UNIVERSIDAD MAYOR
para espíritus emprendedores

Facultad de Ciencias
ESCUELA DE BIOTECNOLOGÍA

Análisis *in silico* de la selección evolutiva de los sitios de glicosilación en proteínas estructurales de *Alfa-* y *Betacoronavirus*

Mauricio Andrés Morales Olavarría

Proyecto de Tesis para optar al título de Biotecnólogo

Tutor: PhD. Juan Pablo Cárdenas Astudillo

Centro de Genómica y Bioinformática, Universidad Mayor

"Esta tesis fue parcialmente apoyada por la infraestructura de cómputo del Centro de Genómica y Bioinformática, Universidad Mayor"

Tesis científica

Santiago - Chile

Año 2020-2021

UNIVERSIDAD MAYOR
para espíritus emprendedores

Facultad de Ciencias
ESCUELA DE BIOTECNOLOGÍA

Análisis *in silico* de la selección evolutiva de los sitios de glicosilación en proteínas estructurales de *Alfa-* y *Betacoronavirus*

Mauricio Andrés Morales Olavarría

Proyecto de Tesis para optar al título de Biotecnólogo

Tutor: Juan Pablo Cárdenas Astudillo

Grados Académicos:

Licenciado en Bioquímica (USACH)

Doctorado en Biotecnología (UNAB, Chile)

Santiago – Chile

Año 2020-2021

INFORMACIÓN DEL LABORATORIO Y/O CENTRO DE INVESTIGACIÓN

Nombre del laboratorio y/o Centro de Investigación: Centro de genómica y Bioinformático de la Universidad Mayor.

Fuente de Apoyo Técnico: Esta tesis fue parcialmente apoyada por la infraestructura de cómputo del Centro de Genómica y Bioinformática, Universidad Mayor.

Fuente de Financiamiento: Esta tesis no contó con ningún tipo de financiamiento.

ABREVIATURAS

ADN	Ácido Desoxirribonucleico
Ala	Alanina
Alfa	*Alfacoronavirus*
ARN	Ácido Ribonucleico
Asn	Asparagina
Asp	Ácido Aspártico
Beta	*Betacoronavirus*
CDC	Centro de Control y Prevención de Enfermedades (*Centers for Disease Control and Prevention*)
CFR	Tasa de Letalidad (*Case Fatality Rate*)
CoV	Coronavirus
CTD	Dominio C-terminal (*C-Terminal Domain*)
Cys	Cisteína
Delta	*Deltacoronavirus*
df	Grados de libertad (*Freedon Degree*)
dN	No sinónimas (*Non-synonymous*)
dN/dS	Tasa no sinónimas/sinónimas (*Rate Non-synonymous/Synonymous*)
dS	Sinónimas (*Synonymous*)
E	Envoltura (*Envelope)*
EDlogo	Logo de Enriquecimiento-Empobrecimiento (*Enrichment Depletion Logo*)
ERGIC	Compartimiento Intermediario Retículo Endoplasmático Golgi (*Endoplasmic Reticulum-Golgi Intermediate Compartment*)
FEL	Probabilidad de Efectos Fijos (*Fixed Effects Likelihood*)
FFT-NS-2	Segundo Alineamiento Progresivo (*Second Progressive Alignment*)
FP	Péptido de Fusión (*Fusion Peptide*)
FUBAR	AppRoximation bayesiana rápida y sin restricciones *(Fast, Unconstrained Bayesian AppRoximation)*
GalNAc	*N-acetilgalactosamina*
GalNAc-tS	Transferasas-GalNAc
Gama	*Gammacoronavirus*

Gln	Glutamina
Gly	Glicina
GPC	Complejo de Glicoproteína
HA	Hemaglutinina
HCoV	Coronavirus de Humanos *(Human Coronavirus)*
HE	Hemaglutinina Esterasa
His	Histidina
HR1/2	*Hepeat Region*
Ile	Isoleucina
Leu	Leucina
lnL	Probabilidad (*Likelihood*)
LRT	Prueba de Razón de Verosimilitud (*Likelihood Ratio Test*)
M	Membrana (*Membrane*)
M0	Modelo de un Radio *(One Ratio Model)*
M1a	Modelo Neutral *(Nearly Neutral Model)*
M2a	Modelo de Selección Positiva *(Positive Selection Model)*
M3	Modelo Discreto *(Discrete Model)*
M7	Modelo de Presión Selectiva Variable con Distribución Beta (*Beta Distributed Variable Selective Pressure Model)*
M8	Modelo de Selección Positiva Beta + (*Beta Plus Positive Selection Model*)
MERS	Síndrome Respiratorio de Oriente Medio (*Middle East respiratory syndrome*)
Met	Metionina
MHV	Virus Hepático Murino
ML	Máxima Verosimilitud *(Maximum Likelihood)*
MPT	Modificación Postraduccional
MSA	Alineamiento Múltiple de Secuencias (*Multiple Sequence Alignment*)
N	Nucleocápside
NA	Neuraminidasa
NCBI	Centro Nacional de Información Biotecnológica (*National Center for Biotechnology Information*)
np	Número de Parámetros
NTD	Dominio N-terminal (*N-terminal Domain*)

OH	Hidroxilo
OMS	Organización Mundial de la Salud
ORF	Marco de Lectura Abierto (*Open Reading Frame*)
PAML	Análisis Filogenético de Máxima Verosimilitud (*Phylogenetic Analysis by Maximum Likelihood*)
PEDV	Virus de Diarrea Porcina Epidémica (*Porcine Epidemic Diarrhea Virus*)
Phe	Fenilalanina
PHEV	Virus de Encefalomielitis Hemaglutinante Porcina (*Porcine Hemagglutinin Encefalitis Virus*)
Pro	Prolina
RBD	Dominio de Unión al Receptor (*Receptor Binding Domain*)
RdRp	Polimerasa dependiente de ARN (*RNA-dependent RNA polymerases*)
S	Espiga (*Spike*)
S1/S2	Subunidad 1/2 (*Subunidad 1/2*)
SARS	Síndrome Respiratorio Agudo Grave (*Severe acute respiratory syndrome*)
SARSr	Síndrome Respiratorio Agudo Grave relacionado (*Severe acute respiratory síndrome related*)
SD	Desviación Estándar (*Standard Deviatio*n)
Ser	Serina
SN°	Suplementaria Número
ssRNA	ARN de hebra simple *(Single Strand RNA)*
Thr	Treonina
TM	Transmembrana (*Transmembrane*)
Trp	Triptófano
Val	Valina
VIH	Virus de la Inmunodeficiencia Humana (*Human immunodeficiency* viruses)

ÍNDICE DE CONTENIDOS

ÍNDICE DE TABLAS

ÍNDICE DE FIGURAS

RESUMEN

Los *Alfacoronavirus* (*Alfa-CoV*) y *Betacoronavirus* (*Beta-CoV*) son patógenos zoonóticos que circulan en un amplio rango de hospederos. En seres humanos, los *Alfa-CoV* ocasionan un cuadro de resfriado común, mientras que algunos *Beta-CoV* han estado involucrados en eventos epidémicos y pandémicos durante las últimas dos décadas. Por lo tanto, nuevos enfoques para enfrentar este tipo de eventos de una manera rápida y eficiente son esenciales.

Aproximadamente dos tercios del genoma de estos patógenos codifica dos marcos de lectura, los cuales se procesan para producir proteínas no estructurales; en cambio, la porción restante contiene información codificante para las proteínas estructurales: espiga (S), membrana (M) y envoltura (E). Estas proteínas cumplen funciones que van desde contribuir al soporte estructural hasta definir el tropismo viral. Estas proteínas sufren modificaciones postraduccionales (MPT) tales como glicosilaciones, las cuales se presentan principalmente como *N*-glicosilaciones, y en menor medida, como *O*-glicosilaciones.

Al igual que muchas proteínas virales, las proteínas estructurales de los CoVs se encuentran bajo la presión selectiva por parte del sistema inmune del huésped, lo cual potencia cambios mayores o menores en los residuos producto de la selección natural. En este estudio se examinaron los residuos de las proteínas estructurales que estén vinculados a sitios de glicosilación (*sequons*) para identificar ganancia de *sequons* producto de la selección natural.

La estrategia de análisis se centró en evaluar secuencias de proteínas estructurales a través de análisis filogenéticos en busca de selección positiva, combinado con la predicción de sitios de glicosilación, lo cual permite evaluar potenciales *sequons* bajo selección positiva. El análisis utilizado no identificó sitios de glicosilaciones directamente bajo selección positiva. Sin embargo, se identificó un patrón consenso de glicosilaciones en la proteína S en los CoVs de los géneros *Alfa-CoV* y *Beta-CoV*, reportado como *O-Follow-N,* se identificó la preferencia del *sequon* del tipo NXT en la proteína S, en contraste con M que posee preferencia por NXS y sitios seleccionados positivamente aledaños a zonas ricas en *N*- y *O*-glicosilaciones consenso en la proteína S y M.

SUMMARY

Alphacoronaviruses (*Alpha-CoV*) and *Betacoronaviruses* (*Beta-CoV*) are zoonotic pathogens that circulate in a wide range of hosts, including humans. In humans, *Alpha-CoV* cause a common cold, while several *Beta-CoV* have been involved in epidemic and pandemic events for the past two decades. Therefore, new approaches to deal with these types of events in a fast and efficient way are critical.

Approximately two thirds of the genome from these pathogens encode two open reading frames, which are processed to generate non-structural proteins, while the remaining portion contains coding information for the structural proteins: spike (S), membrane (M) and envelope (E). These proteins perform functions ranging from contributing to structural support to defining viral tropism. These proteins undergo post-translational modifications (MPT) such as glycosylations, which occur mainly as *N*-glycosylations, and to a lesser time, as *O*-glycosylations.

Like many others viral proteins, the structural proteins of CoVs are under selective pressure from the host's immune system, which enhances major or minor changes in residues resulting from natural selection. In this study, structural protein residues that are linked to glycosylation sites (*sequons*) were examined to identify the gain of *sequons* as a result of natural selection.

The analysis strategy focused on evaluating structural protein sequences through phylogenetic analysis in search of positive selection, combined with the prediction of glycosylation sites, which allows evaluating potential sequons under positive selection. The analysis used did not identify glycosylation sites directly under positive selection. However, a consensus pattern of glycosylations in protein S was identified in CoVs of the genera Alpha-CoV and Beta-CoV, reported as O-Follow-N, the preference of the NXT-type sequon in protein S was identified, in contrast to M, which has a preference for NXS positively selected sites adjacent to areas rich in N- and O-glycosylations consensus in protein S and M.

