

ANÁLISIS DE DATOS PÚBLICOS DE BULK RNA-SEQ

1. Diseño y metodología experimentales.

El estudio asociado al conjunto de datos GSE111003 analiza los mecanismos moleculares implicados en la memoria inmune innata inducida por hemo y otros estímulos en células mieloides humanas y murinas. En el presente trabajo se utilizaron principalmente monocitos humanos obtenidos a partir de sangre periférica de donantes sanos, así como modelos murinos complementarios para validar los hallazgos funcionales.

1.1. Aislamiento celular y condiciones de cultivo

Los monocitos humanos se aislaron a partir de células mononucleares de sangre periférica (PBMCs) obtenidas de buffy coats o sangre fresca de voluntarios sanos, siguiendo protocolos previamente descritos y aprobados por el comité ético correspondiente. Las células se cultivaron en medio RPMI 1640 suplementado con 10% de suero humano pooled, GlutaMAX, piruvato y antibiótico.

Para los experimentos de “training”, los monocitos se incubaron durante 24 h con:

- Medio de cultivo (RPMI) como control negativo.
- 1 µg/mL β-1,3-(D)-glucano.
- Otros porfirinos según el experimento.

Tras 24 h de estimulación, las células se lavaron y se cultivaron nuevamente en medio fresco hasta el día 6. En algunos experimentos, las células fueron reestimuladas con LPS (10 ng/mL) el día 6 para evaluar respuestas secundarias. Los sobrenadantes se recogieron 24 h después para el análisis de citocinas, mientras que las células se lisaron para análisis transcripcional.

Este diseño permite diferenciar entre:

1. Respuestas transcripcionales tempranas (4 h y 24 h).
2. Cambios persistentes asociados a la memoria innata (día 6).
3. Respuesta secundaria tras reestimulación con LPS.

1.2. RNA-seq y análisis bioinformático original

Para el análisis transcriptómico:

- Las lecturas de RNA-seq se alinearon contra el transcriptoma humano Ensembl v68 utilizando Bowtie 1.
- La cuantificación de expresión génica se realizó mediante MMSEQ (Many-against-Many sequence searching).
- La expresión diferencial se determinó mediante DESeq.
- Los genes diferencialmente expresados se definieron con:
 - Fold change > 2.5

- p ajustado < 0.05
- RPKM ≥ 5

Este pipeline refleja una estrategia clásica de análisis de RNA-seq basada en alineamiento a transcriptoma y modelado estadístico de conteos mediante distribución binomial negativa.

1.3. Adaptación al presente análisis

En este informe se ha realizado un subanálisis del conjunto de datos centrado exclusivamente en el estímulo con β -glucano y en los tiempos tempranos (T0, 4 h y 24 h), utilizando como controles las muestras cultivadas en medio (RPMI) en los mismos puntos temporales.

El diseño analizado incluye cinco donantes independientes (HD34, HD37, HD48, HD49 y HD51), cada uno representado en todas las condiciones estudiadas, lo que permite modelar la variabilidad interindividual mediante un diseño estadístico del tipo (~ Donor + Condition).

Este enfoque controla el efecto donante como factor bloque, permitiendo estimar con mayor precisión el efecto específico del estímulo en cada tiempo.

A diferencia del análisis original, que utilizó criterios de filtrado estrictos basados en fold change y RPKM, en este trabajo se ha seguido la metodología docente empleada en clase, incorporando:

- Filtrado mediante filterByExpr() (basado en CPM).
- Normalización inter-muestra mediante TMM (edgeR).
- Análisis exploratorio multivariante (PCA).
- Expresión diferencial mediante DESeq2 con corrección por donante.

2. Descripción de los análisis realizados a partir de la matriz de expresión no normalizada

2.1. Construcción y exploración inicial de la matriz de conteos

A partir de los archivos generados mediante MMSEQ, se construyó una matriz de conteos enteros (genes × muestras), donde cada celda representa el número de lecturas únicas asignadas a un gen en una muestra concreta.

Tras restringir el análisis a las condiciones de interés (T0, RPMI_4h, BG_4h, RPMI_24h y BG_24h), el diseño final incluyó cinco donantes independientes, cada uno representado en todas las condiciones analizadas. Esto permite controlar la variabilidad interindividual en los análisis posteriores.

En una primera inspección, se evaluó la proporción de genes sin expresión en ninguna muestra. Se observó que una fracción considerable de genes presentaba conteos nulos en todas las condiciones, lo cual es esperable al utilizar referencias

transcriptómicas amplias que incluyen transcritos poco o nada expresados en este tipo celular.

