RESEARCH

Respuesta a la Infección por SARS-Cov2

Susana Cano Marín*, Juan Sánchez Rodríguez
, Laura Fernández García
, Mariana Gonzalez Jimenez
and Ana Galiá Caravaca

*Correspondence: scm8u7j@uma.es
ETSI Informática, Universidad de Málaga, Málaga, España
Full list of author information is available at the end of the article

Abstract

En este proyecto realizaremos la coexpresión génica de los datos obtenidos a partir de muestras de células epiteliales de pulmón de diferentes pacientes. De esta forma, estudiaremos la respuesta de estas células a la infección por SARS-CoV2 usando el paquete de R "WGCNA".

Keywords: COVID-19; SARS-CoV2; coexpresión de genes; epitelio; pulmón

1 Introducción

La palabra pandemia ha ocupado un lugar muy importante en nuestras vidas este último año. El virus COVID-19 se ha convertido en la mayor preocupación mundial en la actualidad, no solo tiene consecuencias en la salud de millones de personas alrededor del mundo, sino que también esta situación y esta pandemia mundial ha producido muchos otros efectos negativos. Ningún gobierno, ni organización, ni persona está preparada para sobrellevar una pandemia mundial. Por lo que no solo ha afectado a la salud de miles de personas sino que también ha afectado a la economía, a la sociedad, a la política y las grandes potencias mundiales han actuado a ciegas ya la información sobre el mismo era escasa.

La primera secuencia que se obtuvo del genoma del agente infeccioso pudo ser encontrada en enero del año siguiente. Esto fue crucial para identificar al virus como un coronavirus, encontrando similitudes al coronavirus responsable del Síndrome Agudo Respiratorio Grave (SARS), la enfermedad respiratoria nacida en Asia en 2003, la cual también se convirtió en pandemia. Es por ello que es tan importante el conocimiento de los virus de SARS y MERS. El tamaño de los viriones de SARS-CoV-2 es de aproximadamente 50 a 200 nm de diámetro y su genoma está formado por ARN monocatenario de sentido positivo. La secuencia del betacoronavirus de Wuhan, de aproximadamente treinta mil nucleotidos de longitud, se relacionó por parecido con los betacoronavirus que afectaban a los murciélagos, pero son genéticamente diferenciables de otros coronavirus como el SARS-CoV y el MERS-CoV. Está compuesto de cuatro genes para las proteínas estructurales que caracterizan a los coronavirus, los cuales se identifican mediante las letras S (homotrímero de glicoproteína que forman las puntas de la superficie), E (proteína de bajo tamaño de la envoltura), M (proteína de la matriz que une la envoltura con el núcleo) y N (fosfoproteína de la nucleocápside), además de los marcos de lectura abiertos que codifican proteínas no estructurales en las que encontramos,

las enzimas causantes de su ciclo reproductivo intrahospedero.

Toda esta información sobre el virus era desconocida en diciembre de 2019 cuando apareció en Wuhan , provincia de Hubei (China). Un brote epidémico de lo que se llamaba neumonía por causa desconocida que llegó a afectar a más de 60 personas durante ese mes. Esto es debido a que el coronavirus puede infectar de manera selectiva las mucosas pulmonares o grastrointestinales. La forma de acceder a una célula epitelial es mediante un receptor presente en las superficie del organismo que recibe el nombre de ACE2. Dichos receptores son más comunes ser encontrados en los pulmones, por ello esta enfermedad está considerada de tipo respiratorio. El sistema inmunológico humano contraataca con una respuesta dura, liberando interferones, cuya función es dificultar la replicación del virus dentro de las células epiteliales.

En este trabajo, estudiaremos la respuesta desarrollada en las células del epitelio del pulmón a la infección por SARS-Cov2 mediante el análisis de los perfiles de expresión génica publicados en el dataset GEO GSE147507. Estos perfiles de expresión se analizarán mediante el modelado de redes de coexpresión génica, con el paquete de R WGCNA.