1. INTRODUCCIÓN

1.1 Eventos emergentes: epidemias y pandemias

La capacidad de diferentes virus para cambiar de un hospedero a otro ha posibilitado que a lo largo de la historia la aparición de eventos emergentes como, por ejemplo, aparición de Influenza del tipo aviar y/o porcina (1). Estos se producen cuando nuevas poblaciones de hospederos y reservorios experimentan principalmente cambios demográficos, comportamientos de agregación y/o dispersión, aumento en las tasas de contacto, cambios en las condiciones ambientales, entre otros. Esto está condicionado principalmente por la invasión del ser humano a diversos ecosistemas y la consecuente destrucción de estos, lo cual condiciona el contacto con diversos animales y sus respectivos patógenos (2). Teniendo en cuenta esto, los virus con capacidad zoonótica son aquellos que potencian la aparición de eventos emergentes a través del cambio de hospedero, es decir, pasando de un animal a infectar a humanos.

El Centro de Control y Prevención de Enfermedades (*CDC*, por sus siglas en inglés) define los conceptos de epidemias y pandemias (3), de la siguiente manera:

> "Epidemia: la ocurrencia de más casos de enfermedad, lesión u otra condición de salud de lo esperado en un área determinada o entre un grupo específico de personas durante un período en particular".
>
> "Pandemia: una epidemia que ocurre en un área extensa (varios países o continentes) y que generalmente afecta a una proporción sustancial de la población" (3).

La pandemia más reciente fue declarada en marzo del año 2020, con el caso de un tipo de Coronavirus (CoVs) denominado SARS-CoV-2 (4). El cual se convirtió en el séptimo CoV conocido con capacidad de infectar a humanos y capaz de causar una enfermedad grave (5).

Previamente, se han presentado dos eventos epidémicos de importancia. El primero se produjo en el año 2002-2003 producto del Síndrome Respiratorio Agudo Severo (SARS-CoV-1), en el cual se observaron 8.273 casos con un total de 775 muertes en 37 países (6). Esto corresponde a 9% en la tasa de letalidad (*CFR*, por sus siglas en inglés). Por otro lado, el Síndrome Respiratorio de Oriente Medio (MERS) correspondió al segundo evento epidémico de importancia, ocurrido entre los años

2012 y 2013. Este evento generó 1.621 casos confirmados por laboratorio con un total de 584 decesos (CFR = 36%) en alrededor de 26 países (6). Como se mencionó anteriormente, desde diciembre del 2019 a la fecha (2022) el mundo se ha enfrentado a la pandemia producto del SARS-CoV-2, el cual posee una alta tasa de infectividad, la cual ha ido variando producto de la aparición de nuevas variantes durante el transcurso de la pandemia (7). Esto ha ocasionado que a inicios del 2022 se presenten aproximadamente 328 millones de casos confirmados con más de 5 millones de muertes aproximadamente (CFR= 1,68%), según la Organización Mundial de la Salud (OMS) (8).

Los casos anteriormente mencionados son aquellos que corresponden a eventos emergentes que han generado grandes pérdidas, tanto humanas como socioeconómicas. Sin embargo, un amplio rango de CoVs se encuentran habitualmente en poblaciones de mamíferos (9), por lo que un debido monitoreo es primordial para evitar eventos como estos.

Como otros virus, los CoVs se han podido expandir gracias a mecanismos de zoonosis (6, 10), lo que implica que estos agentes infecciosos tienen posibilidad de infectar a nuevos hospederos pasando desde especies de animales a otras, y así al ser humano. Esta clase de agentes han desarrollado las adaptaciones necesarias para transmitirse entre diferentes hospederos (9, 12), los cuales pueden servir de intermediarios para ser transmitidos al ser humano (2, 6). Adicionalmente, los CoVs, al ser virus que poseen un genoma de ARN, poseen tasas de mutación mayores que los virus de genoma de ADN, a pesar de poseer un sistema de corrección de errores en su maquinaria replicativa (11).

1.2 Consecuencias socioeconómicas

Al presentarse eventos de esta índole un amplio rango de países debe presentar diferentes medidas para evitar pérdidas que se puedan intensificar o simplemente como medidas paliativas, tanto en materia social como económica.

A modo de ejemplo, la primera pandemia del siglo XXI fue provocada por el virus de la Influenza A H1N1, la cual tuvo una duración de 14 meses aproximadamente. Esta pandemia afectó principalmente a niños, seguida de adultos jóvenes (13), ahora bien, también se reportó que la carga ambulatoria que presentó el sistema fue bastante elevada. En donde una mayor carga de casos leves de influenza implicó mayores exigencias del sistema sanitario (14), lo cual también condiciona los costos económicos asociados y el potencial del personal sanitario. En lo que respecta a medidas

sociales-sanitarias para evitar la propagación del virus en ciertos países se optó por el uso de mascarilla, distanciamiento y en ciertos países se llevaron a cabo cuarentenas (14).

Actualmente, la pandemia ocasionada por el SARS-CoV-2 ha provocado que se tomen medidas de cuarentena y distanciamiento social en la gran parte de los países de maneras interrumpidas, dado la rápida tasa de transmisión que presenta el patógeno, lo cual provoca la aparición de nuevas olas de contagios (15).

En general, el primer país en tomar medidas restrictivas fue China (2019) para luego ser acompañada por diferentes gobiernos que a la fecha (2022) han ido cambiando la intensidad de las medidas, sin embargo, en gran medida todas han culminado en generar una crisis económica que se ha traducido en pérdidas de empleos y pérdidas económicas traducidas en millones, a nivel de salud mental el problema se ha incrementado en diferentes rangos etarios y el sistema de salud se ha visto sobrepasado en su capacidad en diferentes periodos de tiempo, dado por falta de personal y/o exceso de individuos enfermos (16, 17).

1.3 Dinámica en la generación de vacunas y nuevas estrategias

Para combatir eventos de esta índole es imperativo la generación de una vacuna efectiva que pueda contribuir a que gran parte de la población genere una respuesta inmunitaria y la consiguiente memoria inmunológica que media la protección contra futuras infecciones (18).

Para llevar a cabo la generación de una vacuna hay un período de 10 a 15 años y un considerable costo económico para obtener una que sea efectiva y se aplique a la población (18). Sin embargo, la dinámica de la pandemia ha condicionado que los tiempos de prueba sean mucho más acotados, por ende, que las dinámicas de investigación también. Los principales enfoques para reducir tiempos y costos se centran en la selección de antígenos o estructuras antigénicas, portadores de adyuvantes y adyuvantes apropiados. Para llevar a cabo estos enfoques la incorporación de metodología y análisis bioinformáticos tienen un gran impacto, por ejemplo, a través de la combinación de tecnologías de ADN recombinante, disponibilidad de información biológica sobre los organismos y el aumento de información genómica en diferentes bases de datos (19, 21).

Teniendo en cuenta estos enfoques, el estudio y comprensión de los candidatos para antígenos son esenciales. Generalmente estas corresponden a macromoléculas, en su mayoría proteínas, las

cuales son capaces de generar una respuesta inmune efectiva. Sin embargo, muchas de estas proteínas se encuentran glicosiladas y/o poseen potenciales sitios de glicosilación (19), los cuales pueden cambiar su posición debido a mutaciones. Esto influye en la antigenicidad e inmunogenicidad, es decir, la capacidad de anticuerpos o receptores de células T de unirse a un antígeno dado y la capacidad de un antígeno de inducir una respuesta inmune adaptativa (20), respectivamente. Por lo tanto, entender la conservación y/o alteraciones que presentan las glicosilaciones es un paso necesario para la dinámica de generación y efectividad de vacunas.

1.4 Familia *Coronaviridae*

La familia *Coronaviridae* se encuentra constituida por la subfamilia *Orthocoronaviridae* que a su vez se divide en cuatro géneros: *Alfacoronavirus (Alfa-CoV)*, *Betacoronavirus (Beta-CoV)*, *Gammacoronavirus (Gamma-CoV)* y *Deltacoronavirus (Delta-CoV).* Los dos últimos circulan principalmente en aves (22), en cambio, *Alfa-CoV* y *Beta-CoV* circulan en un amplio rango de mamíferos (23, 24). Dentro de estos dos últimos géneros nombrados se encuentran 2 causantes de las pasadas epidemias de la década (SARS-CoV-1 y MERS-CoV) y el causante de la actual pandemia, el SARS-CoV-2, todos ellos pertenecientes a los *Beta-CoV* y causantes de un cuadro respiratorio que puede variar de leve a grave (24). Sin embargo, entre ambos géneros *Alfa- y Beta-CoV,* se encuentran variados tipos que comúnmente circulan en humanos (HCoV), como lo son el 229E-HCoV, OC43-HCoV, HKU1-HCoV y NL63-HCoV, las cuales generan síntomas similares a los del resfrío común (23, 24). Dado que estos patógenos circulan en variados hospederos se pueden generar además un amplio rango de síntomas que no sólo se limita al cuadro respiratorio. Por ejemplo, los síntomas en humanos son principalmente respiratorio, en cambio, en animales como cerdos, bovino, entre otros, la manifestación de la enfermedad a través de síntomas es principalmente gastrointestinales (23, 24, 26).

Como se mencionó anteriormente, muchos CoVs circulan en animales que se encuentran en ecosistemas urbanizados, como aquellos que no han sufrido mayor intervención humana. En ocasiones la interacción de animales de uso ganadero con ecosistemas aún no alterados mayoritariamente involucra la interacción con otros animales y aumenta la probabilidad de salto interespecie entre CoVs, lo cual se puede traducir en pérdidas económicas para el sector o riesgo de zoonosis (26, 27). Entre los CoVs que circulan en hospederos animales analizados en este estudio se encuentran HKU-2, 3, 4, 5, 23, 24, SARSr-CoV (relacionados del SARS), Virus de la Hepatitis Murina

(MHV), Virus de la Diarrea Epidémica Porcina (PEDV), Bovino-CoV (BCoV) y Virus de la Encefalomielitis Hemaglutinante Porcina (PHEV).

Sus rasgos zoonóticos aumentan el riesgo de epidemias o pandemias (28). Teniendo en cuenta esto, los *Beta-CoV* causantes de los tres eventos emergentes de la década tienen su posible origen en hospederos animales. Por lo tanto, entender el proceso por el cual ocurren estos eventos de cambio de hospedero contribuyen a estar preparados para enfrentar virus emergentes, tanto por las mutaciones en el genoma y/o intercambios de porciones de genomas virales, las cuales se pueden traducir en cambios estructurales y/o en nuevas variantes (7, 23). Un ejemplo que abarca a la familia *Coronaviridae*, comprende el historial de patrones de recombinación reportado para CoVs aislados, tanto de mamíferos como de aves, en diferentes regiones del genoma (25). En lo que respecta a nuevas variantes, su aparición ha sido producida por la presión selectiva que ejerce el sistema inmune. Por ejemplo, cambios en residuos del dominio de unión al receptor (RBD) posibilito la aparición de la variante B.1.1.7 (Reino Unido, *UK*), la cual permitió interacciones adicionales con el recepto de la célula humana ACE2 (7).

