



**UNIVERSIDAD DE MURCIA**  
**FACULTAD DE BIOLOGÍA**

---

**CARACTERIZANDO LA EXPRESIÓN DE  
RETROVIRUS ENDÓGENOS HUMANOS EN  
EL COVID-19**

---

**AZAHARA MARÍA GARCÍA SERNA**

**TUTORES:**

**DR. SANTIAGO CUEVAS GONZÁLEZ**

**DR. ÁNGEL GIL ESTEBAN**



**UNIVERSIDAD  
DE MURCIA**



---

**TRABAJO DE FIN DE MÁSTER - MÁSTER EN BIOINFORMÁTICA 2023/2024**

# Tabla de contenidos

<b>1</b>	<b>Resumen</b>	<b>4</b>
<b>2</b>	<b>Introducción, antecedentes y objetivos</b>	<b>5</b>
2.1	Introducción	5
2.2	Antecedentes	5
2.2.1	Infección por el coronavirus 2 del síndrome respiratorio agudo grave (SARS-CoV-2) y la enfermedad del coronavirus 19 (COVID-19)	5
2.2.2	La respuesta inmunitaria frente a la infección por SARS-CoV-2	6
2.2.3	Retrovirus endógenos humanos (HERVs)	6
2.3	Objetivos	8
<b>3</b>	<b>Materiales y Métodos</b>	<b>8</b>
3.1	Materiales	8
3.1.1	Muestras analizadas	8
3.1.2	Tecnologías empleadas	8
3.2	Métodos	12
3.2.1	Extracción y secuenciación de ARN	12
3.2.2	Cuantificación del retrotranscriptoma: identificación de HERVs	12
3.2.3	Transcriptoma: expresión de genes en <i>RNA-seq</i>	12
3.2.4	Análisis de expresión diferencial	14
3.2.5	Análisis de coexpresión	14
3.2.6	Análisis funcional de enriquecimiento	14
<b>4</b>	<b>Resultados</b>	<b>14</b>
4.1	Calidad de las secuencias	14
4.2	Identificación de elementos retrotransponibles después de la infección por SARS-CoV-2	15
4.3	Cambios en el transcriptoma tras infección por SARS-CoV-2	18
4.4	Análisis integrado de la expresión diferencial de elementos retrotransponibles y genes	20
<b>5</b>	<b>Discusión</b>	<b>22</b>
5.1	Discusión de resultados en relación a investigaciones previas	22
5.2	Limitaciones	23
5.3	Trabajo futuro	23
<b>6</b>	<b>Conclusiones</b>	<b>24</b>
<b>7</b>	<b>Material suplementario</b>	<b>24</b>
<b>8</b>	<b>Referencias</b>	<b>25</b>
<b>9</b>	<b>Anexos y Material Suplementario</b>	<b>29</b>
9.1	Anexo I. Análisis de calidad de las muestras	29
9.2	Anexo II. elementos retrotransponibles identificados en muestras de pulmón	31
9.3	Anexo III. Evaluación de la red de coexpresión	32
9.4	Anexo IV. Resultados del análisis de enriquecimiento de la red de coexpresión.	32
9.5	ANexo V. Heatmap con elementos genómicos identificados en módulo 3	35
9.6	Otras figuras	35
9.6.1	Distribución de familias HERVs	36

# Índice de figuras

1	Clases y ejemplos de los principales elementos retrotransponibles (TEs). TIR: repeticiones invertidas terminales; LTR, repeticiones directas terminales largas (Long Terminal Repeat); Gag, antígeno de grupo; Prt, proteasa; Pol, polimerasa; Env, envoltura; UTR, región no traducida (untranslated region); ORF, marco de lectura abierto (open reading frame); HR, repetición de hexámero; VNTR, repeticiones en tándem de número variable (variable number tandem repeats); SINE-R, monómero derecho de Alu; SVA, compuesto SINE-R/VNTR/Alu; L, monómero izquierdo; R, monómero derecho. (Ukadike et al. 2021) . . . . .	7
2	Esquema de la metodología empleada para la caracterización de retrovirus endógenos humanos (HERVs) expresados tras sufrir COVID-19 usando datos de *RNA-seq*. (A) Secuenciación de ARN. (B) Cuantificación del retrotranscriptoma. (C) Cuantificación del transcriptoma. (D) Análisis de la expresión diferencial de HERVs y genes. (E) Análisis de coexpresión de HERVs y genes expresados diferencialmente en muestras COVID-19. . . . .	13
3	HERVs y LINE-s diferencialmente expresados (DEH) en muestra de pulmón de pacientes muertos por neumonía aguda tras pasar la COVID-19 (grupo de pacientes COVID-19). (A) Gráfico de volcán que muestra los ERs diferencialmente expresados, sobreexpresados (lado derecho) e infra-expresados (lado izquierdo) en muestras de pulmón del grupo de pacientes COVID-19. Las líneas discontinuas de corte indican un valor de p igual a 0.05 (eje x) y un valor de cambio expresado en log2 (LFC) mayor de 1,5 (eje y). Los ERs sobreexpresados e infraexpresados, con significación estadística (p<0.05), se muestran en color verde y rojo, respectivamente. (B) Distribución de los DEH según su posición génica. (C) Distribución de retrovirus endógenos humanos (HERVs) diferencialmente expresados en muestras COVID-19 agrupados en grupos de familias HERVs. Se representa la relación entre las frecuencias de los HERVs diferencialmente expresados en muestras COVID-19 y la frecuencia de familias HERV en la base de datos de anotación inicial. La línea discontinua roja indica la relación esperada, calculada como el cociente entre el número total de HERVs en la base de datos y el número de DEH. . . . .	17
4	Resultados del análisis de expresión diferencial del transcriptoma de pacientes graves de COVID. (A) Gráfico de volcán que muestra los genes diferencialmente expresados (DEG), sobreexpresados (lado derecho) e infraexpresados (lado izquierdo) en muestra de pulmón de pacientes graves de COVID-19. Las líneas de corte indican un valor de p igual a 0.05 (eje x) y un valor de cambio expresado en log2 (LFC) mayor de 1,5 (eje y). Los genes sobreexpresados e infraexpresados, con significación estadística, se muestra en color verde y rojo respectivamente. (B) Principales rutas biológicas (eje Y) sobrerepresentadas y relacionadas con modificaciones de la cromatina y respuesta inmunitaria, que aparecen en pacientes graves de COVID, y los genes sobreexpresados que contienen (eje X) . . . . .	19
5	Gráfico de circo de los HERVs (DEHs, anillo interno) y genes (DEGs, anillo intermedio) diferencialmente expresados a lo largo del genoma. Los elementos genómicos sobreexpresados en el grupo COVID-19 se muestran en rojo, mientras que los infraexpresados se muestran en azul. . . . .	20
6	Resultados del análisis de coexpresión de genes y elementos retrotransponibles (TEs) en muestras de pulmón. A) Grupos de elementos genómicos (genes Y TEs) identificados en la red de coexpresión. El módulo 0 contiene los elementos genómicos que no se han podido clasificar en ninguno de los otros módulos. B) Asociación fenotípica entre los 5 grupos de elementos genómicos (módulos) hallados y sexo y pertenecer al grupo COVID-19. C) Gráfico de enriquecimiento tipo Manhattan de los módulos 1 y 3 en las bases GO, KEGG, Reactome, Transfac, miRTarBase, Human Protein Atlas, CORUM, Human Phenotype ontology, WikiPathways. D) Distribución de los genes y ERs diferencialmente expresados incluidos en el módulo 3 agrupados por las rutas halladas en el análisis funcional de enriquecimiento empleando la base de datos KEGG. . . . .	21
S1	Distribución del contenido GC por secuencia y el contenido relativo de duplicación encontrado en cada secuencia. Se representan en rojo las secuencias que muestran. A) Resultados del análisis de la distribución del contenido GC en todas las muestras secuenciadas. B) Resultados del análisis de la distribución del contenido GC en las muestras de pacientes secuenciadas controles. C) Resultados del análisis de la distribución del contenido GC en las muestras de pacientes secuenciadas COVID-19. . . . .	30

S2	Distribución del contenido GC por secuencia y el contenido relativo de duplicación encontrado en cada secuencia. Se representan en rojo las secuencias que muestran. A) Resultados del análisis de la distribución del contenido GC en todas las muestras secuenciadas. B) Resultados del análisis de la distribución del contenido GC en las muestras de pacientes secuenciadas controles. C) Resultados del análisis de la distribución del contenido GC en las muestras de pacientes secuenciadas COVID-19. . . . .	32
S3	Gráfico de volcán que muestra los retrovirus endógenos humanos (HERVs) diferencialmente expresados, sobrerregulados (lado derecho) e infraexpresados (lado izquierdo) en muestra COVID-19. Las líneas de corte indican un valor de p igual a 0.05 (eje x) y un valor de cambio expresado en log2 (LFC) mayor de 1,5 (eje y). Los HERVs sobreexpresados e infraexpresados, con significación estadística, se muestra en color verde y rojo respectivamente. . . . .	35
S4	Frecuencias de retrovirus endógenos humanos (HERVs) distribuidos por grupos de familias. (A) Frecuencias de HERVs agrupados por grupos de familias incluidos en la base de datos utilizada para anotar las secuencias identificadas (expresadas en log10). (B) Frecuencias de HERVs agrupados por grupos de familias identificados en los HERVs diferencialmente expresados (DEH) encontrados (expresadas en log10) (C) Relación entre las frecuencias de las familias de HERVs identificadas en los DEH y la frecuencia de familias HERV en la base de datos inicial. La línea discontinua roja indica la relación esperada, calculada como el cociente entre el número total de HERVs en la base de datos y el número de DEH. . . . .	37

## Índice de tablas

1	Resultados tras el alineamiento de las secuencias con Bowtie2. Las variables continuas están expresadas como mediana (rango intercuartílico) . . . . .	15
2	Resultados tras la cuantificación de elementos retrotransponibles en las secuencias analizadas usando el programa Telescope. Las variables continuas están expresadas como mediana (rango intercuartílico) . . . . .	15
S1	Resumen del análisis de los datos de secuenciación de ARN. Análisis descriptivo de los resultados del filtrado y mapeado de las lecturas de la secuenciación de ARN. Lecturas crudas (millones): cantidad total de lecturas crudas, expresadas en millones, obtenidas tras secuenciar las muestras. Este valor se obtiene sumando la cantidad de las lecturas 1 y 2 de cada muestra. Datos crudos: contenido en G en las lecturas, calculado multiplicando las lecturas crudas por la longitud de las secuencias, que es en este caso es 150. Eficacia: medida de la eficacia del filtrado de las lecturas tras eliminar las lecturas con baja calidad y las secuencias de los adaptadores. Este valor se calcula como el ratio entre las lecturas limpias entre las lecturas crudas, expresado en tantos por cien. Q20, Q30: medida de la calidad de las bases en la secuenciación, calculado como el cociente entre el recuento de bases con valor Phred (Q) mayor de 20 o 30, respectivamente, y el recuento de bases total. El valor Phred representa la probabilidad estimada de un error en la base identificada. Cuanto mayor es el valor de Phred, mejor es la calidad de la base secuenciada. GC: contenido de bases G y C referido al contenido total de bases en las lecturas. . . . .	29
S2	Resultado del análisis funcional de enriquecimiento empleando la terminología KEGG y REACTOME sobre los elementos genómicos incluidos en el módulo 3 de la red de coexpresión creada a partir de las expresiones génicas de las muestras de pulmón analizadas. . . . .	32

# 1 Resumen

**Introducción.** La enfermedad del coronavirus 19 (COVID-19) ha provocado más 7 millones de muertes por todo el mundo. Las manifestaciones clínicas de la COVID-19 pueden ser heterogéneas, incluyendo desde síntomas leves como fiebre, hasta graves como neumonía o el síndrome de dificultad respiratoria aguda (SDRA). Está causada por el coronavirus 2 del síndrome respiratorio agudo grave (SARS-CoV-2), el cual es capaz de modificar la regulación de la expresión génica de las células infectadas. Este trabajo pretende evaluar si la COVID-19 induce cambios en los perfiles de expresión de retrovirus endógenos humano (HERVs), que puedan estar relacionados con alteraciones en la regulación de la respuesta inflamatoria la cual está relacionada a la patología de la enfermedad.

**Objetivo.** El objetivo principal de este trabajo es identificar posibles mecanismos moleculares implicados en la severidad de la COVID-19, lo cuales podrían estar correlacionados con el retrotranscriptoma de pulmón y la expresión de HERVs en pacientes graves de COVID-19 fallecidos por neumonía.

**Materiales y métodos:** La metodología propuesta para este trabajo se divide en las siguientes etapas: (1) secuenciación de RNA de muestras de pulmón de 19 pacientes graves de COVID-19 y pulmones sanos, (2) filtrado de lecturas y análisis de calidad de las secuencias, (3A) cuantificación del retrotranscriptoma, alineando con Bowtie2 y cuantificando posteriormente con el programa Telescope, (3B) cuantificación del transcriptoma de las muestras, empleando el programa Salmon, (4) análisis de expresión diferencial de HERVs, LINE-1 y genes, (5) integración de los datos ómicos con la creación de una red de coexpresión entre el transcriptoma y retrotranscriptoma identificado y (6) un análisis funcional de enriquecimiento usando bases de datos como KEGG y Reactome para facilitar la interpretación del análisis. Para ello, se han empleado herramientas bioinformáticas como la plataforma R, y SAMTools.

**Resultados.** Se hallaron 895 HERVs y LINE-1 y 838 genes diferencialmente expresados en muestras de pulmón de pacientes graves de COVID con respecto a muestras de pulmón del grupo control, destacando la sobreexpresión de HERV-K y HERV-W. Por otro lado, se hallaron 376 genes sobreexpresados y relacionados con rutas biológicas implicadas en modificaciones de la cromatina, como la metilación del DNA; así como genes implicados en la respuesta celular a estímulos. Finalmente, se estudió si la expresiones de los elementos genómicos estudiados, elementos retrotransponibles y genes, estaban reguladas de forma conjunta en pacientes graves de COVID-19, identificando un conjunto de 1090 elementos genómicos, 17 genes y 5 ERs que estaban muy correlacionados y se relacionaban con rutas metabólicas asociadas a la respuesta inmunitaria.

**Conclusiones.** Los pacientes graves de COVID-19 con neumonía muestran un aumento en las expresiones de HERVs. Así mismo, estos pacientes muestran aumentada la expresión de genes relacionados con la remodelación y modificación de la cromatina, cambios moleculares que modifican la conformación del material genético y favorecen la transcripción de regiones genómicas que suelen estar inactivas, como las secuencias de HERVs. Además, se identificaron grupos de HERVs y genes que se coexpresan en pacientes graves de COVID-19. Estos cambios están relacionados con alteraciones en la que contribuir a la respuesta inmunitaria exacerbada de pacientes con COVID severo por el poder antigénico que tienen las proteínas que codifican estas secuencias.

**Palabras clave:** COVID-19, retrovirus endógenos humanos, inflamación, *RNA-seq*

## 2 Introducción, antecedentes y objetivos

### 2.1 Introducción

Desde que se detectó el primer caso de la enfermedad del coronavirus 19 (COVID-19) en diciembre de 2019, la enfermedad se ha extendido por todo el mundo, provocando una pandemia que en 4 años ha acumulado más de 770 millones de casos de COVID confirmados (14 millones de personas infectadas en España) y más de 7 millones de muertes por todo el mundo (122 mil muertes en España), según datos de la Organización Mundial de la Salud (OMS) (“COVID-19 Cases | WHO COVID-19 Dashboard” 2024). Variables como la rápida tasa de mutaciones en el virus y la existencia de pacientes asintomáticos con capacidad de transmitir el virus, han influido en la rápida propagación de la enfermedad.

En la fase aguda de la COVID-19, las manifestaciones clínicas de la COVID-19 pueden ser heterogéneas, incluyendo principalmente fiebre y síntomas respiratorios que varían desde síntomas leves como tos seca, dolor de garganta y disnea, hasta síntomas más graves como neumonía o el síndrome de dificultad respiratoria aguda (SDRA); llegando a provocar incluso la muerte (Huang et al. 2020). Sin embargo, el virus ha sido detectado en otros tejidos, como corazón, hígado, riñones, cerebro y muestras de sangre (Puelles et al. 2020); y se han descrito afectaciones multisistémicas (Synowiec et al. 2021) incluyendo complicaciones cardiovasculares, como miocarditis y pericarditis (Fairweather et al. 2023); neurológicas (Harapan and Yoo 2021); gastrointestinales, como diarrea, náuseas y dolor abdominal (Shih and Misdraji 2023); dermatológicas, como erupciones cutáneas; e insuficiencia renal aguda (Zaborska et al. 2024).