2 Materiales y métodos

Para llevar a cabo el modelado de redes de coexpresión génica se ha utilizado el lenguaje R mediante el entorno RStudio. El análisis de coexpresión es un método de biología de sistemas utilizado para decribir los patrones de correlación entre genes en muestras de microarrays. Las redes de correlación facilitan los métodos de cribado de genes basados en redes que se pueden utilizar para identificar posibles biomarcadores o dianas terapeuticas. A continuación, mostraremos y explicaremos los paquetes y funciones esenciales usados para llevar a cabo la realización de nuestro estudio.

WGCNA: Esta es la librería utilizada y en la que nos inspiraremos para el análisis de redes de coexpresión de genes ponderados. Este paquete podemos obtenerlo de Bioconductor instalando BiocManager en nuestro entorno. Hemos utilizado las siguientes funciones:

- enableWGCNAThreads: Estas funciones permiten y deshabilitan subprocesos múltiples para cálculos WGCNA que opcionalmente pueden ser multiproceso, lo que incluye todas las funciones que usan funciones cor o bicor.
- pickSoftThreshold: Análisis de topología libre de escala para múltiples poderes de umbral suave. El objetivo es ayudar al usuario a elegir una potencia de umbral suave adecuada para la construcción de la red.
- adyacencia: Calcula (mediante la correlación o distancia) la adyacencia de la red a partir de datos de expresión dados o de una similitud.
- TOMsimilarity: Cálculo de la matriz de superposición topológica, y la correspondiente disimilitud, a partir de una matriz de adyacencia dada.
- label2colors: Convierte un vector o matriz de etiquetas numéricas en un vector o matriz de colores correspondiente a las etiquetas.
- moduleEigengenes: Calcula los eigengenes del módulo (primer componente principal) de los módulos en un único conjunto de datos determinado.
- mergeCloseModules: Fusiona módulos en redes de expresión génica que están demasiado cerca según lo medido por la correlación de sus genes propios.
- cor: Estas funciones implementan un cálculo más rápido de la correlación de Pearson (ponderada).
- TOMplot: Representación gráfica de la matriz de superposición topológica utilizando un gráfico de mapa de calor combinado con el dendrograma de agrupamiento jerárquico correspondiente y los colores del módulo.
- plotEigengeneNetworks: Esta función traza representaciones de dendrogramas y genes propios de redes de genes propios (consenso). En el caso de redes de genes propios de consenso, la función también traza medidas de preservación por pares entre redes de consenso en diferentes conjuntos.
- exportNetworkToCytoscape: Esta función exporta una red en archivos de lista de nodos y de borde en un formato adecuado para importar a Cytoscape.

cluster: Es el utilizado para hacer el agrupamiento de los datos. Hemos usado las siguientes funciones.

- pam: Esta función realiza un agrupamiento de los datos en k grupos "alrededor de medoides", una versión más robusta de K-means.
- hclust: Esta función realiza un análisis de agrupamiento jerárquico utilizando un conjunto de diferencias para los n objetos que se agrupan. Inicialmente, cada objeto se asigna a su propio grupo y luego el algoritmo procede de forma iterativa, en cada etapa uniendo los dos grupos más similares, continuando hasta que haya un solo grupo. En cada etapa, las distancias entre los conglomerados se vuelven a calcular mediante la fórmula de actualización de disimilitud de Lance-Williams de acuerdo con el método de conglomerado particular que se utilice.

DESeq2: Esta librería la usaremos para el análisis de datos de RNA-seq. Al igual que WGCNA, podemos obtenerla de Bioconductor instalando BiocManager en nuestro entorno. Las funciones usadas son:

- DESeqDataSetFromMatrix: Esta función es una subclase de RangedSummarizedExperiment, que se utiliza para almacenar los valores de entrada, cálculos intermedios y resultados de un análisis de expresión diferencial. La clase DESeqDataSet impone valores enteros no negativos en la matriz de "recuentos" almacenada como el primer elemento en la lista de análisis.
- Counts: La ranura de conteos contiene los datos de conteo como una matriz de valores de conteo de números enteros no negativos, una fila para cada unidad de observación (gen o similar) y una columna para cada muestra.
- Deseqm4: Esta función realiza un análisis predeterminado a través de la estimación de factores de tamaño, a través de la estimación de dispersión y a través del ajuste GLM binomial negativo y estadísticas de Wald.