1.5 Organización estructural de CoVs

Todos los genomas de CoVs comparten características similares, el cual está compuesto por una ARN de hebra positiva (*ssRNA+*), no segmentado, con un largo de 27 a 32 kb (23). Dos tercios del genoma consisten en dos marcos de lectura abierta (*ORF*, por sus siglas en inglés), lo cual codifica para 16 proteínas no estructurales, o nsp1 a la 16 (23, 27). Estas proteínas se encargan principalmente de procesos que están involucrados en la transcripción y replicación viral (23, 29).

Otra característica compartida entre los CoVs es que poseen genes que codifican para cuatro proteínas estructurales: nucleocápside (*nucleocapsid*, N), membrana (*membrane*, M), envoltura (*envelope*, E) y espiga (*spike*, S) (24, 25, 29).

1.6 Proteínas estructurales

Tres de las cuatro proteínas estructurales se encuentran en la superficie del virus y son esenciales para el correcto ensamblaje y función del virión. La nucleoproteína N se encuentra al interior del virus (endovirión), donde cumple la función de asociarse con el genoma de ARN e interactuar con las proteínas de membrana durante el ensamblaje (29). Para el caso de las proteínas de superficie,

todas son glicoproteínas y se encuentran decoradas por *N* y *O*-glicanos (30). Una corresponde a la proteína de membrana M, la cual contribuye principalmente con el soporte estructural de la membrana viral (30). Por otro lado, la proteína de envoltura E se encuentran en menor proporción en la membrana, sin embargo, juega un rol importante en el ensamblaje y liberación del virión (31). Finalmente, la proteína espiga es aquella encargada de unirse al receptor del hospedero y facilitar la unión del virus con la célula, por ende, esta proteína juega un rol importante en la determinación del tropismo del huésped y la capacidad de transmisión (31).

La proteína S corresponde a un trímero, en donde cada monómero se compone de dos subunidades (*Subunit 1 and 2*, S1/S2), en donde S1 posee el dominio de unión al receptor (*Receptor-Binding Domain*, RBD), mientras que S2, se compone de un péptido de fusión (*Fusion Peptide*, FP), las *Heptad Repeat Regions* (HR1/2) y un dominio transmembrana (*Transmembrane Domain*, TM). Por su parte, la proteína M se compone principalmente por un dominio N-terminal (*N-terminal Domain*, NTD) expuesto al exterior del virión y de una región C-terminal, que posee 3 secciones TM y una porción que se localiza al interior del virión. El plegamiento de E corresponde a un NTD, un TM y una porción endovirión (30, 31).

1.7 Modificaciones postraduccionales: glicosilaciones

Las glicosilaciones son modificaciones postraduccionales (PTM) que poseen las proteínas, las cuales puede contribuir con el plegamiento, modulación del reconocimiento inmune y/o influir sobre el tropismo viral (32), en donde una glicoproteína puede llevar un número variable de glicanos en su estructura. Estos se pueden clasificar en dos grandes categorías, los *N* y *O*-glicanos. Esta clasificación se basa principalmente en el enlace que posee el glicano con el aminoácido, en donde los *N*-glicanos corresponde a una cadena de hidratos de carbono unida covalentemente a un residuo de asparagina (*N*), en cambio, los *O*-glicanos se caracterizan por unirse al polipéptido mediante el enlace *N*-acetilgalactosamina (GalNAc) a un grupo hidroxilo (OH) de un residuo de *S* o *T*.

Las glicosilaciones del tipo *N*-glicosilación poseen una secuencia consenso, *sequon* o sitio de glicosilación, la cual corresponde a N/X/S-T, en donde N es el aminoácido asparagina, X es cualquier aminoácido excepto prolina *P* y *S/T* corresponden a serina o treonina, respectivamente. Estos comparten una región común central de pentasacáridos y generalmente se dividen en tres grupos: tipo oligo-manosa, tipo complejos y tipo híbridos. En cambio, como se mencionó los tipos *O*-

glicosilaciones poseen un *sequon* de *S* o *T*, las cuales pueden tener un amplio rango de núcleos estructurales que los representan (30).

1.8 Glicosilaciones en proteínas estructurales de CoVs

Dependiendo del CoV los sitios de glicosilación, tanto del tipo *N* y *O*, pueden variar en número, sitios de glicosilación y del tipo de glicanos presentes. Los glicanos juegan un papel clave en la patogénesis viral al regular el tropismo de la célula huésped y las interacciones con la respuesta inmune del huésped (35). Además, la presencia de glicanos contribuye a evadir el sistema inmune, en ocasiones impidiendo el reconocimiento de epítopos por enmascaramiento. Por lo general, la glicoproteína S se encuentra densamente poblada por *N*-glicanos. Igualmente se han reportado sitios de *O*-glicosilaciones en la estructura de S (33). Para el caso de la glicoproteína E, se han reportado un número y sitios de glicosilación variable de *N*-glicanos, dependiendo del CoV. Finalmente, para el caso de la M los estudios indican la presencia de *O*-glicanos (30).

Las glicosilaciones tienen una relación estrecha con la patogénesis viral, ya que impactan directamente en las funciones de proteínas estructurales y funcionales involucradas en el ciclo viral. Por ejemplo, las glicosilaciones son requeridas para la formación de la progenie e infectividad efectiva de dichos nuevos viriones, la formación de partículas virales y su posterior liberación. Contribuir a la evasión inmune a través de mecanismos como el mimetismo molecular (35, 36), es decir, blindar epítopos de proteínas con glicosilaciones pertenecientes a la célula huésped para evadir el reconocimiento por parte del sistema inmune. Además, pueden estar involucradas en el tropismo viral, dado la composición del glicano, pueden unirse a receptores celular afines con la composición del glicano (18, 38). Algunos de estos impactos serán discutidos más adelante.

Un mejor entendimiento de las posibles variaciones en el historial evolutivo de patrones de los sitios de glicosilación y su respectiva conservación en diferentes CoVs podría arrojar luces sobre la adaptación de los diferentes tipos de CoVs que afectan a humanos y ayudar eventualmente a potenciar el uso racional de este conocimiento para potenciales aplicaciones diagnósticas y/o terapéuticas.

1.9 Evolución molecular y su relación con las glicosilaciones en virus

La naturaleza intrínseca de los virus de ARN es su capacidad de almacenar diversidad genética en un corto periodo de tiempo, dado principalmente por su alta tasa de mutación. A pesar de la

maquinaria de reparación que poseen los CoVs mencionada anteriormente, no quedan exentos a los cambios genéticos de gran magnitud. Por ejemplo, los eventos de recombinación homóloga, es decir, cambios de material genético entre virus en un contexto de coinfección celular, también contribuyen a cambios en la diversidad genética (7, 34).

Un enfoque bastante útil es el estudio de análisis filogenéticos y la inferencia de patrones evolutivos que se pueden presentar entre especies (23, 39). Además, llevar a cabo la medición de la fuerza con la que la selección natural actúa, dado que es uno de los mecanismos principales con los cuales se producen cambios (39), es una herramienta bastante útil. Por ende, analizar los efectos sobre los genomas virales también contribuye a generar conclusiones sobre la naturaleza evolutiva. Por ejemplo, según Forni y colaboradores un enfoque para la medición de la selección natural es el siguiente:

> "la selección natural se estima comúnmente en términos de ω (también conocido como *dN/dS*), es decir, el número observado de diferencias no sinónimas por sitios no sinónimos (*dN*) sobre el número observado de diferencias sinónimas por sitio sinónimo (*dS*)" (23).

A partir de este tipo de análisis y la vía indirecta con la cual la selección natural actúa sobre el ácido nucleico, se toma en cuenta la proporción *dN*/*dS* en una secuencia. Por lo tanto, si $\omega < 1$ significa que los cambios no sinónimos causan una desventaja de aptitud significativa, lo que lleva a una selección purificadora (negativa). Si $\omega = 1$, se interpreta que los cambios no sinónimos son iguales a los sitios no sinónimos, por lo que la selección es neutral. Si $\omega > 1$ entonces el cambio de aminoácidos ha tenido un efecto positivo en la aptitud y, por tanto, hay una selección positiva (39).

Una vez se tenga una idea del tipo de selección que sufre el genoma o ciertos genes, por ejemplo, el de las proteínas estructurales, se puede enfocar el análisis a la conservación de las glicosilaciones durante el tiempo producto del tipo de selección natural. A modo de ejemplo, se han reportado tipos de selección positiva en Influenza A, en donde los análisis sugieren que las ganancias en sitios de *N*-glicosilación son seleccionados positivamente para proteger sitios antigénicos del reconocimiento inmune (40).

1.10 Evolución de CoVs y relación entre ganancia o pérdida de glicosilaciones

En lo que respecta a la evolución de los CoVs, se estima que, a partir de los 4 géneros, se produjo una separación entre *Alfa-CoV* y *Beta-CoV* con los mamíferos y, por otro lado, *Gamma-CoV* y *Delta-*

CoV coevolucionaron con las aves. Ahora bien, centrándose únicamente en los *Alfa-CoV* y *Beta-CoV*, la ganancia y pérdida de genes es un factor importante a tener en cuenta en la aparición de nuevos tipos (41).

En este sentido la selección natural actuando sobre estos procesos juega un rol importante, por ejemplo, el SARS-CoV-1 surgió posiblemente a partir de un proceso de recombinación y selección positiva entre SARSr-CoV en murciélagos, civeta y humano (43). Otro caso corresponde al del MERS-CoV, donde se presume que el posible origen se relaciona al posible hospedador intermedio, que en este caso se presume fueron los camellos. A pesar de que se pudiera pensar de que son lejanos desde un punto de vista filogenético, este CoV posee cercanía filogenética entre el linaje de los HKU-4 y HKU-5 CoV presentes en poblaciones de murciélagos (23, 42). En lo que respecta a la fuerza de la variabilidad genética, los eventos de recombinación y de selección positiva se presume que se presentaron principalmente sobre el gen S (23, 42). Para el SARS-CoV-2, diferentes estudios sugieren cercanía filogenética con el CoV RaGT13 aislado desde murciélagos, sin embargo, los análisis de identidad, selección y recombinación, particularmente en S, sugieren una cercanía con aislados de pangolín (43). Por otro lado, estudios recientes sugieren que la selección natural está actuando de manera purificadora (selección negativa), esto a partir de análisis filogenéticos entre SARS-CoV-2 y RaGT13 (44).

