

Examen final Bioinformática: Análisis de expresión diferencial con DESeq2 y Expression Atlas

Sofía Upegui Del Río; Ingeniería Biotecnológica; Universidad EIA

Descripción del estudio original

En este examen se desarrolló una réplica del artículo de RNA-seq diferencial encontrado en el Expression Atlas: Transcriptional and proteomic insights into the host response in fatal COVID-19 cases, publicado en el año 2020. Este artículo experimental tiene número de acceso E-ENAD-46, correspondiente al European Nucleotide Archive imported data. La página web del Expression Atlas correspondiente al artículo cuenta con los siguientes documentos para los datos crudos, metadata

Raw data:

<https://www.ebi.ac.uk/gxa/experiments-content/E-ENAD-46/resources/DifferentialSecondaryDataFiles.RnaSeq/raw-counts>

Experiment Design:

<https://www.ebi.ac.uk/gxa/experiments-content/E-ENAD-46/resources/ExperimentDesignFile.RnaSeq/experiment-design>

En este estudio se secuenciaron genes humanos de tejidos pulmonares y de colon buscando diferenciar la expresión de estos entre cuerpos postmortem a causa de COVID-19 en el primer brote del virus SARS-CoV-2 en Wuhan, China y tejidos sin infección de resecciones quirúrgicas y biopsias rutinarias para detección de cáncer pulmonar. De esta manera se busca comprender molecularmente la respuesta transcripcional y proteómica del cuerpo humano en casos fatales de COVID-19 al igual que las rutas biológicas alteradas como consecuencia. Los investigadores llegaron a este estudio inspirados por que los pacientes muertos a causa de COVID-19 sufrieron de falla respiratoria, dejando como prueba signos de daño alveolar, formación de membrana hialina, fibrosis e infiltrado inflamatorio a pesar de tener cargas virales realmente bajas al final de su vida.

Diseño experimental y contraste

En el artículo se trata el factor biológico como el estado de salud de los individuos estudiados, **condición**. Para el cual se tienen dos niveles principales a comparar: Estado **covid**, representando los tejidos de los nueve individuos que murieron a causa de COVID - 19 y estado **control**; representando los diez tejidos histológicos sin infección de SARS-CoV-2.

De cada nivel se sacaron dos muestras, secuenciación de tejido pulmonar y secuenciación de tejido de colon. Estas buscan principalmente comparar entre niveles, pero no entre pulmón vs colon lo que las hace una variable más y no una adición a las réplicas biológicas, adicionalmente

estas no se toman como réplicas técnicas al no provenir y comparar del mismo tejido a pesar de venir del mismo individuo representando un nivel del factor, covid o control.

Finalmente, esto nos lleva a concluir que hay una variable biológica con dos niveles, uno con nueve réplicas biológicas (covid) y otro con diez (control). Adicionalmente, hay dos mediciones por réplica una para tejido de colón y otra para tejido pulmonar, lo que nos deja con 38 réplicas en total, 18 para covid y 20 para control. Sin embargo, se utilizan para el trabajo individual **solamente los datos de pulmón** buscando hacer un análisis por partes, pulmón y colon, del estudio (como se hace en el artículo) sin extenderse demasiado, por lo que se hizo una adaptación inicial de la metadata para quedar con dos niveles de factor condición (covid y control) y 19 réplicas para pulmón (9 para covid y 10 para normal).

Resultados principales

Teniendo en cuenta que el cambio en la expresión génica fue muy potente, osea que se tuvieron muchos genes rechazando la hipótesis nula que dicta una indiferencia entre los niveles con $p < 0,05$ se aplicó otro filtro moderado para determinar significancia $abs(Log_2FC) > 0.5$, lo que indicó un cambio significativo en un total de 7717 genes, 3024 genes sobreexpresados y 4693 genes suprimidos. De esta misma manera, al analizar los genes rechazando la hipótesis nula con $padj < 0,05$ y el mismo filtro de $abs(Log_2FC)$, para tomar en cuenta los efectos de las variables en la regresión y llevar conclusiones integrales, se encontraron un total de 5142 genes expresados diferencialmente, de los cuales 2048 fueron de una sobre expresión y 3094 una supresión. De estos últimos criterios para considerar significancia se generaron dos tablas, una de los top tres genes sobreexpresados y top tres reprimidos; y otra de genes de importancia analítica en el artículo: ACE2, TMPRSS2 e IFNG.