DCGL: El utilizado para el análisis de coexpresión diferencial y análisis de regulación diferencial de datos de microarrays de expresión genómica. Hemos utilizado las siguientes funciones.

- qLinkfilter: En esta función los enlaces genéticos con valores 'q' de pares de valores de coexpresión en cualquiera de las dos condiciones superiores al límite se retienen, mientras que los valores de coexpresión de otros enlaces se establecen en cero.
- WGCNA: El 'análisis de red de coexpresión de genes ponderados' pondera los vínculos con los coeficientes de correlación y compara las sumas de los coeficientes de correlación de un gen.

coexnet: es el utilizado para la construcción de la red de coexpresión. Podremos obtenerla de BiocManager. Hemos utilizado las siguientes funciones.

- createNet: Esta, es una función que a partir de una secuencia biológica genera un grafo no direccionado teniendo como palabras vértices, pudiendo esto tener su parámetro de tamaño fijado por el parámetro 'palabra'. Las conexiones entre palabras dependen del parámetro 'paso' que indica la próxima conexión

que se formará.

También se han usado otras librerías de R que nos han facilitado el entendimiento de los resultados, así como la aplicación de estos para el uso de algunas funciones. Algunos son dplyr y base para la manipulación y el manejo de datos; grDevices para la manipulación de gráficos; Stats, para cientas medidas estadísticas; entre otros.

3 Resultados

El primer resultado obtenido corresponde con un análisis de la topología de la red para varias potencias de umbral suave. El panel de la izquierda muestra el ajuste de escala libre, correspondiente al índice (eje y), en función de la potencia de umbral suave, que sería el (eje x). El panel derecho muestra la conectividad media (grados, eje y) en funcion de la potencia de umbral suave (eje x).

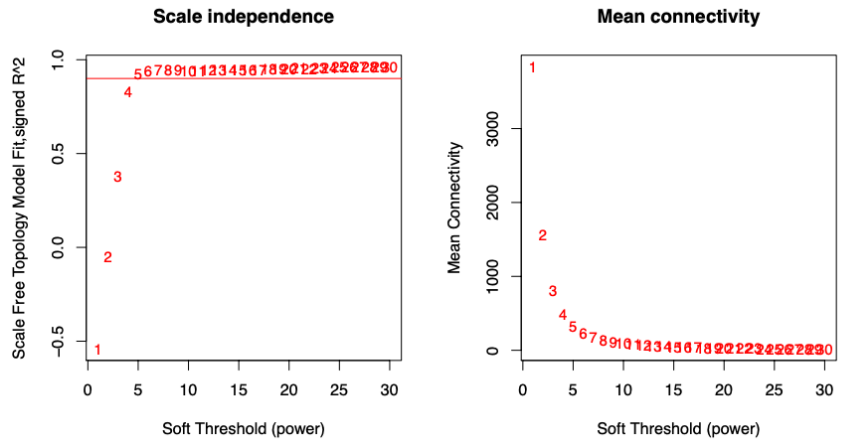

Figure 1 Escala de Independencia

En la siguiente figura podemos ver un dendograma que expresa un clustering TOM disimilitud, y abajo las dos formas de obtener los módulos, primero mediante un árbol de corte dinámico y el otro mediante fusión. Cada color corresponde a un modulo.

