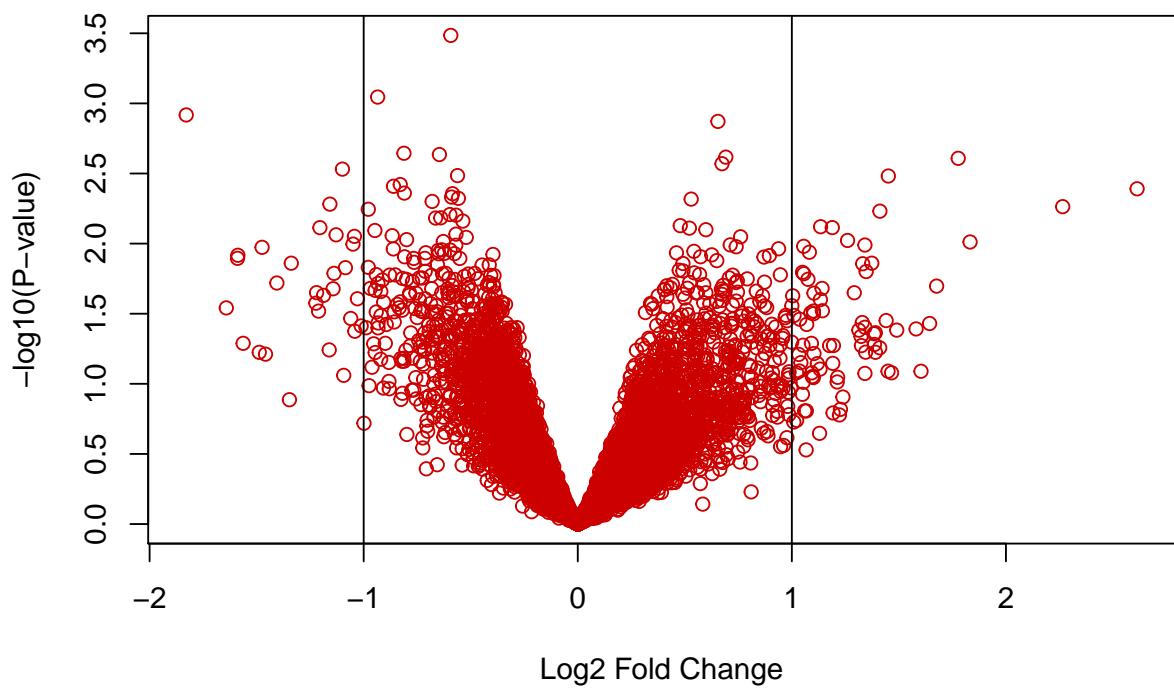


Figura 6: Volcano plot con los genes diferencialmente expresados en la etapa sintomática (90 días).

