PEC1: ANÁLISIS DE DATOS -ÓMICOS

Sonia Doblado Martín

26 de abril, 2020

Índice

1. Resumen

El estudio escogido para el reanálisis es "Expression data for C. elegans blunt force trauma". El objetivo del estudio era averiguar si existen rutas diferenciales en individuos de edad diferente, como respuesta de defensa al aplicarles un estímulo de estrés traumático. Dado que los resultados aún no han sido publicados, se han guardado los resultados de las comparaciones entre todos los grupos de arrays, pero sólo se ha realizado un *"Gene Enrichment Analysis"* a las comparaciones "sin/con lesión" de individuos del día 1, y "sin/con lesión" de individuos del día dos (esto es, UvsP1 y UvsP4). Se han seleccionado estas comparaciones por ser las que nos darán la información sobre las diferentes genes expresados para un mismo estímulo en los diferentes grupos de edad.

Los resultados de este informe coinciden con los del experimento objeto de estudio. Indican que provocar estrés en *C. elegans* vía traumatismo activa rutas de transcripción distintivas y dependientes de edad.

En caso de que se desee reproducir el análisis aquí realizado, el código R empleado se encuentra en el siguiente repositorio **Github**.

2. Introducción

Caenorhabditis elegans es una especie de nematodo comúnmente utilizada como organismo modelo en estudios génicos. En el experimento se estudia la respuesta del organismo frente a estrés mecánico debido a traumatismos, y su relación con la edad del mismo. Es decir, se investigó si un traumatismo induce un programa transcripcional específico y dependiente de la edad. Para ello, se sometió a pruebas muestras de individuos de 1 día y 4 días de edad, a los cuales se les suministró el tratamiento (lesión, 4 lesiones o control). Al cabo de una hora fueron recolectados para procesar su ARN.

Este estudio muestra los resultados de un "Gene Enrichment Analysis", para ser utilizados en la investigación sobre mecanismos de respuesta a estrés diferenciales dependientes de edad en *C. elegans*.

3. Métodos

Para el estudio se utilizó el chip de Affymetrix para la especie *C.elegans* (*high density oligonucleotide array*)(material de soporte disponible online), dando como resultado el dataset *GSE148325* . Dicho dataset ha sido analizado para este ejercicio mediante el software R y Bioconductor(https://www.bioconductor.org/). El código utilizado para este informe se encuentra en el siguiente repositorio [GitHub] (https://github.com/SoniaD89/ADO1).

3.1. Tratamiento de los datos

Para comenzar el nuevo análisis, se obtuvo el dataset *GSE148325* de la base de datos GEO, se leyeron los archivos .CEL del mismo y se procedió a la creación de un archivo llamado *targets* para contener en él todas las variables a analizar. A este archivo, se le añadió la variable `grupo` para facilitar el análisis, donde 1U señala los individuos de 1 día de edad sin lesión provocada (*uninjured*, U), 1P para individuos de la misma edad pero lesionados (*paralyzed*, P), 4U para individuos de 4 días sin lesión, 4P si se lesionaron y 4U4P si esa lesión fue repetida cuatro veces.

La calidad de estos datos sin tratar ha sido analizada mediante la función `arrayQualityMetrics`. Una vez comprobado que los datos son viables, se realizó un análisis de componentes principales, cuyo resultado puede verse en la Figura 1 y se visualizó la variabilidad de intensidad de estos arrays mediante un boxplot (Figura 2).

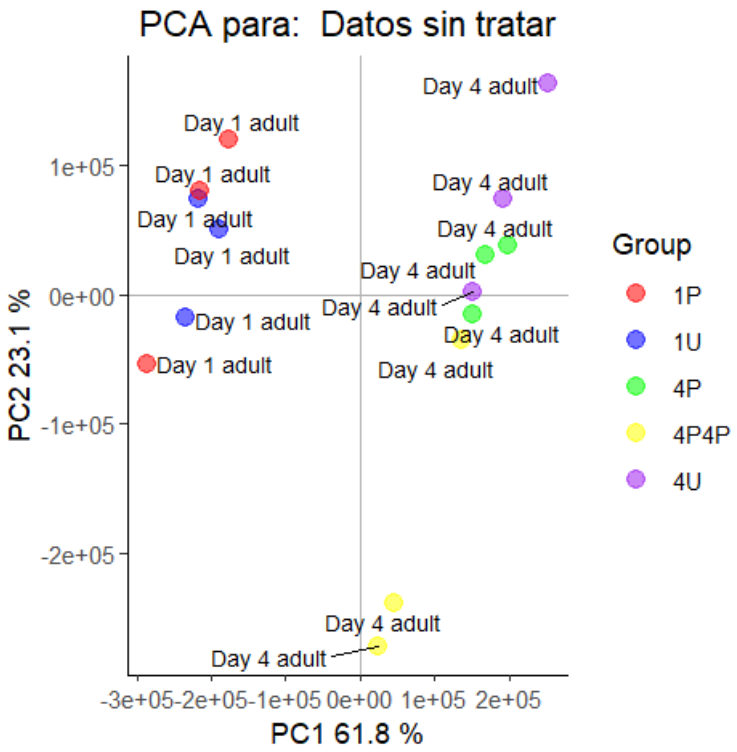

Figura 1: Principales Componentes en datos sin tratar

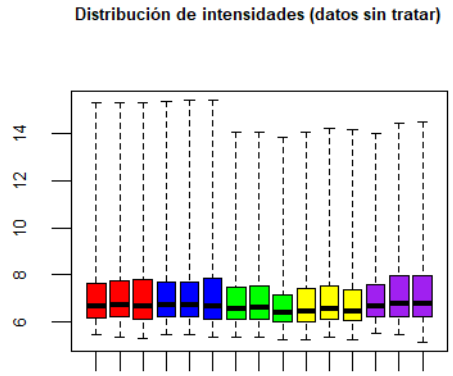

Figura 2: Boxplot con datos sin tratar

En el análisis de los componentes principales (PCA) se observa que el primer componente del PCA es responsable del 61.8 % de la variabilidad de las muestras. Como puede verse en el gráfico de la figura 1, esta variabilidad podría atribuirse ya a la edad, al estar los individuos de edades diferentes claramente separados

(adultos de 4 días a la derecha, adultos de un día a la izquierda).

