Análisis de microarray

Prueba de evaluación contínua 1

Olga Reyes Malia

25/4/2020

Índice

Abstract

Para investigar la influencia del gen MyD88 en la respuesta inmune, ratones MyD88-/- y wild type son infectados con la bacteria *Chlamydia pneumoniae* y se analizan sus transcriptomas pulmonares. Estos datos se reanalizan bioinformáticamente en este estudio, con el uso de R y Bioconductor. Como resultado se obtiene una lista de genes diferencialmente expresados y un conjunto de vías moleculares que se ven afectadas.

Objetivos

Con este estudio se pretende investigar la influencia del gen MyD88 en la inducción de la resupuesta inmune en ratones, ilustrando detallademente el proceso de análisis bioinformático de microarrays. Para ello se reanalizan datos públicos procedentes de otra investigación, en busca de genes diferencialmente expresados.

Materiales y métodos

Este análisis se realiza a partir de los datos publicados por Rodríguez et al. el año 2007 (1). El dataset se encuentra publicado en la base de datos Gene Expression Omnibus (GEO) con el número de acceso GSE6688 (2).

Los datos se generaron con el objetivo de estudiar la influencia del gen MyD88 en la inducción de la resupuesta inmune. Para ello, se infectan ratones con la bacteria *Chlamydia pneumoniae*, la cual provoca pneumonia en humanos y ratones.

En la investigación se realizan 3 experimentos independientes (A, B y C), en los cuales se realiza el mismo procedimiento. Se analiza la expresión génica de muestras de pulmón de ratones con el gen MyD88 desactivado (knock out (MYD)) o no (wild type (WT)), tres días después de ser (CHL) o no ser infectados (MOCK) con la bacteria. Es decir, se utiliza un diseño experimental 2x2, donde el ratón puede ser knock out o wild type y además puede ser infectado o no infectado por la bacteria.

El microarray utilizado para la investigación es Affymetrix Mouse Expression 430A Array.

En el siguiente enlace se encuentra el repositorio github que contiene todos los datos utilizados para este análisis, el código generado y los resultados obtenidos: https://github.com/olrema/Analisis_microarray.

Pipeline

A continuación se explica brevemente el procedimiento que se ha llevado a cabo para el análisis de microarray, utilizando como programa principal R y Bioconductor.

Una vez obtenidos los datos en forma de archivos .CEL y creado el archivo *targets*, se procede a realizar el **primer control de calidad**, utilizando el paquete ArrayQualityMetrics.

Seguidamente se procede a **normalizar** los datos para reducir la variabilidad de las muestras agena a razones biológicas y se realiza un **segundo control de calidad**.

A continuación se realiza el **filtraje no específico**, utilizando la función nsFilter y un punto de corte de 0,75. El paquete de anotaciones utilizado para realizar el filtraje y el resto del análisis es: mouse4302.db (3).

Para **seleccionar los genes diferencialmente expresados** se utiliza el método lineal del paquete limma. La **matriz de contrastes** creada para llevarlo a cabo, consta de las siguientes comparaciones:

- MYDvsWT.CHL: ratones infectados knock out para el gen MyD88 vs ratones infectados wild type.
- MYDvsWT.MOCK: ratones no infectados knock out para el gen vs ratones no infectados wild type.

- INT: ratones infectados vs no infectados.

Con estas comparaciones lo que se pretende es evaluar el efecto de inactivar el gen MyD88 según si hay infección o no, y determinar si hay interacción entre la inactivación del gen y la infección.

Una vez definida la matriz de contrastes, se estima el modelo y se obtiene la lista de genes diferencialmente expresados, utilizando el paquete limma.

Se procede con la **anotación** de los genes para poder identificarlos, especificando el paquete comentado anteriormente.

Para continuar, se realiza una **comparación múltiple** para evaluar cuantos genes se han seleccionado en cada comparación, utilizando la función decideTests, con parámetros p.value=0.1 y lfc=1.

Por último se lleva a cabo el **análisis de significación biológica** para ver qué vías moleculares aparecen más frecuentemente en la lista de genes seleccionados. Para ello se utiliza el paquete clusterProfiler, filtrando por el pvalor ajustado < 0.1 y la función rWikiPathways para descargar el archivo gmt con las vías moleculares.

Resultados

En este apartado se muestran los resultados obtenidos en el estudio. Con el fin de obtener una idea de la distribución de los datos utilizados en el análisis se realizan varios gráficos.

El primer gráfico representado en la **figura 1** muestra la densidad de las 12 muestras. El gráfico de la **figura 2** muestra una posible división de las muestras en grupos.

