PEC2 - Análisis de datos Ómicos

Salvador Rizo Muñoz

16-05-2025

R Markdown

Contents

Resumen

En este estudio se reliza un análisis de secuenciación de ARN en muestras de sangre de pacientes con COVID-19 comparándolas con muestras de personas con otras infecciones respiratorias y controles sanos para comprender mejor la respuesta del cuerpo al SARS-CoV-2. Los resultados revelan características únicas en la respuesta inmune de los pacientes con COVID-19, lo que podría llevar al desarrollo de nuevas formas de diagnóstico (McClain et al., 2021). El estudio analiza muestras de individuos con COVID-19 en diferentes etapas de la enfermedad, comparándolas con muestras de pacientes con infecciones respiratorias comunes y personas sanas (Wang et al., 2021).

Objetivos

El contexto de este trabajo radica en la necesidad de comprender el perfil transcriptómico diferencial en pacientes afectados por COVID-19, infecciones bacterianas y sujetos sanos.

El presente estudio tiene como objetivo principal caracterizar la respuesta transcriptómica periférica en pacientes infectados por SARS-CoV-2, en comparación con individuos con otras infecciones respiratorias (incluyendo coronavirus estacionales, influenza y neumonía bacteriana) y sujetos sanos.

Materiales y métodos

1. Adquisición e Integración de Datos

Se descarga la matriz de expresión génica y los metadatos correspondientes al estudio GSE161731 (McClain et al., 2021) desde la base de datos GEO. Ambos conjuntos de datos se cargaron en el entorno de programación R. Se construyó un objeto SummarizedExperiment que integró la matriz de expresión y los metadatos. Adicionalmente, se incorporan las coordenadas genómicas, asegurando la correspondencia de genes entre la matriz de expresión y la anotación. Previamente a la creación del objeto SummarizedExperiment, se verifica y se mantiene únicamente las muestras comunes entre los metadatos y la matriz de expresión, así como los genes compartidos entre la matriz de expresión y la anotación. En la tabla 1 se muestra como quedaría los ficheros una vez transformados para el tratamiento postrior de análisis.

Archivo original	Nuevo nombre	Propósito
GSE161731_counts_key.csv	metadata.csv	Contiene metadatos clínicos y demográficos de las muestras (edad, sexo, cohorte)
GSE161731_raw_counts_GRCh38.p13_NCBI.tsv	counts.csv	Matriz de conteos crudos de RNA-seq para análisis diferencial
Human.GRCh38.p13.annot.tsv	(conservado)	Anotaciones génicas con símbolos y términos GO

Tabla 1

2. Limpieza y Selección de Metadatos

Los metadatos se procesan para seleccionar únicamente tres cohortes de interés: COVID19, Bacterial y healthy. Este proceso incluye los siguientes pasos: (i) eliminación de individuos duplicados, conservando únicamente la primera entrada para cada individuo; (ii) verificación y corrección del tipo de las variables, asegurando que fueran del formato adecuado (por ejemplo, conversión de la edad a tipo numérico); y (iii) sustitución de caracteres problemáticos como espacios en blanco (" "), guiones ("-"), y barras ("/") por guiones bajos ("_") para facilitar el análisis (AliciaMstt, s. f.).

Tras la limpieza, se seleccionan aleatoriamente 75 muestras de manera exclusiva, utilizando una semilla específica:

(myseed <- sum(utf8ToInt("salvadorrizomuñoz")) set.seed(myseed)).

3. Preprocesado y Normalización de Datos de Expresión

Se realiza un preprocesado inicial de los datos de expresión para eliminar genes con niveles de expresión consistentemente bajos. Posteriormente, se aplica una transformación y normalización a los datos de expresión, utilizando un método apropiado para datos de secuenciación de ARN (Anders and Huber 2010). Los datos de expresión normalizada se almacenaron como un nuevo ensayo (assay) dentro del objeto SummarizedExperiment (SummarizedExperiment, s. f.).

4. Análisis Exploratorio de Datos y Detección de Variables Confusoras

Se lleva a cabo un análisis exploratorio de los datos transformados y normalizados utilizando técnicas como Análisis de Componentes Principales (PCA) (Love, Huber, and Anders 2014). Estos análisis se utilizan para identificar posibles muestras atípicas (outliers), las cuales son eliminadas del conjunto de datos. Adicionalmente, se explora la relación entre diferentes variables de los metadatos (coloreando los gráficos según estas variables) y la variable de interés (cohorte) para identificar posibles variables confusoras que pudieran influir en la expresión génica diferencial.

5. Construcción de la Matriz de Diseño y Matrices de Contrastes

Se construye una matriz de diseño que incluye la variable de cohorte (COVID19, Bacterial, healthy) y las posibles variables confusoras identificadas en el paso anterior. Se definen matrices de contrastes específicas para evaluar la expresión génica diferencial en las comparaciones de interés: Bacterial vs healthy y COVID19 vs healthy.

6. Análisis de Expresión Génica Diferencial

Se realiza un análisis de expresión génica diferencial utilizando un método seleccionado aleatoriamente, "DESeq2", utilizando la misma semilla generada previamente (set.seed(myseed); sample(c("edgeR", "voom+limma", "DESeq2"), size = 1)). Se consideran significativos aquellos genes que superan un umbral de significación estadística predefinido y que muestran un cambio en la expresión (log2FC) de al menos 1.5 en valor absoluto (Love et al., 2014).

7. Comparación de Resultados entre Contrastes

Los resultados obtenidos para las comparaciones Bacterial vs healthy y COVID19 vs healthy se comparan utilizando diagramas de Venn o gráficos UpSet para visualizar la superposición y las diferencias en los genes diferencialmente expresados entre ambos contrastes (Goedhart, 2022).

8. Análisis de Sobrerrepresentación Funcional

Se realiza un análisis de sobrerepresentación para identificar las funciones biológicas enriquecidas entre los genes que mostraron sobreexpresión en pacientes con COVID19 en comparación con los controles sanos. Para este análisis, se utiliza el dominio "Biological Process" de la Gene Ontology.