En lo que respecta a los CoVs que circulan comúnmente en humanos, los eventos evolutivos a los que se han visto sometidos se relacionan principalmente con el gen S. Estos son eventos de recombinación y/o deleciones/inserciones de nucleótidos en dominios de importancia para la glicoproteína S. Por ejemplo, para el caso del 229E-HCoV eventos de deleciones en el gen de la proteína S han ocasionado modificaciones del tropismo viral, por ejemplo, se presentan cambios en la preferencia del receptor de la célula hospedera, afectando un tejido diferente (23, 45). Por otro lado, para el caso de NL63-HCoV y PEDV se han reportado eventos de recombinación que involucran al gen M (45). Estos son algunos de los ejemplos que representan las presiones evolutivas a los cuales están sometidos estos genes de proteínas estructurales.

Unos pocos estudios se han centrado en la comparación entre CoVs y sus sitios de glicosilación. A modo de ejemplo, Watanabe y colaboradores (46) establecieron una comparación entre las proteínas S de SARS-CoV-1, MERS-CoV y HKU1-CoV, en donde se detectaron 23, 23 y 29 sitios de *N*-glicosilación, respectivamente. En este mismo estudio se realizó un análisis de proporción relativa de sustituciones de ω con respecto a las zonas ocupadas por glicanos y aquellas que no en la

estructura de la glicoproteína S. Los resultados sugieren que las zonas sin glicanos poseen valores más altos de ω, en comparación con los protegidos por los glicanos.

En lo que respecta a los efectos de ganancia o pérdidas de glicosilaciones en los CoVs producto de la selección natural no ha sido documentado en gran proporción, sin embargo, como se mencionó anteriormente el caso de virus Influenza A y sus glicoproteínas de superficie, Hemaglutinina (HA) y Neuraminidasa (NA), son un claro ejemplo de cómo las variaciones de las glicosilaciones pueden influir en la protección de sitios antigénicos y evasión inmune a lo largo de la evolución (47). Otro ejemplo es el caso de los Arenavirus del Viejo y Nuevo mundo, donde un estudio sugiere que el agrupamiento de sitios de glicosilación en el complejo de glicoproteína (GPC) es una característica de estos últimos. Esto se puede relacionar con la presencia de epítopos inmunodominantes y de glicosilaciones con importancia durante la historia evolutiva (48).

Finalmente, analizar el tipo de selección que sufren los CoVs y su posible relación con la conservación de sitios de glicosilación podría sugerir de qué manera la selección natural actúa para que las glicosilaciones tengan un rol activo en las proteínas estructurales. Un mejor entendimiento podría generar análisis más predictivos para futuras aplicaciones, por ejemplo, llevando a cabo la selección de ortólogos de las proteínas estructurales y evaluando el grado de selección de residuos expuesto al sistema inmune, lo cual permite seleccionar péptidos que pueden ser usados como candidatos para la generación de vacunas. Esto se discutirá en mayor detalle en la sección implicaciones, recomendaciones y/o proyecciones futuras.

Las principales interrogantes que pretende abordar esta investigación de tesis son las siguientes: ¿Cómo varía la posesión de potenciales sitios de glicosilación entre diferentes miembros de CoVs? ¿Cómo varía el contenido de los *sequons* a través de los diferentes tipos CoVs? ¿Qué tipo de presión evolutiva determina la variación de sitios de glicosilación en los *Alfa-CoV* y *Beta-CoV*?

2. HIPÓTESIS Y OBJETIVOS

2.1 Hipótesis del estudio

Hipótesis nula (H_0): Los sitios de glicosilación en las proteínas de superficie (S, M, E) de las CoVs que infectan seres humanos, no sufren selección positiva, en comparación con sus relativos más cercanos que no infectan humanos.

Hipótesis alternativa (H_1): Los sitios de glicosilación en las proteínas de superficie (S, M, E) de las CoVs que infectan seres humanos, son seleccionados positivamente, en comparación con sus relativos más cercanos que no infectan humanos.

2.2 Objetivos del estudio

Objetivo general: Determinar el tipo de selección natural y su relación con la aparición de sitios de glicosilación en los miembros de los *Alfa-* o *Beta-CoV* detectados en humanos.

Objetivo específico 1: Establecer la relación filogenética de los diferentes linajes de *Alfa* y *Beta-CoV*.

Objetivo específico 2: Determinar la presión evolutiva en los genes de proteínas estructurales de los *Alfa* y *Beta-CoV* a través del análisis de tasas de mutaciones sinónimas y no-sinónimas.

Objetivo específico 3: Determinar los sitios de *N-* y *O-*glicosilaciones en proteínas estructurales de *Alfa* y *Beta-CoV* seleccionados.

Objetivo específico 4: Evaluar las características de los sitios bajo selección positiva o negativa y los sitios de glicosilación de las proteínas estructurales.

3. MATERIALES Y MÉTODOS

3.1 Materiales

Tabla 1. Base de datos utilizadas para recopilar y analizar las secuencias genómicas.

Nombre	Identificador	Referencia
NCBI Virus	https://www.ncbi.nlm.nih.gov/labs/virus/vssi/#/	49, 50
Datamonkey Adaptive Evolution Server	http://www.datamonkey.org/	51

Tabla 2. Programas y algoritmos utilizados para analizar y evaluar los datos.

BioPython	https://biopython.org/	52
Bio Perl	https://bioperl.org/ https://metacpan.org/pod/BioPerl	53
MAFFT	https://mafft.cbrc.jp/alignment/server/	54
AliView	https://ormbunkar.se/aliview/	55
Jalview	https://www.jalview.org/	56
IQ-Tree	http://www.iqtree.org/	57
ModelFinder	http://www.iqtree.org/	58
Ultrafast Bootstrap	http://www.iqtree.org/	59
iTOL	https://itol.embl.de/	60
ORFinder	https://www.ncbi.nlm.nih.gov/orffinder/	61
BLAST	https://blast.ncbi.nlm.nih.gov/Blast.cgi	62
PAL2NAL	http://www.bork.embl.de/pal2nal/	63
PAML y CODEML	http://abacus.gene.ucl.ac.uk/software/paml.html	64
NetNGlyc 1.0	http://www.cbs.dtu.dk/services/NetNGlyc/	65, 66

NetOGlyc 4.0	http://www.cbs.dtu.dk/services/NetOGlyc/	65, 67
GraphPad Prism 8	https://www.graphpad.com/scientific-software/prism/	68
FEL	http://www.datamonkey.org/	69
FUBAR	http://www.datamonkey.org/	70
Logolas	https://github.com/kkdey/Logolas	71
DrawProteins	https://github.com/brennanpincardiff/drawProteins	72

Tabla 3. Lenguajes de programación.

Nombre	**Identificador**
Python 3.9	https://www.python.org/about/
RStudio 1.3.1093	https://rstudio.com/
Strawberry Perl 5.32.0.1	https://www.perl.org/

3.2 Métodos

El flujo de trabajo para cada objetivo específico planteado se puede visualizar en la sección de material suplementario, donde se presentan los diagramas de flujos del 1 al 4.

3.2.1 Adquisición de secuencias, alineamiento múltiple de secuencias (MSA)

Las secuencias de los genomas completos y tres proteínas estructurales (S, M y E) de 19 CoVs diferentes fueron descargadas a partir de la base de datos *NCBI Virus* (49, 50) (Tabla N°1 y SN°1). El listado de los CoVs analizados se encuentra en la Tabla N°4.

Todas los programas y algoritmos utilizados se muestran en la Tabla N°2. Para obtener el genoma completo y las secuencias codificantes de las proteínas estructurales se utilizó la herramienta BioPython (52). En base a esta última se buscaron y seleccionaron las secuencias, tanto nucleotídicas como aminoacídicas, para los genomas y las proteínas estructurales. Para ambos casos se utilizó el formato de Fasta para los archivos de salida. Además, se procedió a obtener los archivos de metadatos en formato CSV, en donde se recopiló información de interés para posteriores filtraciones de datos: identificación de acceso (*Accession*), fecha de lanzamiento (*Release date*), especie (*Species*), género (*Genus*), tipo de secuencia (*Sequence type*), estado completo de secuencia (*Nuc completeness*), localización geográfica (*Geo Location*), hospedero (*Host*) y fecha de colección (*Collection date*).

En paralelo se identificaron aquellos genomas que no se encontraban anotados, para proceder con la herramienta *ORF Finder* (61) para identificar las secuencias codificantes de las proteínas estructurales.

Para confirmar dicha secuencia se procedió a través de análisis *BLAST* (62), utilizando la secuencia de referencia (*NCBI RefSeq*) para confirmar que pertenece al CoVs de interés, por ejemplo, se utilizó NC_045512 (aislado *Wuhan-Hu-1*) para el genoma de referencia del SARS-CoV-2.

Los alineamientos de las secuencias para los genomas completos se realizaron de manera separada de las proteínas estructurales para cada CoVs; en donde para el caso de los genomas se realizó un alineamiento múltiple de los 19 CoVs, en cambio, para las proteínas estructurales se alineó cada

proteína para el conjunto completo de CoVs. La herramienta para llevar a cabo la alineación múltiple de secuencias (*MSA*) fue a través de *MAFFT* (54) en modo de alineamiento progresivo (*FFT-NS-2*). En paralelo, los MSA generados también fueron tratados para proceder al alineamiento de codones a través de *PAL2NAL* (63). Finalmente, se continuó con la obtención de los *MSA* en formato Phylip, con su respectiva visualización y edición a través de *AliView* y *Jalview* (55, 56).

Tabla 4. Listado de los CoVs analizados y las especies en las cuales circulan.