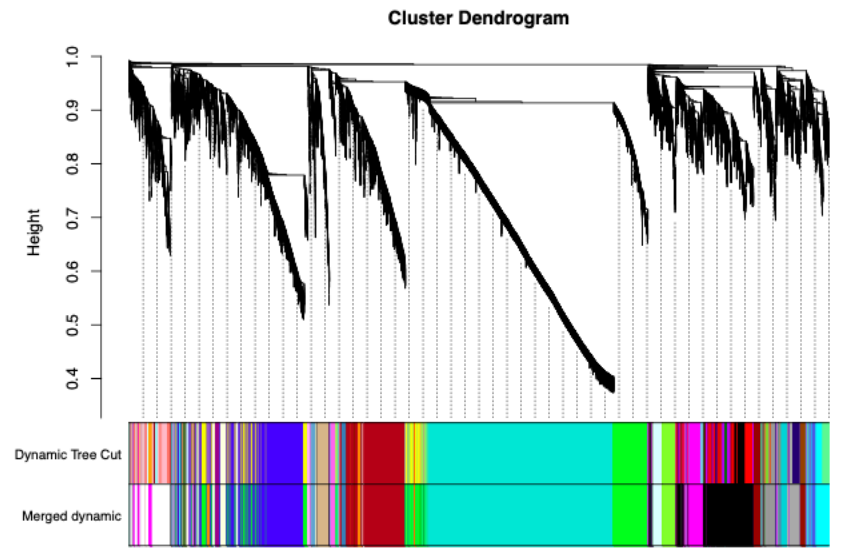

Figure 2 Dendograma de función Merge

En la figura 3 tenemos un mapa de calor que nos expresa la correlación de los módulos expresados mediante corte de árbol dinámico.

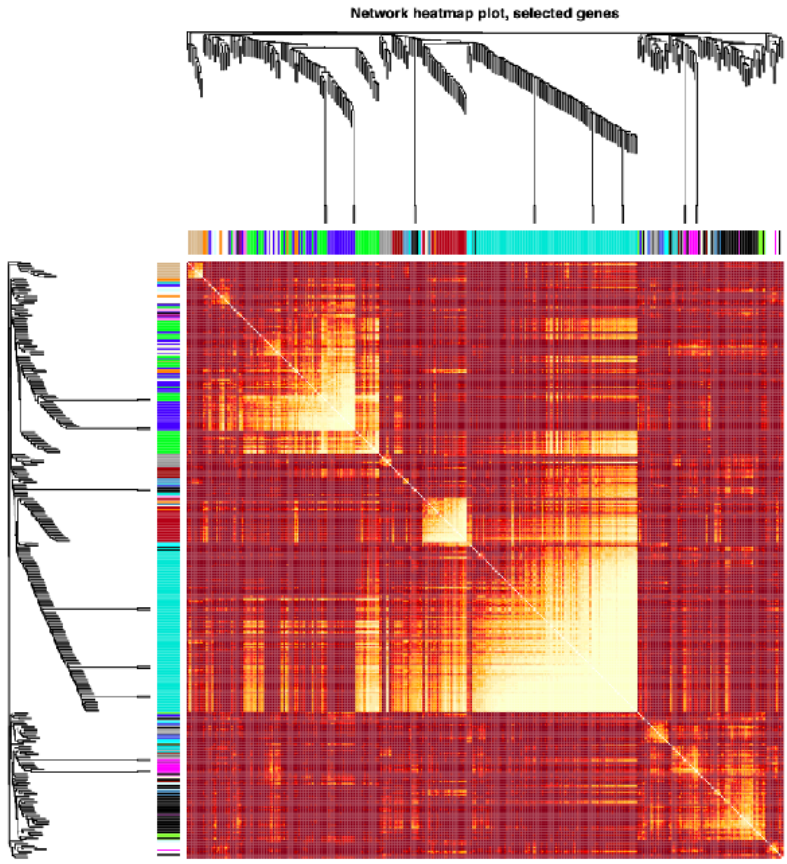

Figure 3 Mapa Calor

En la figura 4 tenemos otro dendograma en el cual están agrupados por los módulos obtenidos anteriormente. A continuación, se encuentra su matriz de adyacencia.