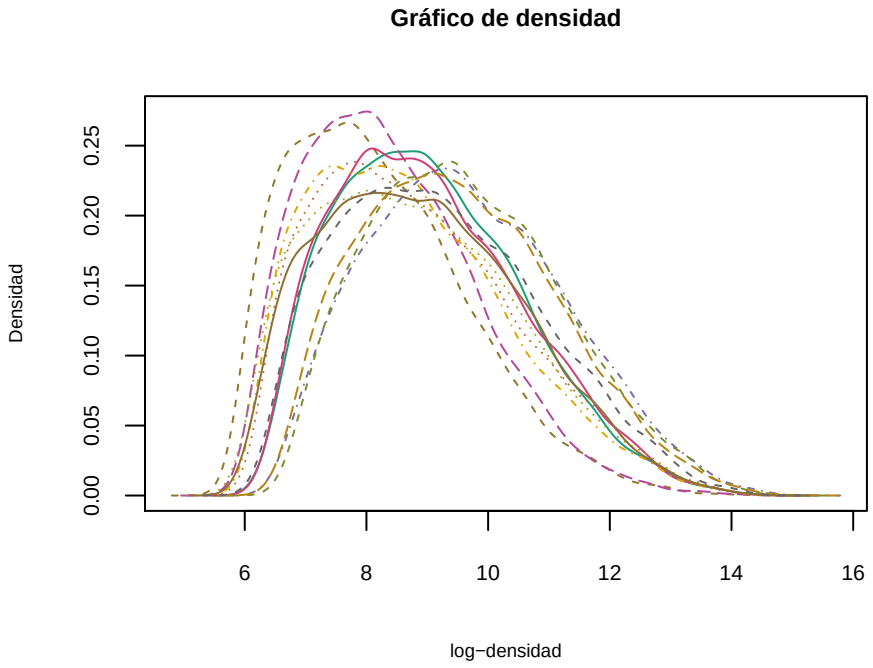

Figura 1: Gráfico de densidad para las 12 muestras del estudio.

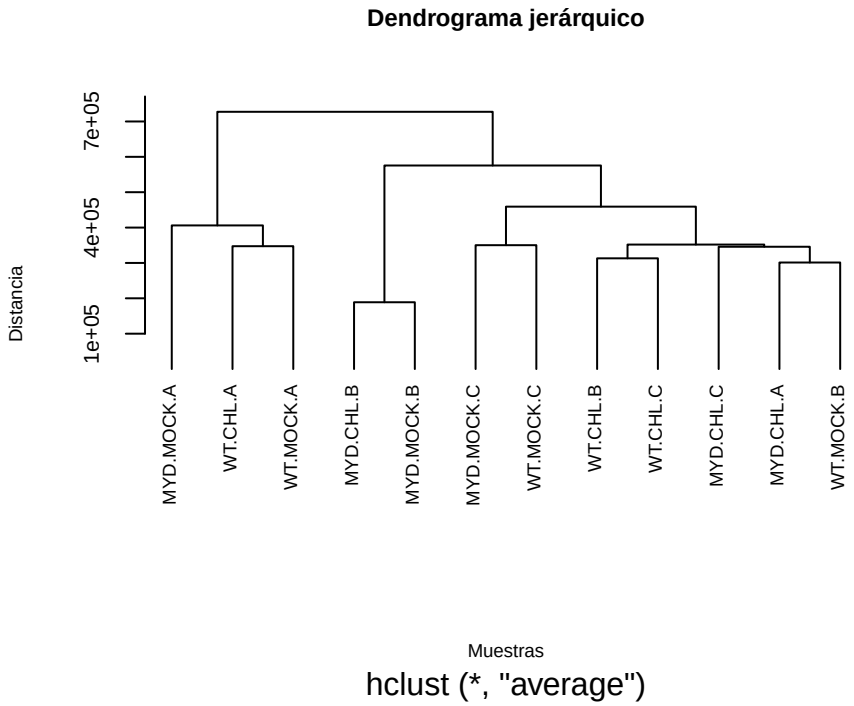

Figura 2: Dendrograma que representa el clúster jerárquico de las muestras, basado en la información de todos los genes.

Control de calidad de los datos crudos

A continuación se muestran los resultados obtenidos del primer control de calidad realizado sobre los datos crudos. En la **Figura 3** se representa un resumen de estos resultados. Como se puede observar, algunas de las muestras aparecen marcadas como problemáticas en el criterio utilizado número 3. Como solo aparecen marcadas en uno de los criterios, se decide que permanezcan en el análisis.

	array	sampleNames	*1	*2	*3	Group	State	Infection	ShortName
☐	1	MYD.CHL.A				MYD.CHL	MYD	CHL	MYD.CHL.A
☐	2	MYD.CHL.B			x	MYD.CHL	MYD	CHL	MYD.CHL.B
☐	3	MYD.CHL.C				MYD.CHL	MYD	CHL	MYD.CHL.C
☐	4	MYD.MOCK.A			x	MYD.MOCK	MYD	MOCK	MYD.MOCK.A
☐	5	MYD.MOCK.B			x	MYD.MOCK	MYD	MOCK	MYD.MOCK.B
☐	6	MYD.MOCK.C				MYD.MOCK	MYD	MOCK	MYD.MOCK.C
☐	7	WT.CHL.A			x	WT.CHL	WT	CHL	WT.CHL.A
☐	8	WT.CHL.B				WT.CHL	WT	CHL	WT.CHL.B
☐	9	WT.CHL.C			x	WT.CHL	WT	CHL	WT.CHL.C
☐	10	WT.MOCK.A			x	WT.MOCK	WT	MOCK	WT.MOCK.A
☐	11	WT.MOCK.B				WT.MOCK	WT	MOCK	WT.MOCK.B
☐	12	WT.MOCK.C			x	WT.MOCK	WT	MOCK	WT.MOCK.C

Figura 3: Tabla resumen resultante del control de calidad de los datos crudos.