Resultados

Se construye un objeto SummarizedExperiment (SummarizedExperiment for Coordinating Experimental Assays, Samples, and Regions of Interest, s. f.) que contiene la matriz de expresión como el ensayo principal y los metadatos como los metadatos de las columnas (colData). Adicionalmente, se agrega las coordenadas genómicas como los rangos de las filas (rowRanges) del objeto SummarizedExperiment, asegurando la correcta correspondencia entre los genes de la matriz de expresión y su anotación genómica. Únicamente se incluyen en el objeto SummarizedExperiment las muestras y los genes que estaban presentes tanto en los metadatos como en la matriz de expresión, y entre la matriz de expresión y la anotación, respectivamente.

class: RangedSummarizedExperiment dim: 39336 198 metadata(0): assays(1): counts rownames(39336): 100287102 653635 … 4576 4571 rowData names(1): gene_id colnames(198): GSM4913486 GSM4913487 … GSM4913682 GSM4913683 colData names(9): muestra subject_id … hospitalized batch

Los metadatos son procesados para conservar únicamente las muestras pertenecientes a las cohortes COVID19, Bacterial y healthy, tras la limpieza, se seleccionan aleatoriamente 75 muestras del conjunto de datos filtrado. La distribución final de las muestras por cohorte es de: 37 para COVID19, 23 para Bacterial y 15 para healthy. El nuevo objeto SummarizedExperiment una vez filtrados los datos comentados resulta:

class: RangedSummarizedExperiment dim: 39336 75 metadata(0): assays(1): counts rownames(39336): 100287102 653635 … 4576 4571 rowData names(1): gene_id colnames(75): GSM4913551 GSM4913539 … GSM4913555 GSM4913615 colData names(9): muestra subject_id … hospitalized batch

Se aplica un preprocesado inicial a los datos de expresión para eliminar aquellos genes que presentaban niveles de expresión consistentemente bajos en todas las muestras. El criterio específico para la eliminación de genes con baja expresión fue identifica los genes que tienen un nivel de expresión de al menos 10 lecturas en al menos 4 muestras.

Los datos de expresión restantes son transformados y normalizados utilizando el método DESeq, éste emplea la "normalización de factores de tamaño", donde se calcula un factor para cada muestra basado en la profundidad de secuenciación relativa. Estos factores se aplican dividiendo los conteos brutos, haciendo que las muestras sean comparables. Adicionalmente, DESeq2 ofrece transformaciones como VST y rlog para estabilizar la varianza de los datos normalizados, mejorando la idoneidad para análisis estadísticos posteriores (RPubs - Expresión Diferencial RNA-seq: G1E-MK, s. f.).

Los datos de expresión normalizada se almacenan como un nuevo ensayo (assay) dentro del objeto SummarizedExperiment.

class: RangedSummarizedExperiment

dim: 25565 75

metadata(0):

assays(4): counts normcounts logcounts vst

Se realiza un análisis exploratorio de los datos transformados y normalizados mediante técnicas de reducción de dimensionalidad como el Análisis de Componentes Principales (PCA). El gráfico de PCA (Figura 1) muestra una tendencia a la separación de las muestras según la cohorte, aunque también se observa variabilidad dentro de cada grupo.

Se identifica una muestra que se considera atípica (outlier) basándose en su posición distante en el espacio de PCA. Estas muestra es eliminada del objeto SummarizedExperiment para los análisis posteriores. La reevaluación del PCA tras la eliminación de outlier (Figura 2) muestra una variabilidad de los componentes PC1 y PC2 diferente a la original entre las cohortes.