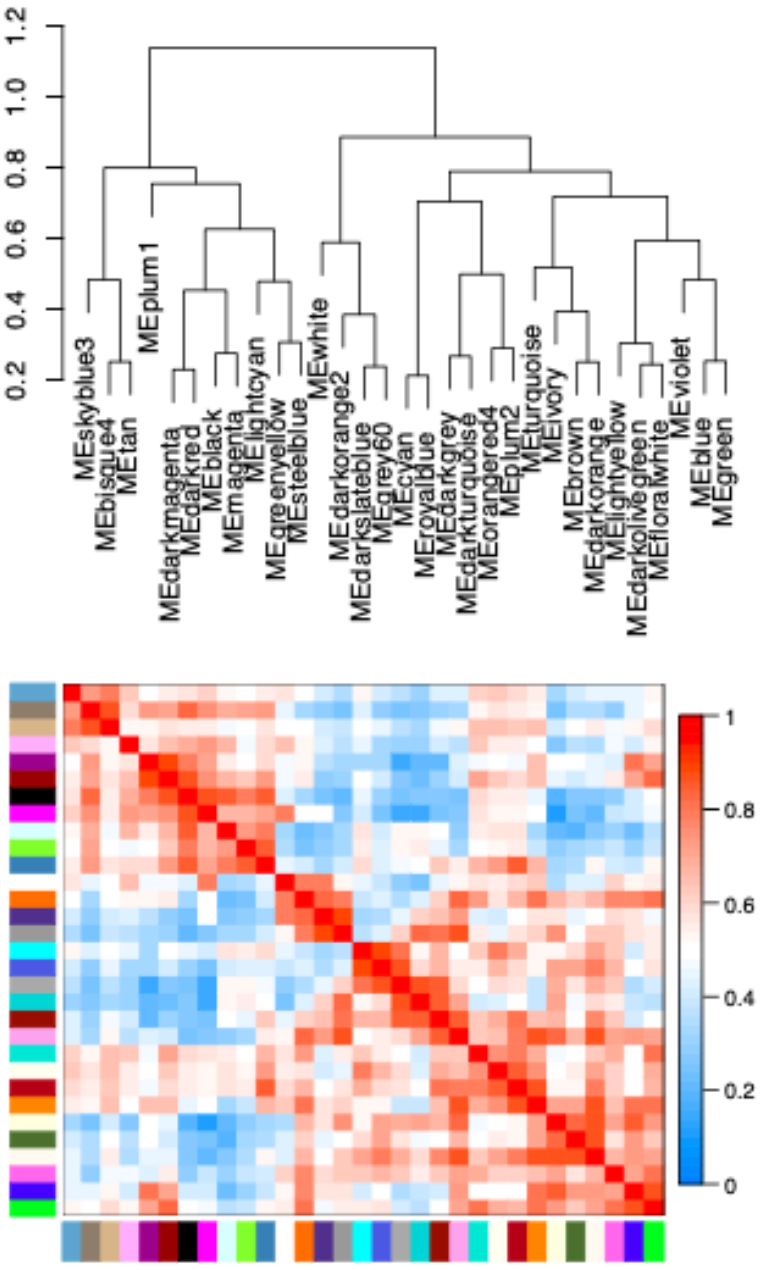

Figure 4 Gráfica de adyacencia

Por último, en la figura 5, mostramos la red de correlación de las muestras y sus genes, en respuesta de las células epiteliales del pulmón a la infección por SARS-CoV2. Nos indican la relación que existe entre las muestras.