Estos genes fueron eliminados antes de continuar con el análisis.

2.2. Filtrado de genes de baja expresión

Para evitar ruido estadístico y mejorar la potencia en el análisis diferencial, se aplicó un filtrado basado en la función `filterByExpr()` del paquete `edgeR`, que selecciona genes con niveles mínimos de expresión considerando la estructura experimental.

Este filtrado se basa en valores de CPM (counts per million), evitando sesgos debidos a diferencias en el tamaño de librería entre muestras.

Tras el filtrado, se retuvo aproximadamente el conjunto de genes con expresión suficiente para análisis posteriores de forma robusta.

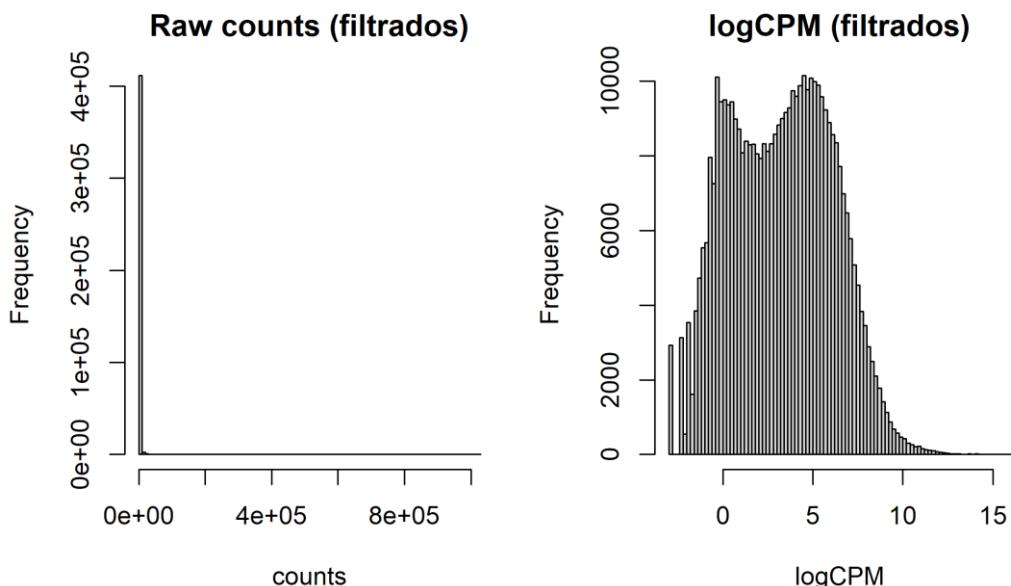


Figura 1. Histograma de conteos crudos y logCPM, antes (izquierda) y después (derecha) del filtrado

El histograma de conteos crudos muestra una distribución fuertemente sesgada hacia valores bajos, característica típica de datos de RNA-seq debido a que la mayoría de los genes se expresan poco y unos pocos genes concentran valores muy elevados. Tras la transformación $\log_2(\text{CPM})$, la distribución se approxima a una forma más simétrica, lo cual facilita análisis multivariantes posteriores.

2.3. Normalización intra-muestra (CPM y log-transformación)

Para corregir diferencias en el tamaño de librería entre muestras, se transformaron los conteos crudos a CPM (counts per million). Esta transformación escala los conteos en función del número total de lecturas por muestra, permitiendo comparaciones relativas entre muestras para un mismo gen.

Posteriormente, se aplicó una transformación $\log_2(\text{CPM})$, que reduce la asimetría de la distribución y estabiliza la varianza. Este paso es especialmente importante para análisis exploratorios PCA, ya que los conteos crudos no siguen una distribución adecuada para este tipo de métodos.

El tamaño de librería fue relativamente homogéneo entre muestras, sin observarse diferencias extremas ni outliers evidentes. Esto sugiere una adecuada técnica de secuenciación y reduce la probabilidad de sesgos importantes derivados de la profundidad de lectura.

2.4. Normalización inter-muestra (TMM)

Además de las diferencias en profundidad de secuenciación, los datos de RNA-seq pueden presentar sesgos de comparación, derivados de la presencia de genes altamente expresados que distorsionan las comparaciones globales entre muestras. Para corregir este efecto, se aplicó el método TMM (Trimmed Mean of M-values) implementado en edgeR.

TMM estima factores de normalización para cada muestra bajo el supuesto que la mayoría de los genes no están diferencialmente expresados. Estos factores permiten ajustar las distribuciones de expresión entre muestras, reduciendo variabilidad técnica no biológica.