Después de la fase aguda de la enfermedad, los síntomas suelen durar aproximadamente 5 días en la población general, aunque este número es muy variable (Sudre et al. 2021). Sin embargo, existen grupos de pacientes en los que los síntomas se prolongan durante semanas o incluso meses, después de la fase aguda, lo que se conoce como “Long COVID” o COVID persistente (“Coronavirus Disease (COVID-19): Post COVID-19 Condition” 2024). Existen diversas hipótesis sobre los mecanismos implicados en el desarrollo del long, que incluyen la desregulación del sistema inmune, procesos de autoinmunidad, alteraciones en la coagulación y alteraciones del endotelio y alteraciones en la señalización neurológica (Davis et al. 2023).

Por otro lado, en pacientes infectados por el virus SARS-CoV-2 se han encontrado aumentados los niveles de HERVs y proteínas de HERVs, las cuales tienen la capacidad de impactar en la respuesta inmunitaria, incrementando el entorno inflamatorio, así como la producción de citocinas proinflamatorias y el daño tisular. Como consecuencia, esta modulación se asocia con un pronóstico clínico desfavorable.”

### 2.2 Antecedentes

#### 2.2.1 Infección por el coronavirus 2 del síndrome respiratorio agudo grave (SARS-CoV-2) y la enfermedad del coronavirus 19 (COVID-19)

La enfermedad del coronavirus 19 (COVID-19) es una enfermedad infecciosa causada por el coronavirus 2 del síndrome respiratorio agudo grave (SARS-CoV-2) que provocó una pandemia en el año 2020 Zhu et al. (2020). El SARS-CoV-2 es un virus encapsulado perteneciente a la familia *Coronaviridae* con un genoma de ácido ribonucleico (ARN) de sentido positivo y cadena sencilla, con una longitud aproximada de 29.9 kb, capaz de infectar a humanos y otros mamíferos (Huang et al. 2020). El virus se transmite por vía respiratoria, principalmente a través de gotículas respiratorias y aerosoles (Synowiec et al. 2021).

La replicación del SARS-CoV-2 está mediada por una ARN polimerasa dependiente de ARN (RdRP), la cual se encarga de replicar su genoma y transcribir ARN subgenómico; y la enzima exorribonucleasa N-terminal (ExoN), encargada de corregir los errores cometidos durante el proceso. La acción conjunta de ambas enzimas, junto con la transcripción discontinua típica en los coronavirus, favorecen un alto ratio de recombinaciones, inserciones, deleciones y mutaciones puntuales facilitando la aparición de variantes genéticas (Carabelli et al. 2023) con diferentes propiedades de transmisibilidad y antigenicidad.

Estudios previos sugieren que el virus SARS-COV-2 podría modificar la regulación de la expresión génica de las células infectadas (Bignon et al. 2022), a través de modificaciones epigenéticas como alteraciones en la metilación del ADN, alteraciones de las histonas y remodelamiento de la cromatina de estas células Kee et al. (2022).

### 2.2.2 La respuesta inmunitaria frente a la infección por SARS-CoV-2

El sistema inmunológico es una compleja red de células y moléculas que trabajan de forma coordinada para defender al huésped frente a patógenos o moléculas reconocidas como extrañas. Principalmente, se distinguen dos tipos de mecanismos inmunológicos: la inmunidad innata, primer mecanismo de defensa activado, que se caracteriza por ser una respuesta rápida e inespecífica; y la inmunidad adaptativa, caracterizada por su acción específica. En la inmunidad innata participan principalmente los macrófagos, neutrófilos, células dendríticas y células natural killer (NK); células que reconocen patrones moleculares asociados a patógenos (PAMPs) a través de receptores de reconocimiento de patrones (PRRs) y secretan moléculas como citoquinas pro-inflamatorias (interferones [IFN] de tipo I, interleucina 6 [IL-6] o IL-1 $\beta$ ) para detener la progresión de la infección viral y para la activación de otras células del sistema inmunitario.

Tras la infección por SARS-COV-2, los PRRs reconocen PAMPs propios del virus que inducen una cascada de procesos celulares, incluyendo activación de genes relacionados con la inmunidad innata, inducción de muerte celular mediada por células inmunitarias y al producción de citoquinas inflamatorias (IL-6, IL-1 $\beta$ , TNF- $\alpha$ ) y quimiocinas (Diamond and Kanneganti 2022). En ocasiones, la producción de citoquinas se intensifica creando mayor nivel inflamación y daño tisular, estado que recibe el nombre de “tormenta de citocinas” (Karki and Kanneganti 2022). En este sentido, aquellos pacientes de COVID-19 con síntomas graves con frecuencia mostraron una respuesta inmune innata disminuida o retrasada, con alteraciones en la expresión génica de los interferones tipo I; y una producción elevada de citoquinas pro-inflamatorias, como la IL-6 (Minkoff and tenOever 2023).

### 2.2.3 Retrovirus endógenos humanos (HERVs)

El genoma humano presenta aproximadamente un 45% de elementos retrotransponibles (ER), un tipo de material genético que puede moverse dentro de un genoma. Los ERs tienen la capacidad de provocar mutaciones genéticas e influir en la expresión genética debido a su capacidad de replicarse e insertarse en diferentes partes del genoma (4 et al. 2001). En base al mecanismo de transposición, los ERs se pueden clasificar en dos grandes tipos (figura 1):

- **Transposones de ADN:** elementos que se mueven directamente como segmentos de ADN de una parte del genoma a otra (transposición conservativa) sin intermediarios de ARN (4 et al. 2001).
- **Retrotransposones:** son elementos genómicos que se transponen a través de un intermediario de ARN (transposición replicativa), el cual es retrotranscrito de nuevo a ADN antes de ser insertado en una nueva ubicación del genoma. Constituyen el 42% del genoma humano y se pueden subdividir en dos subtipos principales según la presencia o no de repeticiones directas terminales largas (LTR, Long Terminal Repeat) flanqueando la región interna codificante:
  - **Retrotransposones con LTR.** Estos elementos tienen secuencias repetitivas largas en sus extremos y constituyen el 8,3% del genoma humano (4 et al. 2001). Dentro de este grupo se encuentran los retrovirus endógenos humanos (HERVs), secuencias virales integradas en el genoma humano.
  - **Retrotransposones no LTR,** que a su vez se dividen en elementos nucleares largos intercalados (LINEs), elementos largos y autónomos que contienen las enzimas necesarias para su propia transposición, destacando el elemento LINE-1; y elementos cortos nucleares intercalados (SINEs), los cuales dependen de las enzimas producidas por los LINEs para su transcripción (4 et al. 2001).

Los retrovirus (de la familia de virus *Retroviridae*) son una familia de virus que tienen la capacidad de convertir su ARN viral en ADN complementario (ADNc) por la presencia de la enzima transcriptasa inversa. Durante el proceso de infección, una enzima denominada integrasa facilita la inserción del ADN viral en el ADN de la célula huésped. Los retrovirus endógenos (Endogenous retroviruses, ERVs) son secuencias virales que derivan de infecciones de retrovirus ancestrales que se integraron en el genoma germinal y se han transmitido verticalmente a través de generaciones, pasando a formar parte de nuestro genoma. En este subgrupo encontramos los retrovirus endógenos humanos (HERVs), los cuales han sido relacionados por múltiples estudios con diversas enfermedades Nali et al. (2022) y alteraciones en la expresión génica (Carter et al. 2022). Éstos se pueden clasificar en clases en función a la homología de secuencia de los elementos con retrovirus exógenos conocidos; la estructura y organización genómica dentro del genoma humano, incluyendo la presencia de genes típicos de retrovirus como *gag*, *pol* y *env*; y por sus características








human genome			
<b>DNA transposons</b>			
Mariner		~1.4 kb	<b>3%</b>
<b>Retrotransposons</b>			
<b>AUTONOMOUS</b>			
HERV		7–9.5 kb	<b>8%</b>
LINE1, L1		6 kb	<b>20%</b>
<b>NON-AUTONOMOUS</b>			
SVA		< 3 kb	<b>0.15%</b>
Alu		300 bp	<b>11%</b>

Figura 1: Clases y ejemplos de los principales elementos retrotransponibles (TEs). TIR: repeticiones invertidas terminales; LTR, repeticiones directas terminales largas (Long Terminal Repeat); Gag, antígeno de grupo; Prt, proteasa; Pol, polimerasa; Env, envoltura; UTR, región no traducida (untranslated region); ORF, marco de lectura abierto (open reading frame); HR, repetición de hexámero; VNTR, repeticiones en tándem de número variable (variable number tandem repeats); SINE-R, monómero derecho de Alu; SVA, compuesto SINE-R/VNTR/Alu; L, monómero izquierdo; R, monómero derecho. (Ukadike et al. 2021)

funcionales, como la capacidad para producir proteínas virales, la expresión de ARN mensajero y la potencial capacidad para formar partículas virales. De forma específica, destacan dos clases de HERVs:

- **HERVs Clase I.** Esta clase incluye HERVs que comparten similitudes estructurales y funcionales con retrovirus gamma, como los gammaretrovirus. Contienen en su estructura genómica los genes víricos *gag*, *pol* y *env*. Además, poseen la capacidad de producir partículas virales. Dentro de esta clase encontramos la familia HERV-H, una de las familias más abundantes de HERVs que parecen estar implicados en la regulación de la expresión génica durante el desarrollo embrionario (Carter et al. 2022).
- **HERVs Clase II.** Similares en estructura y organización a los retrovirus beta (como los betaretrovirus), incluyendo en su estructura genómica los genes víricos *gag*, *pol* y *env*. Algunos integrantes de la familia pueden expresar y formar elementos virales, como la familia HERV-K (HMKL-2), una de las familias más estudiadas y activas de HERVs en el genoma humano. Estos elementos han sido asociados con diversas enfermedades y condiciones, incluyendo ciertos tipos de cáncer, como cáncer de mama o pulmón; y enfermedades autoinmunes (Dervan et al. 2021). Otra familia importante es la HERV-W, cuya expresión también ha sido relacionada con enfermedades como la esclerosis múltiple y otras enfermedades autoinmunes (Nali et al. 2022).
- **Clase III.** Esta clase incluye HERVs que comparten similitudes con los spumavirus, que son un tipo especial de retrovirus. Un ejemplo es HERV-L, la familia más antigua de HERVs en el genoma humano con menor actividad biológica que otras familias de HERVs.

La infección de algunos virus, como el virus de la influenza A (Liu et al. 2022) y SARS-CoV-2 (Petrone et al. 2023), pueden activar la transcripción de HERVs, y los productos resultantes de su transcripción (ADN, ARN y proteínas) pueden ser reconocidos por PRRs, induciendo así una respuesta inmunológica (Duperray et al. 2015) que contribuye al estado de inflamación generado por la infección original.

La expresión génica está modulada a nivel transcripcional por la unión de factores de transcripción a regiones del ADN próximas al gen en cuestión (regiones promotoras, enhancers (potenciadores) o silencers (silenciadores)); y por modificaciones de la cromatina. Además, también destaca la regulación epigenética, un mecanismo de regulación que participa en la modulación de los genes, sin modificar la secuencia de ADN ni ARN. Los mecanismos epigenéticos más destacados son la metilación del ADN, particularmente

en las islas CpG; y modificación de las proteínas que participan en la condensación del ADN (histonas), como metilación y acetilación de histonas (Smith and Meissner 2013). Cambios epigenéticos impactan en la conformación del ADN y han sido relacionados con numerosas patología y enfermedades autoinmunes (Allis and Jenuwein 2016).

La hipótesis de este trabajo es que la expresión de HERVs aumenta la respuesta inflamatoria y contribuye a la severidad de los síntomas en los pacientes graves con neumonía asociada a COVID.

## 2.3 Objetivos

El objetivo principal de este trabajo es identificar mecanismos moleculares implicados en la severidad de la COVID-19. Para ello, se han abordado los siguientes objetivos específicos:

- Caracterización de la expresión diferencial del retrotranscriptoma de pulmón de pacientes graves de COVID-19 fallecidos por neumonía frente a controles sanos.
- Identificación de cambios en el transcriptoma de muestras de pulmón de pacientes graves de COVID-19 fallecidos por neumonía frente a controles sanos.
- Análisis de la correlación entre los cambios en el retrotranscriptoma con los cambios en el transcriptoma de pacientes graves de COVID-19 fallecidos por neumonía.

## 3 Materiales y Métodos

En esta sección se detallan los materiales y métodos empleados para la realización del trabajo.

### 3.1 Materiales

#### 3.1.1 Muestras analizadas

Se analizó el transcriptoma y el retrotranscriptoma de un total de 19 muestras humanas de pulmón. Las muestras de pacientes con COVID-19 (n=9) se obtuvieron de biopsias postmortem realizadas a pacientes graves que habían fallecido a causa de una neumonía aguda asociada a COVID-19. Por otro lado, las muestras del grupo control (n=10) se obtuvieron de biopsias del pulmón sano de pacientes vivos con cáncer en el otro pulmón.

#### 3.1.2 Tecnologías empleadas

##### 3.1.2.1 Herramientas de soporte y lenguajes de programación

**3.1.2.1.1 GitHub** GitHub es una plataforma que permite crear repositorios en los que poder almacenar y compartir archivos con otros usuarios. Además, este repositorio implementa un sistema de control de versiones que permite realizar un seguimiento de los cambios aplicados en los archivos.

Todos los scripts empleados para la realización de este trabajo se encuentran disponibles en un repositorio de [GitHub](https://github.com/AzaharaGS/TFM_HERVs_COVID): [https://github.com/AzaharaGS/TFM\\_HERVs\\_COVID](https://github.com/AzaharaGS/TFM_HERVs_COVID)

**3.1.2.1.2 R version 4.4.0** R es un lenguaje y un entorno de software libre para la manipulación de datos, realización de análisis estadísticos y representación de gráficos (Team 2013). Esta herramienta puede amplificar sus funciones con el uso de extensiones denominadas librerías (*packages*), que están disponibles en el repositorio [CRAN](#) (ComprehensiveR Archive Network), gestionado por la comunidad de R; y en [Bioconductor](#), (3.19), un proyecto de desarrollo de software de libre acceso que utiliza el lenguaje de R y proporciona herramientas análisis de datos biológicos, y especialmente, el análisis de datos genómicos (Huber et al. 2015).

Para la realización de este proyecto, se empleó la versión 4.4.0 de R, a través de la interfaz [RStudio](#) (2024.04.1+748) (RStudio 2016).

A continuación, se listan las librerías paquetes empleadas en este trabajo:

- **knitr 1.47** Esta librerías contiene múltiples herramientas que permiten generar informes en R (Xie 2017). En este trabajo se ha empleado este paquete para crear el informe de la memoria.
- **dplyr 1.1.4** . Se trata de una librerías que proporciona un lenguaje propio para manipular bases de datos (Wickham et al. 2023).
- **readr 2.1.5** y **readxl 1.4.3**. Son librerías que permiten leer o importar bases de datos en formato rectangular, como `.csv` y `.tsv`; y en archivos excel (`.xls` y `.xlsx`), respectivamente.
- **writexl 1.5.0**. Es una librería que permite la escritura de archivos en formato `.xls` y `.xlsx`.
- **gtsummary 1.7.2**: librería que proporciona una manera de resumir tablas comparativas o descriptivas en formato de publicación, sin necesidad de utilizar otros programa para procesarlas.
- **ggplot2 3.5.1** La librería `ggplot2` consiste en un sistema para crear gráficos de forma declarativa, es decir, indicando qué se desea visualizar en el gráfico haciendo uso de una gramática de gráficos. Ha sido empleado en este trabajo para graficar la distribución de las familias de HERVs identificadas en las muestras de pacientes con COVID-19.
- **patchwork 1.2.0**. Esta librería es una extensión del paquete `ggplot` que facilita la combinación de gráficos en una sola figura. En este trabajo fue empleado para crear las figuras compuestas incluidas.
- **TxDb.Hsapiens.UCSC.hg38.knownGene 3.18.0**. Esta librerías contiene bases de datos de anotaciones generadas por UCSC y se ha empleado para crear objetos `TxDd` y agrupar los transcritos cuantificados a nivel de genes.
- **org.Hs.eg.db 3.19.1**. Esta librerías contiene una amplia anotación del genoma humano basada en alineamientos usando identificadores de genes Entrez.
- **Tximport v1.32.0**. La librería `tximport` ayuda a integrar estimaciones de abundancias a nivel de transcritos y convertirlas en matrices de recuentos que podrán ser utilizadas para análisis posteriores usando paquetes estadísticos como `DESeq2` (Soneson, Love, and Robinson 2015). El proceso se realiza en dos etapas: **(1)** asociar los transcritos con los IDs de genes, e **(2)** importar los datos desde los ficheros que contienen las estimaciones a nivel de transcrito. Esta herramienta se ha empleado para integrar los recuentos de transcritos a nivel de genes antes de realizar los análisis posteriores.
- **DESeq2 v1.44.0**. `DESeq2` es una librería de R que proporciona un método de análisis de expresión diferencial de datos de *RNA-seq* (Love, Huber, and Anders 2014). El archivo de entrada debe ser una matriz de recuentos en crudo de *RNA-seq*, recomendando emplear con recuentos a nivel de gen para obtener resultados más precisos y robustos (Soneson, Love, and Robinson 2015). Para el análisis de expresión diferencial, el programa aplica modelos de regresión binomial negativa (distribución típica en datos de conteo) a la matriz de recuentos, donde cada fila representa un gen y cada columna representa una muestra; y normaliza los recuentos aplicando un factor de normalización (`size factors`). Para el análisis de la expresión diferencial, `DESeq2` permite el uso de dos tipos de test: Likelihood Ratio Test (LRT) y el test de Wald. En este trabajo, se aplicó el test de Wald, un test paramétrico que genera p-valores específicos para cada gen, y ajusta estos valores por múltiples comparaciones empleando la tasa de descubrimiento falso (False Discovery Rate, FDR) [Benjamini y Hockberg, 1995]. Por último, el programa genera una tabla con los coeficientes y la magnitud del cambio en la expresión del gen (*Fold Change*) en Log2 (LFC) entre la población a estudio y el control.
- **EnhancedVolcano 1.22.0**. Librería que representa gráficos de volcán donde aparecen representados todos los genes analizados en el análisis de expresión diferencial. Además, permite señalar los grupos de genes que cumplen ciertos criterios de valores de Fold Change y valores de p. Se ha empleado esta librería para representar los resultados de los análisis de expresión diferencial.
- **gprofiler2** (Kolberg et al. 2020). Se trata de una herramienta que permite realizar análisis funcionales de enriquecimiento y la visualización de los mismos en el entorno de R conectando con la herramienta web 'g:Profiler' (<https://biit.cs.ut.ee/gprofiler/>). En este trabajo se ha empleado para facilitar la interpretabilidad de los genes diferencialmente expresados.