En la Figura 2 se comprueba que la distribución de intensidades de los arrays es similar para todas las muestras.

Una vez realizados estos análisis, los datos se normalizaron y se repitieron tanto el análisis de calidad, como el análisis de los componentes principales (Figura 3) y la visualización de la variabilidad (Figura 4).

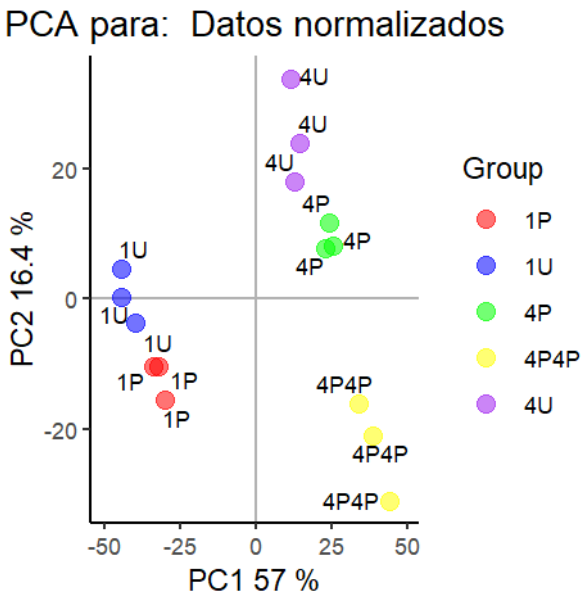

Figura 3: Visualización de los Componentes Principales en datos normalizados

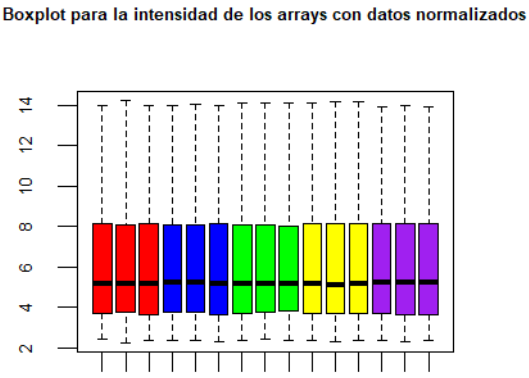

Figura 4: Distribución de las intensidades en datos normalizados

En la figura 3 se observa que la variabilidad debida al primer componente del análisis se ha reducido, pero

aún se distingue la separación de los diferentes grupos de grupos de edad. En la figura 4 se reflejan las variabilidades de intensidades normalizadas, y se comprueba que la normalización se ha realizado correctamente al ser éstas muy similares para todas las muestras.

Con esos datos normalizados, se realizó un test pvca (*Principal Variation Component Analysis*) para eliminar efectos debidos a la manipulación de los arrays en el laboratorio, y representar en un gráfico hasta qué punto cada una de las variables (edad y tratamiento) son responsables de la variabilidad de los arrays (Figura 5). Se observa que un 71,9 % de la variabilidad entre muestras puede atribuirse a la variable edad `age.ch1`, mientras que sólo un 0,9 % se debe a la interacción entre edad y grupo.

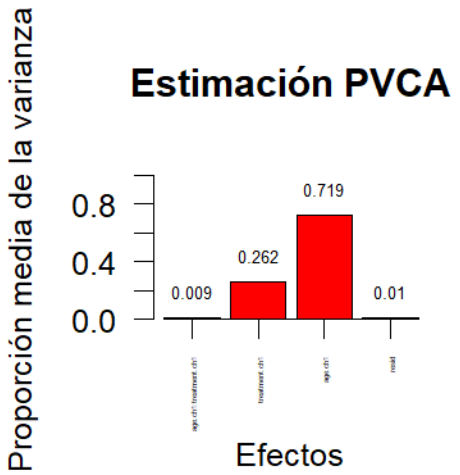

Figura 5: Importancia relativa de los diferentes factores (edad temperatura e interacción entre ellos) que afectan a la expresión génica

El siguiente paso fue realizar un gráfico mostranto la variabilidad de los genes en las muestras (Figura 6) ordenados de menor a mayor variabilidad. Los genes con una variabilidad atribuible al azar y no a expresiones diferenciales entre grupos fueron filtrados mediante la función `nsFilter` del paquete de Bioconductor `genefilter` y la anotación correspondiente a nuestro caso `celegans.db`. Con los 4170 genes que quedaron tras ese filtrado, se realizó una matriz de diseño para posteriormente realizar una matriz de contrastes. Esta matriz de contrastes muestra todas las comparaciones que se quieren realizar, en nuestro caso 1Uvs1P, 4Uvs4P y 4Pvs4P4P, además de la interacción entre el tratamiento y la edad de los individuos muestreados. Con estas comparaciones se puede observar no sólo si hay diferencias en las expresiones de dentro de los grupos muestrales de 1 día y 4 días según tratamiento y poder comprobar la interacción de la edad en la respuesta fisiológica de estudio, sino si hay diferencias entre muestras de individuos de la misma edad lesionados y sin lesionar. Se han añadido las comparaciones entre individuos con un mismo tratamiento pero diferente edad (1Uvs4U y 1Pvs4P) para un posible estudio de diferencias debidas únicamente a la edad. Las comparaciones que se usarán para el estudio de la significación biológica en este informe son 1Uvs1P, 4Uvs4P (desde ahora, UvsP1 y UvsP4 respectivamente) y la interacción entre ambas INT, ya que se entiende que el objetivo es saber la diferencia de las expresiones provocadas por la reacción a un traumatismo entre una edad y otra. Los resultados correspondientes al resto de comparaciones quedarán almacenados en caso de fuesen necesarios para un estudio posterior o futuras publicaciones.