Coronavirus	Especie o Género	Número secuencias
SARS-CoV-1	Humano (*Homo sapiens*) Murciélago (*Rhinolophus spp*) Civeta de Palma (*Paguma larvata*)	90
SARS-CoV-2	Humano (*H. sapiens*) Murciélago (*Rhinolophus spp, Rhinolophus affin*) Pangolin (*Manis javanica*)	127
SARSr-Bat	Murciélago (*Rhinolophus spp, Rhinolophus sinicus*)	16
MERS-CoV	Humano (*H. sapiens*) Murciélago (*Tylonycteris spp* y *Neoromicia spp*)	111
229E-HCoV	Humano (*H. sapiens*)	30
NL63-HCoV		58
OC43-HCoV		99
HKU1-HCoV		38
HKU2-CoV	Murciélago (*R. sinicus*, *Rhinolophus spp*, *Tylonycteris spp, Pipistrellus spp*)	9
HKU3-CoV		12
HKU4-CoV		8
HKU5-CoV		8
229E-CoV-Camel	Camello (*Camelus spp*)	5
HKU23-CoV		9
HKU24-CoV	Rata (*Rattus spp*)	3
MHV-CoV	Ratón (*Mus musculus*)	27
Bovine-CoV	Ganado (*Bos spp*)	41
PEDV-CoV	Cerdo (*Sus spp*)	19
PHEV-CoV		12

3.2.2 Análisis filogenético basado en *Maximum Likelihood* (ML)

Para la obtención de los árboles filogenéticos, se utilizó el método de *maximum likelihood* (ML). Para el árbol de genomas, se generó un alineamiento múltiple con la secuencia completa de todos los 19 grupos de CoVs, utilizando la herramienta *MAFFT* (54). También se generaron alineamientos para cada set de proteínas estructurales. Cada filogenia se ejecutó en el programa IQ-*Tree* (57), evaluando la robustez de los clados a través de *Bootstrapping* (*Bootstrap Support*, BS) de 10.000 réplicas, con el paquete *Ultrafast Bootstrap* (58). Para determinar el mejor modelo de sustitución para cada filogenia de acuerdo con cada tipo de secuencia (ADN genómico, codones y/o aminoácidos), se utilizó la herramienta *ModelFinder* (59). El mejor modelo para el árbol de genoma completo es GTR+F+R10 (modelo general de tiempo reversible con tasas y frecuencia de bases desiguales), mientras que, para las proteínas estructurales, se determinó que los mejores modelos para las secuencias nucleotídicas corresponden a GTR+F+I+G4, TIM2+F+G4 y TN+F+G4, para las proteínas a S, M y E, respectivamente. Todos los árboles generados fueron visualizados y editados con *iTOL* (60).

3.2.3 Análisis para identificar el tipo de selección natural a nivel de secuencias y por sitios a nivel de proteína estructural

Con el objetivo de identificar el tipo de selección dentro de diferentes grupos de secuencias, se utilizó el paquete *PAML* (*Phylogenetic Analysis by Maximum Likelihood*), con su programa CODEML (64). Con el fin de optimizar estimaciones estadísticas, se utilizaron diferentes modelos de sustitución en *CODEML*, seleccionando el mejor modelo para representar las tasas *dN*, *dS* y *ω* (*dN/dS*). El valor *ω* permite estimar el tipo de selección natural actuando a en genes codificantes, en donde si $\omega < 1$ indica que hay una selección negativa, en caso de ser $\omega = 0$ hay una selección neutral y de ser $\omega > 1$ es una selección del tipo positiva (23). Los modelos evaluados incluyeron: M0 ("*One ratio*"), M3 ("*Discrete*"), M1a ("*Nearly Neutral*"), M2a ("*Positive Selection*"), M7 ("*Beta*") y M8 ("*Beta + ω*"). Los mejores modelos fueron seleccionados en función de su *Likelihood Ratio Test* (LRT), utilizando la *Chi-square Test* (χ^2) como medida de significancia estadística. Los números de grados de libertad de cada modelo también fueron considerados. Los modelos de sustitución M7 y M8 fueron los que dieron los mejores resultados, para el caso de M7 no permite la detección de selección del tipo positiva, en contraste con M8, de modo que es posible comparar ambos modelos.

A fin de evitar sesgos comparativos, valores de ω mayores a 10 (provenientes de estimaciones no realistas), fueron filtrados del set de comparaciones. Además, se llevó a cabo el cálculo de las distancias genéticas (según el modelo de *Tajima-Nei*) entre las secuencias utilizando el paquete *Bio::Align::DNAStatistics* de BioPerl (53). Estas distancias fueron comparadas con las tasas *dN*, *dS* y ω.

Para identificar sitios particulares bajo selección positiva o negativa, se utilizó la aplicación *Datamonkey Adaptive Evolution Server* (51), en su módulo de Probabilidad de Efecto Fijo (*Fixed Effects Likelihood*, FEL); este método asume que la presión selectiva es constante a lo largo de toda la filogenia (69). Para la detección de sitios seleccionados también se utilizó el método de Aproximación bayesiana rápida y sin restricciones *(Fast, Unconstrained Bayesian AppRoximation,* FUBAR), el cual se basa principalmente en la detección de selección natural a través de un enfoque Bayesiano (70). En ambos casos, los sitios con algún tipo de selección fueron filtrados de acuerdo a su valor *p* (< 0,05). Ambos métodos fueron evaluados de modo paralelo, a fin de complementarse al evaluar la presencia de algún tipo de selección sobre algún sitio en particular.

3.2.4 Predicción de sitios de glicosilación, representación de logotipos y generación esquemática de proteína S basado en la distribución de glicosilaciones

Para la predicción de los sitios de *N*-glicosilación, se dispuso de la aplicación web *NetNGlyc* (utilizando un valor de corte mínimo de 0.5) (65, 66); para la predicción de las *O*-glicosilaciones se usó *NetOGlyc* (utilizando un valor de corte mínimo de 0.5, según lo recomendado por la herramienta para predecir de manera efectiva sitios de glicosilación) (65, 67). Para cuantificar, graficar y comparar el número de sitios predichos totales y por secuencia, se dispuso de la aplicación *GraphPad Prism* 8 (68). Las glicosilaciones de cada CoVs fueron agrupadas en grupos definidos por el estudio, en donde los agrupamientos se definieron según hospedero, principalmente entre humanos y otros animales. Posteriormente, este agrupamiento contribuyó a definir sitios glicosilados consenso de las proteínas estructurales, las cuales se utilizan para representar un sitio glicosilado en diferentes proteínas estructurales de CoVs relacionados.

La cuantificación de las glicosilaciones para cada CoVs fue distribuida según los grupos relacionados previamente para continuar con la identificación de diferencias significativas entre grupos. Con el fin de chequear el estado de los datos (paramétrica o no paramétrica), se procedió a realizar la

prueba de normalidad. Posteriormente, se procedió a realizar comparación entre grupos para identificar diferencias significativas, en cuyo caso se utilizaron *Mann Whitney U Test* (al comparar dos grupos) y *Kruskal-Wallis test* (al comparar más de dos grupos).

Para determinar cuál es la secuencia que compone cada sitio de glicosilación predichos para cada CoVs, se extrajo cada uno de los *sequons* para las *N* y *O*-glicosilaciones, y los consensos alineados se representaron a través de la herramienta *EDlogo* (*Enrichement Depletion Logo*), del paquete *Logolas* (71). *EDlogo* corresponde a una versión de *Weblogo* (73) que permite representar el nivel de enriquecimiento y agotamiento de las posiciones aminoacídicas en las secuencias al mostrar posiciones específicas.

Con el fin de ejemplificar la distribución de las glicosilaciones en las proteínas S de un grupo de CoVs (SARS-CoV-1, SARS-CoV-2, MERS-CoV, SARSr-RaTG13, SARSr-pangolín y SARSr-BM48-31) se procedió a utilizar la herramienta *DrawProteins* (72), utilizando las predicciones de NetNGlyc y NetOGlyc anteriormente obtenidas.

4. PRESENTACIÓN Y ANÁLISIS DE RESULTADOS

4.1 Análisis filogenético de los genomas y proteínas estructurales de coronavirus

El resultado del análisis para determinar las relaciones filogenéticas de los genomas de los CoVs se puede visualizar en la Figura N°1. En los *Alfa-CoV* el número corresponde a 5 CoVs diferentes (Figura N°1 y SN°1), donde se presentan 2 causantes de cuadros leves en humanos y que corresponden al NL63-HCoV y 229E-HCoV (resaltados en verde azulado oscuro y rosado piel). Se incluyen 2 patógenos que circulan en rumiantes como es el caso del PEDV y 299E-Camello (agrupamiento en color morado y morado claro). Por último, se encuentra el único aislado de murciélago presente en *Alfa-CoV,* el HKU2-CoV (destacado en azul). Para el caso de los *Beta-CoV* el número total corresponde a 14 CoVs, donde se muestran los principales CoVs patógenos que se saben son capaces de infectar a humanos tales como el SARS-CoV-1, SARS-CoV-2 y MERS-CoV. A través de una mirada global de la filogenia correspondiente al genoma completo y al de las proteínas estructurales, se puede mencionar que las diferencias en términos de filogenia entre *Alfa* y *Beta* es mucho más definida para el caso del genoma completo; en cambio, para las proteínas estructurales los agrupamientos varían dependiendo principalmente de la proteína (Figuras N°2). La Figura SN°1 muestra en detalle los agrupamientos de *Beta* y *Alfa-CoV* distribuidos a lo largo de la filogenia.

Para el caso de los *Beta-CoV*, el agrupamiento correspondiente al SARS-CoV-2 y SARSr-CoV (*related*, relacionado) (destacado en color rojo intenso) los posiciona en una cercanía filogenética con secuencias virales aisladas de murciélagos, la cual incluye el tipo de murciélago más cercana al virus pandémico actual, SARSr-CoV-RaTG13 (resaltado en color rojo intenso). Además, en este mismo agrupamiento se encuentran los aislados de pangolín provenientes de la provincia de Guangdong, China. Para el caso del SARS-CoV-1 el análisis determina que las secuencias genómicas se congregan en conjunto con aislados de civeta y murciélago (SARS-bat-BM4831) (colores rojo claro) los cuales han sido reportados como posibles orígenes zoonóticos del virus epidémico del 2002-2003. Para el caso del MERS-CoV, muestra cercanía con aislados de camello y murciélago (MERS-CoV-Neoromicia) (agrupamiento en color verde claro). Para el resto de *Beta-CoV* presentes, los principales aislados corresponden a virus presentes en murciélagos como lo son HKU3, 4 y 5-CoV (agrupados entre SARS-CoV-1 y MERS-CoV en color naranja y amarillo), al igual que con virus capaces de infectar a camellos como el HKU23-CoV (destacado en azul claro). Estos poseen cercanía filogenética con SARS-CoV-1 y MERS-CoV según el resultado generado por el análisis de ML. El análisis también involucró aquellos

CoVs que se sabe circulan en humanos y causan cuadros leves como es el caso del HKU1-HCoV y OC43-HCoV (destacados en colores rosado claro y amarillo claro). Además, dentro de este grupo se encuentran aquellos que afectan principalmente a roedores como lo son MHV-CoV y HKU24-CoV