- **GWENA v1.14.0.** Esta librería permite crear redes de coexpresión y realizar una caracterización completa de los nodulos identificados en dicha red, incluyendo un análisis de enriquecimiento, estudios de asociaciones fenotípicas, identificación de genes hub y análisis de coexpresión diferencial. En este trabajo, hemos utilizado este paquete para realizar crear una red de coexpresión entre los valores de expresión de los elementos retrotransponibles y los genes analizados.
- **circlize 0.4.16** (Gu et al. 2014). Esta librería permite crear figuras circulares para visualizar grandes cantidades de información genómica, permitiendo crear gráficos *circle* para la representación de genomas completos. En este trabajo, se ha empleado para crear un gráfico circular representando las localizaciones de los elementos genómicos diferencialmente expresados entre los dos grupos de estudio.
- **biomaRt 2.60.0.** biomaRt incluye funciones y métodos para la manipulación, visualización y análisis de datos de datos ómicos. En este trabajo se ha empleado para obtener las posiciones de los genes estudiados (Durinck et al. 2005).

Además de las librerías indicadas, se ha empleado la herramienta *R Markdown* (Allaire et al. 2023), una función de RStudio que permite integrar texto con código de R u otros lenguajes de programación como *bash*. En este trabajo se ha empleado para la realización de los análisis y la creación de la presente memoria.

**3.1.2.2 FasQC v0.12.1** FasQC es un programa que permite realizar un análisis de calidad en datos crudos de secuenciación de alto rendimiento. Los análisis por secuencia incluidos son: **(a)** calidad de la secuencia por base en cada posición; **(b)** puntuaciones de calidad; **(c)** contenido de cada base (A,G,T,C), mostrando la proporción de cada base en cada posición del archivo; **(d)** contenido GC, comparando el contenido en la muestra frente a la distribución teórica, **(e)** contenido de bases desconocidas (N); **(f)** distribución de la longitud de las secuencias; **(g)** porcentaje de secuencias duplicadas; **(h)** contenido de secuencias sobrerrepresentadas, **(i)** contenido de secuencias de adaptadores; y **(j)** análisis de calidad de secuencia por mosaico, que muestra la desviación del promedio de la calidad en cada posición de la lectura, mostrando en colores fríos aquellas posiciones donde la calidad es igual o superior al promedio, mientras que los colores cálidos indican calidades peores. Este programa se empleó para analizar la calidad de las secuencias crudas de ARN.

**3.1.2.3 QualiMamp 2.2.2** QualiMap es un programa que permite realizar control de calidad de los datos de secuenciación alineados (Okonechnikov, Conesa, and García-Alcalde 2016). En este trabajo se ha utilizado para realizar el control de calidad del alineamiento con *Bowtie2*.

**3.1.2.4 MultiQC 1.23** MultiQC es una herramienta bioinformática que permite agregar informes de resultados de varias muestras a partir de un sólo informe (Ewels et al. 2016). En este trabajo, se ha empleado para integrar los resultados obtenidos tras realizar el análisis de la calidad de las secuencias crudas de ARN y las secuencias alineadas.

**3.1.2.5 Telescope v1.0.3** Telescope es un programa, escrito en lenguaje de programación Python, que proporciona estimaciones de la expresión de elementos retrotransponibles (ER) en localizaciones genómicas específicas (Bendall et al. 2019). Para ello, emplea un modelo Bayesiano mixto que permite reasignar los fragmentos mapeados de forma ambigua al transcrito más probable, asumiendo que el número de fragmentos generados por un transcrito es proporcional a la cantidad de transcrito presente en la muestra.

Para cuantificar los HERVs presentes en secuencias obtenidas por secuenciación de ARN (*RNA-seq*), el programa requiere la entrada de dos archivos:

- Un archivo con el alineamiento de las lecturas a un genoma de referencia, en formato BAM/SAM y ordenado por nombre de lectura. El alineador empleado para obtener este archivo debe permitir realizar una búsqueda de alineamiento local sensible y reportar múltiples alineamientos válidos por cada fragmento mapeado.
- Un fichero de anotaciones de TE, en formato GTF, que incluya las regiones genómicas que representan a los transcritos de TE.

Tras su ejecución, Telescope genera dos archivos de salida: un archivo que contiene los recuentos de HERVs estimados (`telescope-TE_counts.tsv`) y un archivo que contiene el informe estadístico del proceso (`telescope-stats_report.tsv`). Para este trabajo, se creó un script de bash para procesar todos los ficheros que contenían los estadísticos del proceso (`extracting_telescope_stats_v02.sh`).

**3.1.2.6 Bowtie2 v2.5.3** Bowtie2 es un programa que permite alinear lecturas de secuenciación frente a un genoma de referencia (Langmead and Salzberg 2012). Se recomienda su uso especialmente para alinear secuencias de longitud comprendida entre 50 o 1000 caracteres, y para alinear genomas relativamente grandes como los genomas de mamíferos.

El programa funciona por defecto en el modo de alineamiento *end-to-end*, en el cual debe alinearse la lectura completa. Sin embargo, para este trabajo se ha empleado el modo de alineamiento *local*, en el cual no participan algunos caracteres del final de la lectura mapeada.

Bowtie2 requiere dos ficheros de entrada para su correcto funcionamiento:

- Un fichero que contenga un índice del genoma de referencia generado por el propio programa.
- Un fichero de lecturas de secuenciación, en formato FASTQ.

El archivo que resulta del alineamiento con Bowtie2 es un archivo SAM, dónde cada línea describe un alineamiento, o una lectura si el alineamiento de dicha lectura ha fallado.

**3.1.2.7 Salmon v1.10.2** Salmon es un pseudo-alineador que permite cuantificar directamente las abundancias de los transcritos de las lecturas crudas de *RNA-seq* mediante un método de mapeo denominado *quasi-mapping* (Patro et al. 2017). El método *quasi-mapping* consiste en un procedimiento de inferencia combinado con un modelo de sesgo, que consta de dos fases:

1. Fase de creación del índice a partir del transcriptoma de referencia, para determinar la información de posición y orientación sobre dónde se mapean mejor los fragmentos. En esta fase se requiere un fichero, en formato FASTA, que contenga el transcriptoma de referencia.
2. Fase de cuantificación, en la que los niveles de expresión, estimados inicialmente en los archivos crudos de *RNA-seq*, se refinan empleando la información del índice. Para ello, se necesita un fichero que contenga un índice del genoma de referencia generado por el propio programa; y los ficheros que contengan las lecturas de secuenciación, en formato FASTQ.

Al finalizar, Salmon reporta un archivo de cuantificación denominado *quant.sf*, en el que hay tantas filas como transcritos cuantificados, y 5 columnas que indican el nombre del transcrito incluido en el transcriptoma de referencia, la longitud del transcrito de referencia en nucleótidos, la longitud *efectiva* calculada del transcrito de referencia, teniendo en cuenta todos los factores que pueden influir en la probabilidad de mapear estos fragmentos de transcripción, como el sesgo específico de secuencia y los fragmentos GC, la estimación de las abundancias relativas de los transcritos en unidades de Transcritos Por Millón (TPM) y el número de lecturas mapeadas para cada transcrito que ha sido cuantificado.

**3.1.2.8 Fuentes de datos externas** En este apartado se incluyen las fuentes de datos externas que se han empleado para la realización de este trabajo:

- **Gene Ontology (GO)** (Ashburner et al. 2000). Se trata de una ontología que contiene términos para anotación de genes y proteínas. Los términos están agrupados en tres ontologías independientes: procesos biológicos (BP), funciones moleculares (MF) y componentes celulares (CC). Esta ontología ha sido empleada para analizar las funciones de los elementos genómicos estudiados.
- **Human Phenotype Ontology (HPO)** (Köhler et al. 2014). Esta ontología también ofrece anotaciones de genes y proteínas. Sin embargo, el objetivo de HPO es relacionar fenotipos clínicos con los genes o proteínas identificados.
- **Kyoto Encyclopedia of Genes and Genomes (KEGG)** . Se trata de una colección de bases de datos para el análisis sistemático de las funciones de genes (Kanehisa and Goto 2000).

**Reactome.** Base de datos de vías o rutas biológicas, incluyendo la señalización celular, el metabolismo y el ciclo celular, que proporciona información de procesos y reacciones biológicas en diferentes especies. Se ha empleado para realizar un análisis funcional de enriquecimiento,

## 3.2 Métodos

### 3.2.1 Extracción y secuenciación de ARN

Las muestras de pulmón se conservaron en la solución de estabilización RNAlater® (Thermo Fisher Scientific, Waltham, MA USA) y el ARN fue extraído usando el kit Maxwell® RSC simplyRNA Tissue Kit (Promega, Madison, WI, USA), según las instrucciones del fabricante. La concentración de ARN fue analizada usando el espectrofotómetro NanoDrop-2000 (Thermo Scientific), mientras que el bioanalizador Agilent 2100 (Agilent Technologies) y la electroforesis en gel de agarosa al 1% se emplearon para determinar la integridad y pureza del ARN, respectivamente.

La secuenciación de ARN (*RNA-seq*) fue realizada por la empresa Novogene. De forma resumida, se preparó una biblioteca de ARN de cadena específica después de eliminar el ARN ribosomal. A continuación, todas las muestras fueron secuenciadas en un secuenciador Illumina Novaseq6000 desde ambos extremos (paired-end) y con una longitud de lectura de 150 pares de bases (pb). Por último, los datos crudos de secuenciación se filtraron en base a los siguientes criterios: (1) se eliminaron lecturas que contuviesen secuencias de adaptadores, (2) se eliminaron lecturas en las que el porcentaje de bases desconocidas (N) fuese mayor del 10%, y (3) se eliminaron lecturas en las que más del 50% de las bases tuviesen una calidad inferior a 5.

Finalmente, la calidad de las secuencias en los datos crudos analizados se evaluó empleando el programa FastQC (Andrews et al. 2010).

### 3.2.2 Cuantificación del retrotranscriptoma: identificación de HERVs

Se identificó y cuantificó la presencia de secuencias de HERVs en los datos de *RNA-seq* usando el programa Telescope (Bendall et al. 2019) (Figura 2 B). Primero se empleó el alineador Bowtie2 (Langmead and Salzberg 2012) para alinear las lecturas contra el genoma de referencia (hg38), permitiendo un alineamiento local sensible (`--very-sensitive-local`) que reporta hasta 100 alineamientos (`-k 100`) con un umbral mínimo de puntuación de alineación del 95% o más identidad de secuencia (`--score-min L,0,1.6`) por cada par de fragmentos. A continuación, los archivos bam generados se ordenaron por nombre de lectura empleando la herramienta Samtools (`samtools sort -n`).

A continuación, el programa Telescope v1.0.3 fue empleado para identificar y cuantificar los HERVs usando como archivos de entrada los ficheros bam ordenados junto con un fichero de anotaciones público que incluía las localizaciones de 14968 HERVs, agrupados en 60 familias de HERVs; y 13545 elementos LINE-1 presuntamente activos. Se emplearon los parámetros del modelo y las opciones de ejecución del modelo definidas por defecto, fijando un límite de 100 iteraciones del algoritmo de maximización de expectativas para estimar los parámetros del modelo Bayesiano usado para la reasignación de alineamientos y posterior identificación de los ERs (`-max_iter 100 -theta_prior 200000`). Finalmente, Telescope generó una tabla con la expresión de los ERs identificados expresados como valores de recuentos, la cual se empleó en posteriores análisis.

### 3.2.3 Transcriptoma: expresión de genes en *RNA-seq*

Para analizar los niveles de expresión de genes a partir de los datos de *RNA-seq* se empleó el programa Salmon (Figura 2 C) y el método *quasi-mapping* (Patro et al. 2017) que permite cuantificar directamente las abundancias de los transcritos sobre los ficheros de secuenciación. Para ello, primero se construyó el índice a partir del transcriptoma de referencia, ([Homo\\_sapiens.GRCh38.cdna.all.fa.gz](https://www.ncbi.nlm.nih.gov/assembly/GCF_000001435.1/)); y, a continuación, se cuantificaron los niveles de expresión de los transcritos presentes en los datos de secuenciación crudos. Debido a la observación de distribuciones atípicas en el contenido GC de algunas muestras (fig S1) se aplicó una corrección por contenido GC (`--gcBias`) de las lecturas al realizar la cuantificación del transcriptoma.

Por último, se empleó la función `tximport()` del paquete de R `tximport` y un fichero de cuantificación SF para obtener la matriz de expresión en valores de recuentos a nivel de genes (Soneson, Love, and Robinson 2015).

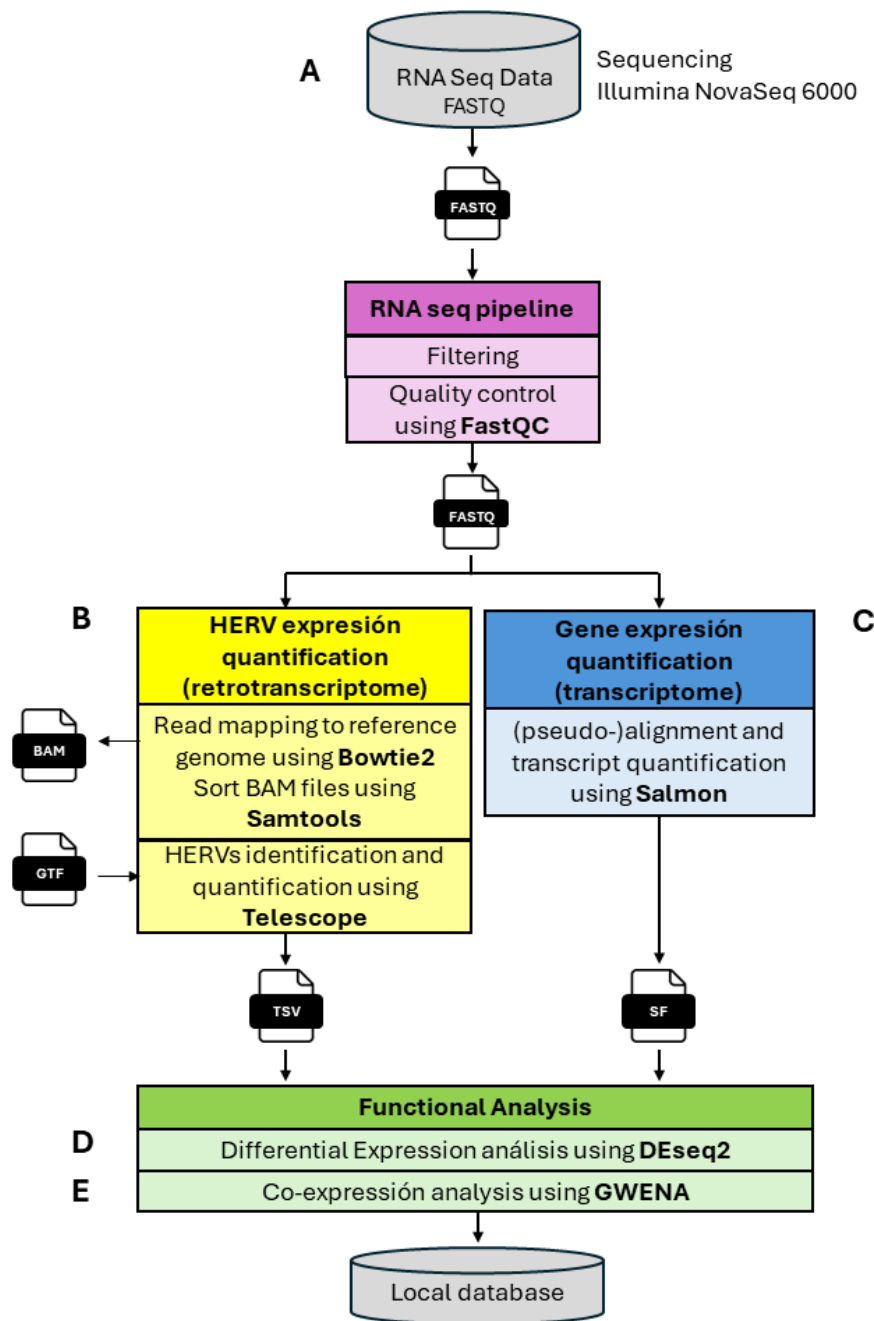


Figura 2: Esquema de la metodología empleada para la caracterización de retrovirus endógenos humanos (HERVs) expresados tras sufrir COVID-19 usando datos de \*RNA-seq\*. (A) Secuenciación de ARN. (B) Cuantificación del retrotranscriptoma. (C) Cuantificación del transcriptoma. (D) Análisis de la expresión diferencial de HERVs y genes. (E) Análisis de coexpresión de HERVs y genes expresados diferencialmente en muestras COVID-19.