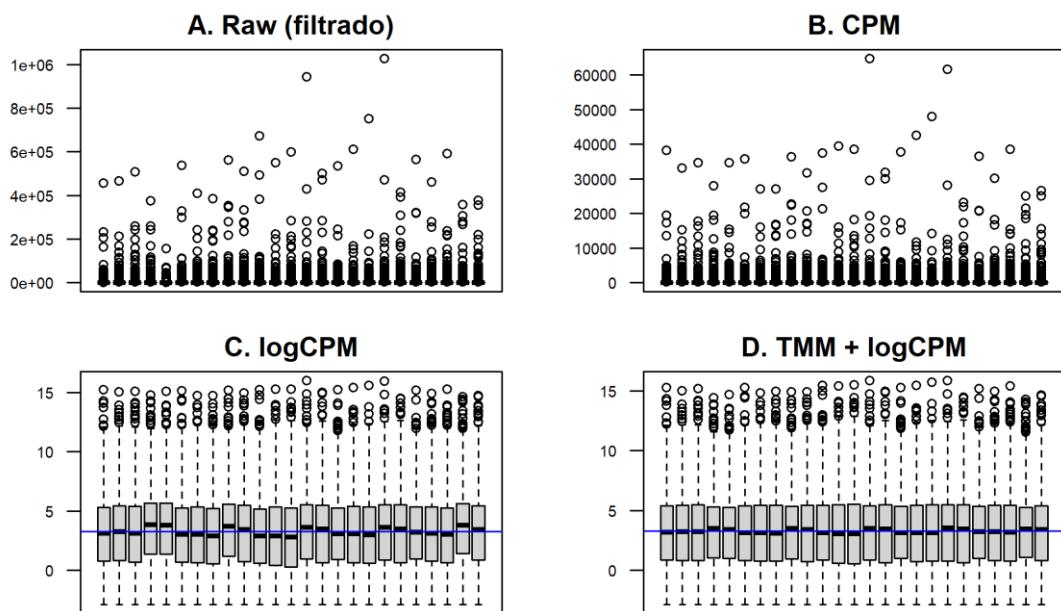


Figura 2. Boxplots comparativos (Raw vs CPM vs logCPM vs TMM + logCPM)

Tras aplicar TMM, las distribuciones de expresión entre muestras se alinean de forma más homogénea, especialmente en términos de mediana y dispersión global. Esto indica que se ha reducido la variabilidad técnica asociada a diferencias de composición entre bibliotecas.

En conjunto, los pasos de filtrado y normalización garantizan que las diferencias observadas en los análisis posteriores reflejen principalmente cambios biológicos asociados a las condiciones experimentales y no artefactos técnicos.

2.5. Análisis multivariante: PCA

Con el objetivo de identificar las principales fuentes de variabilidad en el conjunto de datos, se realizó un análisis de componentes principales (PCA) utilizando la matriz TMM-normalizada (logCPM) y posteriormente escalada por gen. El escalado permite que todos los genes contribuyan de forma comparable al análisis, independientemente de su nivel absoluto de expresión.

La figura 3 muestra la representación de las muestras en el espacio definido por las dos primeras componentes principales. La PC1 explica el 33,6% de la variabilidad total, mientras que la PC2 explica el 18%, sumando en conjunto más del 50% de la variabilidad del dataset.