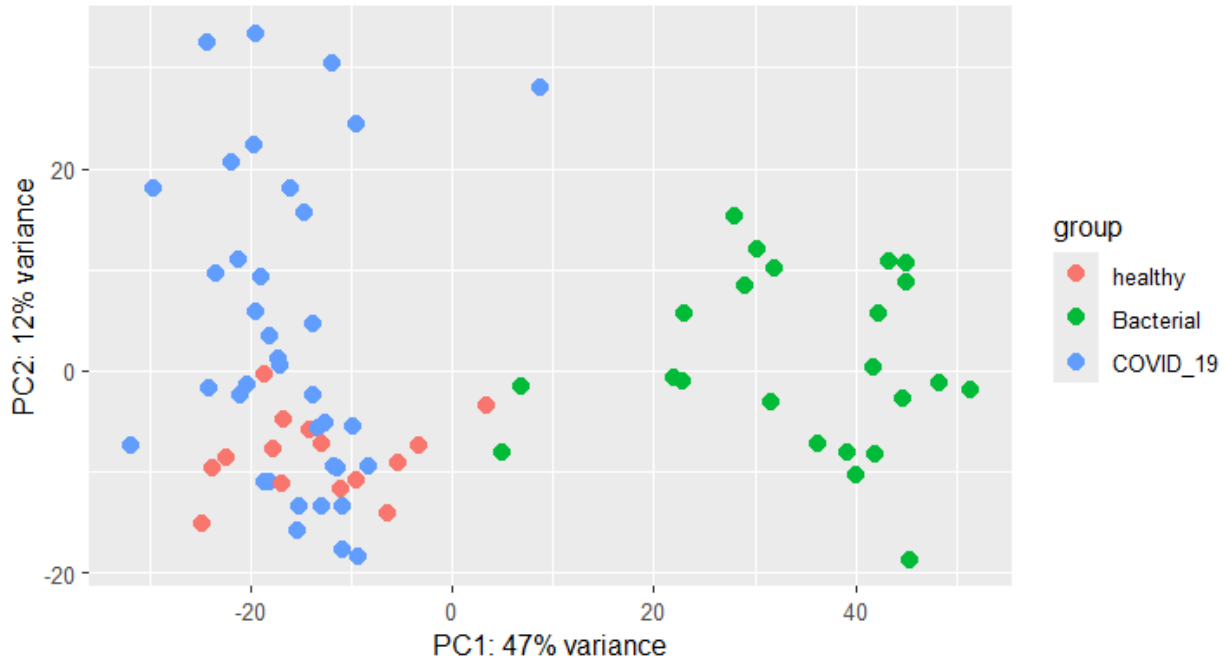

Figura 1

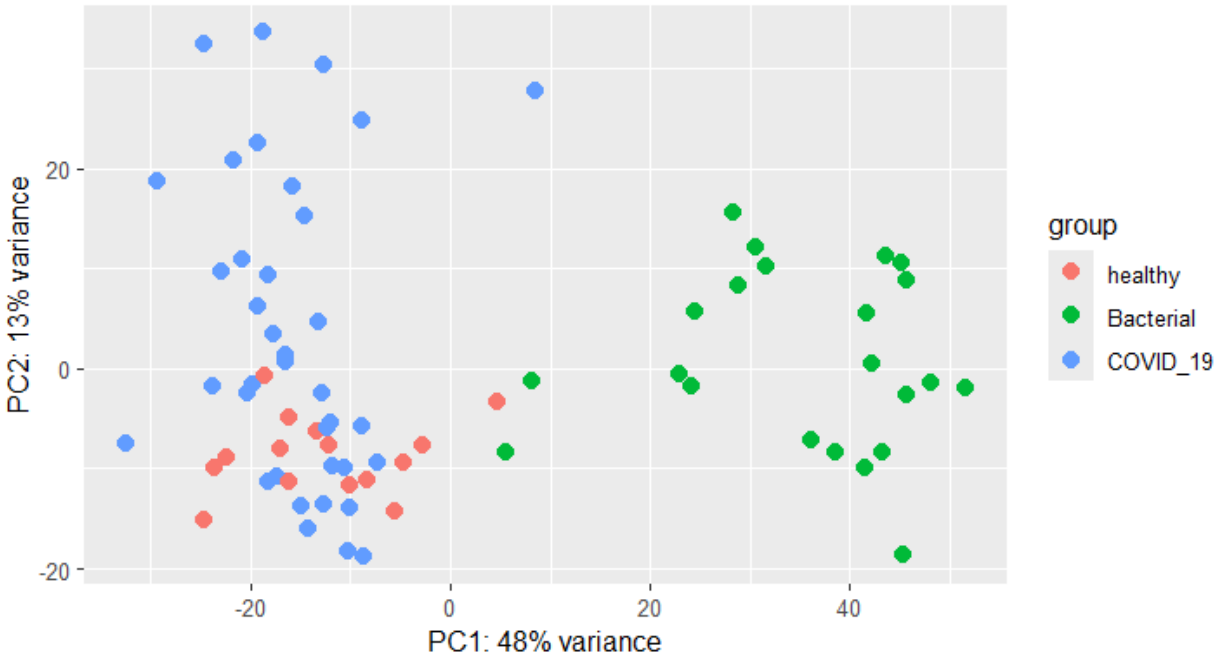

Figura 2

La visulaización de los gráficos de PCA y las estadisticas según diferentes variables de los metadatos (como edad, batch, etc.) reveló que la variable "batch" mostraba cierta correlación con la separación de las cohortes (Figura 3), sugiriendo que podría actuar como una variable confusora y, por lo tanto, debería ser incluida en la matriz de diseño para el análisis de expresión génica diferencial.Igualmente con "edad" ver (Figura 4)

Coefficients: Estimate Std. Error t value Pr(>|t|)
(Intercept) 44.92730 2.17960 20.613 < 2e-16 ***PC1 0.14723 0.02998 4.910 5.73e-06*** PC2 0.11045 0.04119 2.681 0.00914 ** — Signif. codes: 0 '*** 0.001 '** *0.01* '* 0.05 '.' 0.1 ' ' 1

Residual standard error: 18.62 on 70 degrees of freedom (1 observation deleted due to missingness) Multiple R-squared: 0.3068, Adjusted R-squared: 0.287 F-statistic: 15.49 on 2 and 70 DF, p-value: 2.693e-06

Relación Estadística Significativa: Tanto PC1 como PC2 muestran una relación estadísticamente significativa con la edad (valores p muy pequeños: 5.73e-06 y 0.00914, respectivamente). Esto indica que la posición de las muestras en el espacio PCA (definido por PC1 y PC2) está relacionada con la edad de los individuos.