### 3.2.4 Análisis de expresión diferencial

El análisis de expresión diferencial entre muestras de pulmón de pacientes graves de COVID-19 fallecidos por neumonía aguda y muestras de pulmón sano del grupo control se realizó empleando el paquete DESeq2 (Figura 2 D). Primero, se importaron las matrices de recuentos de la expresión de HERVs y genes empleando las funciones `DESeqDataSetFromMatrix()` y `DESeqDataSetFromTximport()`, respectivamente, e indicando la fórmula del diseño del análisis. A continuación, se filtraron las secuencias que presentaban recuentos iguales o superiores a 10 en, al menos, 9 muestras (tamaño del grupo COVID-19); y se aplicaron modelos de regresión binomial negativa y el test de significación de Wald para evaluar la expresión diferencial entre las muestras de pacientes con COVID-19 y control, empleando la función `DESeq()`. Por último, se construyó una tabla con los resultados empleando la función `results()`. Se consideraron diferencialmente expresados aquellos HERVs y genes si el valor de la *p* ajustada era inferior a 0.05 y el LFC en valor absoluto era mayor que 1.5.

### 3.2.5 Análisis de coexpresión

Para analizar la existencia de conjuntos de elementos genómicos (genes y TEs) que se coexpresen en pacientes graves de COVID-19, se creó una red de coexpresión para las expresiones génicas de elementos retrotransponibles y genes empleando el paquete de R GWENA (Lemoine et al. 2021) (Figura 2 E). Seguidamente, se analizó la estructura de la red de coexpresión obtenida y se evaluó si existía correlación entre los grupos de genes altamente correlacionados entre sí y los pacientes graves de COVID-19. Para estos análisis, se decidió coger toda la población en vez de sólo los sujetos con COVID-19 debido al bajo número de pacientes analizados.

Primero, se normalizaron los recuentos génicos utilizando la normalización por transformación estabilizadora de varianza (VST) (Anders and Huber 2010), la cual transforma la matriz de recuentos en una matriz con varianza constante a lo largo del rango de valores medios. A continuación, se filtraron los elementos reteniendo el 70% de los elementos y que además poseían un recuento mayor a 9 en la suma de todos los pacientes.

Para construir la red se utilizó la correlación de spearman y se permitió un valor de corte de 0.8 para el coeficiente de determinación. La red obtenida consistió en un grafo de grado de escala libre con nodos unidos con distintas intensidades. A continuación, se analizó la existencia de grupos de elementos genómicos que estuvieran relacionados fuertemente entre sí en la red obtenida. Estos grupos son denominados módulos, y su alta correlación sugiere que trabajan de forma conjunta para realizar un conjunto de funciones. Para la identificación de los módulos, se empleó el método de clustering jerárquico y aprendizaje no supervisado.

### 3.2.6 Análisis funcional de enriquecimiento

Para simplificar la interpretabilidad de los resultados de la red de coexpresión y dar un sentido biológico a la identificación de los módulos de la red de coexpresión asociados con la patología de estudio, se realizó un análisis funcional de enriquecimiento sobre la lista de elementos genómicos incluidos en los módulos de la red. Para ello, se empleó la función `bio_enrich()` del paquete GWENA (Lemoine et al. 2021).

## 4 Resultados

Esta sección muestra los resultados obtenidos tras realizar el análisis de la expresión diferencial de genes y retrovirus endógenos humanos (HERVs) en muestras de pulmón de pacientes con neumonía aguda que habían pasado la COVID-19 frente a muestras de pulmón sano.

### 4.1 Calidad de las secuencias

Las secuencias obtenidas tras la secuenciación de ARN por pares de extremos de las muestras de pulmón analizadas mostraron un contenido promedio de 137 millones de lecturas. El 93.36% (IQR: 92.93%, 93.77%) de las lecturas mostró un Q30 (tabla S1), lo que indicó alta calidad de secuencia. Sin embargo, se identificó un contenido de bases GC en las secuencias de las muestras de pacientes graves de COVID-19 (43.74



[IQR: 43.67, 44.34]%) significativamente menor que en las secuencias del grupo control (46.13 [IQR: 45.41, 48.04]%) (tabla S1), con perfiles que no seguía una distribución normal (Figura S1 A).

## 4.2 Identificación de elementos retrotransponibles después de la infección por SARS-CoV-2

Tabla 1: Resultados tras el alineamiento de las secuencias con Bowtie2. Las variables continuas están expresadas como mediana (rango intercuartílico)

Parámetros	Total, N = 19	Control, N = 9	COVID, N = 10	p-value
Lecturas analizadas (Millones)	135 (130, 142)	135 (123, 142)	135 (131, 142)	0.6
Lecturas alineadas (Millones)	115 (89, 122)	117 (106, 124)	115 (72, 119)	0.7
Lecturas alineadas (%)	88 (86, 89)	87 (87, 88)	88 (57, 89)	0.2
Ratio de error de alineamiento	0.39 (0.36, 0.45)	0.45 (0.42, 0.47)	0.37 (0.35, 0.38)	0.002

Se alinearon un promedio de 115 millones de lecturas por muestra, de las que 115 consiguieron alinearse (88%) (tabla 1). Después, se cuantificaron los ERs expresados en las muestras de pulmón empleando el programa Telescope. De las 53 268 653 (IQR: 47 121 292, 54 788 983) lecturas alineadas correctamente, se identificaron un promedio de 14,614 (IQR: 13 487, 15 069) secuencias de ERs en las muestras analizadas (Tabla 2). Asimismo, en las secuencias de los pacientes graves de COVID-19 se identificó un número más alto de secuencias de ERs (14 951 [IQR:14 679, 15 262]) que en las del grupo control (13 529 [IQR: 13 024, 14 311]).

Tabla 2: Resultados tras la cuantificación de elementos retrotransponibles en las secuencias analizadas usando el programa Telescope. Las variables continuas están expresadas como mediana (rango intercuartílico)

Parámetros	Overall, N = 19	Control, N = 9	COVID, N = 10	p-value
Fragmentos totales procesados	68 429 143 (66 410 792, 71 161 487)	67 645 324 (66 134 345, 71 000 875)	68 963 528 (66 689 838, 71 048 783)	0.6
Lecturas pareadas alineadas	53 268 653 (47 121 292, 54 788 983)	52 241 741 (47 422 616, 54 151 647)	53 439 841 (12 766 349, 55 049 984)	0.14
Lecturas pareadas alineadas a ubicaciones distintas del genoma	9 805 488 (8 869 520, 12 662 033)	9 937 791 (9 299 166, 12 448 392)	9 614 832 (8 667 842, 48 400 348)	0.11
Fragmentos no alineados	4 706 009 (4 402 758, 6 533 982)	5 004 224 (4 398 654, 5 838 808)	4 655 361 (4 416 270, 6 599 691)	0.5
Alineamientos únicos sin ambigüedad	55 507 835 (50 776 256, 58 112 033)	48 487 601 (47 084 764, 55 507 835)	55 884 017 (54 993 938, 58 309 173)	0.030
Alineamientos ambiguos	8 104 490 (6 927 066, 10 035 373)	10 319 793 (8 982 318, 11 388 457)	6 927 066 (6 471 421, 7 717 822)	0.007
Superposiciones con regiones únicas	2 361 658 (2 169 018, 2 515 640)	2 159 434 (1 915 038, 2 346 467)	2 434 240 (2 376 778, 2 548 581)	0.014
Superponen con regiones ambiguas	1 391 533 (1 292 604, 1 650 493)	1 405 808 (1 326 795, 1 722 679)	1 386 018 (1 277 130, 1 580 130)	0.7
Lecturas que no han sido asignadas a ninguna secuencia de TE conocida	1 334 621 (1 238 644, 1 600 887)	1 355 512 (1 273 895, 1 657 007)	1 328 387 (1 222 907, 1 535 464)	0.8
TEs con cero lecturas	4 791 (4 565, 5 076)	4 800 (4 755, 4 993)	4 607 (4 390, 5 192)	0.9

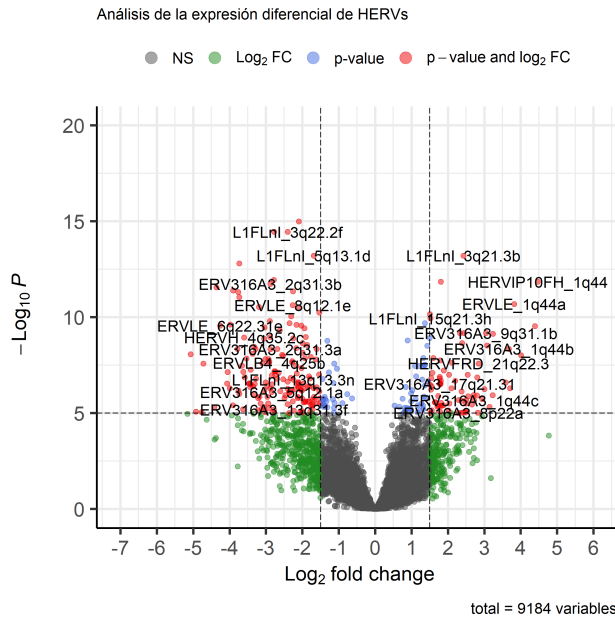
Parámetros	Overall, N = 19	Control, N = 9	COVID, N = 10	p-value
TEs no detectados	5 187 (4 979, 5 889)	6 256 (5 519, 6 422)	5 034 (4 701, 5 119)	0.002
TEs con alguna lectura	14 614 (13 487, 15 069)	13 529 (13 024, 14 311)	14 951 (14 679, 15 262)	0.010

Para evaluar la expresión diferencial de HERVs en pacientes graves de COVID-19, se analizó la expresión diferencial de 9184 elementos retrotransposones en 19 muestras de pulmón. Se identificaron 895 (9.75%) HERVs y elementos LINE-1 potencialmente activos diferencialmente expresados (DEH) en muestras de pulmón de pacientes graves de COVID-19 con respecto a muestras de pulmón del grupo control (Figura 3 A);

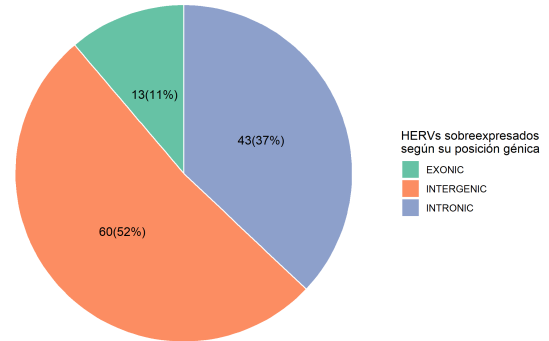
HERVs estaban sobreexpresados, el de estos elementos se localiza en regiones intergénicas del genoma (Figura 3 B). La tabla que contiene la lista completa de todos los DEH está disponible en el repositorio. Dicha tabla incluye los valores de expresión media en todas las muestras analizadas, los valores de cambio en el grupo de pacientes con COVID-19 frente al grupo control, los valores de p ajustados, las localizaciones génicas de cada uno y la familia a la que pertenecen.

Dentro de los HERVs sobreexpresados, se encontró que los grupos de familias HERV-K y Harlequin estaban sobrerrepresentada en los HERVs sobreexpresados en muestras de pulmón de pacientes graves de COVID-19 (Figura 3 C). También se observó gran representación de las familias HERV-H,HERV-W y ERV1.

A



B



C

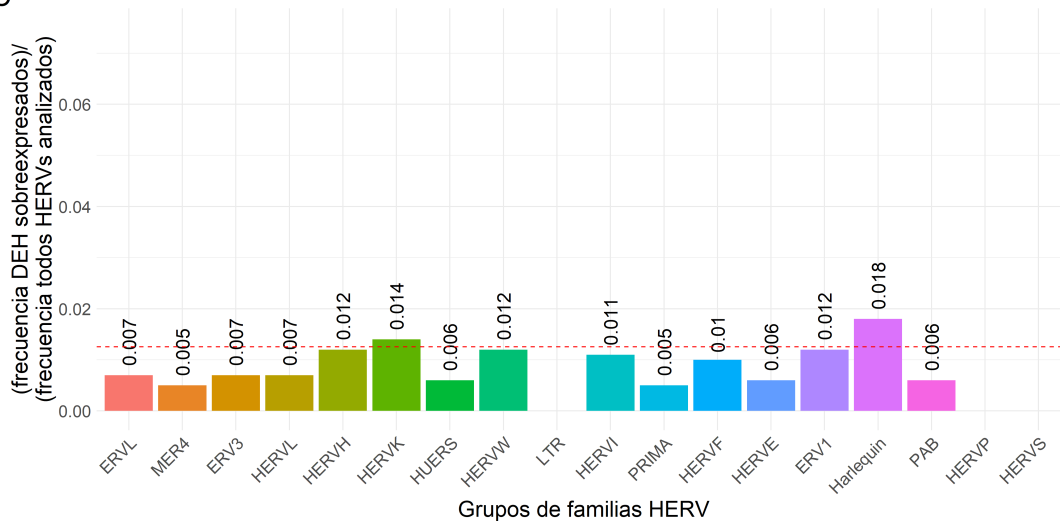
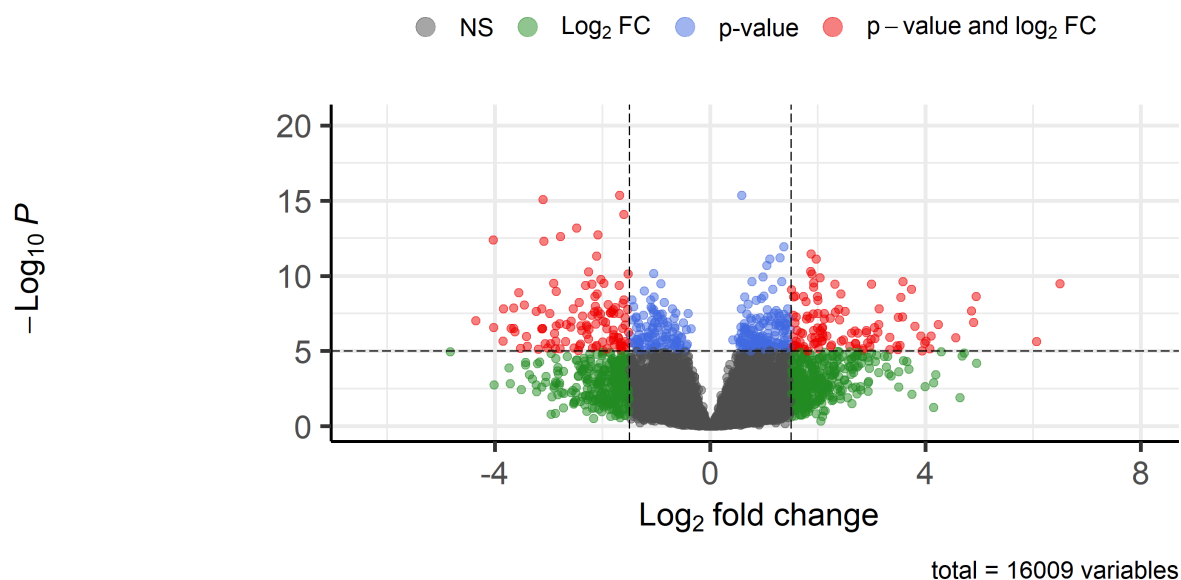


Figura 3: HERVs y LINE-s diferencialmente expresados (DEH) en muestra de pulmón de pacientes muertos por neumonía aguda tras pasar la COVID-19 (grupo de pacientes COVID-19). (A) Gráfico de volcán que muestra los ERs diferencialmente expresados, sobreexpresados (lado derecho) e infra-expresados (lado izquierdo) en muestras de pulmón del grupo de pacientes COVID-19. Las líneas discontinuas de corte indican un valor de p igual a 0.05 (eje x) y un valor de cambio expresado en log2 (LFC) mayor de 1,5 (eje y). Los ERs sobreexpresados e infraexpresados, con significación estadística ( $p < 0.05$ ), se muestran en color verde y rojo, respectivamente. (B) Distribución de los DEH según su posición génica. (C) Distribución de retrovirus endógenos humanos (HERVs) diferencialmente expresados en muestras COVID-19 agrupados en grupos de familias HERVs. Se representa la relación entre las frecuencias de los HERVs diferencialmente expresados en muestras COVID-19 y la frecuencia de familias HERV en la base de datos de anotación inicial. La línea discontinua roja indica la relación esperada, calculada como el cociente entre el número total de HERVs en la base de datos y el número de DEH.