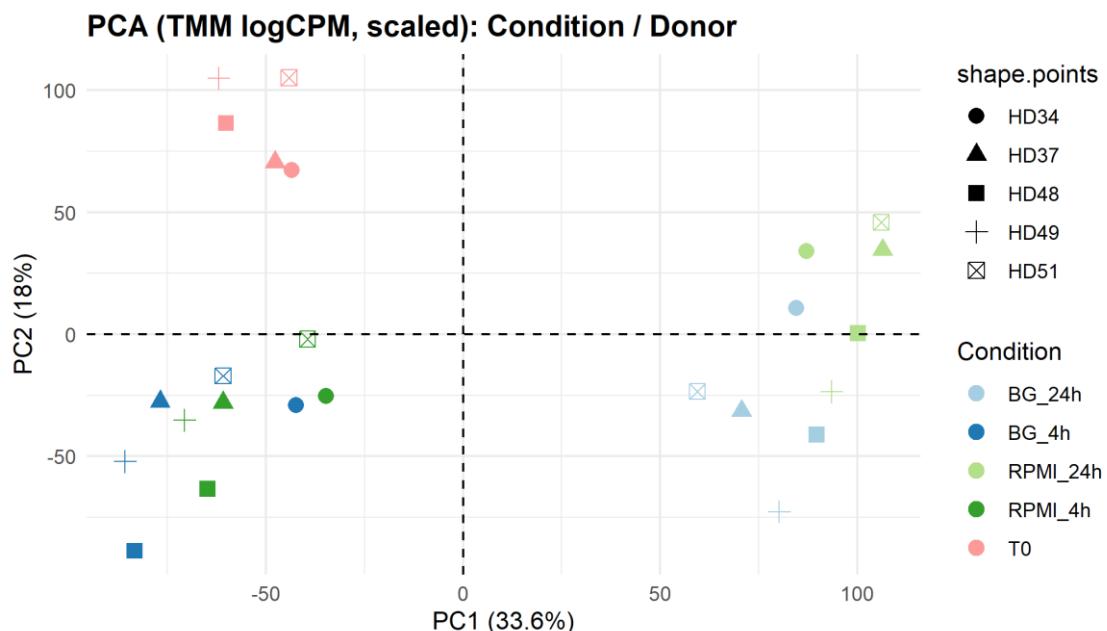


Figura 3. PCA (PC1 vvs PC2)

Se observa una separación clara a lo largo de PC1 entre las muestras correspondientes a T0 y las muestras correspondientes a 4h y 24h, tanto en condiciones RPMI como BG. Esto sugiere que la mayor fuente de variabilidad en el conjunto de datos está asociada al tiempo de estimulación y al proceso de activación celular respecto al estado basal.

En segundo lugar, la PC2 parece capturar diferencias adicionales relacionadas tanto con el tiempo (4h vs 24h) como con el tipo de estímulo (BG vs RPMI). Aunque

la separación entre BG y RPMI no es tan marcada como la observada con T0, se aprecia una tendencia a la diferenciación entre ambas condiciones dentro de cada punto temporal, con una mayor diferenciación en 24h.

Respecto a los donantes, las muestras no se organizan principalmente según el donante, sino que tienden a agruparse en función de la condición experimental. Esto sugiere que el efecto biológico asociado al estímulo y al tiempo de exposición tiene un peso mayor en la variabilidad global.

2.6. Matrices de distancia y correlación

Para complementar el PCA, se calcularon matrices de distancia euclídea y correlación de Pearson entre muestras. Estas aproximaciones permiten evaluar la similitud global entre muestras desde dos perspectivas distintas: la distancia absoluta en el espacio transcriptómico y la correlación lineal entre perfiles de expresión.

- Distancia euclídea.

La matriz muestra una clara estructuración de las muestras en bloques definidos principalmente por el tiempo y la condición experimental. Se observan tres agrupaciones principales correspondientes a T0, 4h y 24h.

Las muestras T0 presentan distancias menores entre sí y mayores respecto a las muestras estimuladas, lo que indica que el estado basal constituye un perfil transcriptómico diferenciado respecto da las condiciones post-estimulación.