En la **figura 4** se puede observar también la distribución de intensidades de las muestras antes de normalizarlas en forma de diagrama de cajas. Cada color representa un tipo de muestra.

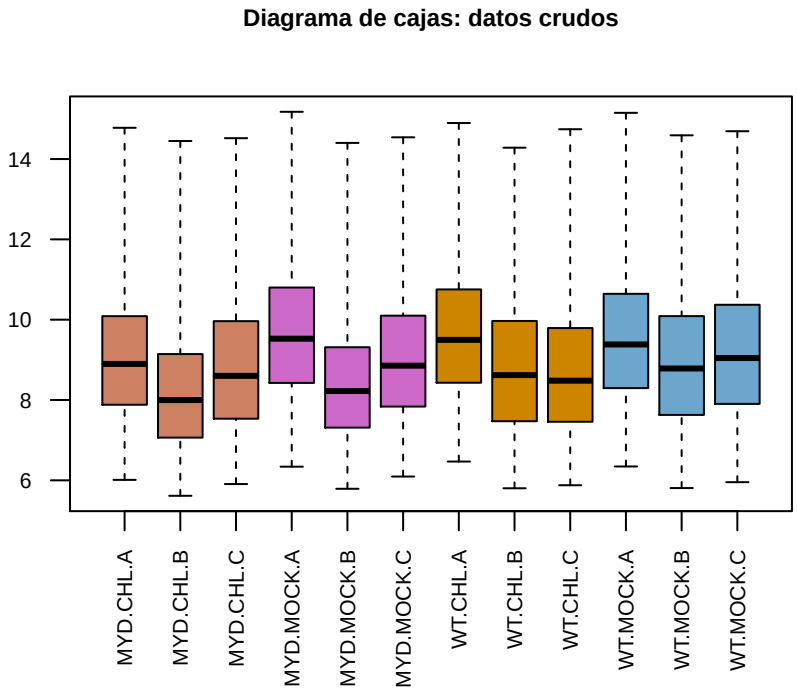

Figura 4: Distribución de intensidades de los datos crudos.

Control de calidad de los datos normalizados

Los siguientes resultados muestran el mismo control de calidad realizado anteriormente, ahora con los datos normalizados. Tanto en la **figura 5** como en la **figura 6** se observa una mejora clara de la calidad de los datos. Esto indica que la normalización se ha realizado adecuadamente y los datos estan preparados para ser filtrados.

	array	sampleNames	*1	*2	*3	Group	State	Infection	ShortName
☐	1	MYD.CHL.A				MYD.CHL	MYD	CHL	MYD.CHL.A
☐	2	MYD.CHL.B				MYD.CHL	MYD	CHL	MYD.CHL.B
☐	3	MYD.CHL.C				MYD.CHL	MYD	CHL	MYD.CHL.C
☐	4	MYD.MOCK.A				MYD.MOCK	MYD	MOCK	MYD.MOCK.A
☐	5	MYD.MOCK.B				MYD.MOCK	MYD	MOCK	MYD.MOCK.B
☐	6	MYD.MOCK.C				MYD.MOCK	MYD	MOCK	MYD.MOCK.C
☐	7	WT.CHL.A				WT.CHL	WT	CHL	WT.CHL.A
☐	8	WT.CHL.B				WT.CHL	WT	CHL	WT.CHL.B
☐	9	WT.CHL.C				WT.CHL	WT	CHL	WT.CHL.C
☐	10	WT.MOCK.A				WT.MOCK	WT	MOCK	WT.MOCK.A
☐	11	WT.MOCK.B				WT.MOCK	WT	MOCK	WT.MOCK.B
☐	12	WT.MOCK.C				WT.MOCK	WT	MOCK	WT.MOCK.C

Figura 5: Tabla resumen resultante del control de calidad de los datos normalizados.