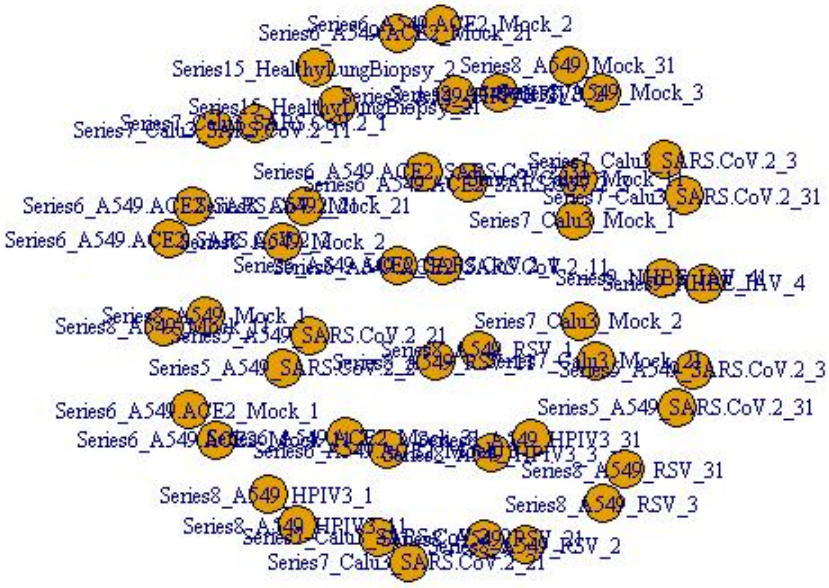

Figure 5 Red de Correlación

4 Discusión

Para la correcta realización de nuestro estudio, al tener un número de datos tan grande, se ha requerido realizar una serie de filtrados. Tras una profunda revisión de los datos, nos dimos cuenta que habıa muchos valores nulos, que en este caso se corresponden con valores 0, por lo que primero comenzamos eliminando aquellas columnas que tengan más de un 51% de 0. El numero de variables se ha reducido en torno a un 70%

Una vez realizado el filtrado, dado que todavía tenemos una enorme cantidad de valores 0, sustituiremos los 0 por la media de las columnas. A continuación, ya podremos realizar nuestro estudio.

Comenzaremos realizando un Sample Clustering, el cuál es un método de muestreo donde se crea mútiples grupos de genes de una población donde son indicativos de características homógeneas y tienen la misma probabilidad de ser parte de la muestra. Se encarga de agrupar. Después realizamos la escapa de independencia, el valor 3 indica que se encuentra en torno al 50% indicaría que tiene ese porcentaje de probabilidades de seguir un modelo de escala libre topológica. Como podemos observar, sigue una relación inversa con la conectividad media, ya que en

la conectividad media, conforme aumenta el umbral suave disminuye la media de la conectividad.

Hemos usado WGCNA para obtener los datos de expresión que han sido obtenidos mediante la separación de los clusters y a cada cluster, le aplicamos el análisis de coexpresión y vemos cuales tiene mayor correlación con la coexpresión del cluster 1.

Después, pasamos a hacer un clustering jerérquico basado de la matriz de medición de superposición topológica (TOM) para los datos de expresión génica. La gráfica de medida de superposición topológica muestra grupos de genes (módulos) altamente interconectados. Los genes se asignaron a módulos nombrados por los colores debajo del dendrograma utilizando el método de corte de árbol dinámico.

Una vez que hemos obtenido los módulos, realizaremos un mapa de calor que nos mostrará la correlación de cada uno de los módulos. Cuanto mayor es la correlación, más claro será el color. También obtenemos un mapa de adyacencia, cuanto más se acerca al color rojo más adyacentes son los módulos.

Por último, realizamos una red de correlación que nos indica la relación de los genes.

5 Conclusiones

Basándonos en los resultados obtenidos y en las diferentes fuentes, podemos llegar a la conclusión de que los módulos más largos y comunes, tanto la obtención por corte de árbol dinámico como por fusión, son aquellos que se sobreexpresan o activan a la infección por SARS-CoV2 en las células epiteliales del pulmón. Siendo uno de los genes más activados el IL-6 indicando una infección, trauma o hemorragia en nuestras muestras. Además, también podemos concluir que ciertas citoquinas se muestran elevadas y los niveles de interferones tipo I y tipo III (IFN-I, III) tienen un bajo nivel de activación.

Para finalizar, nos hemos dado cuenta que el epitelio, además de tener un papel estructural, es un claro indicador de la respuesta inmune.

Abreviaciones
SARS: Síndrome Agudo Respiratorio Severo
COVID: Coronavirus Disease
ACE2: Célula epitelial

Disponibilidad de datos y materiales
Debéis indicar aquí un enlace a vuestro repositorio de github.

Contribución de los autores
S.C.M: Flujo de trabajo en bash, organización, coordinación y revisión de código.
J.S.R: Código, bibliografía y redacción.
A.G.C: Flujo de trabajo en bash, redacción y organización.
M.G.J: Código y redacción.
L.F.G: Código y redacción.

Author details
ETSI Informática, Universidad de Málaga, Málaga, España.