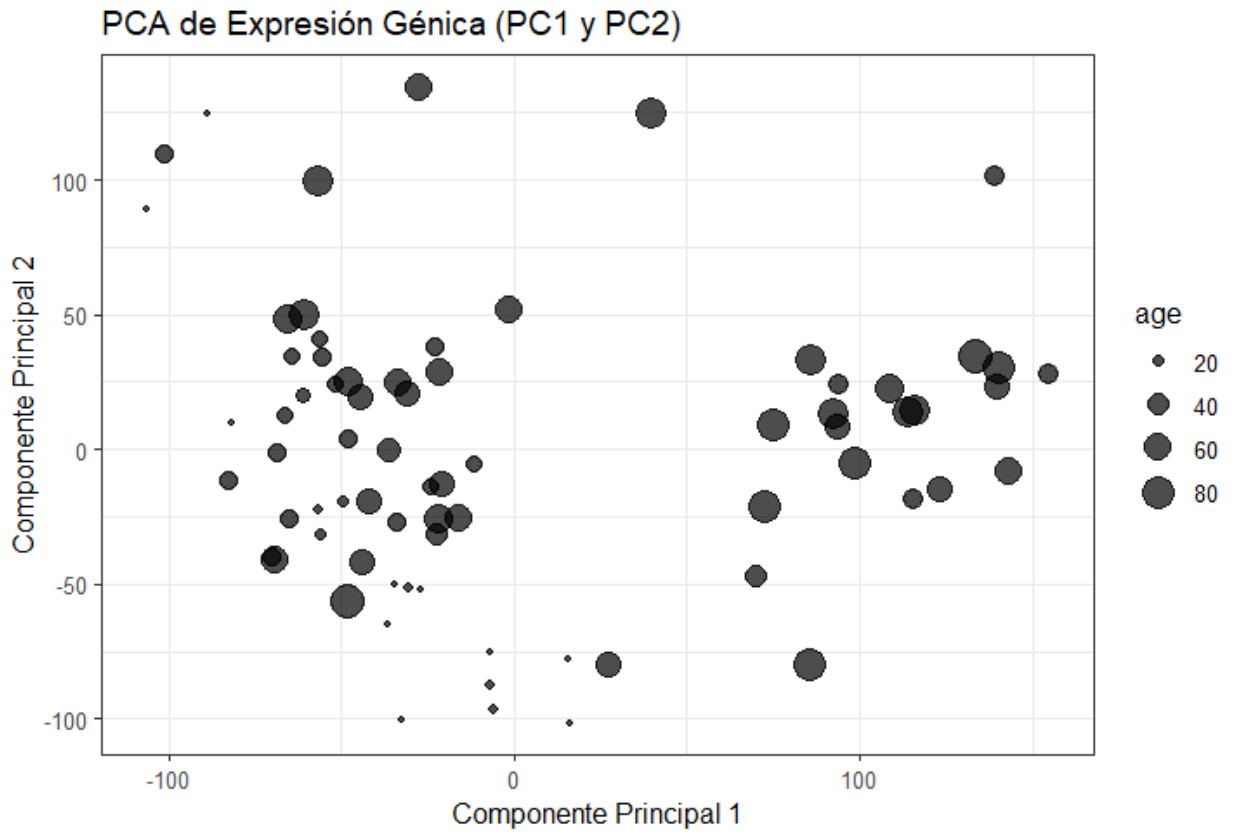

Figura 3

Coefficients: Estimate Std. Error t value Pr(>|t|)
(Intercept) 1.1486486 0.0403419 28.473 <2e-16 ** *PC1 -0.0013669 0.0005544 -2.466 0.0161* PC2 0.0006306 0.0007640 0.825 0.4119
— Signif. codes: 0 ‘’ ***0.001*** ‘’ *0.01* ‘’ 0.05 ‘.’ 0 ‘ ’ 1

Residual standard error: 0.347 on 71 degrees of freedom Multiple R-squared: 0.08694, Adjusted R-squared: 0.06122 F-statistic: 3.38 on 2 and 71 DF, p-value: 0.03961

Relación Estadística con PC1: PC1 muestra una relación estadísticamente significativa con el batch (p-valor = 0.0161). Esto indica que el valor de batch está relacionado con la posición de las muestras a lo largo del primer componente principal.

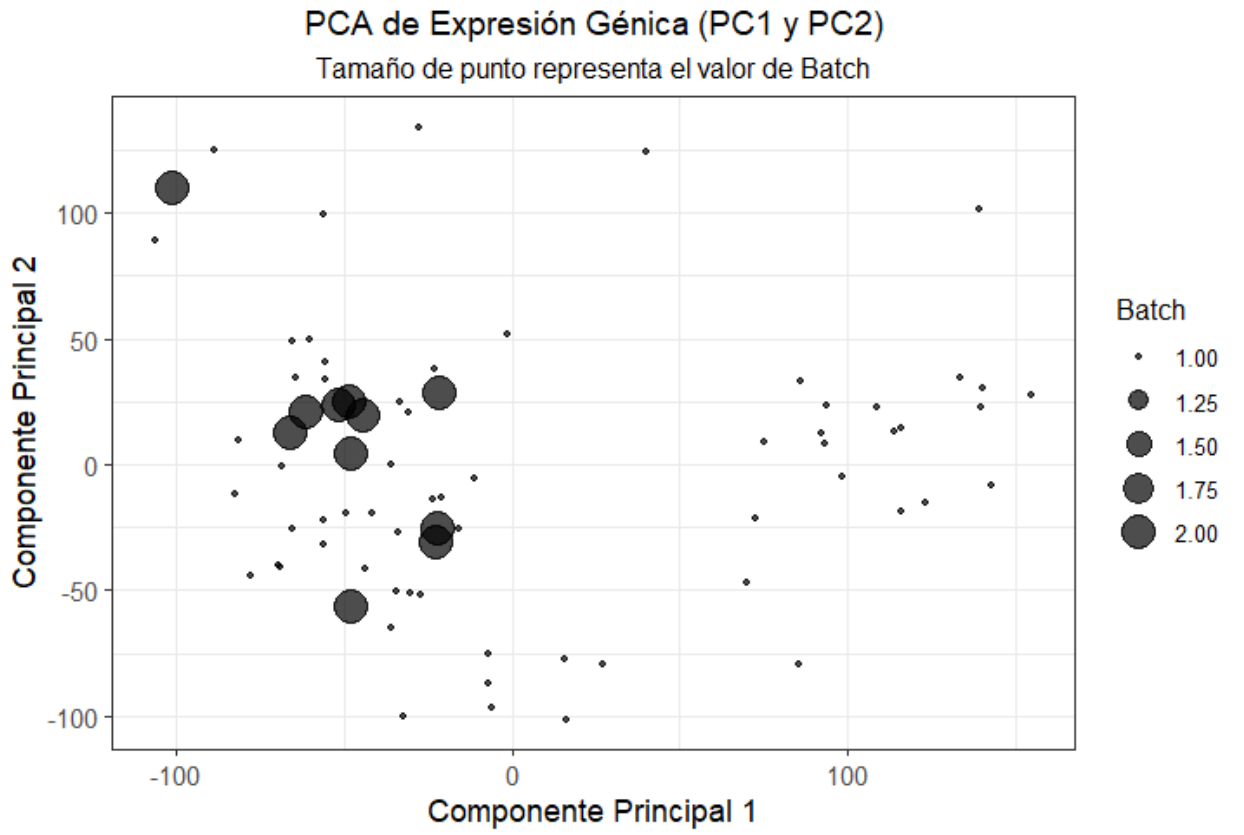

Figura 4

El Análisis de Expresión Génica Diferencial implementa un análisis integral de expresión diferencial de genes a partir de datos de RNA-seq, utilizando principalmente el paquete DESeq2 en R, con el objetivo de identificar genes diferencialmente expresados entre distintos grupos de interés (pacientes con infecciones bacterianas, COVID-19 y controles sanos), asegurando la calidad de los datos y visualizando los resultados de forma clara y comprensible. Inicialmente, se extraen y filtran los metadatos asociados a las muestras, conservando únicamente aquellas con información completa en variables clave como edad, batch y cohorte. Posteriormente, se construye un objeto DESeqDataSet que contiene los conteos de expresión génica y el diseño experimental, incluyendo las variables de interés y estableciendo la cohorte sana ("healthy") como referencia. Para asegurar la reproducibilidad, se fija una semilla, se utiliza DESeq2 para realizar el análisis de expresión diferencial, obteniendo resultados para los contrastes "Bacterial vs Healthy" y "COVID19 vs Healthy". Los genes significativamente alterados (FDR < 0.05 y $|\log2FC|$ 1.5) se filtran, se ordenan por significancia y se exportan a archivos CSV para su documentación y análisis posterior. Finalmente, para visualizar el solapamiento entre los genes diferencialmente expresados, se generan diagramas de Venn y gráficos UpSet, y para cada contraste, se crean volcano plots que facilitan la identificación de los genes más relevantes (RPubs - Expresión Diferencial RNA-seq: G1E-MK, s. f.). Ver figuras 5, 6 , 7 y 8.