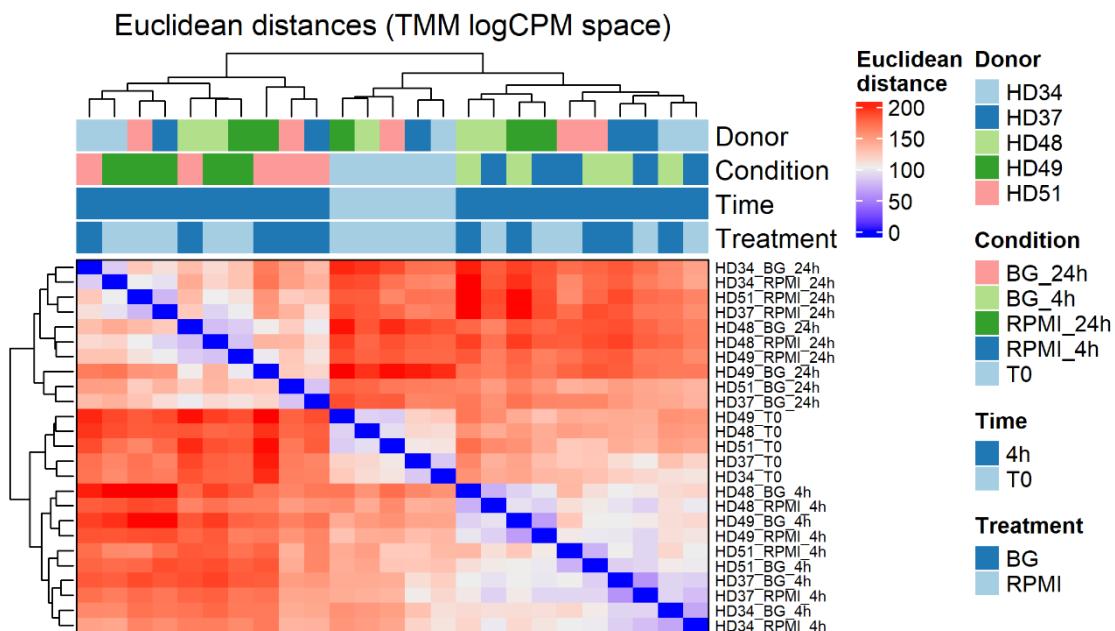


Figura 4. Heatmap de distancias euclídeas

- Correlación de Pearson

La matriz de correlación de Pearson confirma estos resultados desde una perspectiva complementaria. Las muestras pertenecientes a la misma condición presentan correlaciones más altas entre sí, mientras que las correlaciones disminuyen al comparar muestras de distintos tiempos o estados experimentales.

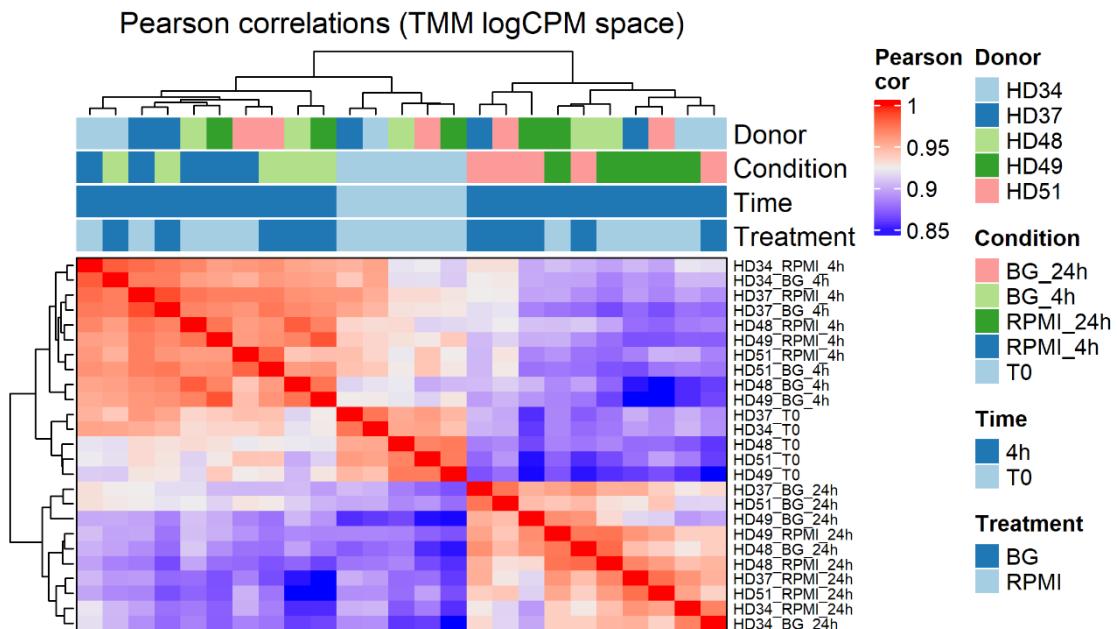


Figura 5. Heatmap de correlaciones de Pearson

Las muestras correspondientes a T0 muestran menor correlación con las muestras estimuladas, reforzando la idea de que el estímulo induce cambios transcriptómicos globales relevantes.

2.7. Análisis de expresión diferencial

La expresión diferencial se evaluó mediante DESeq2 utilizando un modelo que incluye el efecto donante como factor bloque:

$$\sim \text{Donor} + \text{Condition}$$

Este diseño permite controlar la variabilidad interindividual y evaluar específicamente el efecto del estímulo dentro de cada punto temporal.

Se realizaron los siguientes contrastes principales:

- BG_4h vs RPMI_4h
- BG_24h vs RPMI_24h

Se consideraron genes diferencialmente expresados aquellos con *adjusted p-value* < 0.1.

- **Contraste temprano: BG 4h vs RPMI 4h**

En el punto temprano (4 h), de un total de 16.604 genes con expresión detectable, se identificaron:

- 200 genes sobreexpresados (1,2%)
- 127 genes reprimidos (0,76%)

En conjunto, aproximadamente un 2% del transcriptoma mostró cambios significativos a las 4 h tras la estimulación con β -glucano. Estos resultados indican que, aunque ya existe una respuesta detectable en fases tempranas, el efecto transcriptómico global es todavía relativamente moderado.