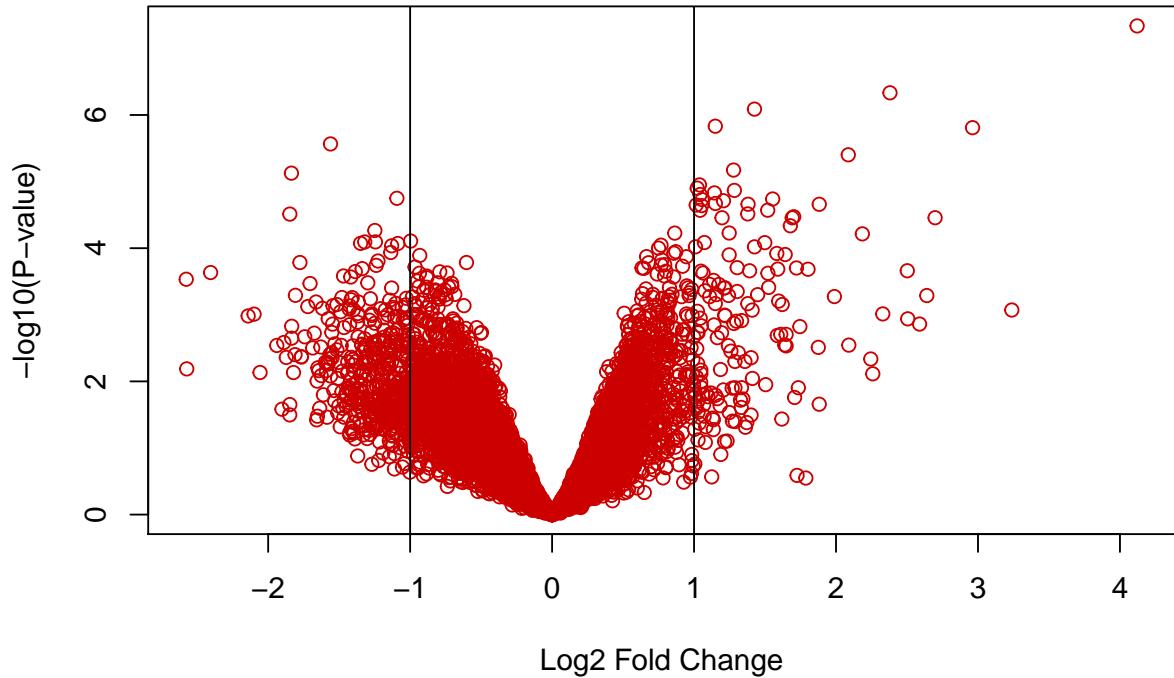


Figura 7: Volcano plot con los genes diferencialmente expresados en la etapa final (120 días).

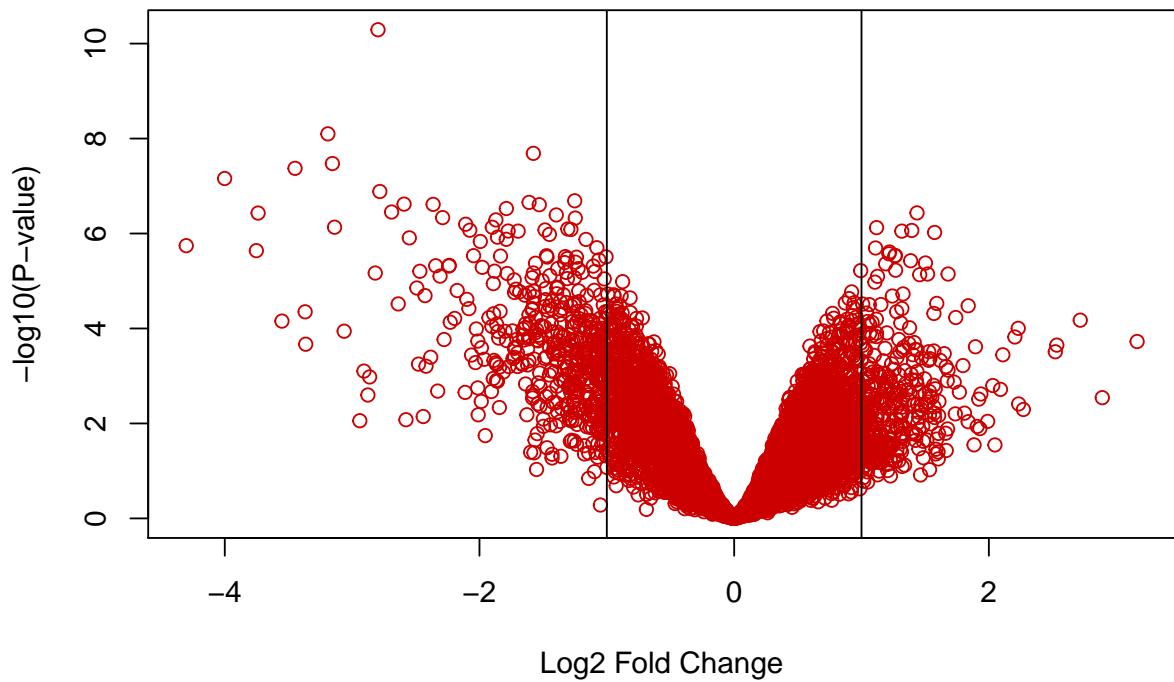


Figura 8: Volcano plot con los genes diferencialmente expresados entre las etapas presintomática (60 días) y final (120 días).