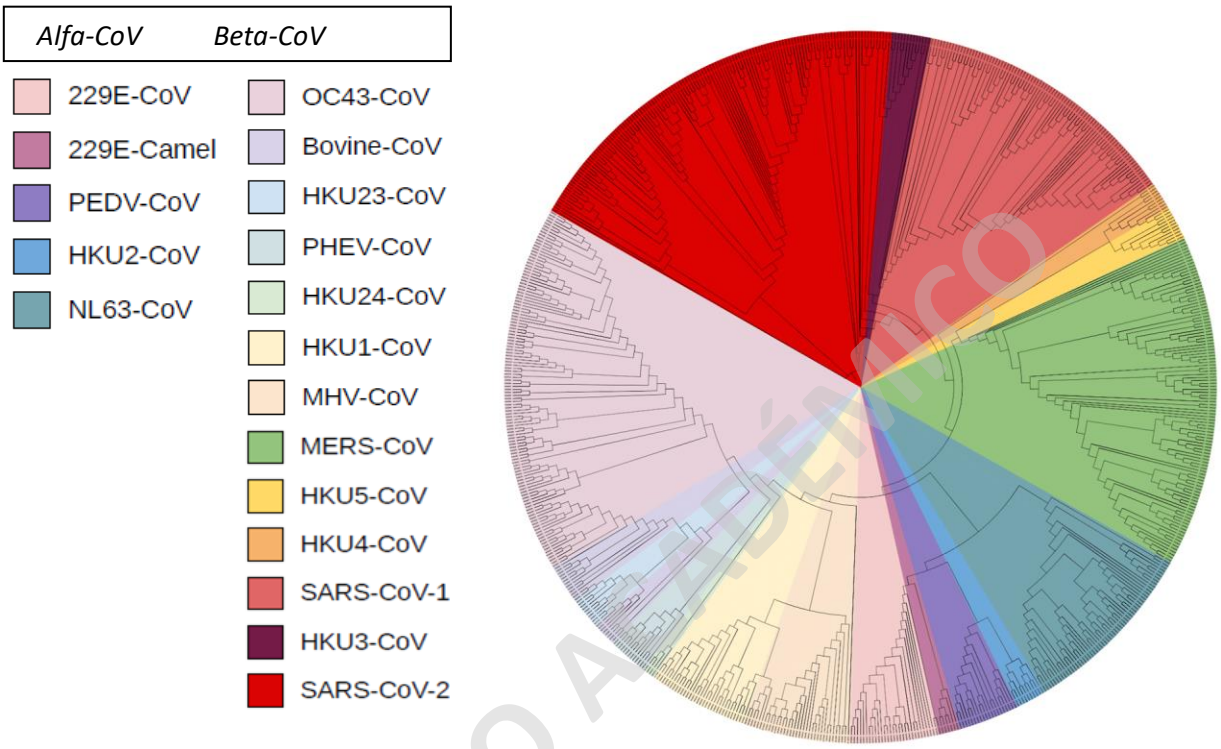

Figura 1. Árbol filogenético del genoma completo.

Árbol filogenético de *Alfa* y *Beta-CoV* correspondiente al alineamiento del genoma completo. El árbol representa un total de 19 tipos de CoVs, en donde el rango de colores representa el clado al cual pertenece cada uno. El árbol posee un soporte *bootstrap* de 10.000 réplicas cada uno (soporte no mostrado en los nodos). Se destacan los CoVs con mayor presencia de secuencias correspondientes a CoVs que circulan (o circularon) en humanos como OC43, NL63, MERS, SARS1 y SARS2.

(colores naranjo claro y verde claro). Por último, el agrupamiento de virus que circulan principalmente en rumiantes corresponde al PHEV-CoV y Bovino-CoV (resaltados en colores verde azulado claro y morado claro).

En lo que respecta a las proteínas estructurales (S, M y E) los tres análisis filogenéticos presentan incongruencia filogenética en comparación al análisis para el genoma completo y entre las mismas proteínas. Esto ha sido reportado previamente dado que las proteínas estructurales han sido foco de eventos de recombinación en el paso para los diferentes CoVs analizados. Por ejemplo, la Figura N°2 muestra que para el caso de la proteína S, la cercanía filogenética cambia en relación aislados de pangolín y el SARS-CoV-2, donde SARS-CoV-RaTG13 ya no es el más cercano al SARS-CoV-2. En particular, los aislados correspondientes al SARS-CoV-1 poseen una cercanía con los aislados relacionados al SARS-CoV-2, principalmente para la proteína S. En el caso de las proteínas M y E, el cambio es menos evidente, dado que los agrupamientos se encuentran definidos entre los CoV severos, al igual que sus cercanos presentes en murciélagos, es decir, HKU-3, 4, 5 y SARSr-CoV.

Para el caso del 229E-CoV y 229E-Camello su relación filogenética se mantiene a lo largo de las tres filogenias representadas. Esta conservación de la filogenia también se puede observar para el caso de HKU23-CoV, Bovino-CoV y PHEV. Estos CoVs mencionados corresponden a los que comúnmente circulan en rumiantes.

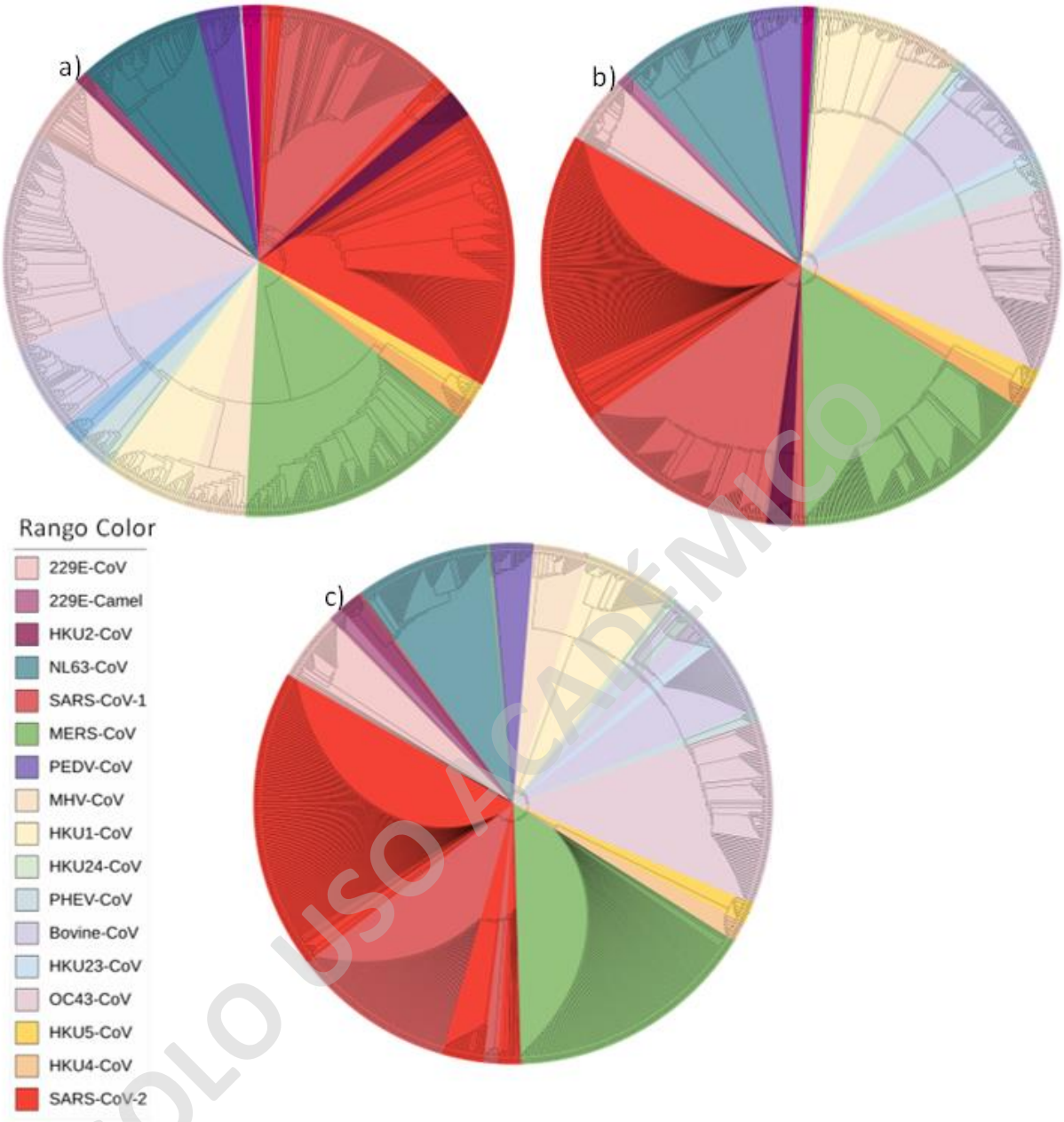

Figura 4. Árbol filogenético de las proteínas estructurales.

Árbol filogenético de *Alfa* y *Beta-CoV* correspondiente al alineamiento de las proteínas estructurales S (a), M (b) y E (c). Los árboles poseen un soporte *bootstrap* de 10.000 réplicas cada uno (soporte no mostrado en los nodos). Una de las características más relevantes de los árboles mostrados corresponde al fenómeno de incongruencia filogenética, por lo cual los árboles difieren en topología dado el alineamiento de las proteínas estructurales.

4.2 Análisis de la selección natural en las secuencias de las proteínas estructurales para los diferentes tipos de CoVs

Para evaluar el tipo de selección presente en las secuencias de los CoVs analizados, se realizó la estimación de los valores *dS*, *dN* y *ω*, por medio del programa CODEML del paquete PAML, utilizando alineamientos para los genes estructurales de los 727 genomas de los diferentes tipos de CoV secuenciados. Esto involucró la evaluación de los diferentes modelos de sustitución de codones para cada grupo de proteínas estructurales. Tal evaluación incluyó el cálculo de parámetros definidos (*np*), valores de verosimilitud estimados (*lnL*), *p-value* y grados de libertad (*df*) para cada combinación de modelos (Tablas SN°2, 3 y 4). Para estos casos, cada parámetro tiene un uso diferente: los *np* sirven para verificar cuál modelo es más enriquecido; los *lnL* para comparar entre modelos y *df* para definir el grado de restricciones a los modelos. A partir de esto, se pudo determinar que la mejor combinación de modelos para los datos correspondió al M7 y M8 (donde M7 asume una distribución tipo *β* e impone la ausencia de selección positiva, mientras que M8 asume distribución *β,* pero permite la existencia de valores *ω > 1*), sin embargo, entre ambos el mejor para representar los datos obtenidos corresponde al M8 según el LRT. Para determinar el mejor modelo para estimar selección en las proteínas estructurales en la Tabla SN°2, 3 y 4 muestran los valores con los cuales se procedió a determinar *2ΔlnL* para cada comparación y seleccionar el más adecuado según el resultado de la comparación de sus *lnL*.