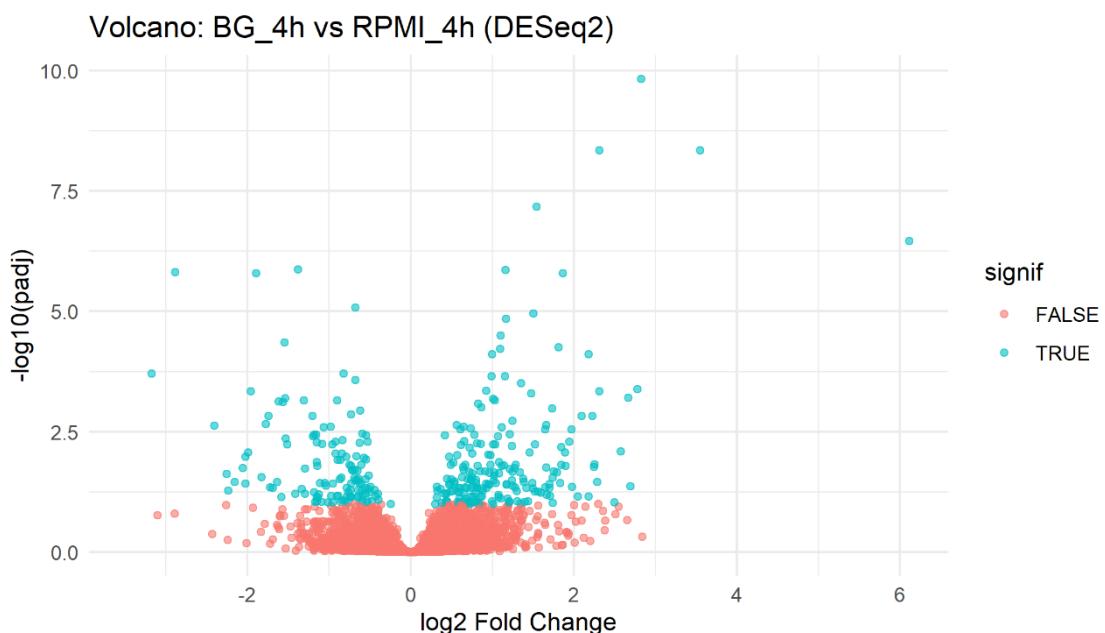


Figura 6. Volcano plot BG 4h vs RPMI 4h

El volcán plot muestra una distribución relativamente concentrada alrededor de $\log_{2}FC \approx 0$, con un número moderado de genes superando el umbral de significación.

La mayoría de los cambios presentan magnitudes moderadas de \log_{2} fold change, aunque algunos genes alcanzan valores superiores a ± 2 . Esto sugiere que la respuesta transcripcional temprana al β -glucano es específica pero todavía limitada en extensión.