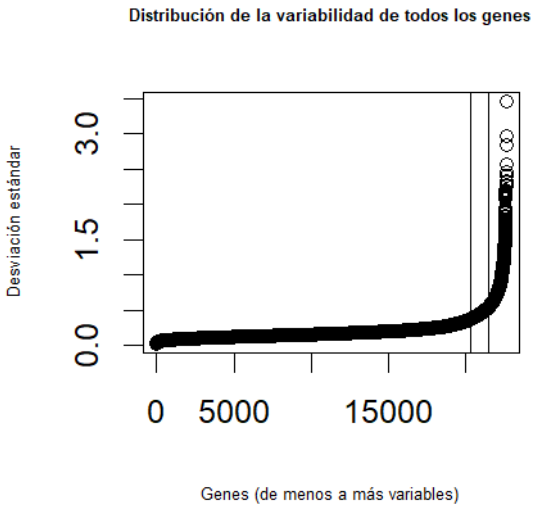

Figura 6: Variabilidad de genes. Las líneas verticales representan los percentiles 90 y 95. Los genes están ordenados de menos a más variables

Una vez estas matrices han quedado definidas, se procede a estimar el modelo y estimar los contrastes, así como a realizar los tests de significancia adecuados. En este caso, se ha utilizado el paquete `limma`, que utiliza análisis de Bayes empírico para combinar una estimación de la variabilidad basada en la matriz completa con estimaciones individuales, basadas éstas en valores individuales, dando lugar así a estimaciones de error mejoradas (Smyth 2004). Este análisis provee de estadísticos comunes como *Fold Change* o *p-value*. Para intentar mitigar la aparición de falsos positivos, el p-valor se ajustó utilizando el método de Benjamini y Hochberg.

3.2. Expresión diferencial

Para obtener una lista de los genes expresados diferencialmente, se ha utilizado la función `topTable` del ya mencionado paquete `limma`, para cada una de las comparaciones. Las listas ofrecen los genes ordenados de menor a mayor p-valor, lo que puede considerarse como de mayor a menor expresión diferencial. Dicha lista incluye también otros estadísticos como `logFC`(media de la diferencia entre grupos), `AveExpr` (expresión media de los genes en la comparación), `t`(estadístico tipo t-test para la comparación), el p-valor ajustado y el estadístico `B` (probabilidad logaritmica de que el gen esté diferencialmente expresado contra que no lo esté). Dado que en esta lista los genes vienen nombrados según el criterio del fabricante (en este caso, Affymetrix), el siguiente paso será el de la anotación de los genes. Se añadieron a los datos los identificadores de Entrez Gene y Gene Symbol. El paquete de anotación utilizado para ello ha sido `celegans.db`.

Una vez tenemos los genes anotados, procedemos a la visualización de los resultados de la expresión diferencial obtenidos. En este informe se muestran varios ejemplos de visualización. Para mostrar una visión general de la expresión diferencial se realiza un `volcanoplot` para la comparación UvsP1 y otro para la comparación UvsP4 (Figura **??**). Para ver cuántos genes han sido seleccionados en una o varias de las comparaciones, se obtiene una tabla mediante la función `decideTest`y se realiza un diagrama de Venn mediante la función `VennDiagram`(Figura 9). Para una visualización más detallada, se elaboraron dos mapas de calor(con y sin clustering) mediante la función `heatmap.2` (Figura 10 (Figura 11 del paquete `gplots`, en el que cada una de las columnas representa a un grupo muestral. Para elaborar este mapa se utilizaron los genes seleccionados en los pasos previos.

3.3. Importancia biológica

Por último, para ayudar a interpretar la importancia biológica de los resultados del experimento, se realizó un test `enrichgo` (`Gene  Enrichment Analysis`) (Guangchuang 2018) mediante el paquete `clusterProfiler` (Guangchuang 2020). Para representar los resultados de este test, se han utilizado diagramas de barras mediante la función `barplot`, un diagrama en forma de red especificando qué genes son los que actúan mediante la función `cnetplot` y un gráfico `emapplot`, que pueden ayudar a la interpretación de los resultados aquí obtenidos. Se han calculado los resultados para las ontologías `BP`(Biological Process), `CC`(Cellular Component) y `MF` (Molecular Function), para tener una visión más general de la importancia biológica de los resultados, pero aquí sólo se mostrarán los resultados correspondientes a `BP`al considerarse la ontología de interés. El criterio de selección de genes diferencialmente expresados ha sido un p-valor ajustado menor de 0.05, salvo para el caso de `MF`, en el que se ha elevado ese valor a 0.15 para poder realizar el análisis y obtener resultados en todas las comparaciones. Se seleccionan todos los genes que tienen al menos una anotación en la base de datos GO.

Las figuras resultantes de los análisis han sido guardadas en archivos .png y .pdf para su uso posterior en otros estudios o publicaciones.

4. Resultados

Los genes seleccionados para cada comparación según la función `decideTests` son:

```
       UvsP4 UvsP1 PvsPP4 U1vsU4 P1vsP4 INTedad
Down     171     4     18    209    829     462
NotSig  3982  4044   3710   3636   2963    3682
Up        17   122    442    325    378      26
```

En las Figura 7 y 8 se observan resaltados los 4 genes principales para cada una de las comparaciones UvsP4 y UvsP1. Se puede ver lo reflejado en la tabla anterior: en la respuesta de los individuos de 4 días intervienen muchos más genes "*down-regulated*" que en la respuesta de los individuos de 1 día de edad. En esta última comparación ocurre al contrario, hay una mayor cantidad de genes "*up-regulated*". Además, los genes resaltados son distintos entre ambas comparaciones.