La distribución de los valores *dN*, *dS* y *ω* obtenidos se representó en una comparación a pares, distinguiendo a comparaciones inter- o intragrupos de acuerdo a la clasificación de cada tipo de CoV (ver Tabla 4). Para el caso de tipos de CoVs humano-específicas, se hizo una diferenciación con respecto a los virus causantes de cuadros leves y virus causantes de enfermedades severas. Los resultados de dichas comparaciones pueden ser visualizados en la Figura N°3 y 4, en formato gráfico de dispersión (*Scatterplot*), o en las FigurasN°5 y 6 para una visualización en formato gráfico de cajas (*Boxplot*). Para el caso de las comparaciones a pares de *ω-dN y ω-dS* para grupos severos y de resfriado común, se observan diferentes agrupamientos de diferentes tipos de CoVs, en ambos casos, se observan agrupamientos con valores de *ω* con valores entre 0 y 1. Para evidenciar la frecuencia de distribución de los valores *ω* obtenidos se dispuso de la visualización de histogramas para comparaciones intra- e intergrupos (Figura SN°2), al igual que para cada proteína estructural (Figura SN°3). Tal como se puede observar, el conjunto de valores *ω* se concentra principalmente entre 0 y 1, para comparaciones intergrupo. Al tratarse de las proteínas estructurales, S es aquella

que concentra valores entre 0 y 1, a diferencia de las proteínas M y E, las cuales tienen una distribución mucho más homogénea.

Para el caso de la proteína S (Figura N°5), cerca del 25% de valores *ω* fue detectado como selección positiva. Cabe destacar que dentro de este 25% se distribuye entre valores que van de 1 y 2 para *ω,* además cerca del 50% se encuentra cercano a 0. En cambio, para el caso de las proteínas M y E se evidencia que los valores se distribuyen en más de un 75% entre los valores que van desde 1 a 2,5. Incluso, cerca del 25% posee valores que son incluso 5. Para caso de *dN* y *dS* la distribución mencionada sugiere ser diferente, donde para el caso de S los valores se distribuyen entre 0,3-0,9 y 0-5,8, para *dN* y *dS*, respectivamente. Esto varía para el caso de M y E en donde cerca del 50% de valores se distribuye entre 0 y 0,8. Además, la Figura N°6 representa cómo es la distribución inter e intragrupo, tanto de *dN, dS* y *ω.* Según la data representada la distribución de valores para *ω* en la comparación intergrupo para S cerca del 20% se distribuye entre 1 y 2,4 en donde esto contrasta con el caso de M y E, dado que más del 75% de valores *ω* se distribuye entre 1,5 y 5. Para el caso de *dN* y *dS* la comparación intergrupo varía entre ambos tipos de mutación, manteniéndose para S, M y E, donde el 50% para *dN* se encuentra bajo el umbral de 1, al igual que para M y E. Según la comparación intragrupos en los 3 casos se cumple que los valores se encuentran bajo 0,5. Ahora bien, se debe recalcar que para el caso de la Figura N°5 se presenta un alto número de valores atípicos, especialmente para el caso de M y E. Cuando se comparan la distribución de estos valores fuera del rango esperable, podría ser producto de la poca variabilidad en las secuencias de las proteínas estructurales mencionadas. Esto se puede evidenciar en la Figura N°6, en donde la comparación intragrupos posee una menor cantidad de valores fuera del rango esperable, en comparación con los intergrupos en donde se visualiza una distribución más amplia para las tasas *dN, dS* y *ω.*

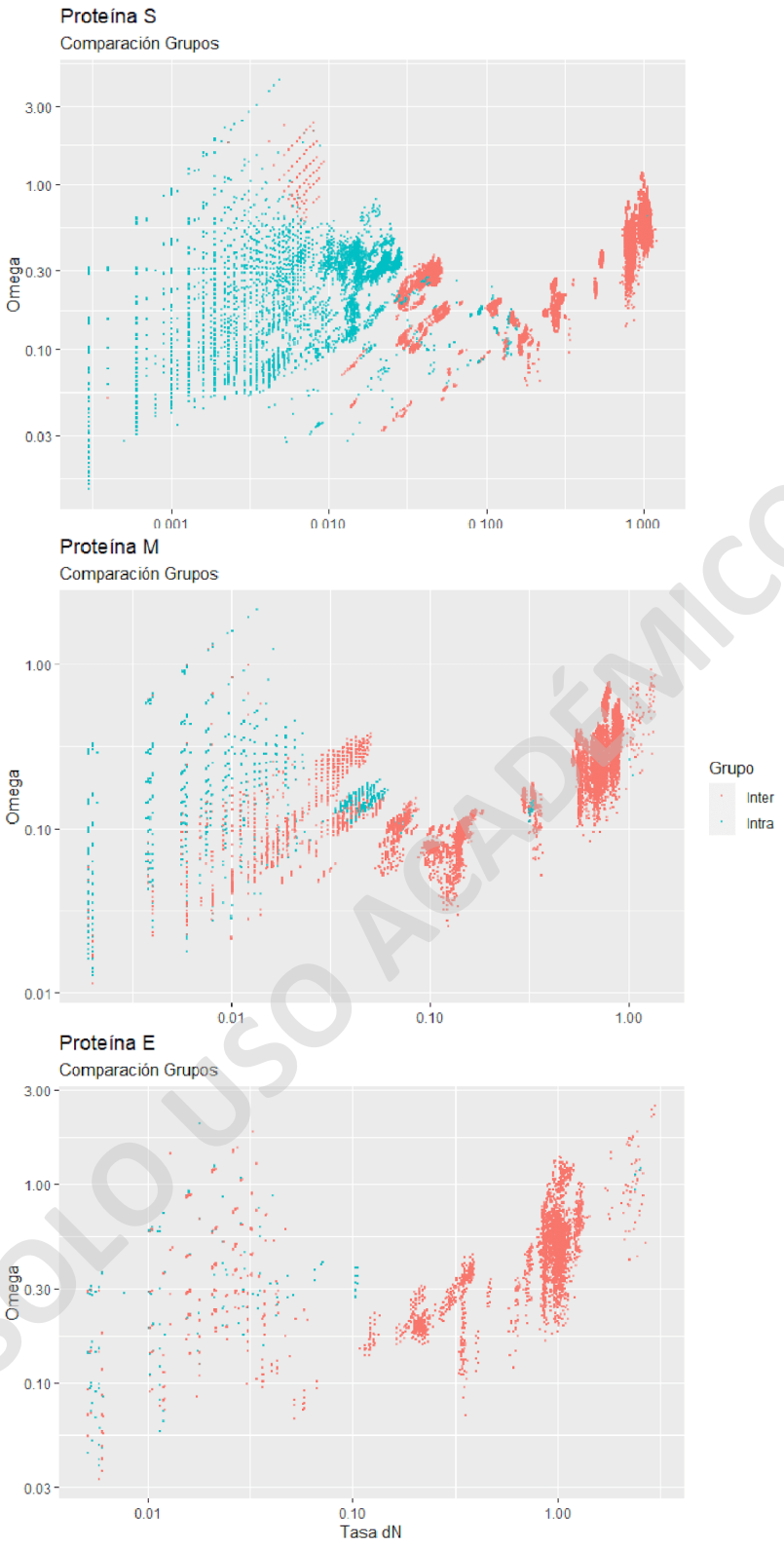

Figura 7. Comparación tasa *dN* versus *ω* en proteínas estructurales.

Estimación de selección natural en comparación a pares utilizando el modelo M8 para las proteínas estructurales S, M y E. Donde se muestra principalmente la distribución de valores estimados según la comparación entre la tasa *ω* y *dN,* dicha comparación muestra los valores intra e intergrupos.

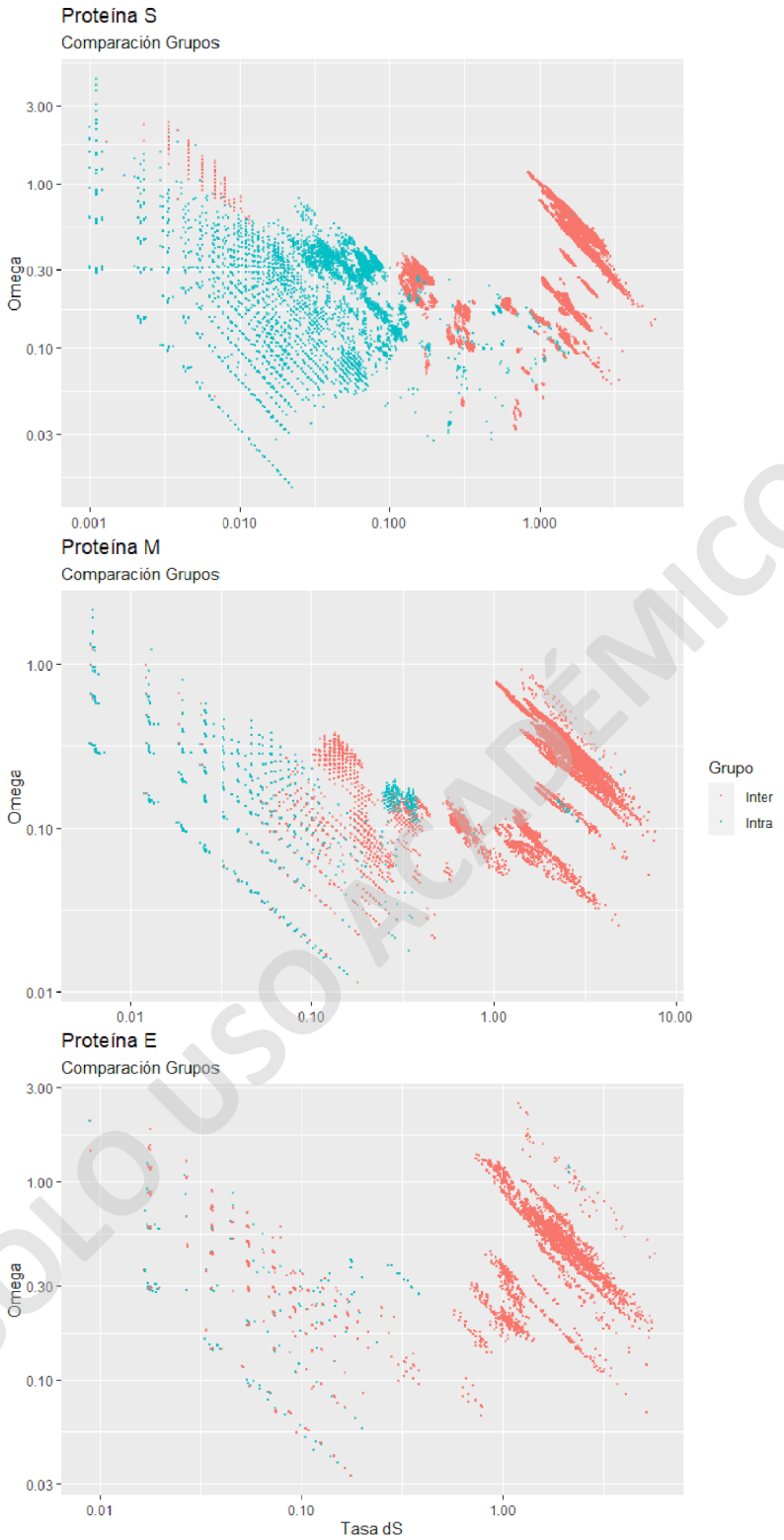

Figura 10. Comparación tasa *dS* versus *ω* en proteínas estructurales.