### **4.3 Cambios en el transcriptoma tras infección por SARS-CoV-2**

Tras observar diferencias de expresión en el retrotranscriptoma de muestras de pulmón de pacientes COVID-19, se evaluó si la infección por SARS-CoV-2 también afectaba a la expresión génica. Después de analizarse la expresión de 16009 genes, se identificaron 838 (5.23%) genes diferencialmente expresados (DEG) en muestras de pulmón de pacientes graves de COVID-19 frente al grupo control, mostrando 376 genes sobreexpresados (Figura 4 A). Al analizar las principales rutas representadas en estos genes, se identificaron rutas biológicas relacionadas con modificaciones de la cromatina, como la metilación del DNA; así como genes implicados en la respuesta celular a estímulos (Figura 4 A).

A



B

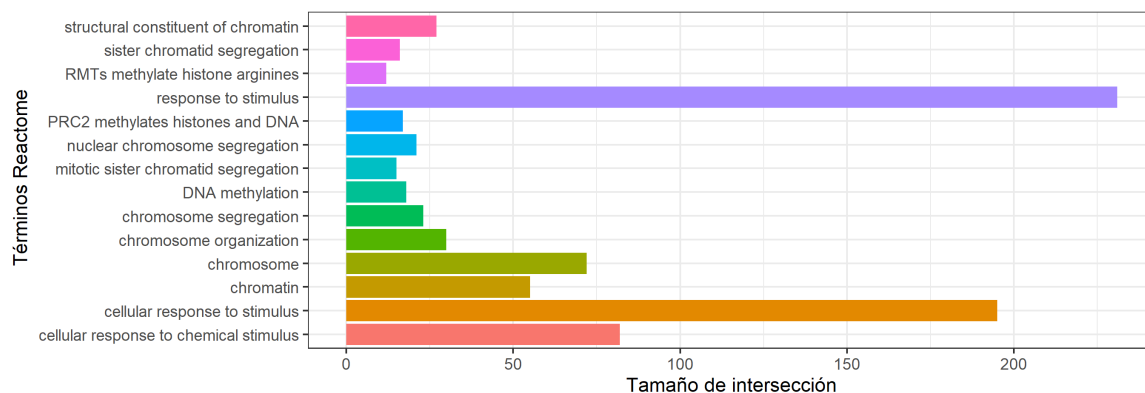


Figura 4: Resultados del análisis de expresión diferencial del transcriptoma de pacientes graves de COVID. (A) Gráfico de volcán que muestra los genes diferencialmente expresados (DEG), sobreexpresados (lado derecho) e infraexpresados (lado izquierdo) en muestra de pulmón de pacientes graves de COVID-19. Las líneas de corte indican un valor de p igual a 0.05 (eje x) y un valor de cambio expresado en log2 (LFC) mayor de 1,5 (eje y). Los genes sobreexpresados e infraexpresados, con significación estadística, se muestra en color verde y rojo respectivamente. (B) Principales rutas biológicas (eje Y) sobrerepresentadas y relacionadas con modificaciones de la cromatina y respuesta inmunitaria, que aparecen en pacientes graves de COVID, y los genes sobreexpresados que contienen (eje X)

#### 4.4 Análisis integrado de la expresión diferencial de elementos retrotransponibles y genes

Los elementos genómicos (genes y elementos retrotransponibles) diferencialmente expresados en muestras de pulmón de pacientes graves de COVID-19 estaban localizados en todos los cromosomas del genoma, y en regiones cromosómicas próximas entre sí (Figura 5). Resultó llamativo encontrar principalmente sobreexpresión de ERs y genes en el cromosoma 20.

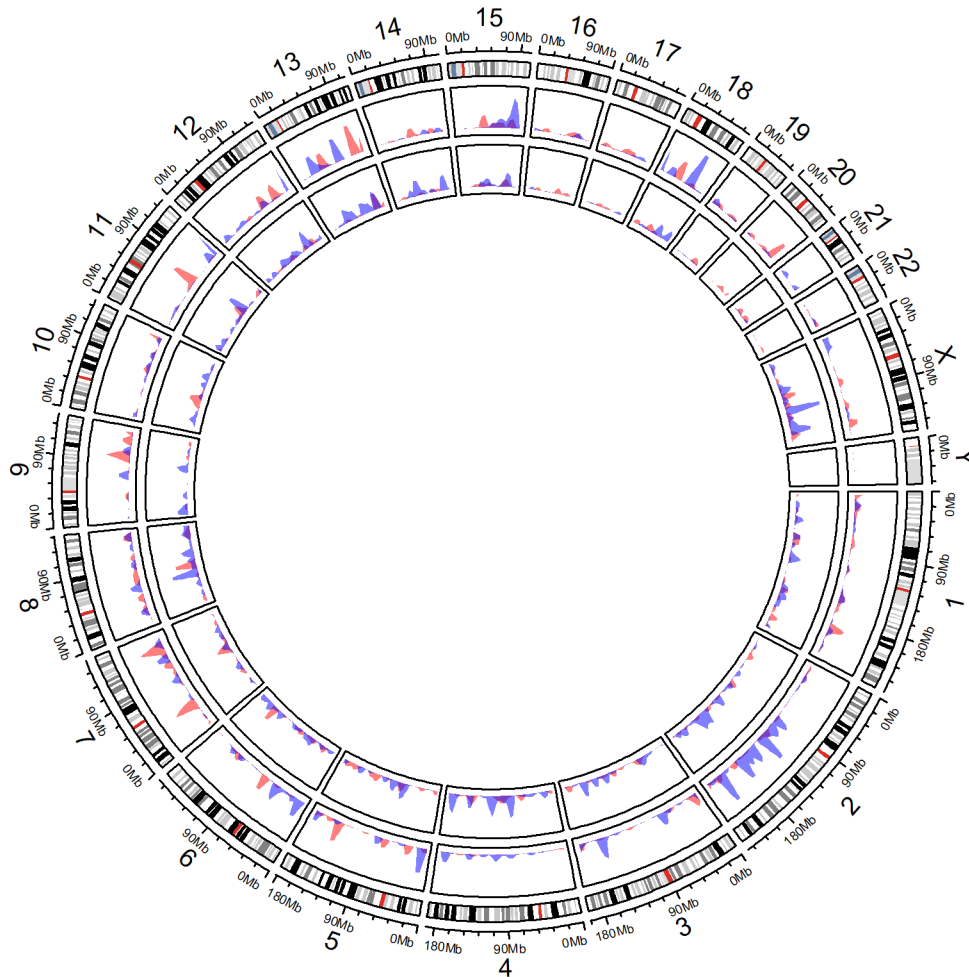


Figura 5: Gráfico de circo de los HERVs (DEHs, anillo interno) y genes (DEGs, anillo intermedio) diferencialmente expresados a lo largo del genoma. Los elementos genómicos sobreexpresados en el grupo COVID-19 se muestran en rojo, mientras que los infraexpresados se muestran en azul.

Para evaluar si la expresiones de los elementos genómicos estudiados, elementos retrotransponibles y genes, estaban reguladas de forma conjunta en pacientes graves de COVID-19, se realizó un análisis de coexpresión. Sin embargo, dado el bajo número de muestras analizadas y para asegurar la libertad de la red creada, se decidió crear una matriz de coexpresión utilizando todas las muestras y evaluar la existencia de grupos de elementos genómicos diferencialmente expresados en pacientes graves de COVID-19 sobre esa matriz. Así, se creó una red con 9486 secuencias genómicas, compuesta por 5 grupos de elementos genómicos que estaban relacionados fuertemente entre sí (Figura 6 A). A continuación, se evaluó la correlación de los perfiles de expresión de cada módulo con variables fenotípicas, hallando dos grupos cuyas expresiones

estaban correlacionadas significativamente en pacientes graves de COVID-19: módulo 1 (4397 elementos genómicos) (Figura 6 B), que contenía 175 DEGs y 46 DEHs ; y módulo 3 (1090 elementos genómicos), el cual incluía 17 DEGs y 5 DEHs. También se observó correlación con el sexo, aunque sin llegar a ser estadísticamente significativa.

Por último, se analizó la participación en rutas biológicas conocidas de los elementos genómicos desregulados en pacientes graves de COVID-19 que presentaban alta correlación aplicando un análisis funcional de enriquecimiento. Se escogió el módulo 3 (Figura 6 C) para realizar el análisis funcional de enriquecimiento por su correlación con las expresiones en pacientes graves de COVID-19 y por estar enriquecido en rutas biológicas relacionados con enfermedades infecciosas como la enfermedad del COVID-19, así como rutas propias de la infección por SARS-COV (tabla S2). Los resultados del análisis de enriquecimiento sobre los elementos diferencialmente expresados fueron limitados, destacando la identificación de la vía de presentación y procesamiento de antígeno, así como la vía de señalización de estrógenos. Además, el conjunto de elementos génicos también estaba enriquecido en elementos relacionados con la unión a desacetilasa de histonas.

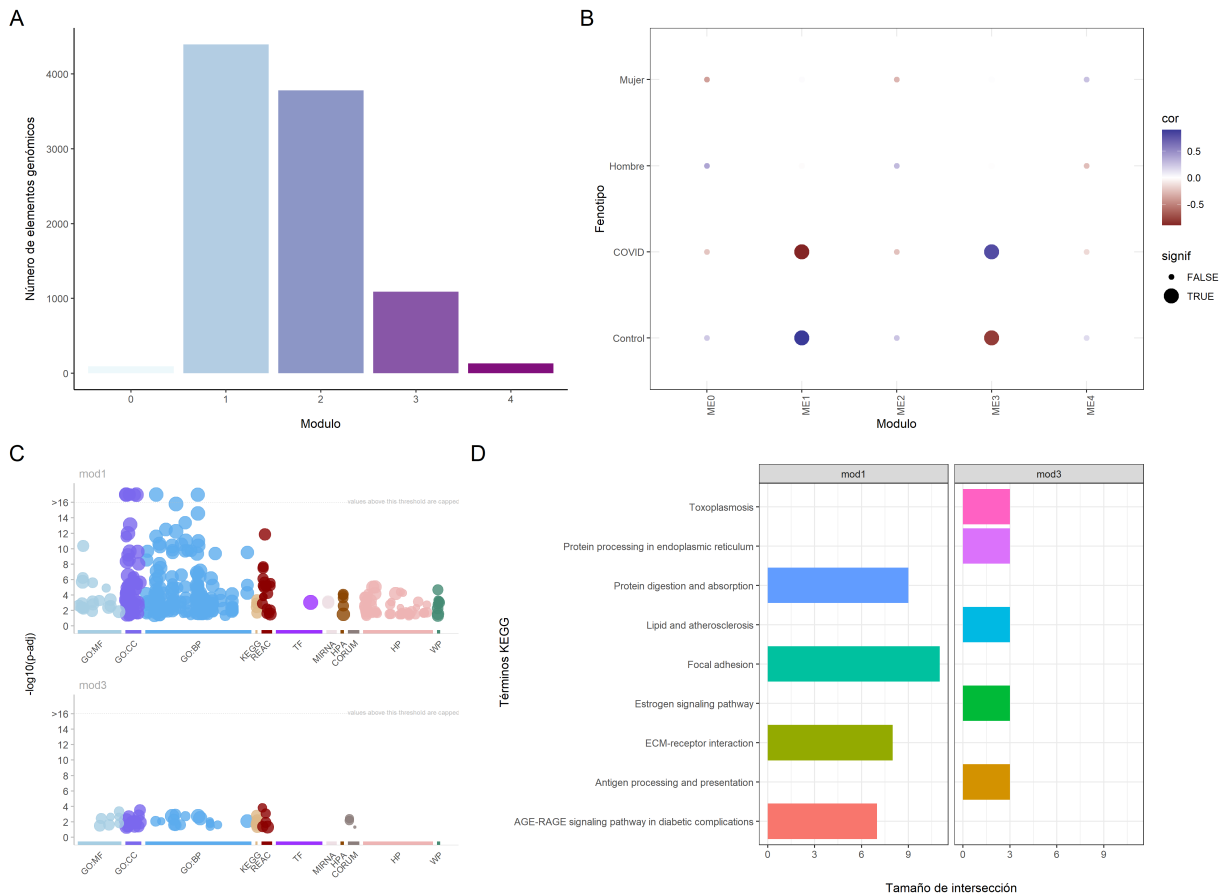


Figura 6: Resultados del análisis de coexpresión de genes y elementos retrotransponibles (TEs) en muestras de pulmón. A) Grupos de elementos genómicos (genes Y TEs) identificados en la red de coexpresión. El módulo 0 contiene los elementos genómicos que no se han podido clasificar en ninguno de los otros módulos. B) Asociación fenotípica entre los 5 grupos de elementos genómicos (módulos) hallados y sexo y pertenecer al grupo COVID-19. C) Gráfico de enriquecimiento tipo Manhattan de los módulos 1 y 3 en las bases GO, KEGG, Reactome, Transfac, miRTarBase, Human Protein Atlas, CORUM, Human Phenotype ontology, WikiPathways. D) Distribución de los genes y ERs diferencialmente expresados incluidos en el módulo 3 agrupados por las rutas halladas en el análisis funcional de enriquecimiento empleando la base de datos KEGG.

## 5 Discusión

Hace cuatro años se identificó el primer caso de COVID-19 en Wuhan (China), y desde entonces, más de 770 millones de personas han sido infectadas por todo el mundo. A pesar de los avances en la investigación de la enfermedad (se han publicado más de 400 000 artículos científicos publicados) y el desarrollo de vacunas (Feikin et al. 2022), la mortalidad asociada a la COVID-19 continúa siendo significativa. La infección por SARS-CoV-2 induce en algunos pacientes una respuesta inmunitaria exacerbada, caracterizada por una tormenta de citoquinas (Karki and Kanneganti 2022). Las principales hipótesis relacionadas con los mecanismos subyacentes a esta respuesta desregulada sugieren que existe una desregulación en los genes implicados en la respuesta inflamatoria (Wong and Perlman 2022). Sin embargo, de forma paralela, durante la COVID-19, se podrían inducir cambios conformacionales en la cromatina que favorecerían la transcripción de HERVs (X. Guo, Zhao, and You 2024), los cuales son reconocidos por el sistema inmunitario como antígenos exógenos. Esta reactivación de HERVs podría contribuir a la inflamación sistémica y a la gravedad de la enfermedad en algunos pacientes.

En este trabajo se ha evaluado la existencia de mecanismos moleculares implicados en la severidad de la COVID-19, analizando el impacto de la expresión de HERVs y su relación con el transcriptoma de pulmón de pacientes graves de COVID-19. Estudios previos han identificado cambios en los perfiles de expresión de HERVs tras la infección por SARS-CoV-2 que podrían estar asociados con la patogénesis y severidad del COVID-19 X. Guo, Zhao, and You (2024).

En este sentido, los pacientes graves de COVID-19 mostraron un aumento en la expresión de HERVs (Figura 3 A), destacando la presencia HERV-K, HERV-W y ERV1 (Figura 3 C). También se identificó un elevado número de elementos LINE-1 (Koch 2022). Así mismo, los pacientes graves de COVID-19 mostraron sobreexpresados genes relacionados con mecanismos epigenéticos, como la metilación de histonas (Figura 4).