References

1 Bioconductor - coexnet. (n.d.). Retrieved February 14, 2021, from https://bioconductor.org/packages/release/bioc/html/coexnet.html
2 Bioconductor - DESeq2. (n.d.). Retrieved February 14, 2021, from https://bioconductor.org/packages/release/bioc/html/DESeq2.html
3 Bioconductor - RTCGAToolbox. (n.d.). Retrieved February 14, 2021, from https://bioconductor.org/packages/release/bioc/html/RTCGAToolbox.html
4 Blanco-Melo, D., Nilsson-Payant, B. E., Liu, W. C., Møller, R., Panis, M., Sachs, D., Albrecht, R. A., & tenOever, B. R. (2020). SARS-CoV-2 launches a unique transcriptional signature from in vitro, ex vivo, and in vivo systems. In bioRxiv. bioRxiv. https://doi.org/10.1101/2020.03.24.004655
5 factoextra package — R Documentation. (n.d.). Retrieved February 14, 2021, from https://www.rdocumentation.org/packages/factoextra/versions/1.0.3
6 Función NbClust — Documentación R. (n.d.). Retrieved February 14, 2021, from https://www.rdocumentation.org/packages/NbClust/versions/3.0/topics/NbClust
7 Lokugamage, K. G., Yoshikawa-Iwata, N., Ito, N., Watts, D. M., Wyde, P. R., Wang, N., Newman, P., Kent Tseng, C. Te, Peters, C. J., & Makino, S. (2008). Chimeric coronavirus-like particles carrying severe acute respiratory syndrome coronavirus (SCoV) S protein protect mice against challenge with SCoV. Vaccine, 26(6), 797–808. https://doi.org/10.1016/j.vaccine.2007.11.092
8 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). https://doi.org/10.1186/s13059-014-0550-8
9 paquete dplyr — Documentación R. (n.d.). Retrieved February 14, 2021, from https://www.rdocumentation.org/packages/dplyr/versions/0.7.8
10 Peter Langfelder, A., Hor-, S., Cai, C., Dong, J., Miller, J., Song, L., Yip, A., & Zhang Maintainer Peter Langfelder, B. (2020). Package "WGCNA" Title Weighted Correlation Network Analysis. https://doi.org/10.2202/1544
11 Pinto, B. G. G., Oliveira, A. E. R., Singh, Y., Jimenez, L., Gonçalves, A. N. A., Ogava, R. L. T., Creighton, R., Peron, J. P. S., & Nakaya, H. I. (2020). ACE2 expression is increased in the lungs of patients with comorbidities associated with severe COVID-19. In medRxiv. medRxiv. https://doi.org/10.1101/2020.03.21.20040261
12 Respuesta inflamatoria y apoptosis en la lesión pulmonar aguda. (n.d.). Retrieved February 14, 2021, from http://scielo.isciii.es/scielo.php?script=sci_arttext&pid=S0210-56912006000600003
13 RPubs - Identifying Coexpressed Genes. (n.d.). Retrieved February 15, 2021, from https://rpubs.com/mikelapika/491660
14 RPubs - WGCNA Tutorial. (n.d.). Retrieved February 15, 2021, from https://rpubs.com/natmurad/WGCNA

15 WGCNA/script_WGCNAwithoutTrait.R at master · paytonyau/WGCNA. (n.d.). Retrieved February 15, 2021, from https://github.com/paytonyau/WGCNA/blob/master/script_WGCNAwithoutTrait.R

16 Yang, J., Yu, H., Liu, B.-H., Zhao, Z., Liu, L., Ma, L.-X., Li, Y.-X., Li, Y.-Y., Bao, M., & Liu, H. (2015). Package "DCGL" Type Package Title Differential Co-expression Analysis and Differential Regulation Analysis of Gene Expression Microarray Data. https://doi.org/10.1371/journal.pone.0079729

17 Zhang, B., & Horvath, S. (2005). A general framework for weighted gene co-expression network analysis. Statistical Applications in Genetics and Molecular Biology, 4(1). https://doi.org/10.2202/1544-6115.1128