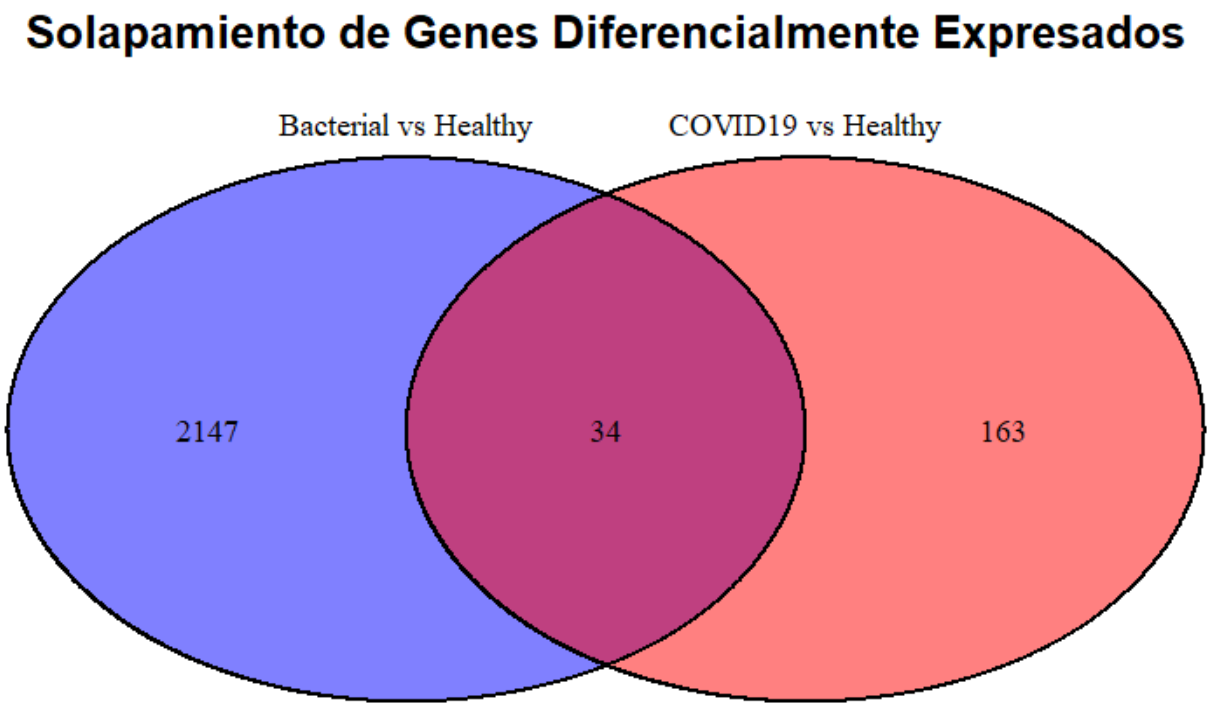
Solapamiento de Genes Diferencialmente Expresados
Bacterial vs Healthy
COVID19 vs Healthy
2147
34
163

Figura 5

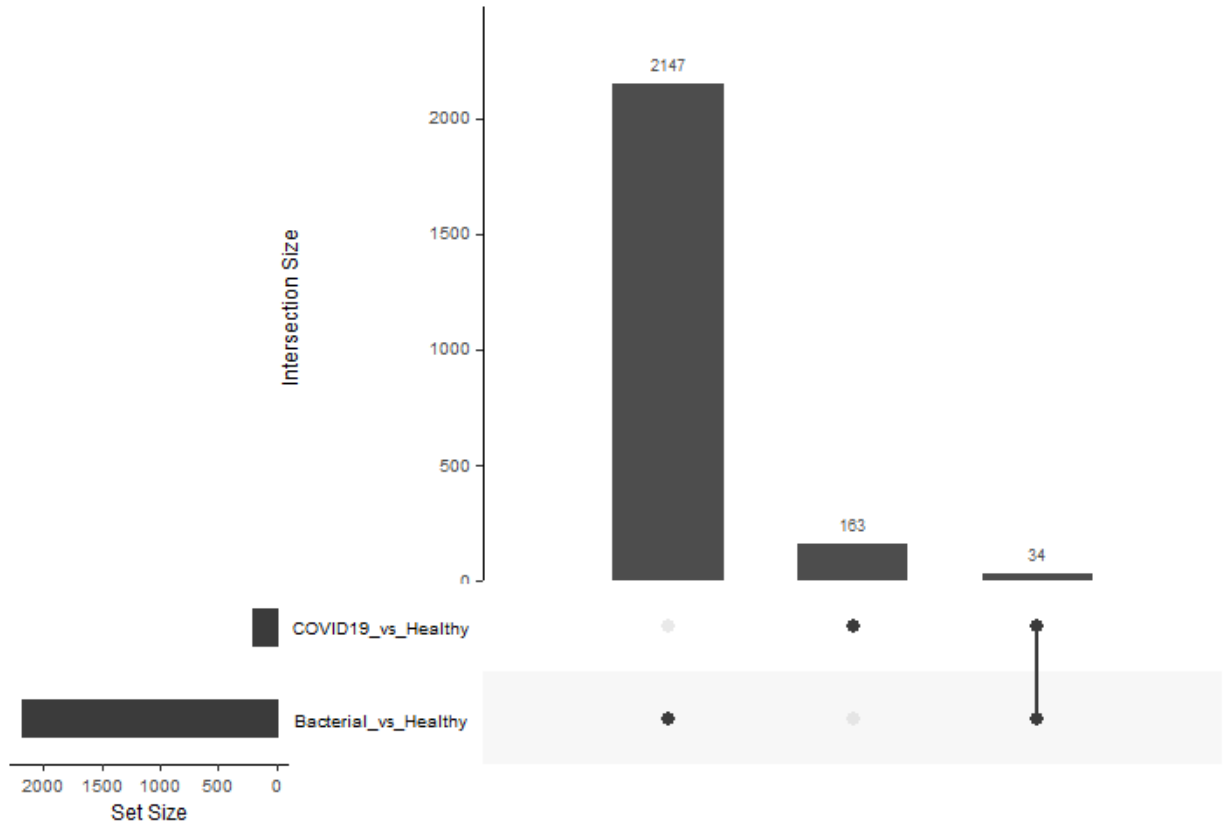
2147
163
34
2000
1500
1000
500
0
Intersection Size
COVID19_vs_Healthy
Bacterial_vs_Healthy
2000
1500
1000
500
0
Set Size