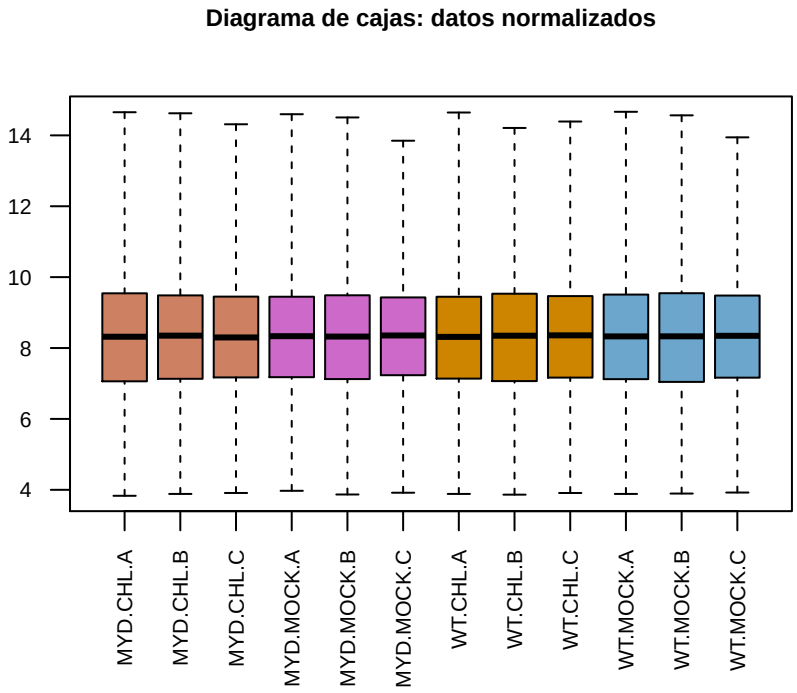

Figura 6: Distribución de intensidades de los datos normalizdos

Filtraje no específico

A continuación aparece el informe de los resultados del filtraje realizado.

```
## $numDupsRemoved
## [1] 8225
##
## $numLowVar
## [1] 9744
##
## $numRemoved.ENTREZID
## [1] 1460
##
## $feature.exclude
## [1] 13
```

Como resultado, queda un total de 3248 genes para analizar.

Identificación de genes diferencialmente expresados

Las siguientes tablas muestran parte del listado de genes expresados diferencialmente en cada comparación realizada. En concreto se muestran los genes con el menor p-valor y por tanto, con una expresión diferencial mayor.

Para cada gen se muestra la siguiente información:

- LogFC: diferencia media entre los grupos.

- AveExpr: expresión promedio de los genes
- t: estadístico t
- P.Value: prueba valor p.
- adj.P.Val: valor p ajustado.
- B: prueba estadística B.

Tabla 1: Genes que cambian su expresión en ratones infectados según si son knock out para el gen MyD88 o no.

	logFC	AveExpr	t	P.Value	adj.P.Val	B
1419561_at	-4.036089	8.484462	-18.703193	0e+00	0.0000003	14.171415
1427381_at	-3.905535	9.032549	-13.080580	0e+00	0.0000082	10.537956
1449984_at	-3.732388	7.192287	-12.896862	0e+00	0.0000082	10.386679
1420330_at	-3.152260	8.509696	-12.834314	0e+00	0.0000082	10.334598
1419482_at	-2.137534	8.852913	-11.766041	0e+00	0.0000186	9.397343
1451161_a_at	-1.809038	9.589791	-9.715821	3e-07	0.0001448	7.311533

Tabla 2: Genes que cambian su expresión en ratones no infectados según si son knock out para el gen MyD88 o no.

	logFC	AveExpr	t	P.Value	adj.P.Val	B
1456182_x_at	4.5865350	7.468588	9.600806	0.0000003	0.0009953	-0.0698446
1427351_s_at	-3.0198592	10.740436	-6.267841	0.0000299	0.0485158	-1.0277221
1416306_at	3.1539103	9.282170	4.828273	0.0003367	0.3645235	-1.7706895
1460285_at	0.7843039	9.254150	4.191720	0.0010723	0.8707482	-2.1914634
1417851_at	-1.7051450	8.713466	-3.937570	0.0017243	0.9900856	-2.3762816
1418095_at	-0.7289167	6.618324	-3.906260	0.0018290	0.9900856	-2.3997015

Tabla 3: Genes que cambian su expresión según si los ratones están infectados o no

	logFC	AveExpr	t	P.Value	adj.P.Val	B
1419561_at	-3.909063	8.484462	-12.808926	0.0e+00	0.0000337	9.492719
1427381_at	-4.050144	9.032549	-9.591841	3.0e-07	0.0003722	6.811973
1449984_at	-3.889980	7.192287	-9.504507	3.0e-07	0.0003722	6.724286
1419482_at	-2.180580	8.852913	-8.487395	1.2e-06	0.0008412	5.633622
1420330_at	-2.933702	8.509696	-8.446012	1.3e-06	0.0008412	5.586411
1451798_at	-3.062571	8.797627	-6.356210	2.6e-05	0.0140687	2.879441

Los resultados obtenidos en las tablas anteriores se aprecian de forma más visual en las gráficas siguientes (**figura 7, 8, 9**). Estas gráficas contienen resaltados los símbolos de los genes candidatos a tener una expresión diferencial en cada una de las tres comparaciones.