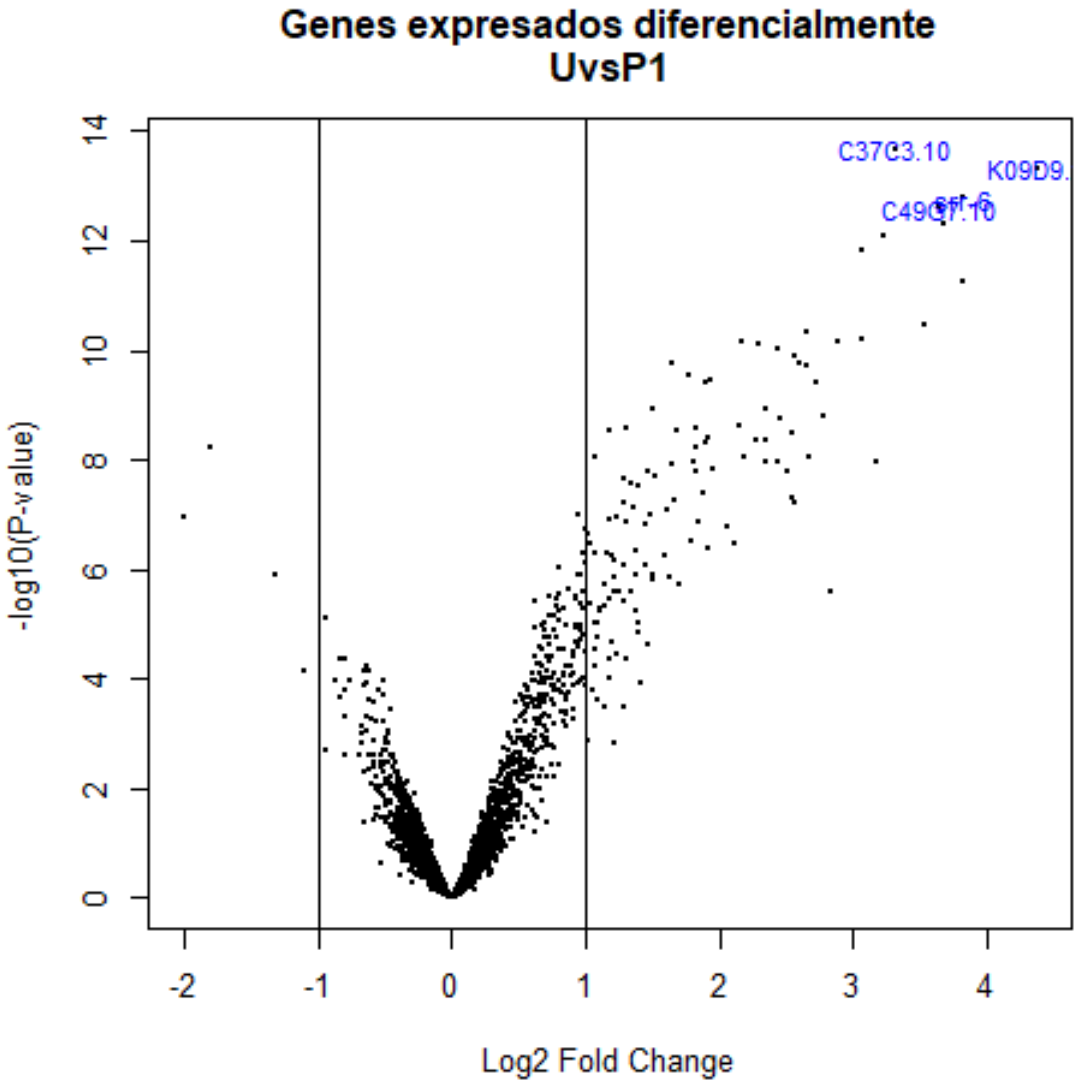

Figura 7: Volcanoplot para UvsP1. Resaltados aparecen los 4 primeros genes seleccionados