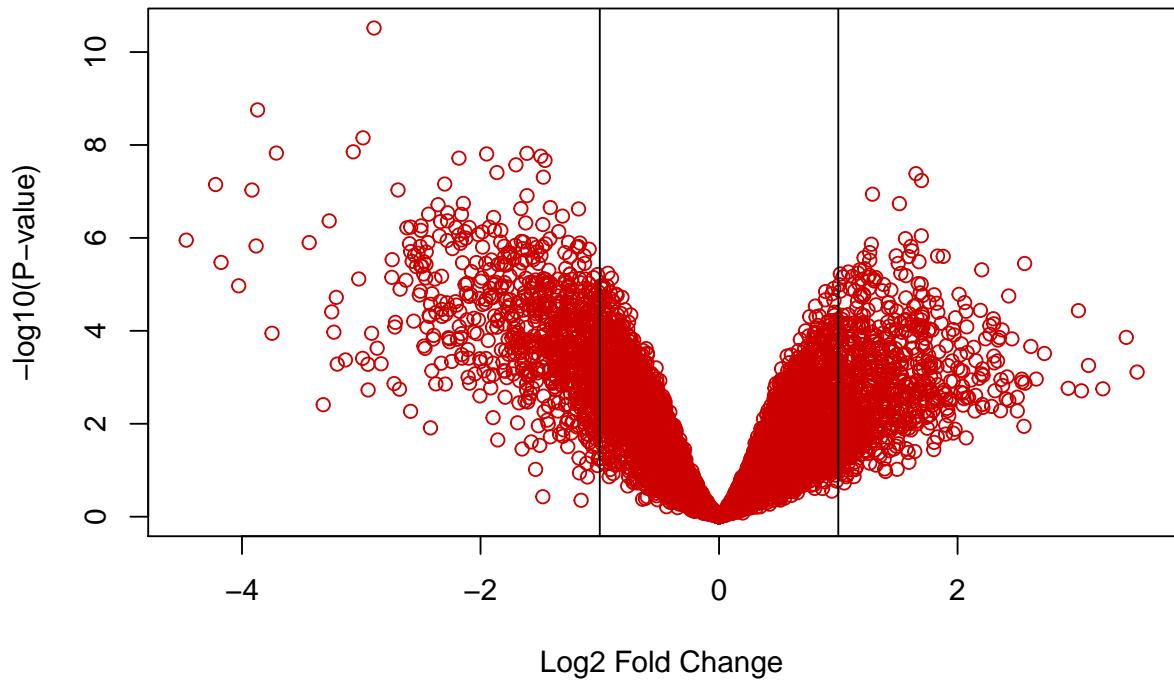


Figura 9: Volcano plot con los genes diferencialmente expresados entre las etapas sintomática (90 días) y final (120 días).