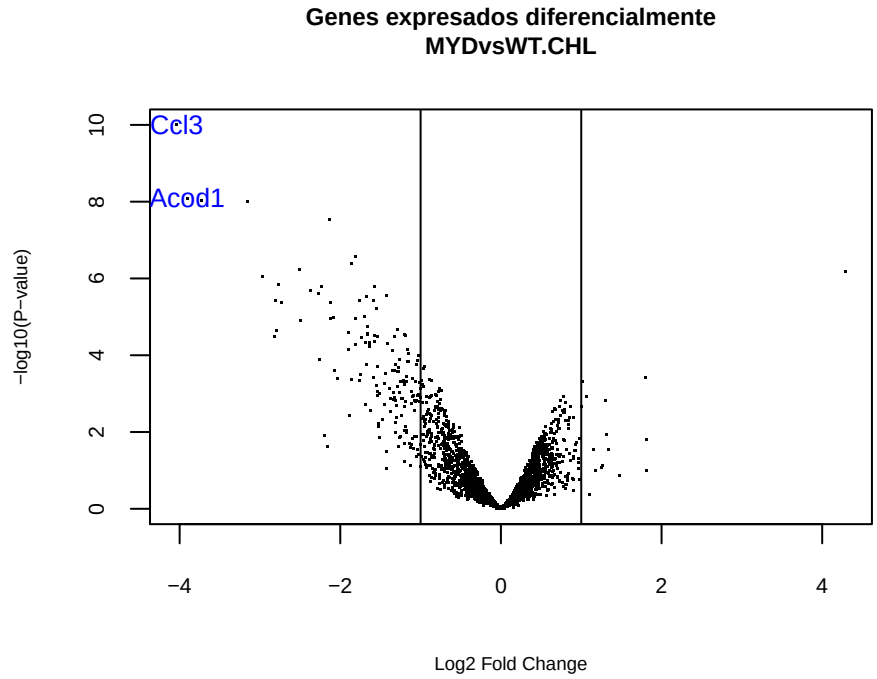

Figura 7: Volcanoplot que muestra 2 de los genes diferencialmente expresados en ratones infectados según si son knock out o wild type.

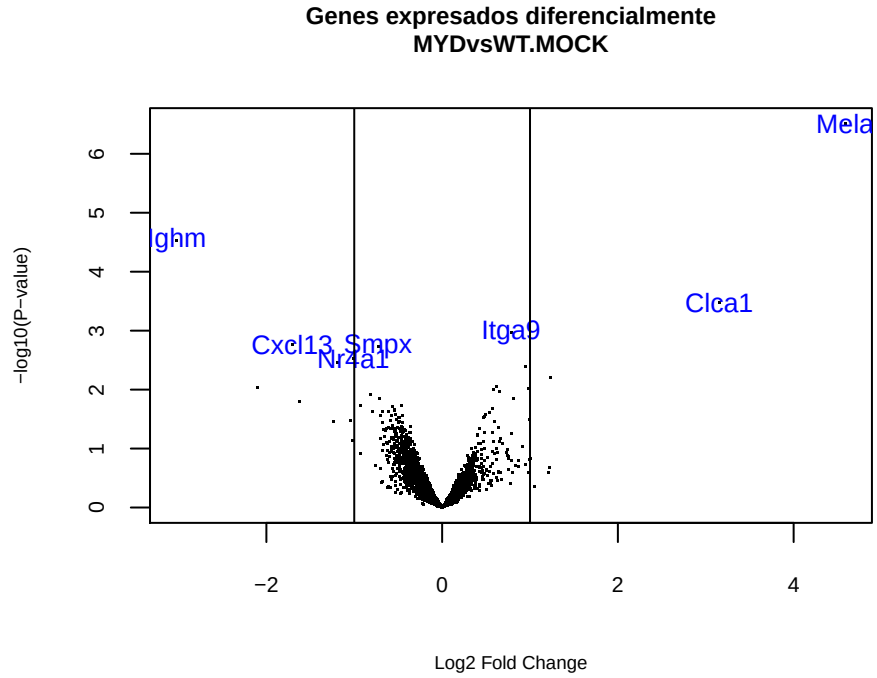

Figura 8: Volcanoplot que muestra 7 de los genes diferencialmente expresados en ratones no infectados según si son knock out para el gen MyD88 o no.

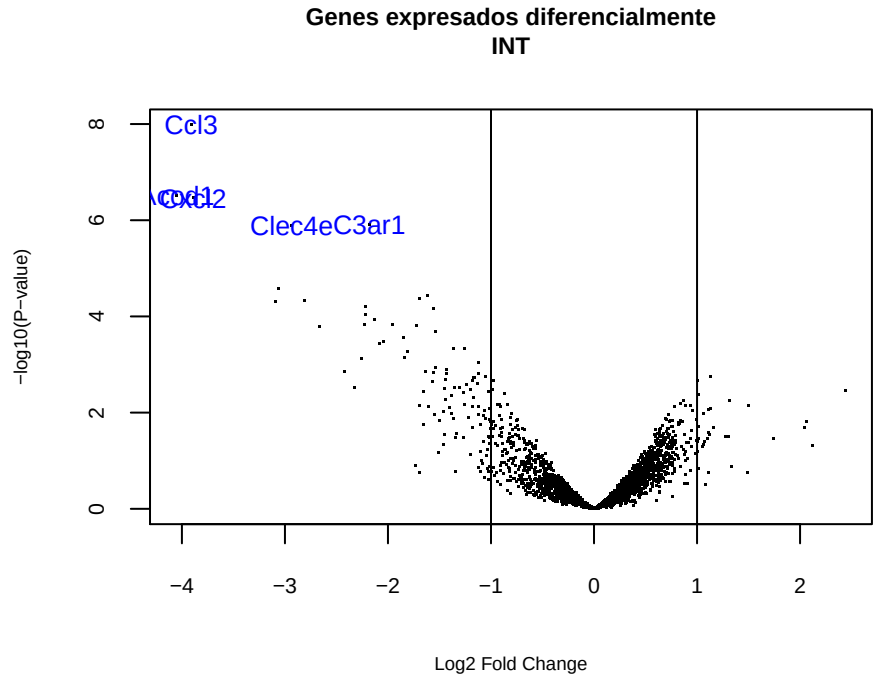

Figura 9: Volcanoplot que muestra 5 de los genes diferencialmente expresados en la comparación entre ratones infectados y no infectados.