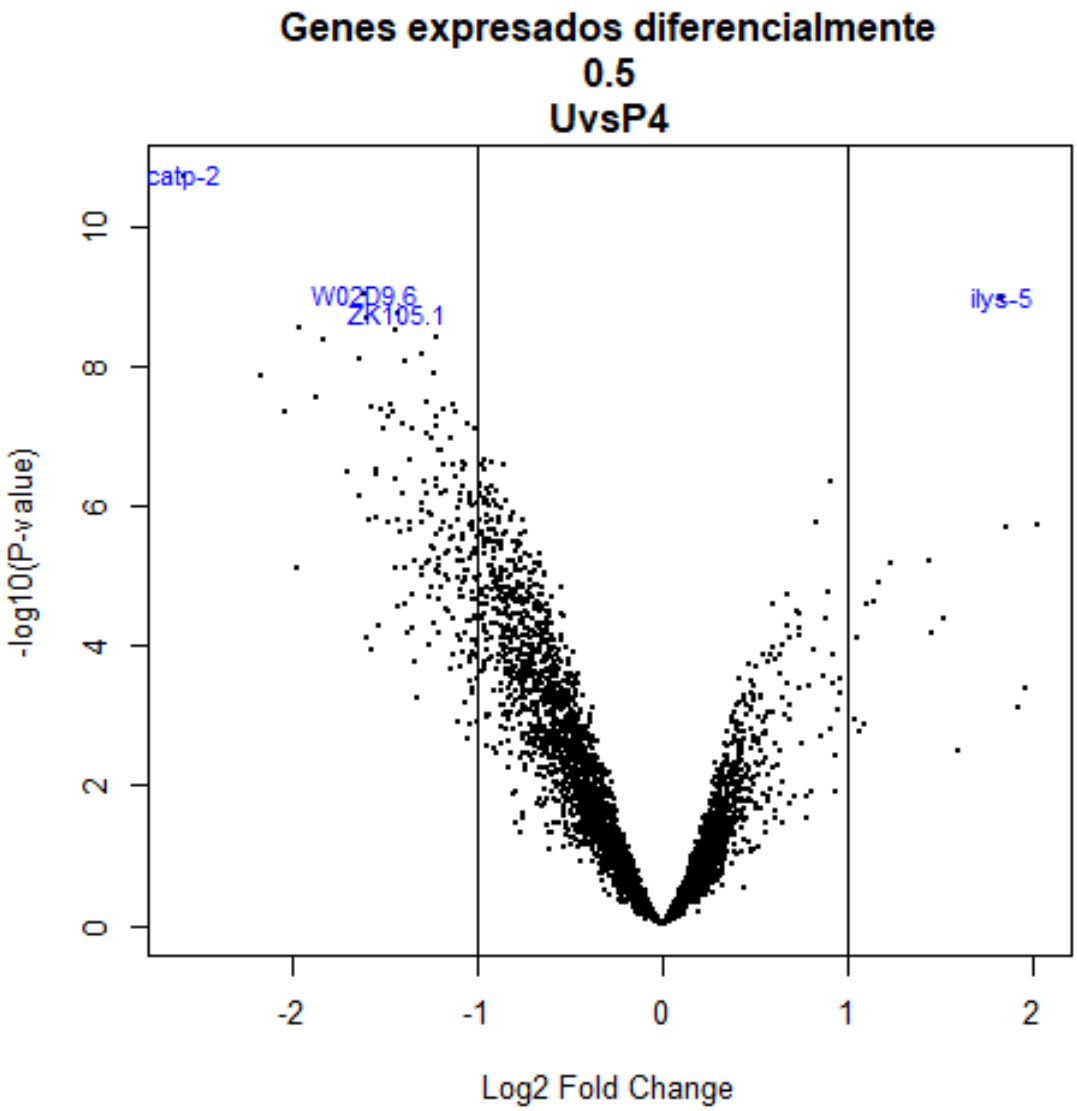

Figura 8: Volcanoplots para UvsP4. Resaltados aparecen los 4 primeros genes seleccionados

En la Figura 9 Se observa que de la lista de seleccionados, sólo 22 genes son compartidos entre las comparaciones UvsP1 y UvsP4. Un número mucho más elevado es compartido con la comparación PvsPP4, lo que podría indicarnos que una vez ejercido el primer traumatismo, el individuo puede acostumbrarse a la señal de estrés reduciendo las diferencias entre los individuos de edad 1 y edad 4.

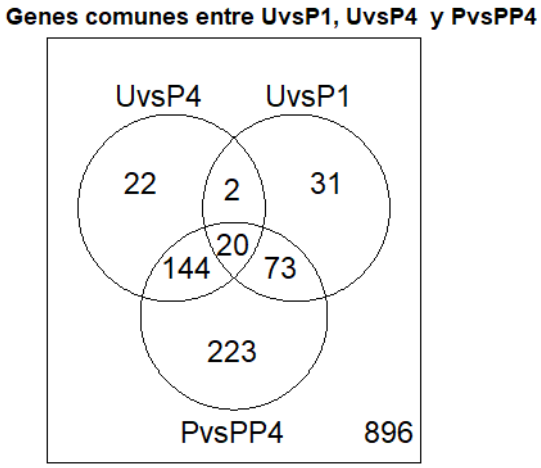

Figura 9: Diagrama de Venn que muestra los genes compartidos entre las comparaciones UvsP1

En los gráficos resultado de `heatmap`, se observan columnas claramente diferentes para cada uno de los grupos muestrales, tanto en el gráfico sin agrupamiento como en el gráfico con él.