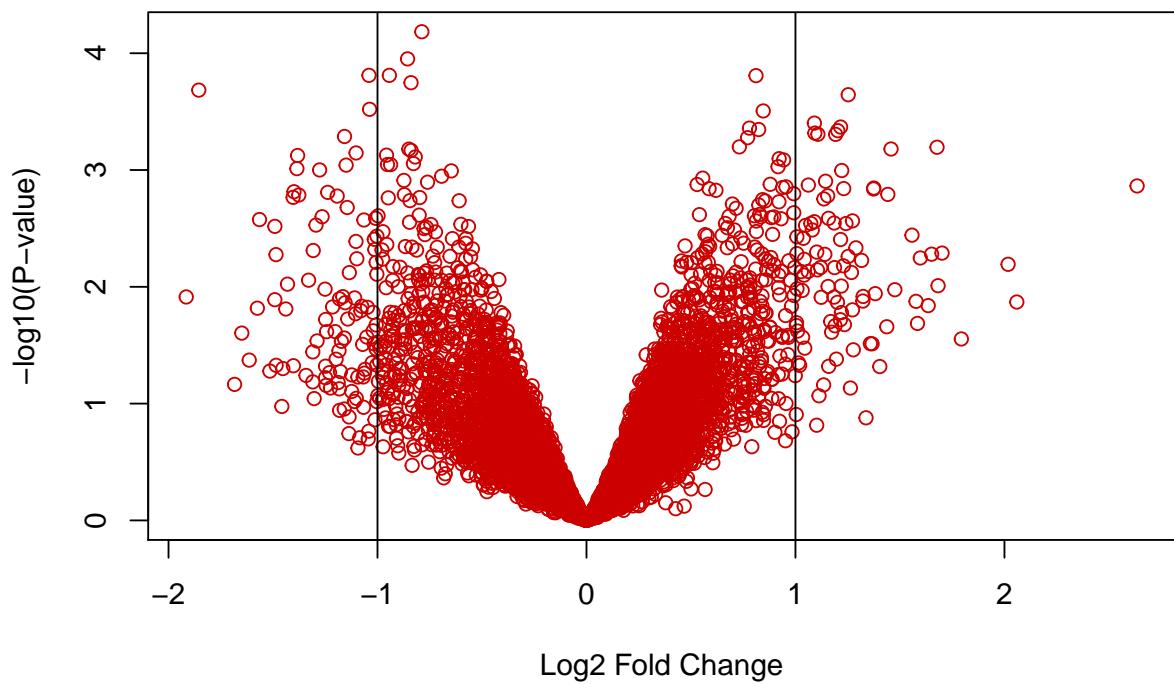


Figura 10: Volcano plot con los genes diferencialmente expresados entre las etapas presintomática (60 días) y sintomática (90 días)

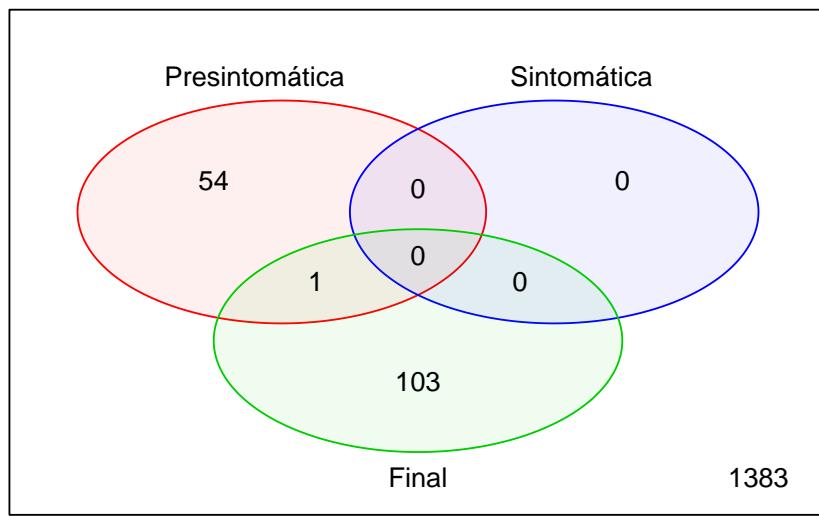


Figura 11: Diagrama de Venn con los genes diferencialmente expresados entre las etapas presintomática, sintomática y final (60, 90 y 120 días).

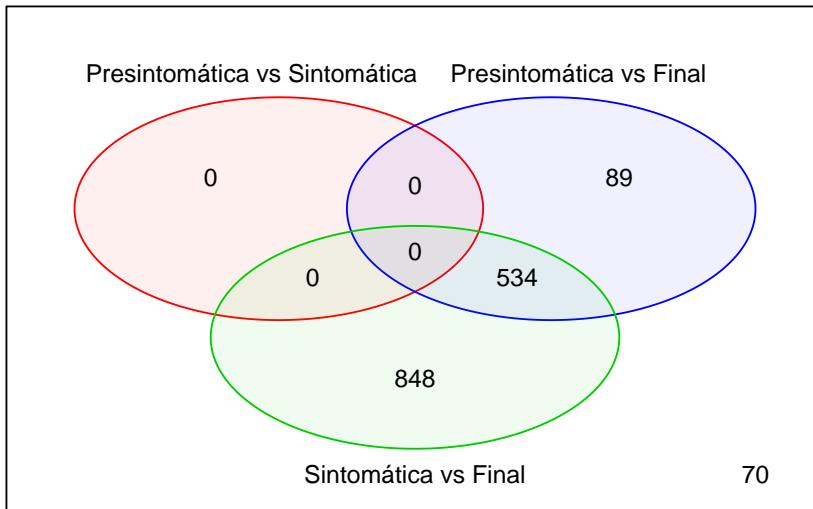


Figura 12: Diagrama de Venn con los genes diferencialmente expresados entre las comparaciones entre etapas.