Anotación de los resultados

Las **tablas 4, 5 y 6** contienen la anotación de algunos de los genes seleccionados anteriormente. Si se desea obtener más información, en los archivos de resultados adjuntos se encuentra la anotación completa de cada comparación realizada, con una columna adicional que contiene el nombre completo y la descripción de cada gen (*topAnnotated*_comparación).

Tabla 4: Anotación de los genes que cambian su expresión en ratones infectados.

PROBEID	SYMBOL	logFC	AveExpr	t	P.Value	adj.P.Val	B
1415673_at	Psph	0.1177870	7.169913	0.4687976	0.6470358	0.8656426	-6.594931
1415676_a_at	Psmb5	0.0210047	10.783048	0.1154445	0.9098700	0.9740059	-6.703543
1415683_at	Nmt1	-0.0403194	10.701729	-0.1824978	0.8580281	0.9606680	-6.692955
1415694_at	Wars	-0.5456767	10.461282	-2.2128770	0.0455556	0.2705018	-4.512359
1415695_at	Psma1	-0.1005872	9.586011	-0.3714355	0.7163408	0.8990243	-6.637761
1415698_at	Golm1	-0.0386347	10.294030	-0.2221640	0.8276638	0.9533361	-6.684461

Tabla 5: Anotación de los genes que cambian su expresión en ratones no infectados.

PROBEID	SYMBOL	logFC	AveExpr	t	P.Value	adj.P.Val	B
1415673_at	Psph	-0.2161215	7.169913	-0.8601737	0.4053921	0.9926301	-4.807934
1415676_a_at	Psmb5	-0.2069272	10.783048	-1.1372980	0.2760926	0.9926301	-4.652269

PROBEID	SYMBOL	logFC	AveExpr	t	P.Value	adj.P.Val	B
1415683_at	Nmt1	-0.2323719	10.701729	-1.0517840	0.3121977	0.9926301	-4.703921
1415694_at	Wars	-0.1673811	10.461282	-0.6787788	0.5092698	0.9926301	-4.889916
1415695_at	Psma1	-0.2176704	9.586011	-0.8037850	0.4360791	0.9926301	-4.835220
1415698_at	Golm1	-0.1023699	10.294030	-0.5886644	0.5662489	0.9926301	-4.924167

Tabla 6: Anotación de los genes que cambian su expresión según si los ratones están infectados o no.

PROBEID	SYMBOL	logFC	AveExpr	t	P.Value	adj.P.Val	B
1415673_at	Psph	0.3339085	7.169913	0.9397246	0.3646182	0.8611680	-5.864418
1415676_a_at	Psmb5	0.2279319	10.783048	0.8858227	0.3919226	0.8710431	-5.912497
1415683_at	Nmt1	0.1920524	10.701729	0.6146782	0.5494533	0.9053127	-6.115155
1415694_at	Wars	-0.3782956	10.461282	-1.0847713	0.2978750	0.8455014	-5.723025
1415695_at	Psma1	0.1170832	9.586011	0.3057173	0.7646961	0.9581297	-6.260647
1415698_at	Golm1	0.0637351	10.294030	0.2591549	0.7996041	0.9616858	-6.274269

Comparación múltiple

Los resultados obtenidos de la comparación múltiple se resumen en la **tabla 7**, la cual se encuentra representada gráficamente en la **figura 10**.

"Down" hace referencia a los genes que se han visto *down* regulados, "NotSig" a los que no tienen una expresion diferencial significaiva y, por último, "Up" hace referencia a aquellos genes que se han visto sobreexpresados.

Estos resultados muestran que la mayor parte de genes diferencialmente expresados encontrados en el estudio, forman parte del grupo infectado de ratones. Así mismo, los ratones que no han sido infectados presentan únicamente dos genes expresados diferencialmente.

Tabla 7: Resumen de la comparación múltiple.