	Gene.Name	Gene.ID	log2FoldChange	pvalue	padj	Regulacion
17042	KRT6A	ENSG00000205420	10.9243945	4.966664e-28	4.005946e-24	UP
21886		ENSG00000231683	9.3833328	2.824657e-09	2.649156e-07	UP
13758	FDCSP	ENSG00000181617	8.4844835	9.434568e-04	8.027785e-03	UP
12243	JMJD1C	ENSG00000171988	-0.5002220	2.672577e-03	1.784939e-02	DOWN
3746	GNPTAB	ENSG00000111670	-0.5025331	1.470657e-03	1.131494e-02	DOWN
5912	AFDN	ENSG00000130396	-0.5148501	8.886271e-03	4.345617e-02	DOWN

Tabla 1. Resumen de genes con diferencias estadísticamente significativas entre covid y control. Rangos de corte de $padj < 0,05$ y $abs(Log_2FC) > 0.5$.

De la Tabla 1 resaltan especialmente los genes de sobreexpresión, que cuentan con un Log_2FC mucho más alto que el de los reprimidos, lo que indica que divergen del control mucho más y generan una reacción abrumadora en el cuerpo. Específicamente, aquí se encuentra KRT6A, que codifica para la queratina tipoII y se expresa especialmente en la diferenciación celular y curación de heridas (Keratin 6A, 2025), por lo que su sobreexpresión descontrolada podría explicar el engrosamiento epitelial e inflamación en pulmones. Por otro lado, se encuentra la sobreexpresión de FDCSP, sustancia de células dendríticas para la presentación de antígenos a las células tipo B que buscan el control y activación de la respuesta inmune del cuerpo

(Follicular Dendritic Cell Secreted Protein, 2025), lo que indica una respuesta inmune en pacientes con covid descontrolada teniendo en cuenta la baja carga viral de los pacientes a la hora de la muestra. Este proceso nos permite entender mejor la reportada formación de NETs (neutrophil extracellular traps) en el artículo los cuales se forman por acumulación y muerte de células del sistema inmune que forman barreras que al acumularse limitan el funcionamiento apropiado de la respiración (Wu et al., 2020).

	Gene.ID	baseMean	log2FoldChange	lfcSE	stat	pvalue	padj	Gene.Name
3725	ENSG00000111537	44.317183	0.5919812	0.9067917	0.6528304	0.5138656096	0.696309915	IFNG
5886	ENSG00000130234	1.964323	2.5050931	1.4312389	1.7502970	0.0800670861	NA	ACE2
14174	ENSG00000184012	304.597572	-1.6691086	0.4903200	-3.4041208	0.0006637739	0.006215687	TMPRSS2

Tabla 2. Resumen de estadísticas de genes de interés en el artículo, con rangos de corte de $\text{padj} < 0,05$ y $\text{abs}(\text{Log}_2FC) > 0.5$.

De esta información fue claro que de los últimos tres genes solo uno fue de significancia, TMPRSS2, teniendo su valor padj menor a 0.5 y su $\text{abs}(\text{Log}_2FC) > 0.5$. Específicamente, esta última estadística nos indica que el gen sufre de una supresión al exponerse a infecciones de SARS-CoV-2. Por parte de IFNG, no se ve un cambio significativo desde ninguno de los dos cortes, lo que da a entender que la infección por covid realmente no tiene gran efecto sobre este gen. Adicionalmente, es importante destacar el resultado del gen ACE2, que a pesar de no clasificar como significativo por pval o padj , su valor para Log_2FC es significativo, mostrando un aumento en la expresión del gen ante la infección así esta sea pequeña, lo que se comprueba en los conteos crudos de este gen en el pulmón expresados gráficamente en la figura 1.