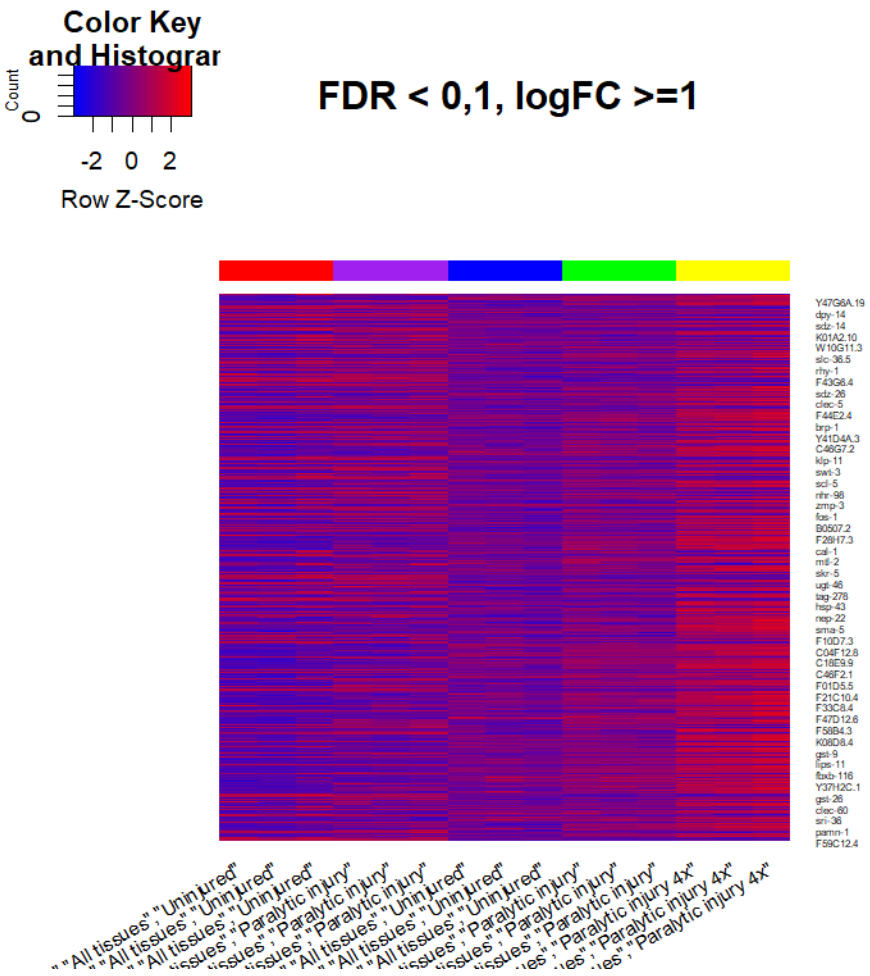

Figura 10: Heatmap para genes expresados diferencialmente sin agrupamiento, para FDR <0,1 y logFC >=1

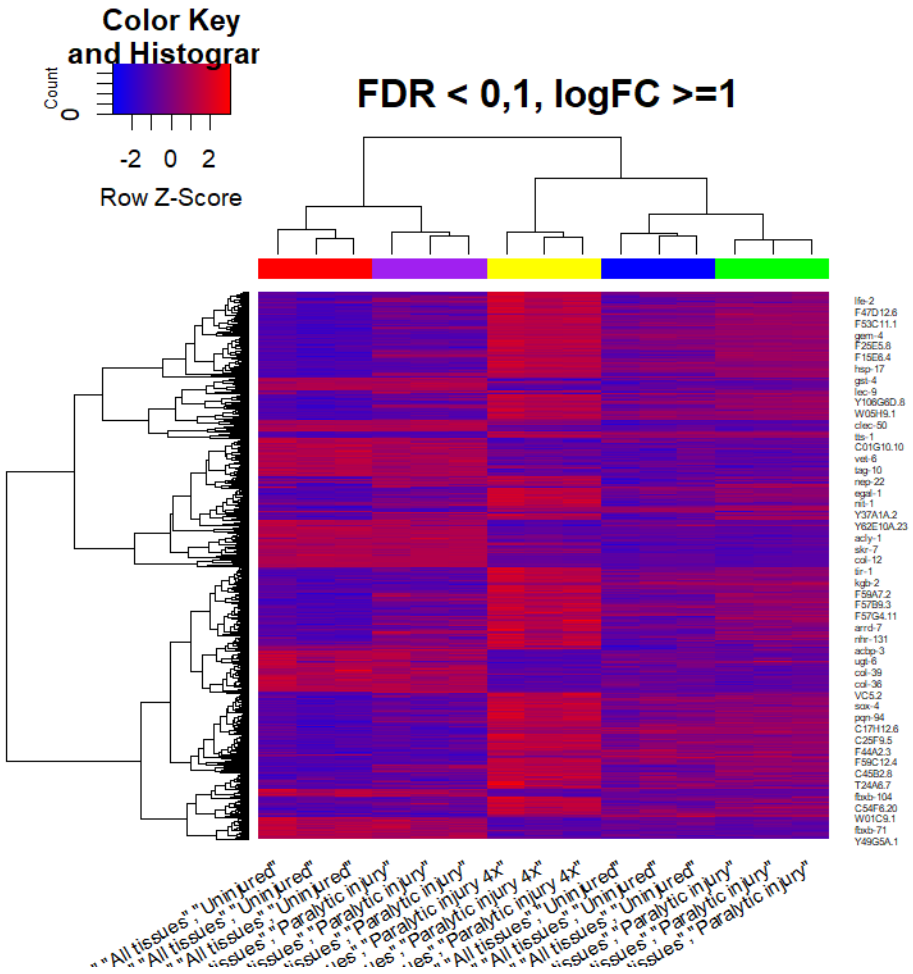

Figura 11: Heatmap para genes expresados diferencialmente con agrupamiento, para FDR <0,1 y logFC >=1

4.0.1. Resultados del "Gene Enrichment Analysis"

Los genes selecionados para este análisis en cada una de las comparaciones son:

```
  day1   day4 day4x4      U      P    INT
   284   1030   1400   2054   2854    785
```

Sólo se utilizarán los correspondientes a day1 (UvsP1), day4 (UvsP4) e INT (interacción entre ambos). Para ceñirme al objetivo del estudio, sólo se muestran aquí los resultados del análisis realizado con enrichGO para procesos biológicos (Figuras 12 y 13). En la Figura 12 aparecen cinco procesos biológicos enriquecidos, la respuesta inmune innata, la respuesta inmune, el proceso del sistema inmune, la respuesta de defensa y la respuesta a estímulos bióticos:

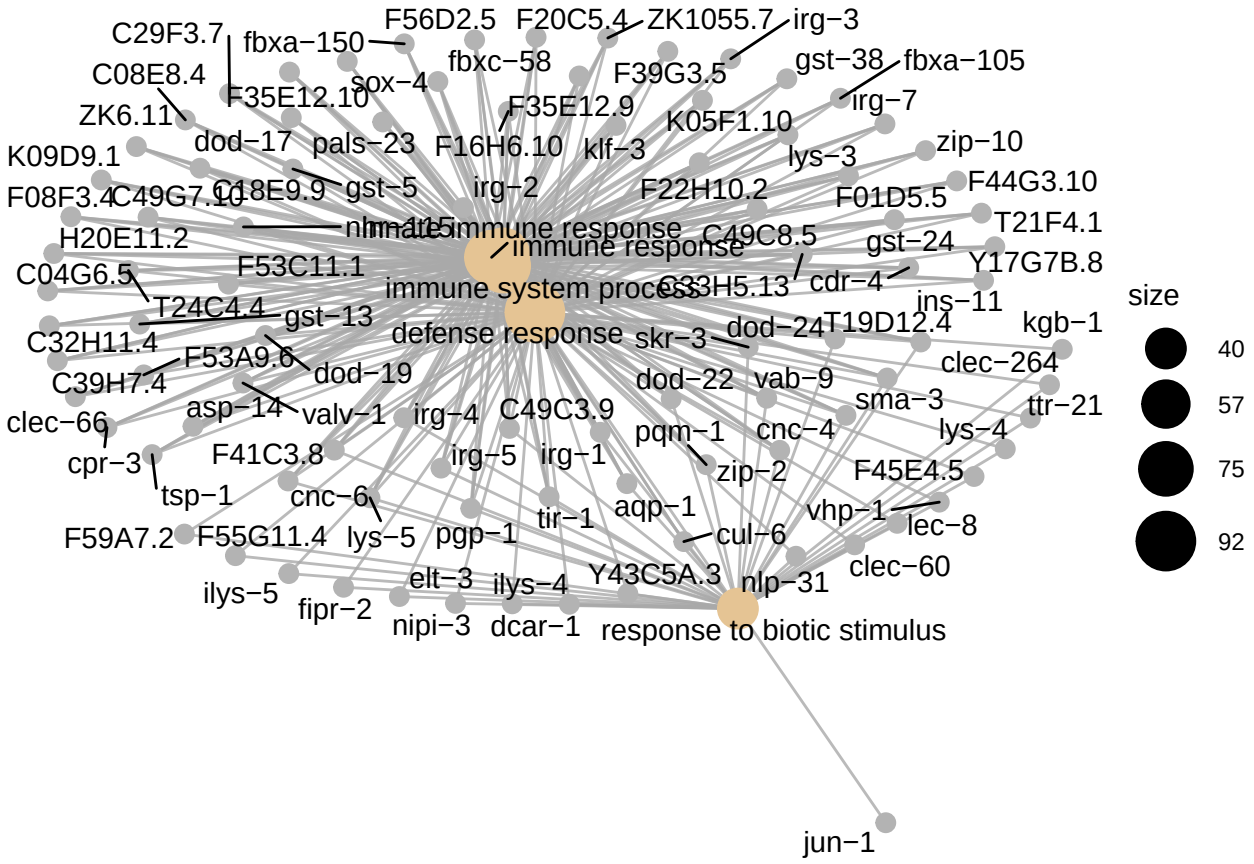

Figura 12: Network obtenida mediante el análisis enrichGO en la lista obtenida de la comparación entre UvsP1 y UvsP4

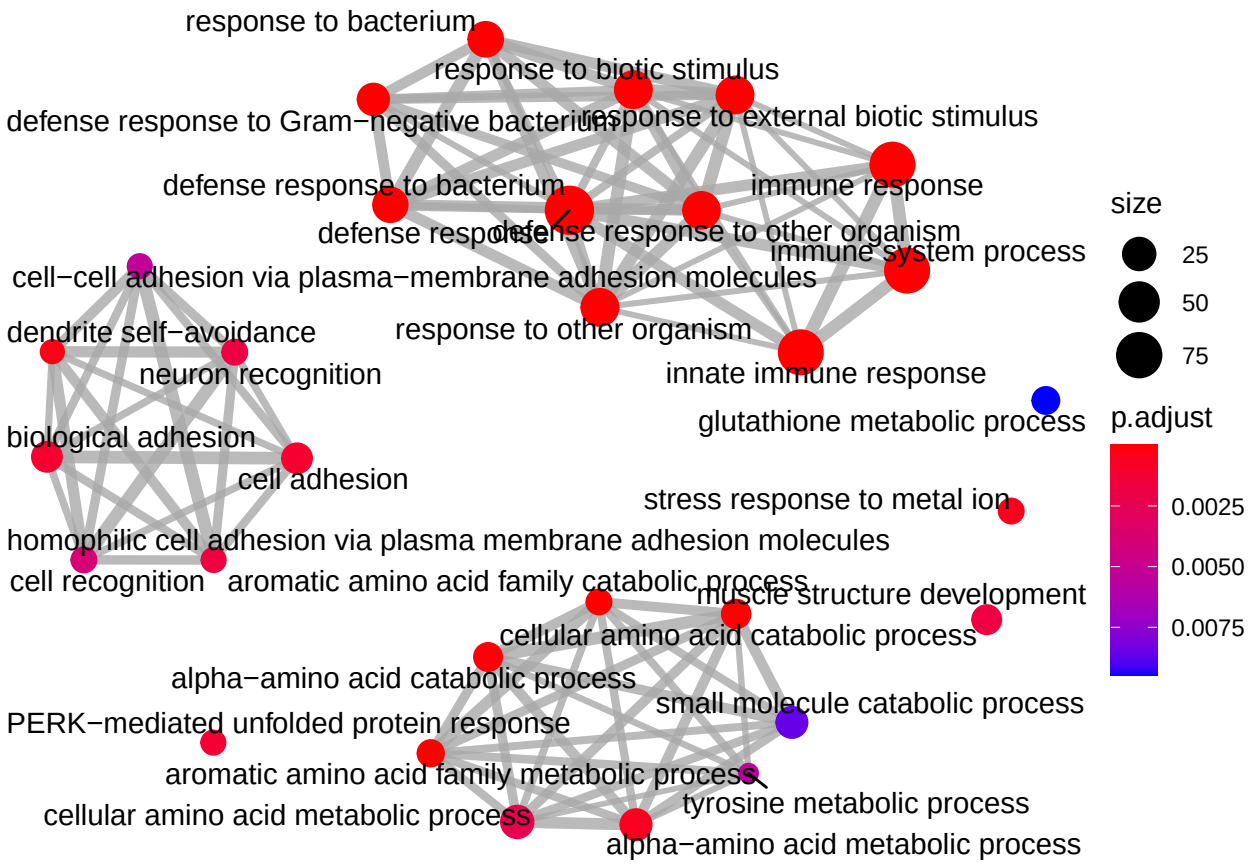

Figura 13: 'Enrichment Map' mediante el análisis enrichGO en la lista obtenida de la comparación entre UvsP1 y UvsP4