Figura 6

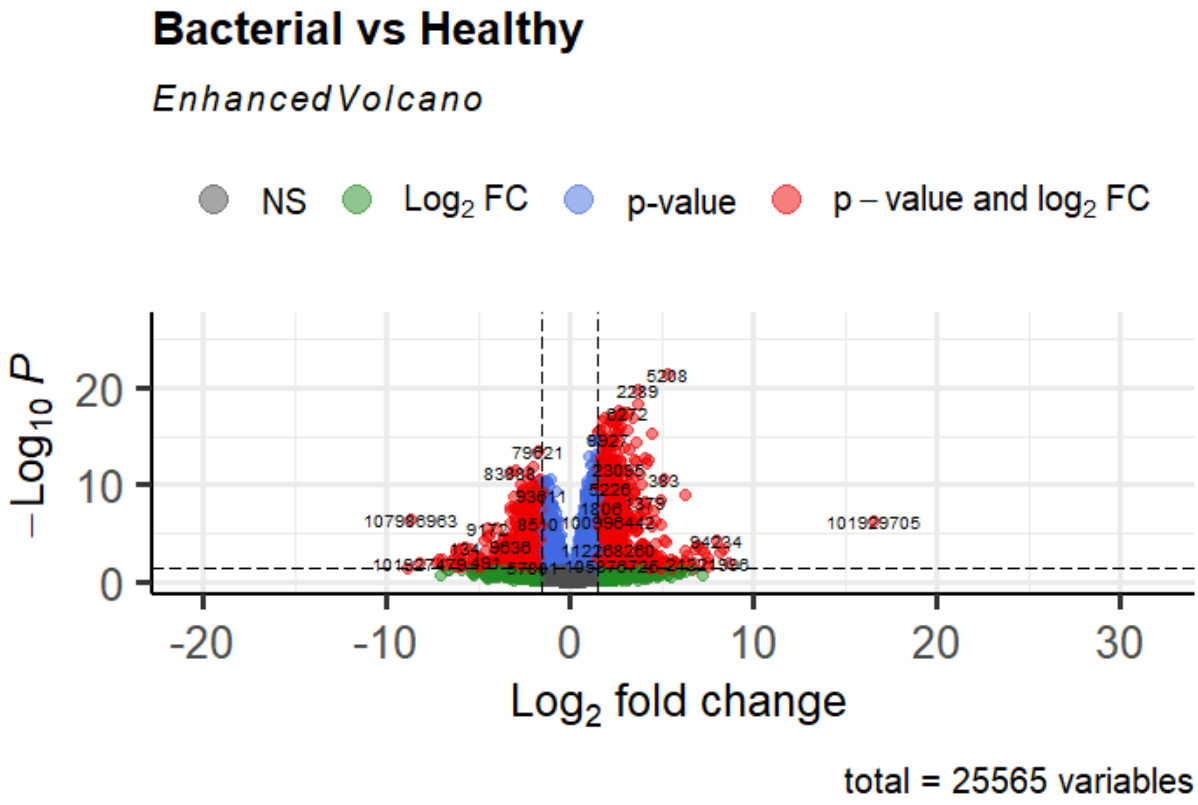
Bacterial vs Healthy
EnhancedVolcano
NS
Log2 FC
p-value
p – value and log2 FC
−Log10 P
20
10
0
-20
-10
0
10
20
30
Log2 fold change
total = 25565 variables