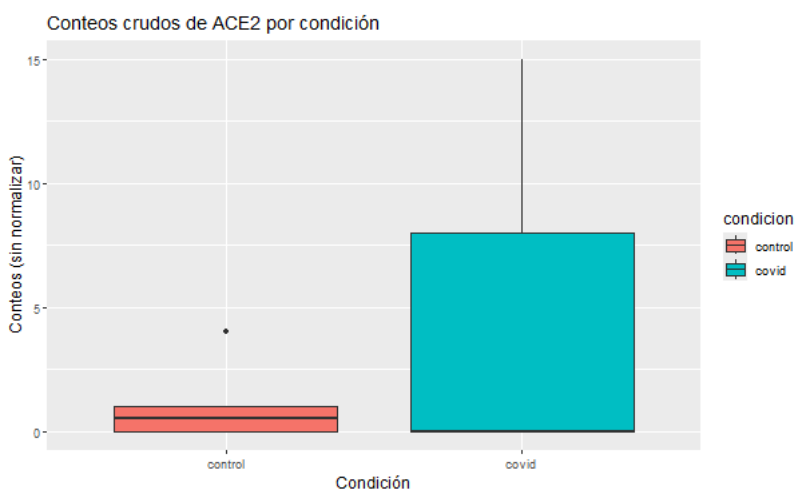
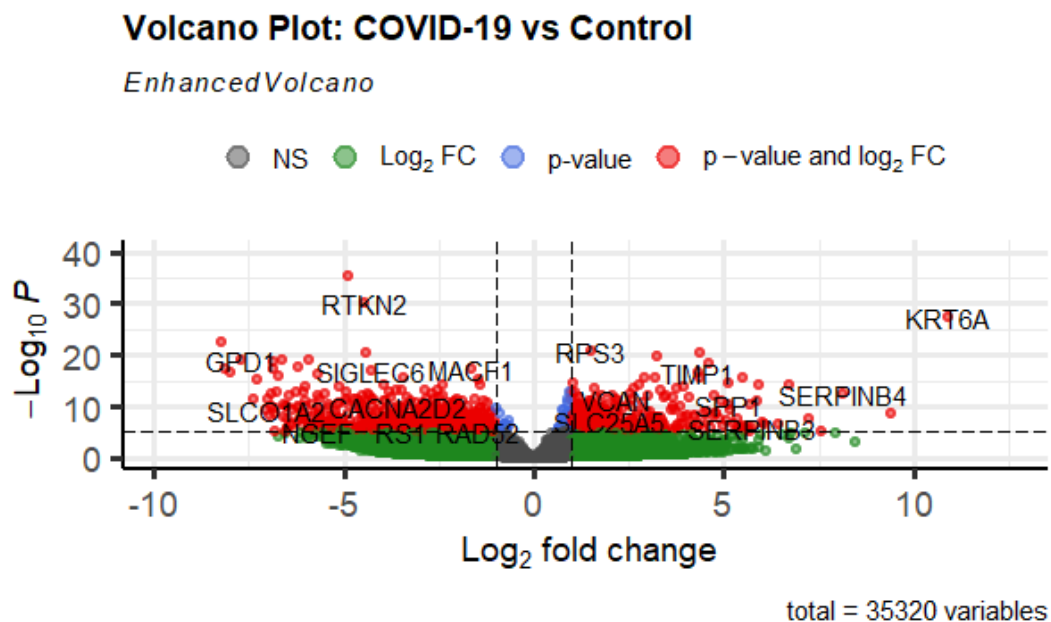


Figura 1. Diagrama de cajas y bigotes para conteos crudos de ACE2 por condición en pulmón.

Estos resultados coinciden con el artículo analizado que muestra una subida de este gen cuatro veces por encima del control. ACE2 (Angiotensin converting enzyme) es la receptora principal de la infección viral de este tipo en las células, común en infecciones virales por la reproducción del virus que busca generar más receptores. Sin embargo, el análisis se pone interesante al

estudiar la proteasa TMPRSS2, que codifica por una espícula del SARS-CoV-2 que le permite entrar fácilmente a las células a infectar. Su expresión en estas etapas tardías de enfermedad fue menor que en los controles, lo que indica que la infección estaba bajando a pesar de que los pacientes presentaban síntomas severos o que el virus encontró una manera diferente de entrar a la célula, lo que apunta de nuevo a que lo que lleva al estado crítico a los pacientes puede no ser la infección sino su respuesta inmune y a la razón por la que los antivirales comunes no tendrían por qué funcionar teniendo en cuenta su acción sobre este tipo de espículas. Por parte del gen IFNG, que es un homodímero que se une a un interferón activador de la respuesta celular ante virus y bacterias, encontramos pocos cambios en su expresión (*Interferon Gamma*, 2025), lo que indica que las células no respondieron ante la infección como lo harían normalmente, evidenciando la capacidad de este virus de escabullirse y evadir la respuesta celular.



Figuras 2. Volcano plot de resultados de DESeq2, siendo los puntos grises no significativos, los verdes solo para el eje x, los azules solo para el eje y y los rojos para ambos.