4.1. Sumario de resultados

Para los individuos de un día de edad, se han encontrado 20 procesos biológicos mediante el análisis de "enrichGO", mientras que para los individuos de cuatro días de edad se han encontrado 59. Es interesante observar que, aunque la diferencia en el número de procesos biológicos señalados por el análisis es notable entre ambas comparaciones, los 4 más importantes para ambas comparaciones son los mismos (Cuadros 1 y 2).

Cuadro 1: Primeras filas y columnas de resultados de enrichGO para la comparación UvsP1

	Description	GeneRatio	BgRatio	pvalue	p.adjust
GO:0006952	defense response	38/142	366/9266	8.70448695314621e-22	6.63281905829741e-19
GO:0006955	immune response	28/142	245/9266	3.50831230746877e-17	1.10674214900122e-14
GO:0002376	immune system process	28/142	247/9266	4.3572525551229e-17	1.10674214900122e-14
GO:0045087	innate immune response	27/142	239/9266	1.86020382160753e-16	3.54368828016235e-14

Cuadro 2: Primeras filas y columnas de resultados de enrichGO para la comparación UvsP4

	Description	GeneRatio	BgRatio	pvalue	p.adjust
GO:0006952	defense response	92/510	366/9266	1.8865113669616e-37	2.89579494828606e-34
GO:0045087	innate immune response	73/510	239/9266	9.4153309746098e-36	6.73536808807016e-33
GO:0002376	immune system process	74/510	247/9266	1.31635858398765e-35	6.73536808807016e-33
GO:0006955	immune response	73/510	245/9266	6.02383266111223e-35	2.31164578370182e-32

5. Discusión

Los resultados de este informe coinciden con los obtenidos en el experimento original; existen diferencias en las expresiones génicas de individuos de 1 y 4 días de edad, sometidos a un mísmo estímulo de estrés. Tanto el tamaño muestral como el diseño experimental son suficientes para observar diferencias en la expresión génica de los diferentes grupos. Aún así, se recomienda aumentar el número de muestras, principalmente para los grupos 1U, 1P, U4 y 4P. Además, sería conveniente crear el grupo 1P4P, esto es, individuos de 1 día de edad estimulados 4 veces. Así podría estudiarse también la habituación de *C.elegans* a estímulos de estrés y posibles cambios de la misma respecto a la edad del individuo.

Por último, se aconseja obtener muestras recogidas en tiempos variables. En el presente estudio, se recogieron las muestras de RNA una hora después de realizar el tratamiento. Si se recogiesen muestras, por ejemplo, 15min, 30min y 45min después del tratamiento, el estudio sería mucho más preciso a la hora de decidir qué rutas son las que efectivamente se han activado por el tratamiento recibido, y no por otras posibles señales que estén recibiendo del medio externo (cambio de temperatura, vibraciones por movimientos en el laboratorio, interacción con otros individuos si los hubiera,...). Además permitiría el estudio del tiempo de reacción de *C.elegans* desde la recepción del estímulo hasta su respuesta mde modificación de expresión génica.

Bibliografía

Braeckman, Bart P, Koen Houthoofd, y Jacques R Vanfleteren. 2002. «Assessing metabolic activity in aging Caenorhabditis elegans: concepts and controversies». *Aging cell* 1 (2): 82-88.

Bushel, Pierre. 2013. *pvca: Principal Variance Component Analysis (PVCA).*

Carlson, Marc. 2017. *celegans.db: Affymetrix celegans annotation data (chip celegans).*

Davis, Sean, y Paul Meltzer. 2007. «GEOquery: a bridge between the Gene Expression Omnibus (GEO) and BioConductor». *Bioinformatics* 14: 1846-7.

Guangchuang, Yu. 2018. *clusterProfiler: universal enrichment tool for functional and comparative study.* https://yulab-smu.github.io/clusterProfiler-book/index.html.

———. 2020. *Package clusterprofiler.* http://bioconductor.riken.jp/packages/release/bioc/manuals/clusterProfiler/man/clusterProfiler.pdf.

Hastings, Janna, Abraham Mains, Bhupinder Virk, Nicolas Rodriguez, Sharlene Murdoch, Juliette Pearce, Sven Bergmann, Nicolas Le Novère, y Olivia Casanueva. 2019. «Multi-omics and genome-scale modeling reveal a metabolic shift during C. elegans aging». *Frontiers in molecular biosciences* 6: 2.

Smyth, Gordon K. 2004. «Linear Models and Empirical Bayes Methods for Assessing Differential Expression in Microarray Experiments». *Statistical Applications in Genetics and Molecular Biology* 3 (1): 1-25. https://doi.org/10.2202/1544-6115.1027.

Warnes, Gregory R., Ben Bolker, Lodewijk Bonebakker, Robert Gentleman, Wolfgang Huber Andy Liaw, Thomas Lumley, Martin Maechler, etal. 2016. *gplots: Various R Programming Tools for Plotting Data.* https://CRAN.R-project.org/package=gplots.