De forma conjunta, en los perfiles del retrotranscriptoma y el transcriptoma de pulmón analizados, se hallaron grupos de ERs y genes implicados en procesos biológicos relacionados con infecciones por SARS-CoV, cuyas expresiones se correlacionaban con los perfiles de expresión en pacientes graves de COVID-19, sugiriendo una posible regulación conjunta de estos elementos en pulmón en relación con la infección por SARS-CoV-2. Dentro de ese grupo de elementos genómicos, se identificaron 17 genes y 5 ERs con perfiles de expresión alterados en pacientes graves de COVID-19 que estaban relacionados con modificaciones epigenéticas de la cromatina; y procesos inmunológicos como la degranulación de neutrófilos y presentación y procesamiento de antígenos (Figura 6). Estos resultados sugieren que la infección por SARS-CoV-2 podría inducir cambios en la conformación del DNA que induzcan la expresión de HERVs.

### 5.1 Discusión de resultados en relación a investigaciones previas

Numerosos estudios han reportado alteraciones en la expresión de HERVs en pacientes con COVID-19 X. Guo, Zhao, and You (2024), destacando el aumento de la expresión de HERV-K Charvet et al. (2023), HERV-W (Petrone et al. 2023) y ERV1 (X. Guo, Zhao, and You 2024). La expresión de estos HERVs podría estar relacionada con la desregulación de la respuesta inmunitaria frente al COVID-19 X. Guo, Zhao, and You (2024). Pacientes con una respuesta severa a la infección del COVID-19 mostraron aumentada la expresión de elementos ERV1 enriquecidos en procesos biológicos relacionados modificaciones de histonas (X. Guo, Zhao, and You 2024). Tras la infección por SARS-CoV-2, las proteínas del virus interactúan con diversos componentes de la célula hospedadora, entre los que destacan las enzimas responsables de las modificaciones de histonas, provocando a su vez alteraciones en la cromatina que afectan a la expresión génica (Guarnieri et al. 2024). En este sentido, leucocitos de pacientes con COVID-19 mostraron un aumento en la expresión de la proteína HERV-W ENV (Charvet et al. 2023), la es reconocida por las células inmunitarias Giménez-Orenga et al. (2022) y desencadena la producción de citoquinas inflamatorias Y. Guo et al. (2022).

Además, la expresión de HERVs también se ha relacionado con mayor severidad de la enfermedad y peor pronóstico X. Guo, Zhao, and You (2024). Un aumento en la expresión de HERV-K han sido asociados con severidad de la COVID-19 y la muerte prematura causada por la misma (Petrone et al. 2023).

Por otro lado, alteraciones en los perfiles de expresión de elementos retrotransponibles han sido asociados a alteraciones en la expresión de genes relacionados con la COVID-19 y la infección por SARS-CoV-2. Así,



la expresión de HERV-H ha sido relacionada con cambios en las expresiones de los genes de interferones de tipo I y II en pacientes pediátricos con COVID-19 (Tovo et al. 2021).

Estudios previos han reportado aumentos en la expresión de genes relacionados con modificaciones de histonas en pacientes de COVID-19 severo Grandi et al. (2023), sugiriendo que los cambios en la expresión génica podrían estar mediados por modificaciones epigenéticas tras la infección por SARS-CoV2. HERV-H, una de las familias más abundantes en el genoma humano, ha sido relacionada con la regulación genética (Carter et al. 2022), modificaciones en la organización estructural del genoma (Grandi et al. 2023), y la configuración de los dominios de asociación topológica (TAD) en el genoma (Zhang et al. 2019). Por otro lado, existen moléculas derivadas de la transcripción de elementos HERV en el genoma humano con capacidad inmunomoduladora, como algunas moléculas de ARN producidas tras la transcripción de HERV-k y las proteína HML-2 Env y Gag, resultando en la activación de la vía canónica NK- $\kappa\beta$  y aumento de la producción de interferon, respectivamente (Dervan et al. 2021); y HERV-H, que ha sido relacionada con procesos de regulación génica que relacionados con alteraciones en la respuesta inmunitaria (Grandi et al. 2023).

De forma conjunta, la evidencia científica sugiere que la infección por SARS-CoV-2 podría desencadenar cambios moleculares que impacten en la expresión génica, desregulando la respuesta inmunológica frente al virus y desencadenando procesos biológicos relacionados con la severidad y patogenicidad de la enfermedad que persisten en el tiempo tras la resolución de la infección.

## 5.2 Limitaciones

Este trabajo presenta algunas limitaciones. El diseño del estudio supone una limitación para los resultados, ya que la propia comparación entre muestras de sujetos vivos y muertos podría conllevar diferencias significativas en la expresión de elementos genómicos. Sin embargo, las biopsias de pulmón de pacientes muertos se cogieron en las primeras horas para evitar la degradación del tejido. Además, también se observó la presencia de genes y HERVs sobre-expresados en las muestras de pacientes muertos respecto a los vivos. Para corroborar estos resultados, debería repetirse el análisis con muestras del grupo control tomadas también en pulmones de pacientes muertos. Sin embargo, la realización de estudios clínicos en humanos es compleja, por varias razones: (A) sobrecarga asistencial del personal clínico, quienes se encargan de tomar las muestras biológicas de los pacientes; (B) dificultad de encontrar pacientes cumplan con los criterios de inclusión parahomogeneizar la población en cuanto a variables relevantes como edad y sexo; (c) así la complejidad de lo trámites burocráticos propios de la investigación con muestras humanas.

Los análisis genómicos deben realizarse con un número alto de muestras para aumentar la potencia estadística del análisis y obtener estimadores más robusto. En este trabajo se ha empleado un número reducido de muestras analizadas, lo que podría haber afectado a los resultados del análisis y a la precisión de los estimadores. Debido a esto, no se ha evaluado la influencia de terceras variables en los análisis de asociaciones.

El reducido tamaño muestral también ha impactado en los métodos empleados para la creación de la matriz de coexpresión, la cual se ha tenido que crear con todos los sujetos y evaluar *a posteriori* la relación entre los módulos y la patología de estudio. Además, la red de coexpresión mostró un poder de superior al recomendado, lo que podría deberse al número insuficiente de muestras analizadas, siendo 20 el número mínimo recomendado. Sin embargo, la red de coexpresión creada mostró un valor de umbral suave de 0,8 indicando que la red creada se aproximaba a una topología de red libre de escala (figura S2)

En conjunto, los hallazgos de este trabajo sugieren que la infección por SARS-CoV-2 podría inducir cambios epigenéticos, afectando a la expresión de elementos transpoibles como HERVs, contribuyendo a la severidad de la enfermedad y a la persistencia de los síntomas.

## 5.3 Trabajo futuro

Este trabajo constituye una prueba piloto para evaluar el papel de los HERVs en la inflamación asociada a la enfermedad del coronavirus 19. Teniendo en cuenta los resultados obtenidos y las limitaciones expuestas previamente, las líneas de trabajo para el futuro son:

- Realizar estudios de metilación en biopsias de pacientes infectados por SARS-CoV-2 en parénquima

pulmonar para verificar el efecto de la infección sobre los patrones de metilación y el grado de condensación del DNA.

- Realizar un análisis de expresión diferencial comparando una muestra mayor de pacientes graves de COVID-19 fallecidos con neumonía aguda y pacientes sanos que hayan fallecido por otras causas no relacionadas con patologías respiratorias.
- Evaluar la patogenicidad de los elementos retrotransponibles diferencialmente expresados en pacientes graves de COVID-19, introduciendo los péptidos codificados por las secuencias de los ERs más interesantes identificados en nuestro análisis y evaluando la respuesta inmunitaria desencadenada *in vitro*.

## 6 Conclusiones

A lo largo de este trabajo se ha analizado el impacto de la COVID-19 en la expresión de retrovirus endógenos humano (HERVs), y cómo la expresión de esos HERVs se relaciona la disrupción de la respuesta inmunitaria frente a SARS-CoV-2. Las conclusiones obtenidas se enumeran a continuación:

1. Los pacientes de COVID-19 fallecidos por neumonía asociada a la enfermedad presentan incrementos en las expresiones de HERVs, destacando la sobreexpresión de HERV-K y HERV-W. Estos HERVs modulan la respuesta inmunitaria y están relacionados con la severidad del COVID-19.
2. La infección por SARS-CoV-2 podría inducir la expresión de genes relacionados con modificaciones epigenéticas, y con la respuesta celular a cambios en el estado o la actividad de la célula.
3. La expresión de HERVs en pacientes con SARS-CoV-2 también podría estar relacionada con cambios en la expresión de genes en pacientes fallecidos por neumonía aguda asociada a COVID-19.

Finalmente, es importante tener en cuenta que las técnicas bioinformáticas empleadas en este trabajo han demostrado ser potentes herramientas para identificar patrones y conjuntos de elementos correlacionados entre sí dentro de grandes conjuntos de datos como son los obtenidos en las técnicas ómicas. Estas técnicas permiten el manejo de datos masivos, aportando un enfoque global. Sin embargo, existen limitaciones inherentes a estos métodos como la dependencia de la calidad e integridad de los datos analizados y la influencia de la elección de los parámetros aplicados en cada técnica.

Para asegurar la robustez y poder validar los hallazgos obtenidos en el análisis de datos masivos usando técnicas bioinformáticas, es importante completarlos con determinaciones moleculares experimentales, como la PCR y ensayos de expresión de proteínas.

## 7 Material suplementario

Todo el material suplementario se encuentra incluido en los anexos que acompañan este documento y en el repositorio creado.

## 8 Referencias

- 4, US DOE Joint Genome Institute: Hawkins Trevor 4 Branscomb Elbert 4 Predki Paul 4 Richardson Paul 4 Wenning Sarah 4 Slezak Tom 4 Doggett Norman 4 Cheng Jan-Fang 4 Olsen Anne 4 Lucas Susan 4 Elkin Christopher 4 Uberbacher Edward 4 Frazier Marvin, RIKEN Genomic Sciences Center: Sakaki Yoshiyuki 9 Fujiyama Asao 9 Hattori Masahira 9 Yada Tetsushi 9 Toyoda Atsushi 9 Itoh Takehiko 9 Kawagoe Chiharu 9 Watanabe Hidemi 9 Totoki Yasushi 9 Taylor Todd 9, Genoscope, CNRS UMR-8030: Weissenbach Jean 10 Heilig Roland 10 Saurin William 10 Artiguenave Francois 10 Brottier Philippe 10 Bruls Thomas 10 Pelletier Eric 10 Robert Catherine 10 Wincker Patrick 10, Institute of Molecular Biotechnology: Rosenthal André 12 Platzer Matthias 12 Nyakatura Gerald 12 Taudien Stefan 12 Rump Andreas 12 Department of Genome Analysis, GTC Sequencing Center: Smith Douglas R. 11 Doucette-Stamm Lynn 11 Rubenfield Marc 11 Weinstock Keith 11 Lee Hong Mei 11 Dubois JoAnn 11, Beijing Genomics Institute/Human Genome Center: Yang Huanming 13 Yu Jun 13 Wang Jian 13 Huang Guyang 14 Gu Jun 15, et al. 2001. "Initial Sequencing and Analysis of the Human Genome." *Nature* 409 (6822): 860–921.
- Allaire, JJ, Yihui Xie, Jonathan McPherson, Javier Luraschi, Kevin Ushey, Aron Atkins, Hadley Wickham, Joe Cheng, Winston Chang, and Richard Iannone. 2023. "Rmarkdown: Dynamic Documents for r, 2021." *R Package Version 2*.
- Allis, C David, and Thomas Jenuwein. 2016. "The Molecular Hallmarks of Epigenetic Control." *Nature Reviews Genetics* 17 (8): 487–500.
- Anders, S, and W Huber. 2010. "Differential Expression Analysis for Sequence Count Data." *Nat. Prec.*
- Andrews, Simon et al. 2010. "FastQC: A Quality Control Tool for High Throughput Sequence Data." Cambridge, United Kingdom.
- Ansari, Shabnam, Nidhi Gupta, Rohit Verma, Oinam N Singh, Jyoti Gupta, Amit Kumar, Mukesh Kumar Yadav, et al. 2023. "Antiviral Activity of the Human Endogenous Retrovirus-r Envelope Protein Against SARS-CoV-2." *EMBO Reports* 24 (7): e55900.
- Ashburner, Michael, Catherine A Ball, Judith A Blake, David Botstein, Heather Butler, J Michael Cherry, Allan P Davis, et al. 2000. "Gene Ontology: Tool for the Unification of Biology." *Nature Genetics* 25 (1): 25–29.
- Balestrieri, Emanuela, Antonella Minutolo, Vita Petrone, Marialaura Fanelli, Marco Iannetta, Vincenzo Malagnino, Marta Zordan, et al. 2021. "Evidence of the Pathogenic HERV-w Envelope Expression in t Lymphocytes in Association with the Respiratory Outcome of COVID-19 Patients." *EBioMedicine* 66.
- Bendall, Matthew L, Miguel De Mulder, Luis Pedro Iñiguez, Aarón Lecanda-Sánchez, Marcos Pérez-Losada, Mario A Ostrowski, R Brad Jones, et al. 2019. "Telescope: Characterization of the Retrotranscriptome by Accurate Estimation of Transposable Element Expression." *PLoS Computational Biology* 15 (9): e1006453.
- Bhat, Swati, Praveen Rishi, and Vijayta D Chadha. 2022. "Understanding the Epigenetic Mechanisms in SARS CoV-2 Infection and Potential Therapeutic Approaches." *Virus Research* 318: 198853.
- Bignon, Emmanuelle, Stéphanie Grandemange, Elise Dumont, and Antonio Monari. 2022. "How SARS-CoV-2 Alters the Regulation of Gene Expression in Infected Cells." *Journal of Physical Chemistry Letters* 14 (13): 3199–3207. <https://doi.org/10.1021/acs.jpclett.3c00582>.
- Carabelli, Alessandro M, Thomas P Peacock, Lucy G Thorne, William T Harvey, Joseph Hughes, Sharon J Peacock, Wendy S Barclay, Thushan I De Silva, Greg J Towers, and David L Robertson. 2023. "SARS-CoV-2 Variant Biology: Immune Escape, Transmission and Fitness." *Nature Reviews Microbiology* 21 (3): 162–77.
- Carter, Thomas A, Manvendra Singh, Gabrijela Dumbović, Jason D Chobirko, John L Rinn, and Cédric Feschotte. 2022. "Mosaic Cis-Regulatory Evolution Drives Transcriptional Partitioning of HERVH Endogenous Retrovirus in the Human Embryo." *Elife* 11: e76257.
- Charvet, Benjamin, Joanna Brunel, Justine Pierquin, Mathieu Iampietro, Didier Decimo, Nelly Queruel, Alexandre Lucas, et al. 2023. "SARS-CoV-2 Awakens Ancient Retroviral Genes and the Expression of Proinflammatory HERV-w Envelope Protein in COVID-19 Patients." *Iscience* 26 (5).
- "Coronavirus Disease (COVID-19): Post COVID-19 Condition." 2024. [https://www.who.int/news-room/questions-and-answers/item/coronavirus-disease-\(covid-19\)-post-covid-19-condition](https://www.who.int/news-room/questions-and-answers/item/coronavirus-disease-(covid-19)-post-covid-19-condition).
- "COVID-19 Cases | WHO COVID-19 Dashboard." 2024. <https://data.who.int/dashboards/covid19/cases>.
- Davis, Hannah E, Lisa McCorkell, Julia Moore Vogel, and Eric J Topol. 2023. "Long COVID: Major Findings, Mechanisms and Recommendations." *Nature Reviews Microbiology* 21 (3): 133–46.
- Dervan, Eoin, Dibyangana D Bhattacharyya, Jake D McAuliffe, Faizan H Khan, and Sharon A Glynn. 2021. "Ancient Adversary–HERV-k (HML-2) in Cancer." *Frontiers in Oncology* 11: 658489.
- Diamond, Michael S, and Thirumala-Devi Kanneganti. 2022. "Innate Immunity: The First Line of Defense