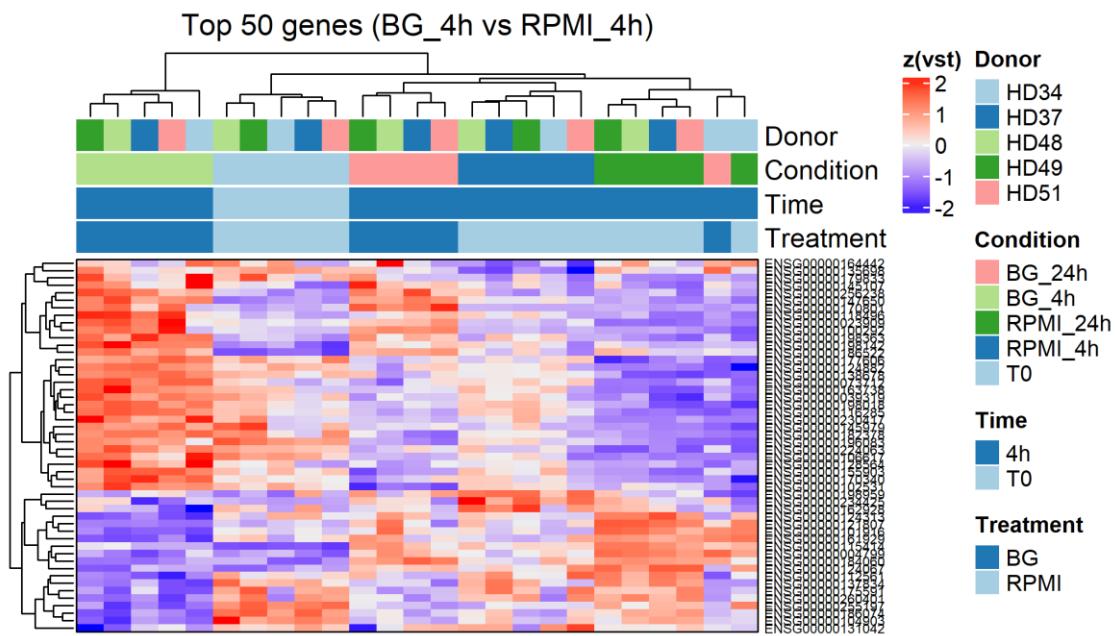


Figura 7. Heatmap de los 50 genes más significativos (BG 4h vs RPMI 4h)

El heatmap de los 50 genes más significativos muestra una separación clara entre BG 4h y RPMI 4h. Las muestras se agrupan principalmente por condición experimental más que por donante, lo que indica que el efecto del estímulo ya es detectable a las 4 horas.

- **Contraste temprano: BG 24h vs RPMI 24h**

En contraste, a las 24 h se observó un aumento notable en el número de genes diferencialmente expresados. En este punto temporal se identificaron:

- 2015 genes sobreexpresados (12%).
- 1710 genes reprimidos (10%).

En total, aproximadamente un 22% del transcriptoma analizado mostró cambios significativos.

Este incremento sustancial en el número de genes regulados sugiere que la respuesta transcripcional inducida por β -glucano se intensifica con el tiempo, generando una reprogramación más amplia del perfil de expresión génica.

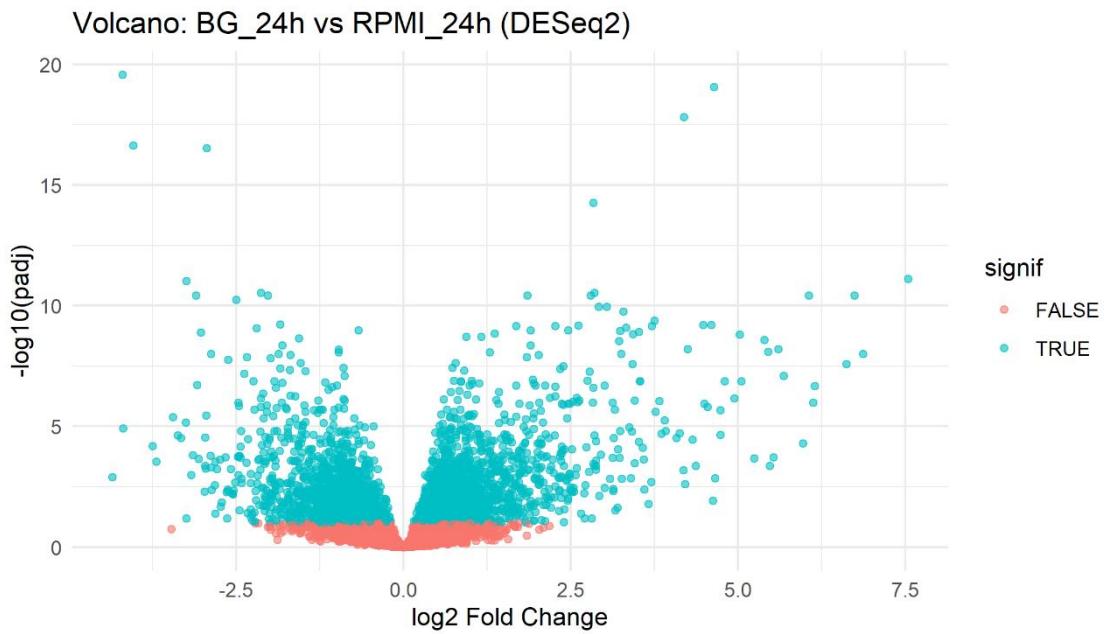


Figura 8. Volcano plot BG 24h vs RPMI 24h.

El volcano plot correspondiente a 24 h muestra una expansión clara de puntos significativos tanto en dirección positiva como negativa. Se observan numerosos genes con log2 fold change elevados (algunos superiores a 5–7), junto con valores de significación muy altos ($-\log_{10}$ padj elevados).