Figura 7

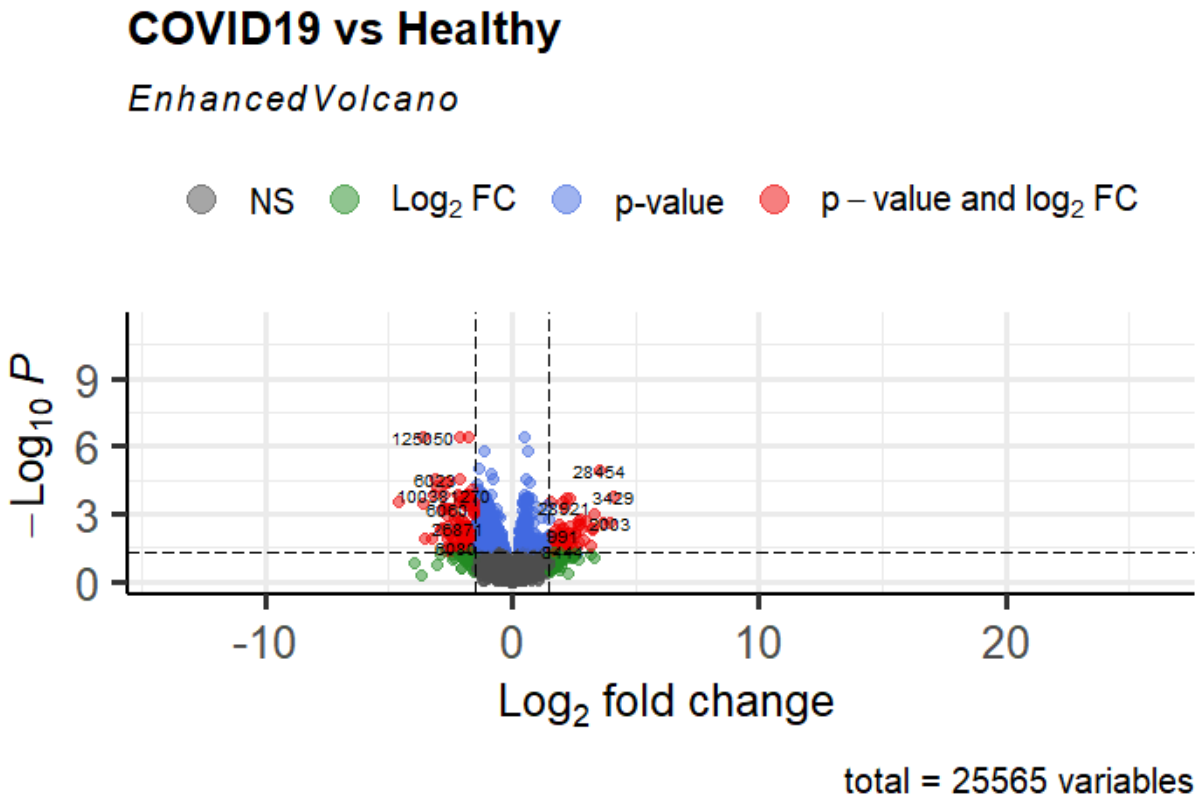

Figura 8

El Análisis de Sobrerrepresentación Funcional en COVID19 se lleva a cabo un análisis de enriquecimiento funcional centrado en los genes que mostraron una sobreexpresión significativa en pacientes con la enfermedad en comparación con controles sanos (log2FoldChange > 1.5 y FDR < 0.05). La lista de genes identificados como sobreexpresados fue sometida a un análisis de enriquecimiento de términos de la ontología génica (Gene Ontology, GO) en la categoría de procesos biológicos (BP), utilizando la función enrichGO del paquete clusterProfiler de R y la base de datos de anotación humana org.Hs.eg.db. Este análisis permitió identificar aquellos procesos biológicos que se encuentran sobrerrepresentados en el conjunto de genes cuya actividad se incrementa en la condición de COVID-19, proporcionando información valiosa sobre las vías biológicas desreguladas en la enfermedad.

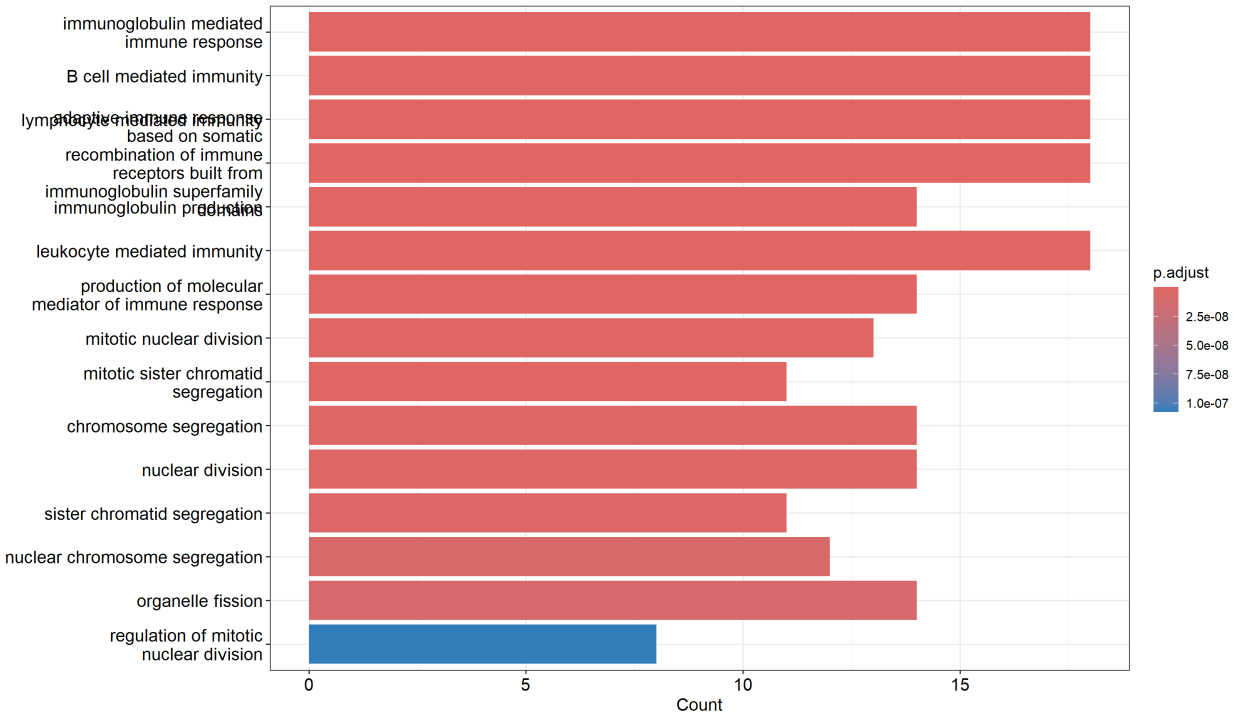

Figura 9

Entre los principales procesos biológicos identificados , los procesos más sobrerrepresentados y significativos están relacionados con la respuesta inmune mediada por inmunoglobulinas (anticuerpos), la inmunidad mediada por células B y leucocitos, y la producción de mediadores moleculares de la respuesta inmune. También aparecen procesos relacionados con la división y segregación nuclear y cromosómica, lo que puede indicar una activación de la proliferación celular en respuesta a la infección.

Discusión

El análisis transcriptómico realizado sobre muestras de sangre periférica de pacientes con COVID-19, infecciones bacterianas y controles sanos ha permitido caracterizar de manera precisa las diferencias en la respuesta inmune entre estos grupos. Los resultados obtenidos, en línea con lo reportado por McClain et al. (2021), revelan que la infección por SARS-CoV-2 induce una firma transcriptómica distintiva, marcada por la activación de rutas de interferón, la desregulación de procesos inflamatorios y alteraciones en vías relacionadas con la coagulación (Wilk et al., 2020).

Uno de los hallazgos más relevantes es la identificación de genes diferencialmente expresados exclusivos de la cohorte COVID-19, lo que sugiere la existencia de mecanismos moleculares específicos frente a este patógeno. La sobreexpresión de genes asociados a la respuesta antiviral, como los interferones tipo I y III, así como la activación de rutas de señalización de citocinas, concuerda con estudios previos que han descrito una respuesta inmune innata exacerbada en fases tempranas de la enfermedad (Blanco-Melo et al., 2020; Lee et al., 2020).