- Against SARS-CoV-2." *Nature Immunology* 23 (2): 165–76.
- Duperray, Alain, Delphin Barbe, Gilda Raguenez, Babette B Weksler, Ignacio A Romero, Pierre-Olivier Couraud, Hervé Perron, and Patrice N Marche. 2015. "Inflammatory Response of Endothelial Cells to a Human Endogenous Retrovirus Associated with Multiple Sclerosis Is Mediated by TLR4." *International Immunology* 27 (11): 545–53.
- Durinck, Steffen, Yves Moreau, Arek Kasprzyk, Sean Davis, Bart De Moor, Alvis Brazma, and Wolfgang Huber. 2005. "BioMart and Bioconductor: A Powerful Link Between Biological Databases and Microarray Data Analysis." *Bioinformatics* 21 (16): 3439–40.
- Ewels, Philip, Måns Magnusson, Sverker Lundin, and Max Källér. 2016. "MultiQC: Summarize Analysis Results for Multiple Tools and Samples in a Single Report." *Bioinformatics* 32 (19): 3047–48.
- Fairweather, DeLisa, Danielle J Beetler, Damian N Di Florio, Nicolas Musigk, Bettina Heidecker, and Leslie T Cooper Jr. 2023. "COVID-19, Myocarditis and Pericarditis." *Circulation Research* 132 (10): 1302–19.
- Feikin, Daniel R, Melissa M Higdon, Laith J Abu-Raddad, Nick Andrews, Rafael Araos, Yair Goldberg, Michelle J Groome, et al. 2022. "Duration of Effectiveness of Vaccines Against SARS-CoV-2 Infection and COVID-19 Disease: Results of a Systematic Review and Meta-Regression." *The Lancet* 399 (10328): 924–44.
- Giménez-Orenga, Karen, Justine Pierquin, Joanna Brunel, Benjamin Charvet, Eva Martín-Martínez, Hervé Perron, and Elisa Oltra. 2022. "HERV-w ENV Antigenemia and Correlation of Increased Anti-SARS-CoV-2 Immunoglobulin Levels with Post-COVID-19 Symptoms." *Frontiers in Immunology* 13: 1020064.
- Grandi, Nicole, Maria Chiara Erbi, Sante Scognamiglio, and Enzo Tramontano. 2023. "Human Endogenous Retrovirus (HERV) Transcriptome Is Dynamically Modulated During SARS-CoV-2 Infection and Allows Discrimination of COVID-19 Clinical Stages." *Microbiology Spectrum* 11 (1): e02516–22.
- Gu, Zuguang, Lei Gu, Roland Eils, Matthias Schlesner, and Benedikt Brors. 2014. "Circize" Implements and Enhances Circular Visualization in R."
- Guarnieri, Joseph W, Jeffrey A Haltom, Yentli E Soto Albrecht, Timothy Lie, Arnold Z Olali, Gabrielle A Widjaja, Sujata S Ranshing, Alessia Angelin, Deborah Murdock, and Douglas C Wallace. 2024. "SARS-CoV-2 Mitochondrial Metabolic and Epigenomic Reprogramming in COVID-19." *Pharmacological Research* 204: 107170.
- Guo, Xuefei, Yang Zhao, and Fuping You. 2024. "Identification and Characterization of Endogenous Retroviruses Upon SARS-CoV-2 Infection." *Frontiers in Immunology* 15: 1294020.
- Guo, Yaolin, Caiqin Yang, Yongjian Liu, Tianyi Li, Hanping Li, Jingwan Han, Lei Jia, et al. 2022. "High Expression of HERV-k (HML-2) Might Stimulate Interferon in COVID-19 Patients." *Viruses* 14 (5): 996.
- Harapan, Biyan Nathanael, and Hyeon Joo Yoo. 2021. "Neurological Symptoms, Manifestations, and Complications Associated with Severe Acute Respiratory Syndrome Coronavirus 2 (SARS-CoV-2) and Coronavirus Disease 19 (COVID-19)." *Journal of Neurology* 268 (9): 3059–71.
- Huang, Chaolin, Yeming Wang, Xingwang Li, Lili Ren, Jianping Zhao, Yi Hu, Li Zhang, et al. 2020. "Clinical Features of Patients Infected with 2019 Novel Coronavirus in Wuhan, China." *The Lancet* 395 (10223): 497–506.
- Huber, Wolfgang, Vincent J Carey, Robert Gentleman, Simon Anders, Marc Carlson, Benilton S Carvalho, Hector Corrada Bravo, et al. 2015. "Orchestrating High-Throughput Genomic Analysis with Bioconductor." *Nature Methods* 12 (2): 115–21.
- Jansz, Natasha, and Geoffrey J Faulkner. 2021. "Endogenous Retroviruses in the Origins and Treatment of Cancer." *Genome Biology* 22 (1): 147.
- Kanehisa, Minoru, and Susumu Goto. 2000. "KEGG: Kyoto Encyclopedia of Genes and Genomes." *Nucleic Acids Research* 28 (1): 27–30.
- Karki, Rajendra, and Thirumala-Devi Kanneganti. 2022. "Innate Immunity, Cytokine Storm, and Inflammatory Cell Death in COVID-19." *Journal of Translational Medicine* 20 (1): 542.
- Kee, John, Samuel Thudium, David M Renner, Karl Glastad, Katherine Palozola, Zhen Zhang, Yize Li, et al. 2022. "SARS-CoV-2 Disrupts Host Epigenetic Regulation via Histone Mimicry." *Nature* 610 (7931): 381–88.
- Kitsou, Konstantina, Anastasia Kotanidou, Dimitrios Paraskevis, Timokratis Karamitros, Aris Katzourakis, Richard Tedder, Tara Hurst, et al. 2021. "Upregulation of Human Endogenous Retroviruses in Bronchoalveolar Lavage Fluid of COVID-19 Patients." *Microbiology Spectrum* 9 (2): e01260–21.
- Koch, Benjamin Florian. 2022. "SARS-CoV-2 and Human Retroelements: A Case for Molecular Mimicry?" *BMC Genomic Data* 23 (1): 27.
- Köhler, Sebastian, Sandra C Doelken, Christopher J Mungall, Sebastian Bauer, Helen V Firth, Isabelle

- Bailleul-Forestier, Graeme CM Black, et al. 2014. "The Human Phenotype Ontology Project: Linking Molecular Biology and Disease Through Phenotype Data." *Nucleic Acids Research* 42 (D1): D966–74.
- Kolberg, Liis, Uku Raudvere, Ivan Kuzmin, Jaak Vilo, and Hedi Peterson. 2020. "Gprofiler2— an r Package for Gene List Functional Enrichment Analysis and Namespace Conversion Toolset g:profiler." *F1000Research* 9 (ELIXIR) (709).
- Langmead, Ben, and Steven L Salzberg. 2012. "Fast Gapped-Read Alignment with Bowtie 2." *Nature Methods* 9 (4): 357–59.
- Lemoine, Gwenaëlle G, Marie-Pier Scott-Boyer, Bathilde Ambroise, Olivier Périn, and Arnaud Droit. 2021. "GWENA: Gene Co-Expression Networks Analysis and Extended Modules Characterization in a Single Bioconductor Package." *BMC Bioinformatics* 22 (1): 267.
- Liu, Hengyuan, Valter Bergant, Goar Frishman, Andreas Ruepp, Andreas Pichlmair, Michelle Vincendeau, and Dmitrij Frishman. 2022. "Influenza a Virus Infection Reactivates Human Endogenous Retroviruses Associated with Modulation of Antiviral Immunity." *Viruses* 14 (7): 1591.
- Love, Michael I, Wolfgang Huber, and Simon Anders. 2014. "Moderated Estimation of Fold Change and Dispersion for RNA-Seq Data with DESeq2." *Genome Biology* 15: 1–21.
- Minkoff, Judith M, and Benjamin tenOever. 2023. "Innate Immune Evasion Strategies of SARS-CoV-2." *Nature Reviews Microbiology* 21 (3): 178–94.
- Nali, Luiz H, Guilherme S Olival, Horácio Montenegro, Israel T da Silva, Emmanuel Dias-Neto, Hugo Naya, Lucia Spangenberg, Augusto C Penalva-de-Oliveira, and Camila M Romano. 2022. "Human Endogenous Retrovirus and Multiple Sclerosis: A Review and Transcriptome Findings." *Multiple Sclerosis and Related Disorders* 57: 103383.
- Okonechnikov, Konstantin, Ana Conesa, and Fernando García-Alcalde. 2016. "Qualimap 2: Advanced Multi-Sample Quality Control for High-Throughput Sequencing Data." *Bioinformatics* 32 (2): 292–94.
- Patro, Rob, Geet Duggal, Michael I Love, Rafael A Irizarry, and Carl Kingsford. 2017. "Salmon Provides Fast and Bias-Aware Quantification of Transcript Expression." *Nature Methods* 14 (4): 417–19.
- Petrone, Vita, Marialaura Fanelli, Martina Giudice, Nicola Toschi, Allegra Conti, Christian Maracchioni, Marco Iannetta, et al. 2023. "Expression Profile of HERVs and Inflammatory Mediators Detected in Nasal Mucosa as a Predictive Biomarker of COVID-19 Severity." *Frontiers in Microbiology* 14: 1155624.
- Puelles, Victor G, Marc Lütgehetmann, Maja T Lindenmeyer, Jan P Sperhake, Milagros N Wong, Lena Allweiss, Silvia Chilla, et al. 2020. "Multiorgan and Renal Tropism of SARS-CoV-2." *New England Journal of Medicine* 383 (6): 590–92.
- RStudio, Team. 2016. "RStudio: Integrated Development for r. Boston, MA: RStudio." *Inc.[Google Scholar]*.
- Shih, Angela R, and Joseph Misdraji. 2023. "COVID-19: Gastrointestinal and Hepatobiliary Manifestations." *Human Pathology* 132: 39–55.
- Simula, Elena Rita, Maria Antonietta Manca, Marta Noli, Somaye Jasemi, Stefano Ruberto, Sergio Uzzau, Salvatore Rubino, Pietro Manca, and Leonardo A Sechi. 2022. "Increased Presence of Antibodies Against Type I Interferons and Human Endogenous Retrovirus w in Intensive Care Unit COVID-19 Patients." *Microbiology Spectrum* 10 (4): e01280–22.
- Simula, Elena Rita, Ignazio Roberto Zarbo, Giannina Arru, Elia Sechi, Rossella Meloni, Giovanni Andrea Deiana, Paolo Solla, and Leonardo Antonio Sechi. 2023. "Antibody Response to HERV-k and HERV-w Envelope Epitopes in Patients with Myasthenia Gravis." *International Journal of Molecular Sciences* 25 (1): 446.
- Smith, Zachary D, and Alexander Meissner. 2013. "DNA Methylation: Roles in Mammalian Development." *Nature Reviews Genetics* 14 (3): 204–20.
- Soneson, Charlotte, Michael I Love, and Mark D Robinson. 2015. "Differential Analyses for RNA-Seq: Transcript-Level Estimates Improve Gene-Level Inferences." *F1000Research* 4.
- Sudre, Carole H, Benjamin Murray, Thomas Varsavsky, Mark S Graham, Rose S Penfold, Ruth C Bowyer, Joan Capdevila Pujol, et al. 2021. "Attributes and Predictors of Long COVID." *Nature Medicine* 27 (4): 626–31.
- Synowiec, Aleksandra, Artur Szczepański, Emilia Barreto-Duran, Laurensius Kevin Lie, and Krzysztof Pyrc. 2021. "Severe Acute Respiratory Syndrome Coronavirus 2 (SARS-CoV-2): A Systemic Infection." *Clinical Microbiology Reviews* 34 (2): 10–1128.
- Team, R Core. 2013. "R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing." (No Title).
- Tovo, Pier-Angelo, Silvia Garazzino, Valentina Dapra, Giulia Pruccoli, Cristina Calvi, Federica Mignone, Carla Alliaudi, et al. 2021. "COVID-19 in Children: Expressions of Type I/II/III Interferons, TRIM28, SETDB1, and Endogenous Retroviruses in Mild and Severe Cases." *International Journal of Molecular Sciences* 22 (14):

7481.

- Wang, Ruoyu, Joo-Hyung Lee, Jieun Kim, Feng Xiong, Lana Al Hasani, Yuqiang Shi, Erin N Simpson, et al. 2023. "SARS-CoV-2 Restructures Host Chromatin Architecture." *Nature Microbiology* 8 (4): 679–94. <https://doi.org/10.1038/s41564-023-01344-8>.
- WHO, TWHO. 2020. "WHO Director-General's Opening Remarks at the Media Briefing on COVID-19-11 March 2020." *Geneva, Switzerland*, 3–5.
- Wickham, Hadley, Romain François, Lionel Henry, Kirill Müller, and Davis Vaughan. 2023. *Dplyr: A Grammar of Data Manipulation*. <https://dplyr.tidyverse.org>.
- Wong, Lok-Yin Roy, and Stanley Perlman. 2022. "Immune Dysregulation and Immunopathology Induced by SARS-CoV-2 and Related Coronaviruses—Are We Our Own Worst Enemy?" *Nature Reviews Immunology* 22 (1): 47–56.
- Xie, Yihui. 2017. *Dynamic Documents with r and Knitr*. Chapman; Hall/CRC.
- Zaborska, Monika, Maksymilian Chruszcz, Jakub Sadowski, Tomasz Klaudel, Michał Pelczarski, Anna Sztangreciak-Lehun, and Rafał Jakub Bułdak. 2024. "The Most Common Skin Symptoms in Young Adults and Adults Related to SARS-CoV-2 Virus Infection." *Archives of Dermatological Research* 316 (6): 292.
- Zhang, Yanxiao, Ting Li, Sebastian Preissl, Maria Luisa Amaral, Jonathan D Grinstein, Elie N Farah, Eugén Destici, et al. 2019. "Transcriptionally Active HERV-h Retrotransposons Demarcate Topologically Associating Domains in Human Pluripotent Stem Cells." *Nature Genetics* 51 (9): 1380–88.
- Zhu, Na, Dingyu Zhang, Wenling Wang, Xingwang Li, Bo Yang, Jingdong Song, Xiang Zhao, et al. 2020. "A Novel Coronavirus from Patients with Pneumonia in China, 2019." *New England Journal of Medicine* 382 (8): 727–33.

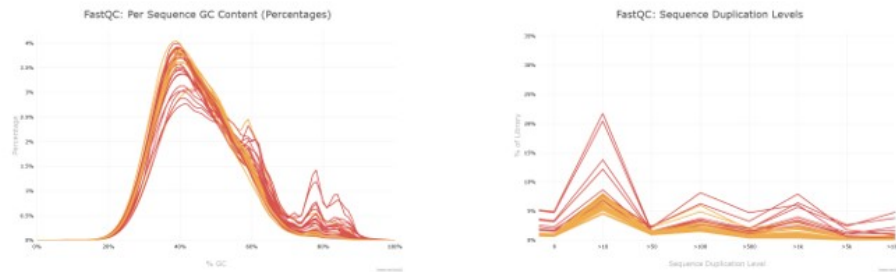
## 9 Anexos y Material Suplementario

### 9.1 Anexo I. Análisis de calidad de las muestras

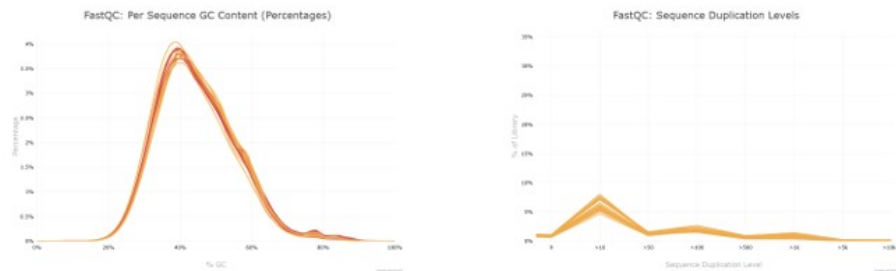
Tabla S1: Resumen del análisis de los datos de secuenciación de ARN. Análisis descriptivo de los resultados del filtrado y mapeado de las lecturas de la secuenciación de ARN. Lecturas crudas (millones): cantidad total de lecturas crudas, expresadas en millones, obtenidas tras secuenciar las muestras. Este valor se obtiene sumando la cantidad de las lecturas 1 y 2 de cada muestra. Datos crudos: contenido en G en las lecturas, calculado multiplicando las lecturas crudas por la longitud de las secuencias, que es en este caso es 150. Eficacia: medida de la eficacia del filtrado de las lecturas tras eliminar las lecturas con baja calidad y las secuencias de los adaptadores. Este valor se calcula como el ratio entre las lecturas limpias entre las lecturas crudas, expresado en tantos por cien. Q20, Q30: medida de la calidad de las bases en la secuenciación, calculado como el cociente entre el recuento de bases con valor Phred (Q) mayor de 20 o 30, respectivamente, y el recuento de bases total. El valor Phred representa la probabilidad estimada de un error en la base identificada. Cuanto mayor es el valor de Phred, mejor es la calidad de la base secuenciada. GC: contenido de bases G y C referido al contenido total de bases en las lecturas.