3.5. Significación biológica

Tras el análisis de enriquecimiento realizado, bajo las categorías de GO, en cada una de las comparaciones significativas realizadas, se muestran las primeras cinco categorías ordenadas por su significancia en las Tablas 5, 6, 7 y 8. Para observar el resto de entradas en cada tabla, acúdase a los archivos CSV del repositorio.

Tabla 5: Primeras cinco categorías de GO más significativas resultantes del análisis de enriquecimiento en la etapa presintomática (60 días).

GO Id.	Descripción	Gene ratio
GO:0031647	regulation of protein stability	37/878
GO:2001233	regulation of apoptotic signaling pathway	46/878
GO:2001234	negative regulation of apoptotic signaling pathway	31/878
GO:0050821	protein stabilization	25/878
GO:2001242	regulation of intrinsic apoptotic signaling pathway	24/878

Tabla 6: Primeras cinco categorías de GO más significativas resultantes del análisis de enriquecimiento en la etapa final (120 días).

GO Id.	Descripción	Gene ratio
GO:0007005	mitochondrion organization	62/1060
GO:0072594	establishment of protein localization to organelle	50/1060
GO:0006914	autophagy	47/1060
GO:0061919	process utilizing autophagic mechanism	47/1060
GO:0006091	generation of precursor metabolites and energy	46/1060

Tabla 7: Primeras cinco categorías de GO más significativas resultantes del análisis de enriquecimiento entre las etapas presintomática (60 días) y final (120 días).

GO Id.	Descripción	Gene ratio
GO:0007005	mitochondrion organization	141/3241
GO:0031346	positive regulation of cell projection organization	137/3241
GO:1901214	regulation of neuron death	112/3241
GO:0001558	regulation of cell growth	124/3241
GO:0045666	positive regulation of neuron differentiation	130/3241

Tabla 8: Primeras cinco categorías de GO más significativas resultantes del análisis de enriquecimiento entre las etapas sintomática (90 días) y final (120 días).

GO Id.	Descripción	Gene ratio
GO:0007005	mitochondrion organization	191/4437
GO:0031346	positive regulation of cell projection organization	180/4437
GO:0045666	positive regulation of neuron differentiation	177/4437
GO:0019693	ribose phosphate metabolic process	173/4437
GO:0051656	establishment of organelle localization	178/4437

4. Repositorio GitHub

En el siguiente enlace del repositorio del proyecto en GitHub, se encuentran disponibles los resultados y el código utilizado:

https://github.com/edsantor/PEC_1

Referencias

- [1] Ferraiuolo L, Heath PR, Holden H, Kasher P, Kirby J, Shaw PJ. Microarray analysis of the cellular pathways involved in the adaptation to and progression of motor neuron injury in the SOD1 G93A mouse model of familial ALS. *The Journal of Neuroscience : The Official Journal of the Society for Neuroscience* 2007;27:9201–19. <https://doi.org/10.1523/JNEUROSCI.1470-07.2007>.
- [2] Turner MR, Talbot K. Motor neurone disease is a clinical diagnosis. *Practical Neurology* 2012;12:396–7. <https://doi.org/10.1136/practneurol-2012-000374>.
- [3] Rosen DR. Mutations in Cu/Zn superoxide dismutase gene are associated with familial amyotrophic lateral sclerosis. *Nature* 1993;364:362. <https://doi.org/10.1038/364362c0>.
- [4] Yoshihara T, Ishigaki S, Yamamoto M, Liang Y, Niwa J-i, Takeuchi H, et al. Differential expression of inflammation- and apoptosis-related genes in spinal cords of a mutant SOD1 transgenic mouse model of familial amyotrophic lateral sclerosis. *Journal of Neurochemistry* 2002;80:158–67. <https://doi.org/10.1046/j.0022-3042.2001.00683.x>.
- [5] Irizarry RA, Hobbs B, Collin F, Beazer-Barclay YD, Antonellis KJ, Scherf U, et al. Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics (Oxford, England)* 2003;4:249–64. <https://doi.org/10.1093/biostatistics/4.2.249>.
- [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] Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B (Methodological)* 1995;57:289–300.