	MYDvsWT.CHL	MYDvsWT.MOCK	INT
Down	122	1	27
NotSig	3121	3246	3221
Up	5	1	0

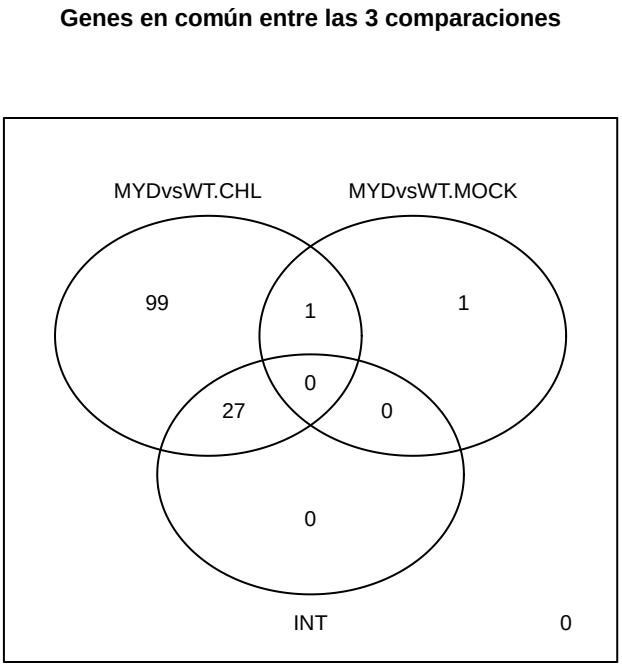

Figura 10: Diagrama de venn mostrando los genes diferencialmente expresados en común entre las 3 comparaciones realizadas.

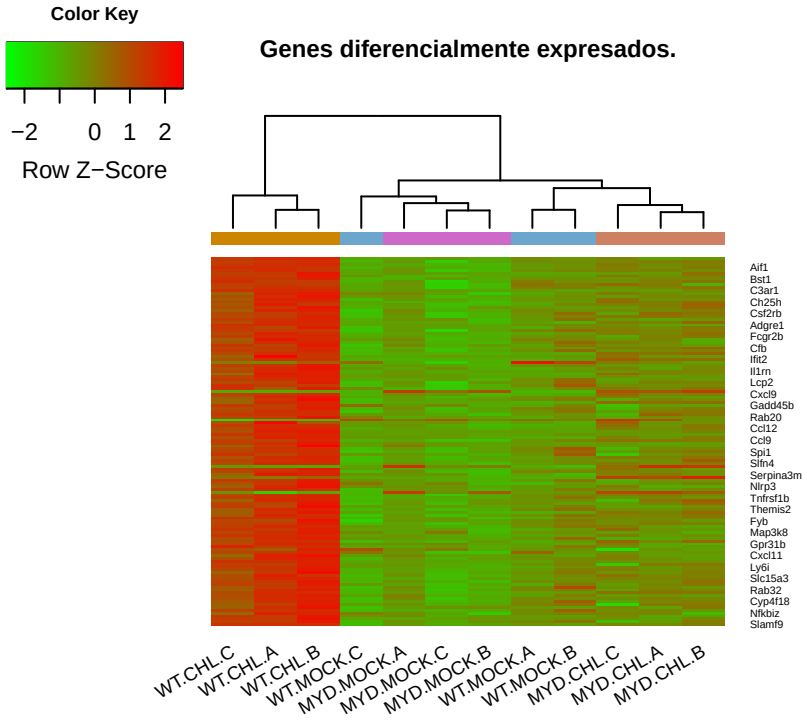

Figura 11: Heatmap para los genes seleccionados como diferencialmente expresados en alguna de las tres comparaciones, agrupados por muestras.

La **figura 11** refleja en forma de heatmap los genes que se han seleccionado como diferencialmente expresados anteriormente.

Análisis de significación biológica

La **tabla 8** resume el número de genes de cada comparación que se van a tener en cuenta para realizar el análisis de significación biológica. Tal y como se observa, la única comparación que tiene una cantidad de genes óptimos para ser incluida en el análisis es la primera, correspondiente a la comparación entre ratones infectados knock out para el gen MYD y ratones infectados wild type.

Tabla 8: Número de genes a estudiar en cada comparación.

	x
MYDvsWT.CHL	217
MYDvsWT.MOCK	2
INT	27

Como resultado del análisis de significación se obtiene la **tabla 9** (que aparece completa en los archivos adjuntos como *clusterProfiler.Results.MYDvsWT.CHL*) y la **figura 12** , que resume la red de las diferentes vías encontradas en el análisis. Son un total de 11 vías encontradas en la comparación mencionada, entre las cuales destaca la via de señalización de la quimiocina.

Tabla 9: Parte de los resultados del clusterProfiler para la comparación entre ratones knock outs y wild types infectados.

	Description	GeneRatio	BgRatio	pvalue	p.adjust
WP2292	Chemokine signaling pathway	18/99	191/4612	0.0000001	0.0000057
WP3626	Microglia Pathogen Phagocytosis Pathway	9/99	41/4612	0.0000001	0.0000057
WP1253	Type II interferon signaling (IFNG)	8/99	34/4612	0.0000004	0.0000111
WP3625	TYROBP Causal Network	7/99	58/4612	0.0002082	0.0040406
WP2432	Spinal Cord Injury	9/99	99/4612	0.0002349	0.0040406
WP3632	Lung fibrosis	7/99	61/4612	0.0002867	0.0041088