El volcano plot muestra una gran cantidad de genes diferencialmente significativos y un balance entre la cantidad de los suprimidos y los sobreexpresión, sin embargo, estos últimos parecen alejarse más del control por sus altos valores en $\log_2 FC$. Adicionalmente se ve una importante significancia por valor p (altos en la tabla) por parte de los genes RTKN2 y KRT6A. Para hablar un poco de los genes subexpresados, aquí destaca el MACF1 (microtubule actin crosslinking factor 1) y el GPD1 (glycerol-3-phosphate dehydrogenase 1) que suprimidos generan deficiencias en la estructura del citoesqueleto y organización celular; y un cambio en la producción energética celular respectivamente, sosteniendo la vulnerabilidad en la que se

encuentran los pulmones infectados en pacientes con COVID-19 (Glycerol-3-Phosphate Dehydrogenase 1, 2025b) & (Microtubule Actin Crosslinking Factor 1, 2025).

Comparación con Expresión Atlas

Bajo los mismos criterios de $\text{padj} < 0.05$ y $\text{abs}(\text{Log}_2\text{FC}) > 0.5$ se encontraron en el Expression Atlas 11620 genes significativos, un número mayor que los 5142 genes significativos en el análisis propio. De esta búsqueda se encontraron resultados similares a los reportados en el artículo y análisis propio, principalmente el gen ACE2 también mostró insignificancia diferencial nula entre las condiciones de tejido pulmonar, mientras que el TMPRSS2 mostró significancia con un Log_2FC de -1.6 y un p adj de 7.9301×10^{-3} , valores muy similares a los encontrados en R y en concordancia con las conclusiones del artículo. x

Conclusión

Este ejercicio permitió replicar el proceso de diferenciación en RNA-seq de tejidos pulmonares en personas fallecidas por COVID-19 y biopsias para análisis de cáncer sin infección por SARS-CoV-2. De esta manera, y en congruencia con los resultados reportados en Expression Atlas y el artículo de estudio, se detectó con el software DESeq y los criterios $\text{padj} < 0.05$, $|\log_2\text{FC}| > 0.5$, una importante sobreexpresión de genes en las muestras pulmonares con COVID-19, relacionados especialmente con inflamaciones y debilitación celular, así como una supresión de genes de alerta celular a infecciones virales. Adicionalmente, se identificaron el cambio en genes de interés como ACE2 y TMPSSR2 que dan una posible razón por la que los pacientes con COVID-19 mueren por su propia respuesta inmune y no directamente por una alta carga viral.

Bibliografía

Bioinformagician. (2025). DESeq2 Basics Explained | Differential Gene Expression Analysis |

Bioinformatics 101. In *Youtube.com*. https://www.youtube.com/watch?v=0b24mpzM_5M

Biostatsquid. (2022, November 2). *How to interpret a volcano plot*. Biostatsquid.com - Easy

Bioinformatics and Biostatistics. <https://biostatsquid.com/volcano-plot/>

follicular dendritic cell secreted protein. (2025). NCBI.

<https://www.ncbi.nlm.nih.gov/datasets/gene/260436/>

Genomics Team. (2020). *Experiment < Expression Atlas < EMBL-EBI*. Ebi.ac.uk.

<https://www.ebi.ac.uk/gxa/experiments/E-ENAD-46/Downloads>

glycerol-3-phosphate dehydrogenase 1. (2025). NCBI.

<https://www.ncbi.nlm.nih.gov/datasets/gene/2819/>

interferon gamma. (2025). NCBI. <https://www.ncbi.nlm.nih.gov/datasets/gene/3458/>

keratin 6A. (2025). NCBI. <https://www.ncbi.nlm.nih.gov/datasets/gene/3853/>

microtubule actin crosslinking factor 1. (2025). NCBI.

<https://www.ncbi.nlm.nih.gov/datasets/gene/23499/>

Piper, M. (2017, May 12). *Set up and overview for gene-level differential expression analysis*.

Introduction to DGE - ARCHIVED.

https://hbctraining.github.io/DGE_workshop/lessons/01_DGE_setup_and_overview.html

Winge, M. C. G., Bradley, M., & E. Björck. (2014). Impaired wound healing and cheilitis in a

Pachyonychia Congenita K6a family. *Journal of the European Academy of Dermatology and Venereology*, 29(1), 185–187. <https://doi.org/10.1111/jdv.12386>

Wu, M., Chen, Y., Xia, H., Wang, C., Tan, C. Y., Cai, X., Liu, Y., Ji, F., Xiong, P., Liu, R., Guan,

Y., Duan, Y., Kuang, D., Xu, S., Cai, H., Xia, Q., Yang, D., Wang, M.-W., Chiu, I. M., &

Cheng, C. (2020). Transcriptional and proteomic insights into the host response in fatal COVID-19 cases. *Proceedings of the National Academy of Sciences*, 117(45),

28336–28343. <https://doi.org/10.1073/pnas.2018030117>