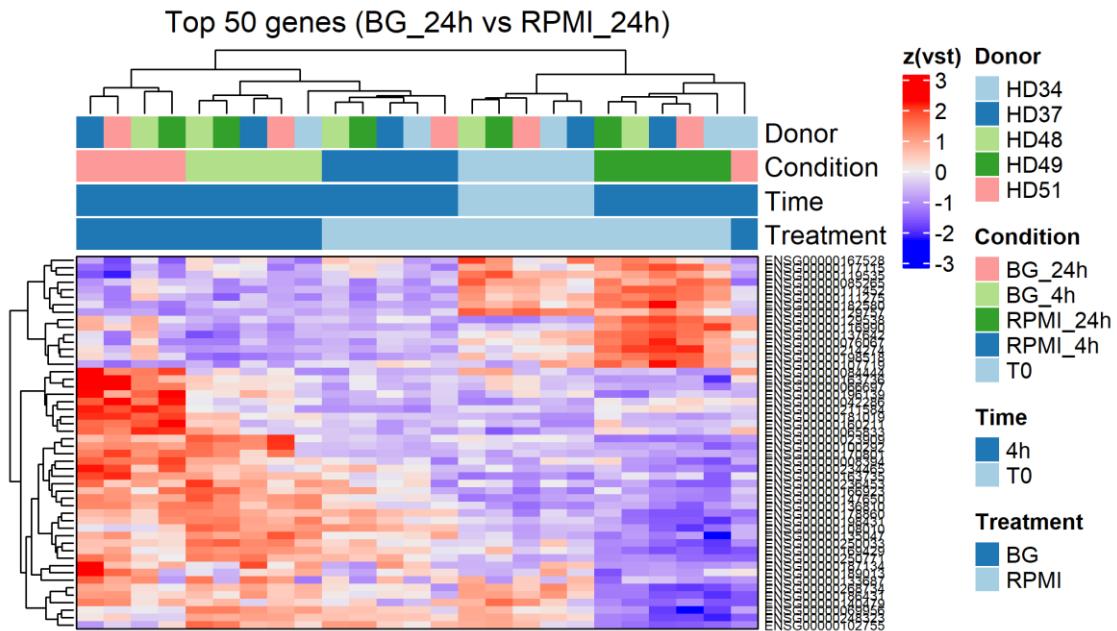


Figura 9. Heatmap de los 50 genes más significativos (BG 24h vs RPMI 24h)

El heatmap muestra una separación muy marcada entre BG_24h y RPMI_24h. Las muestras se agrupan claramente por condición experimental, con una homogeneidad interna notable dentro de cada grupo.

Se distinguen dos grandes bloques, siendo la separación mucho más definida que en el contraste de 4 h, lo que confirma que la respuesta al estímulo se intensifica con el tiempo.

3. Métodos – Tabla de recursos

Software y herramientas bioinformáticas utilizadas

Todos los análisis bioinformáticos realizados en este trabajo fueron implementados en el entorno R (versión 4.5.2) bajo Windows 11 x64. Se emplearon paquetes del ecosistema Bioconductor y CRAN para el preprocesamiento, normalización, análisis multivariante y expresión diferencial. A continuación se detallan los principales recursos computacionales utilizados.

| Recurso | Versión | Fuente | Uso en el análisis |
|----------------|--------------------|--|--|
| R | 4.5.2 (2025-10-31) | R Foundation for Statistical Computing | Entorno principal de análisis estadístico |
| RStudio | 2026.01.0+392 | Posit Software | Entorno de desarrollo |
| DESeq2 | 1.50.2 | Bioconductor | Análisis de expresión diferencial |
| edgeR | 4.8.2 | Bioconductor | Filtrado de baja expresión y normalización TMM |
| limma | 3.66.0 | Bioconductor | Dependencias y soporte estadístico |
| ComplexHeatmap | 2.26.1 | Bioconductor | Visualización de heatmaps |
| circlize | 0.4.17 | CRAN | Soporte gráfico para heatmaps |
| ggplot2 | 4.0.2 | CRAN | Visualización gráfica (PCA, volcano plots) |
| ggpubr | 0.6.2 | CRAN | Composición de figuras |
| factoextra | 1.0.7 | CRAN | Extracción de varianza explicada en PCA |
| data.table | 1.18.2.1 | CRAN | Lectura eficiente de archivos mmseq |
| dplyr | 1.2.0 | CRAN | Manipulación de datos |
| stringr | 1.6.0 | CRAN | Procesamiento de cadenas |
| purrr | 1.2.1 | CRAN | Iteración funcional |
| tibble | 3.3.1 | CRAN | Manejo de data frames |
| rmarkdown | 2.30 | CRAN | Generación del informe en PDF |
| knitr | 1.51 | CRAN | Ejecución reproducible del código |