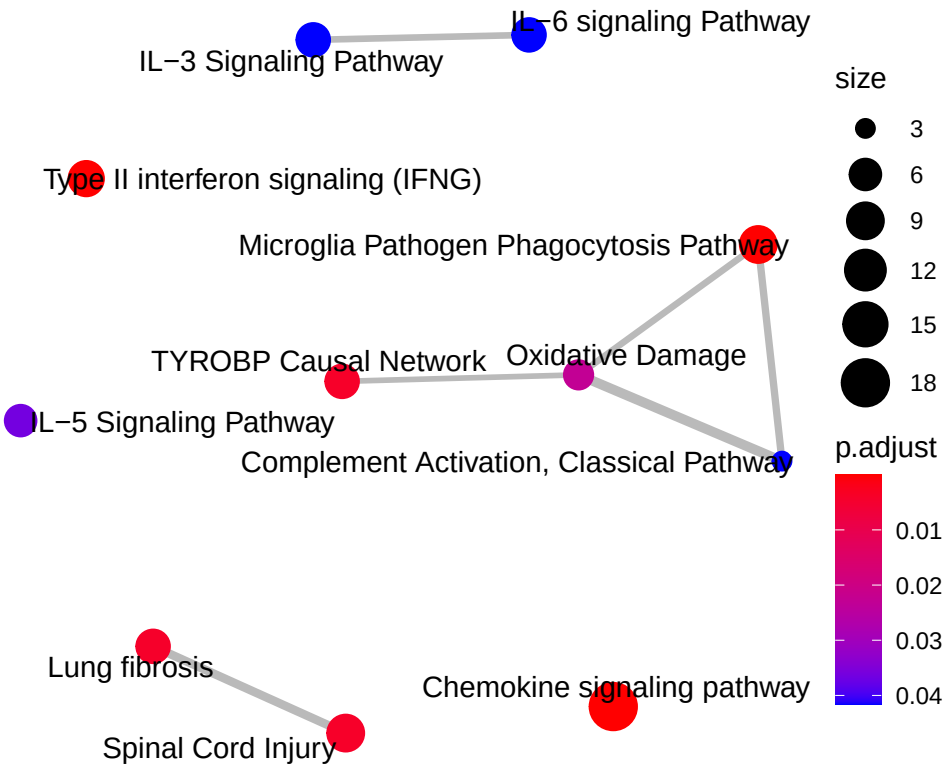

Figura 12: Red obtenida del análisis de significación biológica para la comparación KO y WT en ratones infectados.

Discusión

Comparando los resultados obtenidos con el estudio original se aprecian algunas diferencias. La principal diferencia es el número de genes diferencialmente expresados que se encuentran. En el estudio original se encuentran 378 genes, mientras que en este caso se encuentran 3121. Este aspecto podría mejorarse, afinando más los filtrajes, así como el valor de lfc o fdr.

Otra diferencia remarcable es que la mayoría de genes seleccionados se encuentran upregulados en el estudio original, mientras que en este análisis la mayoría son downregulados. Esto puede deberse a que las matrices de contrastes realizadas en cada caso sean diferentes. Dependiendo de como se hagan las comparaciones, puedes obtener un resultado u otro, significando lo mismo biológicamente.

Finalmente, si observamos los heatmaps obtenidos, son muy similares. En ambos casos la mayoría de genes diferencialmente expresados se situan en el grupo de ratones infectados.

Una de las limitaciones encontradas a la hora de realizar el análisis ha sido el paquete de anotaciones. Al inicio del análisis se pretendría utilizar el paquete Affymoe4302Expr (4), pero no era posible cargarlo en el programa, así que se decidió cambiarlo.

Conclusión

Como conclusión final podemos extraer que el gen MyD88 es clave para la respuesta inmune, puesto que en ratones infectados que lo contienen, la expresión de otros genes secundarios se ve afectada. En cambio, en ratones infectados que no tienen el gen, el número de genes expresados diferencialmente disminuye mucho.

Por último, el análisis de pathways indica claramente que la mayoría de genes que se han visto expresados diferencialmente están involucrados en la vía de señalización de las quimiocinas, la cual es crucial en las respuestas inflamatorias y el control de tráfico de leucocitos (5).

En resumen, este análisis nos permite comprobar que realmente el gen MyD88 es importante para activar la respuesta inmune en ratones ya que su ausencia provoca que, no inicien la respuesta inmune contra la bacteria con la que han sido infectados.

Bibliografía

(1) Rodríguez, N., Mages, J., Dietrich, H., Wantia, N., Wagner, H., Lang, R., & Miethke, T. (2007). MyD88-dependent changes in the pulmonary transcriptome after infection with Chlamydia pneumoniae. Physiological Genomics, 30(2), 134–145.doi:10.1152/physiolgenomics.00011.2007

(2) GEO Accession viewer (2020). https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE6688

(3) GEO Accession viewer (2020). http://bioconductor.org/packages/release/data/annotation/html/mouse4302.db.html

(4) GEO Accession viewer (2020). http://bioconductor.org/packages/release/data/experiment/html/Affymoe4302Expr.html

(5) R&D Systems 2020. Chemokine Signaling Pathways. https://www.rndsystems.com/pathways/chemokine-signaling-pathways