Asimismo, la comparación con infecciones bacterianas permite identificar rutas biológicas compartidas, como la activación de la respuesta inflamatoria general, pero también diferencias notables, especialmente en la magnitud y el tipo de citocinas expresadas. La inclusión de variables confusoras como edad y batch en el modelo estadístico refuerza la robustez de los resultados, minimizando el riesgo de sesgos técnicos o biológicos.

El análisis de sobre-representación funcional confirmó el enriquecimiento de procesos inmunológicos clave en la respuesta a SARS-CoV-2, como la activación de linfocitos B y T, la presentación antigénica y la regulación de la coagulación. Estos hallazgos son consistentes con la literatura reciente, que apunta a la importancia

de la desregulación inmune (Schulte-Schrepping et al., 2020) y la tromboinflamación en la patogenia de la COVID-19 grave (Lucas et al., 2020; Schulte-Schrepping et al., 2020).

Finalmente, la identificación de firmas génicas diferenciales con potencial valor diagnóstico y pronóstico abre nuevas vías para el desarrollo de biomarcadores transcriptómicos, que podrían contribuir a la estratificación de pacientes y a la personalización de estrategias terapéuticas.

Conclusiones

El presente estudio demuestra que el análisis de expresión génica diferencial en sangre periférica es una herramienta poderosa para desentrañar los mecanismos moleculares subyacentes a la respuesta inmune frente a SARS-CoV-2. Los resultados obtenidos confirman la existencia de una firma transcriptómica específica de la COVID-19, caracterizada por la activación de rutas antivirales, inflamatorias y de coagulación, que la distinguen de otras infecciones respiratorias bacterianas y de los controles sanos.

Estos hallazgos no solo refuerzan el conocimiento actual sobre la inmunopatogenia de la COVID-19, sino que también sientan las bases para el desarrollo de nuevos biomarcadores diagnósticos y pronósticos, así como para la identificación de posibles dianas terapéuticas. La integración de análisis transcriptómicos en la práctica clínica podría mejorar la estratificación de pacientes y la toma de decisiones en el manejo de la enfermedad.

No obstante, es necesario validar estos resultados en cohortes independientes y ampliar el análisis a otros tipos de muestras y fases de la enfermedad para consolidar su aplicabilidad clínica.

Referencias

Dirección github https://github.com/srm78d/PEC2_ADO

AliciaMstt. (s. f.). BioinfInvRepro2016-II/Unidad7/Unidad 7 Introducción a R con un enfoque bioinformático.md at master · AliciaMstt/BioinfInvRepro2016-II. GitHub. https://github.com/AliciaMstt/BioinfInvRepro2016-II/blob/master/Unidad7/Unidad%207%20Introducci%C3%B3n%20a%20R%20con%20un%20enfoque%20bioinform%C3%A1tico.md

Anders, Simon, and Wolfgang Huber. 2010. "Differential expression analysis for sequence count data." Genome Biology 11 (10): R106+. https://doi.org/10.1186/gb-2010-11-10-r106.

Arunachalam, P. S., Wimmers, F., Mok, C. K. P., et al. (2020). Systems biological assessment of immunity to mild versus severe COVID-19 infection in humans. Science, 369(6508), 1210-1220. https://doi.org/10.1126/science.abc6261

Blanco-Melo, D., Nilsson-Payant, B. E., Liu, W.-C., et al. (2020). Imbalanced Host Response to SARS-CoV-2 Drives Development of COVID-19. Cell, 181(5), 1036-1045.e9. https://doi.org/10.1016/j.cell.2020.04.026

Goedhart, A. H. A. J. (2022, 14 septiembre). Chapter 10 Venn, Euler, and upSet diagrams | Visualization in R workshop. https://a-h-b.github.io/R_base_vis_course/venn-euler-and-upset-diagrams.html

Lee, J. S., Park, S., Jeong, H. W., et al. (2020). Immunophenotyping of COVID-19 and influenza highlights the role of type I interferons in development of severe COVID-19. Science Immunology, 5(49), eabd1554. https://doi.org/10.1126/sciimmunol.abd1554

Love, M. I., Huber, W., & Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biology, 15(12), 550. https://doi.org/10.1186/s13059-014-0550-8

Lucas, C., Wong, P., Klein, J., et al. (2020). Longitudinal analyses reveal immunological misfiring in severe COVID-19. Nature, 584(7821), 463-469. https://doi.org/10.1038/s41586-020-2588-y

McClain, M. T., Constantine, F. J., Henao, R., Liu, Y., Tsalik, E. L., Burke, T. W., Steinbrink, J. M., Petzold, E., Nicholson, B. P., Rolfe, R., Kraft, B. D., Kelly, M. S., Saban, D. R., Yu, C., Shen, X., Ko, E.

M., Sempowski, G. D., Denny, T. N., Ginsburg, G. S., & Woods, C. W. (2021). Dysregulated transcriptional responses to SARS-CoV-2 in the periphery. Nature communications, 12(1), 1079. https://doi.org/10.1038/s41467-021-21289-y

RPubs - Expresión diferencial RNA-seq: G1E-MK. (s. f.). https://rpubs.com/diana_trujillo/expres_difRNA-seq

Schulte-Schrepping, J., Reusch, N., Paclik, D., et al. (2020). Severe COVID-19 Is Marked by a Dysregulated Myeloid Cell Compartment. Cell, 182(6), 1419-1440.e23. https://doi.org/10.1016/j.cell.2020.08.001

SummarizedExperiment. (s. f.). Bioconductor. https://bioconductor.org/packages/release/bioc/html/SummarizedExperiment.html

Wang, L., Balmat, T. J., Antonia, A. L., Constantine, F. J., Henao, R., Burke, T. W., Ingham, A., McClain, M. T., Tsalik, E. L., Ko, E. R., Ginsburg, G. S., DeLong, M. R., Shen, X., Woods, C. W., Hauser, E. R., & Ko, D. C. (2021). An atlas connecting shared genetic architecture of human diseases and molecular phenotypes provides insight into COVID-19 susceptibility. Genome medicine, 13(1), 83. https://doi.org/10.1186/s13073-021-00904-z

Wilk, A. J., Rustagi, A., Zhao, N. Q., et al. (2020). A single-cell atlas of the peripheral immune response in patients with severe COVID-19. Nature Medicine, 26(7), 1070-1076. https://doi.org/10.1038/s41591-020-0944-y