Parámetros	Total, N = 191	Controles, N = 91	COVID, N = 101	p-value2
Lecturas crudas (millones)	137 (133, 142)	135 (132, 142)	138 (133, 142)	0.6
Datos crudos	20.50 (19.90, 21.35)	20.30 (19.80, 21.30)	20.65 (20.00, 21.33)	0.6
Lecturas limpias (millones)	131 (128, 139)	131 (127, 139)	133 (129, 138)	0.6
Eficacia (%)	97.02 (96.48, 97.36)	97.05 (96.60, 97.59)	96.91 (96.39, 97.09)	0.3
Q20 (%)	97.57 (97.35, 97.72)	97.48 (97.32, 97.80)	97.60 (97.47, 97.69)	>0.9
Q30 (%)	93.36 (92.93, 93.77)	93.15 (92.93, 93.92)	93.38 (93.05, 93.60)	0.7
GC (%)	44.78 (43.74, 45.94)	46.13 (45.41, 48.04)	43.74 (43.67, 44.34)	0.002

### A) Todo



### B) Controles



### C) COVID

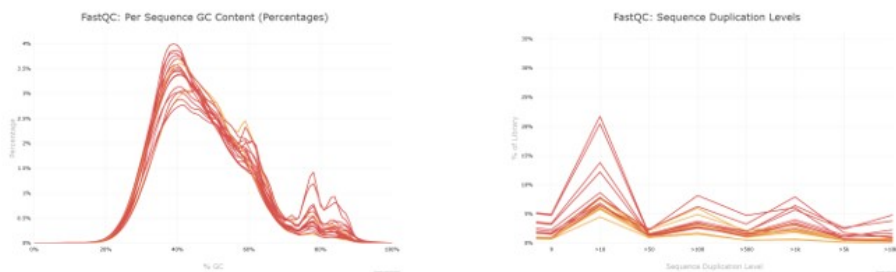


Figura S1: Distribución del contenido GC por secuencia y el contenido relativo de duplicación encontrado en cada secuencia. Se representan en rojo las secuencias que muestran. A) Resultados del análisis de la distribución del contenido GC en todas las muestras secuenciadas. B) Resultados del análisis de la distribución del contenido GC en las muestras de pacientes secuenciadas controles. C) Resultados del análisis de la distribución del contenido GC en las muestras de pacientes secuenciadas COVID-19.



## **9.2 Anexo II. elementos retrotransponibles identificados en muestras de pulmón**

El anexo IV se corresponde con el archivo `HERVs_COVID__n19_univariate.csv` depositado en el repositorio creado para este trabajo: [https://github.com/AzaharaGS/TFM\\_HERVs\\_COVID](https://github.com/AzaharaGS/TFM_HERVs_COVID)

### 9.3 Anexo III. Evaluación de la red de coexpresión

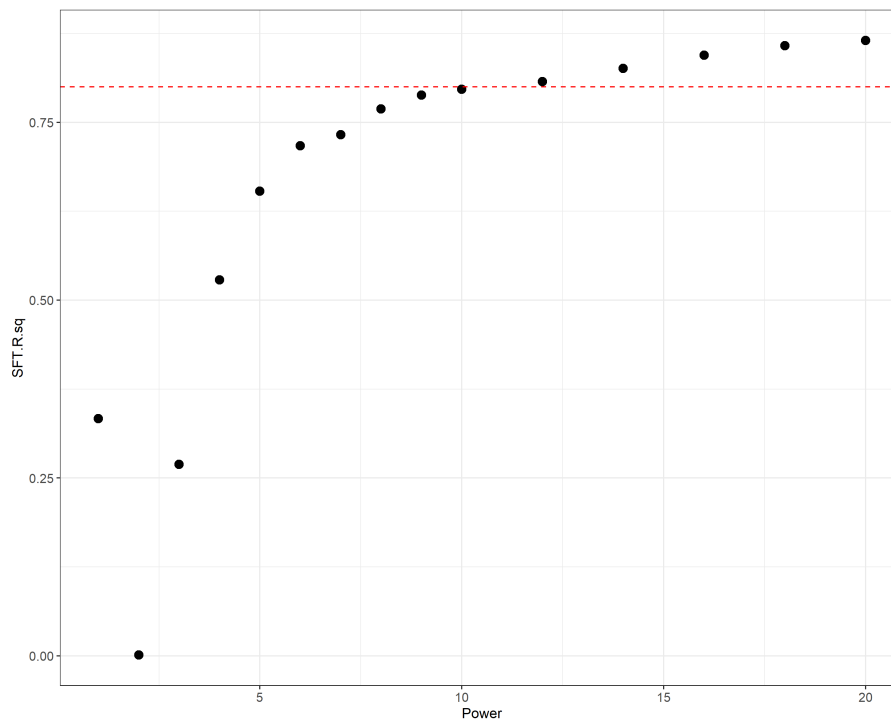


Figura S2: Distribución del contenido GC por secuencia y el contenido relativo de duplicación encontrado en cada secuencia. Se representan en rojo las secuencias que muestran. A) Resultados del análisis de la distribución del contenido GC en todas las muestras secuenciadas. B) Resultados del análisis de la distribución del contenido GC en las muestras de pacientes secuenciadas controles. C) Resultados del análisis de la distribución del contenido GC en las muestras de pacientes secuenciadas COVID-19.

### 9.4 Anexo IV. Resultados del análisis de enriquecimiento de la red de coexpresión.

El anexo IV se corresponde con el archivo `enrichment_mod3complete_KEGG_REACT.csv` depositado en el repositorio creado para este trabajo: [https://github.com/AzaharaGS/TFM\\_HERVs\\_COVID](https://github.com/AzaharaGS/TFM_HERVs_COVID)

Tabla S2: Resultado del análisis funcional de enriquecimiento empleando las terminología KEGG y REACTOME sobre los elementos genómicos incluidos en el módulo 3 de la red de coexpresión creada a partir de las expresiones génicas de las muestras de pulmón analizadas.

query	p_value	term_size	intersec- query_size	union_size	source	term_name
3	0.00000007	621	175	796	REAC	Disease
3	0.00000214	621	168	789	REAC	Metabolism of proteins
3	0.00000099	621	120	741	REAC	Infectious disease
3	0.00000078	621	114	735	REAC	Cellular responses to stress
3	0.00000079	621	114	735	REAC	Cellular responses to stimuli
3	0.00041771	621	104	725	REAC	Developmental Biology
3	0.00000080	621	103	724	REAC	Viral Infection Pathways
3	0.00000070	621	101	722	REAC	Metabolism of RNA
3	0.00000057	621	83	704	REAC	Nervous system development
3	0.00000054	621	81	702	REAC	Axon guidance
3	0.00000029	621	71	692	REAC	Translation

query	p_value	term_size	intersection_size	source	term_name
3	0.0000000231	473	65	KEGG	Coronavirus disease - COVID-19
3	0.0000000370	621	64	REAC	Metabolism of amino acids and derivatives
3	0.0000000217	621	61	REAC	Signaling by ROBO receptors
3	0.0000000102	621	59	REAC	Formation of a pool of free 40S subunits
3	0.0000000112	621	59	REAC	L13a-mediated translational silencing of Ceruloplasmin expression
3	0.0000000113	621	59	REAC	GTP hydrolysis and joining of the 60S ribosomal subunit
3	0.0000000120	621	59	REAC	Cap-dependent Translation Initiation
3	0.0000000120	621	59	REAC	Eukaryotic Translation Initiation
3	0.0000000182	621	59	REAC	Major pathway of rRNA processing in the nucleolus and cytosol
3	0.0000000192	621	59	REAC	rRNA processing in the nucleus and cytosol
3	0.0000000202	621	59	REAC	rRNA processing
3	0.0000000116	621	58	REAC	Nonsense Mediated Decay (NMD) enhanced by the Exon Junction Complex (EJC)
3	0.0000000116	621	58	REAC	Nonsense-Mediated Decay (NMD)
3	0.0000000152	621	58	REAC	Influenza Infection
3	0.0000000171	621	58	REAC	Regulation of expression of SLITs and ROBOs
3	0.0000000094	621	57	REAC	Eukaryotic Translation Elongation
3	0.0000000117	621	57	REAC	Selenoamino acid metabolism
3	0.0000000133	621	57	REAC	Influenza Viral RNA Transcription and Replication
3	0.0000000157	621	57	REAC	Cellular response to starvation
3	0.0000105459	621	57	REAC	SARS-CoV Infections
3	0.0000000153	473	56	KEGG	Ribosome
3	0.0000000102	621	56	REAC	Response of EIF2AK4 (GCN2) to amino acid deficiency
3	0.0000000090	621	55	REAC	Peptide chain elongation
3	0.0000000090	621	54	REAC	Viral mRNA Translation
3	0.0000000094	621	54	REAC	Selenocysteine synthesis
3	0.0000000094	621	54	REAC	Eukaryotic Translation Termination
3	0.0000000096	621	54	REAC	Nonsense Mediated Decay (NMD) independent of the Exon Junction Complex (EJC)
3	0.0000000113	621	54	REAC	SRP-dependent cotranslational protein targeting to membrane
3	0.0000045292	621	43	REAC	SARS-CoV-2 Infection
3	0.00000008197	621	35	REAC	SARS-CoV-2-host interactions
3	0.0000000140	621	33	REAC	SARS-CoV-1 Infection
3	0.0000000095	621	29	REAC	SARS-CoV-1-host interactions
3	0.0000000052	621	25	REAC	Formation of the ternary complex, and subsequently, the 43S complex
3	0.0000000059	621	25	REAC	Ribosomal scanning and start codon recognition
3	0.0000000059	621	25	REAC	Translation initiation complex formation
3	0.0000000060	621	25	REAC	Activation of the mRNA upon binding of the cap-binding complex and eIFs, and subsequent binding to 43S
3	0.0062294168	473	23	KEGG	Protein processing in endoplasmic reticulum
3	0.0209060170	473	22	KEGG	Tight junction
3	0.0000000051	621	21	REAC	SARS-CoV-2 modulates host translation machinery
3	0.0000000037	621	20	REAC	SARS-CoV-1 modulates host translation machinery
3	0.029722876	473	13	KEGG	Pancreatic cancer
3	0.001264713	621	7	REAC	NF-kB is activated and signals survival
3	0.007244316	621	7	REAC	p75NTR signals via NF-kB
3	0.021335013	621	6	REAC	p75NTR recruits signalling complexes
3	0.035532814	621	6	REAC	IRAK1 recruits IKK complex upon TLR7/8 or 9 stimulation
3	0.035532814	621	6	REAC	Attenuation phase



## 9.5 ANexo V. Heatmap con elementos genómicos identificados en módulo 3

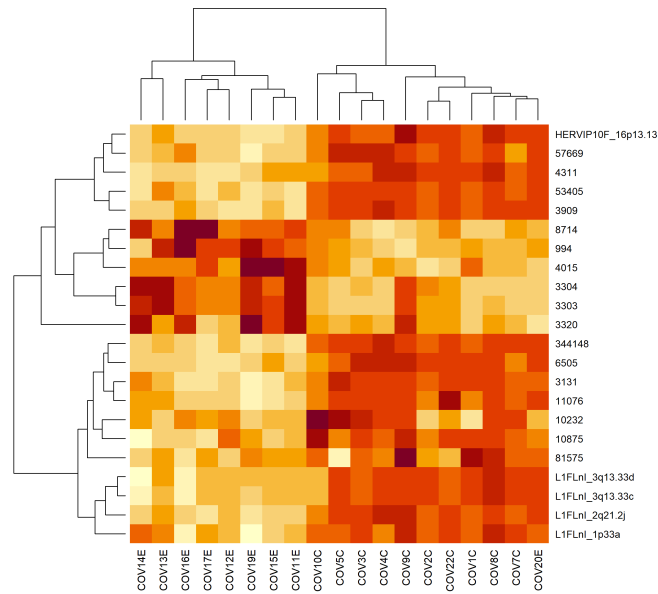


Figura S3: Gráfico de volcán que muestra los retrovirus endógenos humanos (HERVs) diferencialmente expresados, sobreexpresados (lado derecho) e infraexpresados (lado izquierdo) en muestra COVID-19. Las líneas de corte indican un valor de  $p$  igual a 0.05 (eje  $x$ ) y un valor de cambio expresado en  $\log_2$  (LFC) mayor de 1,5 (eje  $y$ ). Los HERVs sobreexpresados e infraexpresados, con significación estadística, se muestra en color verde y rojo respectivamente.

## 9.6 Otras figuras

### **9.6.1 Distribución de familias HERVs**

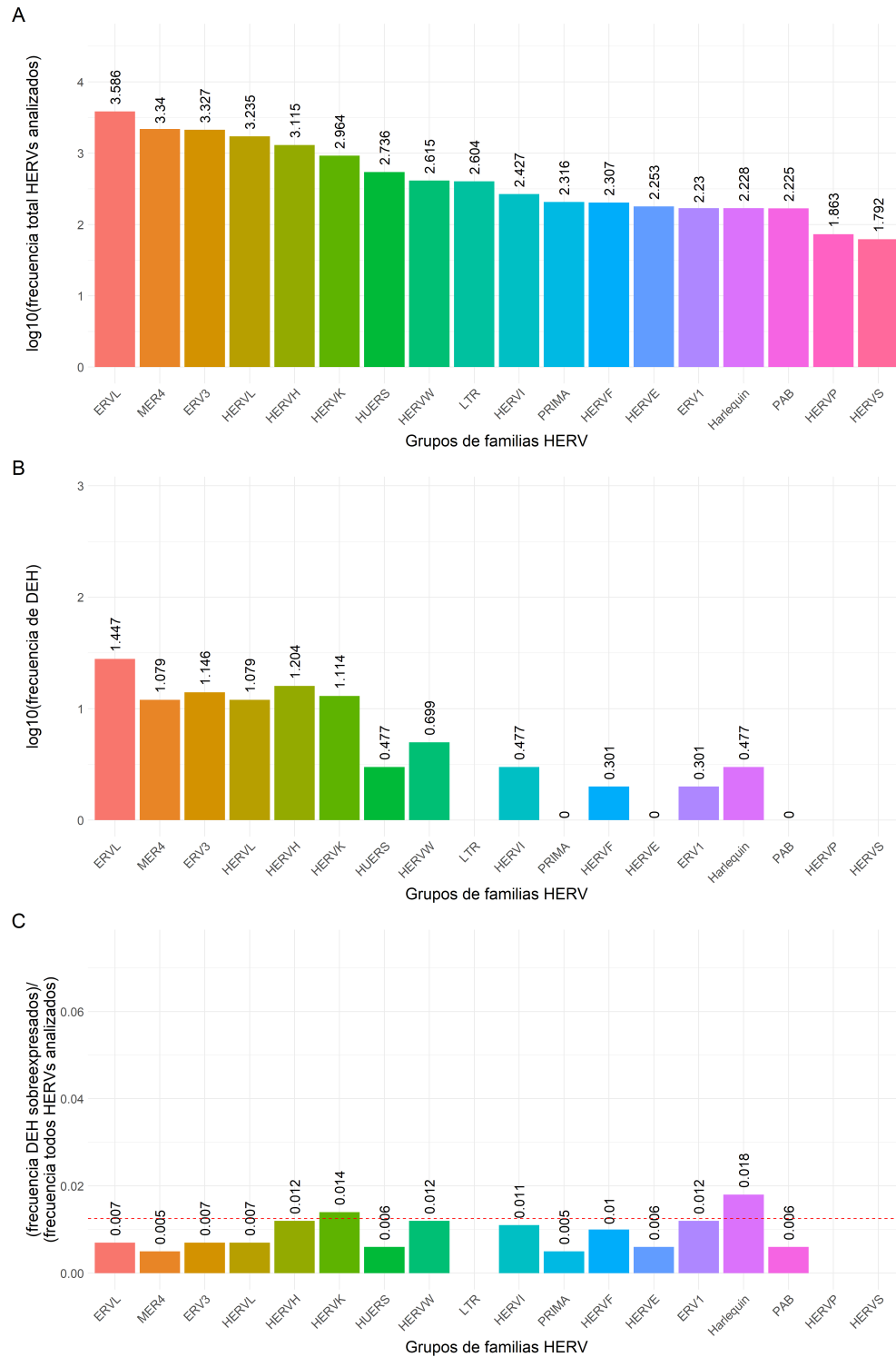


Figura S4: Frecuencias de retrovirus endógenos humanos (HERVs) distribuidos por grupos de familias. (A) Frecuencias de HERVs agrupados por grupos de familias incluidos en la base de datos utilizada para anotar las secuencias identificadas (expresadas en log10). (B) Frecuencias de HERVs agrupados por grupos de familias identificados en los HERVs diferencialmente expresados (DEH) encontrados (expresadas en log10) (C) Relación entre las frecuencias de las familias de HERVs identificados en los DEH y la frecuencia de familias HERV en la base de datos inicial. La línea discontinua roja indica la relación esperada, calculada como el cociente entre el número total de HERVs en la base de datos y el número de DEH.