Estimación de selección natural en comparación a pares utilizando el modelo M8 para las proteínas estructurales S, M y E. Donde se muestra principalmente la distribución de valores estimados según la comparación entre la tasa *ω* y *dS,* dicha comparación muestra los valores intra e intergrupos.

Los resultados revelaron que el promedio del valor *ω* para la proteína S fue de 0,43, en cambio, para M y E los valores fueron de 1,84 y 1,63, respectivamente. Los valores promedios de *dN y dS* para cada proteína estructural se pueden visualizar en la Tabla N°5. Como se puede esperar, estos resultados sugieren que a nivel de proteínas estructurales se presenta una selección del tipo mayoritariamente neutral y en menor cantidad positiva. Utilizando un total 256.830, 144.013 y 149.393 comparaciones pareadas para S, M y E, respectivamente, la distribución de las comparaciones inter- e intragrupo para las proteínas estructurales (Figura N°5), en la cual se estimaron un total de 0,84 % bajo selección positiva para el caso de S, en contraste con M y E, cuyo porcentaje ronda el 60% en la comparación de pares según lo evidenciado en la figura mencionada (N°5) y Tabla N°5.

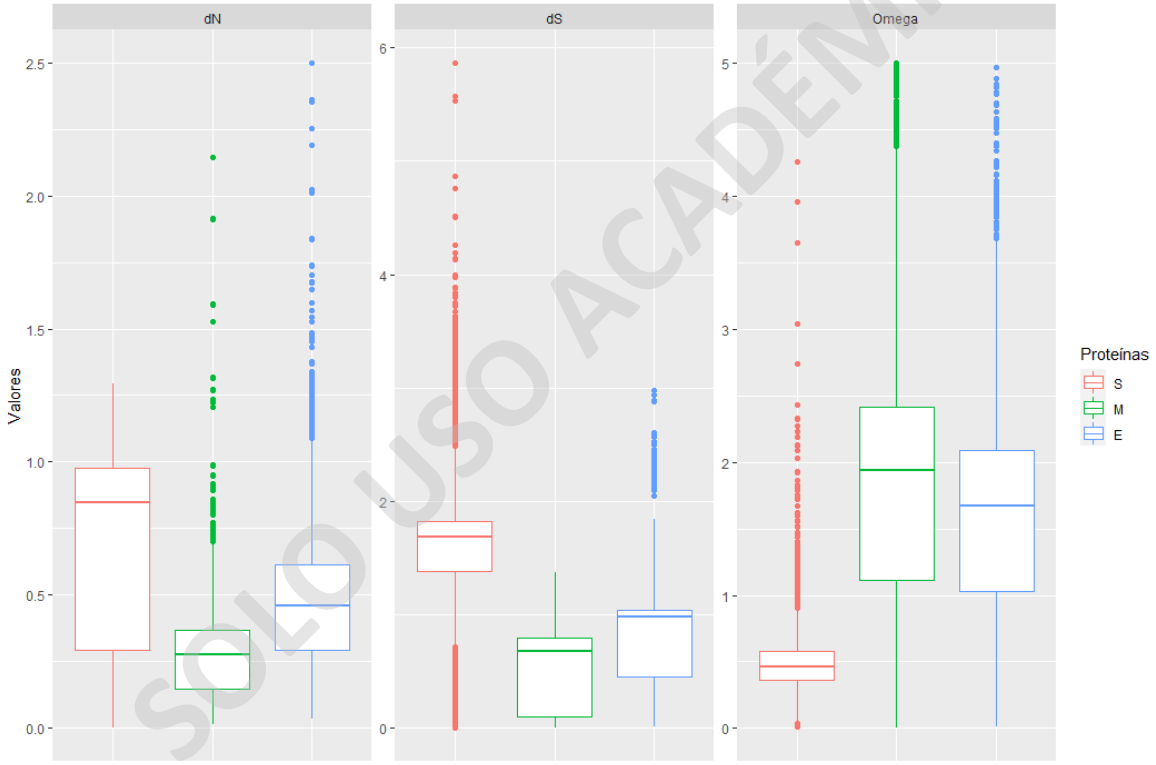

Figura 13 Distribución de tasas de selección para proteínas estructurales.

Gráfico de caja *(Boxplot)* para representar la distribución de las tasas estimadas para las proteínas estructurales de cada CoVs correspondiente al *dN, dS* y *ω.* Valores obtenidos a través del paquete CODEML, utilizando el modelo M8.

Para una estimación de la distancia evolutiva a través de las secuencias de proteínas estructurales de los tipos de CoVs severos, se procedió a calcular la distancia entre secuencias según el modelo de *Tajima-Nei* (la cual asume una igualdad en las tasas de transversión y transición, es decir, de

sustituciones entre nucleótidos de bases de distinto o de igual tipo (sean purinas o pirimidinas), respectivamente) (68). Los valores de distancia fueron contrastados con la estimación de ω (Figura SN°4), donde se evidencian las distancias estimadas para ortólogos de la proteína S. En particular, se calcularon para las secuencias nucleotídicas de los tipos SARS-CoV-2, SARSr-RaTG13-CoV, SARS-CoV-1, SARSr-BM4831-CoV, MERS-CoV y MERSr-Neoromicia-CoV. La relación de comparación entre ω y distancia *Tajima-Nei* mostró dos tipos de agrupamientos. El primero está caracterizado por poseer distancias relativamente muy cortas, el cual se refiere a la agrupación de ortólogos proveniente de distintas variantes del SARS-CoV-2, el cual corresponde a comparaciones entre miembros del mismo grupo. El segundo agrupamiento, posee grupos con mayor distancia genética entre sí y corresponden a comparaciones entre SARS-CoV-2 y otros grupos.

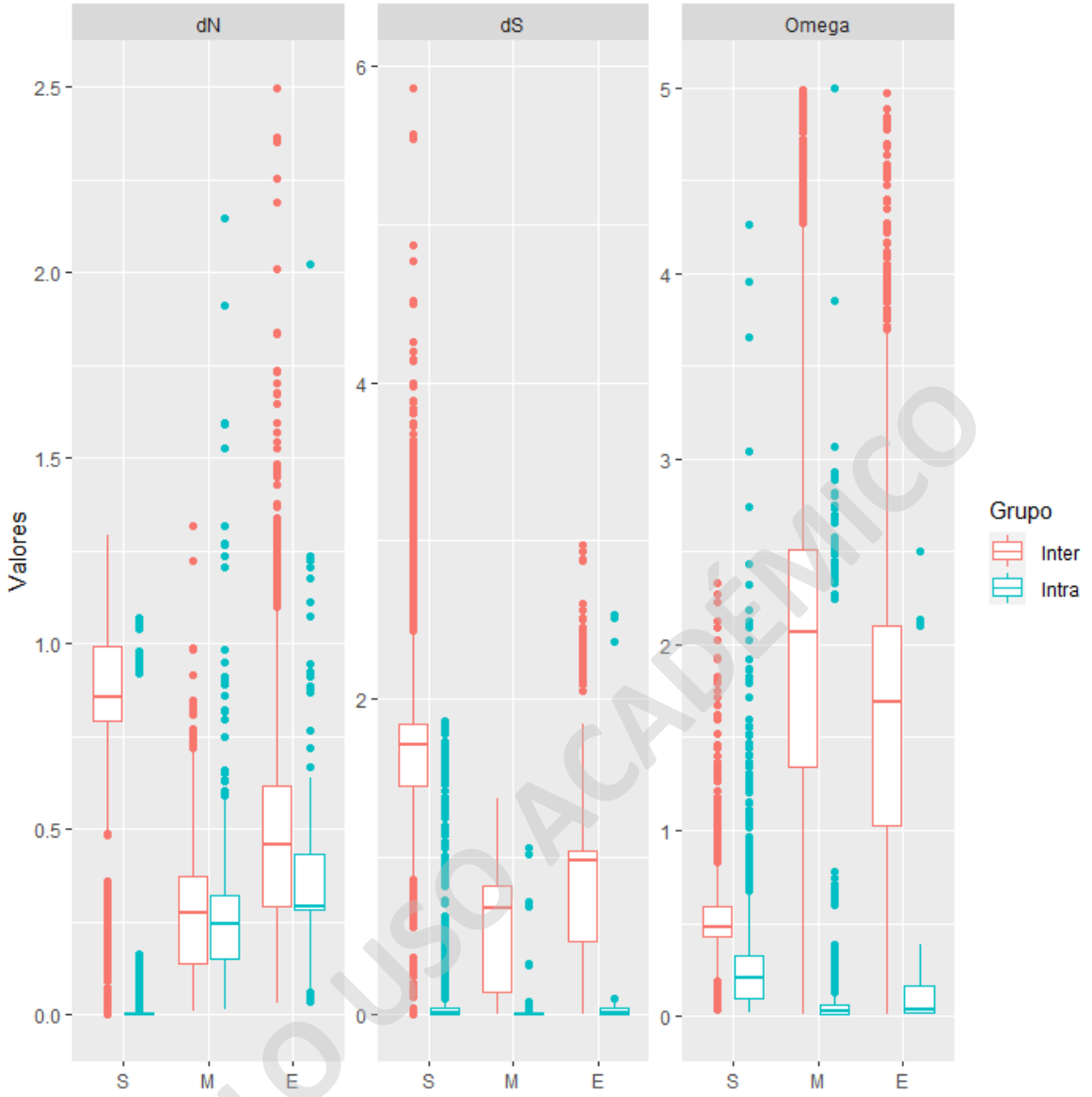

Figura 16. Distribución de tasas de selección en comparaciones intragrupo.

Gráfico de cajas *(Boxplot)* de la distribución de valores estimados para las proteínas estructurales en la distribución intra e intergrupo correspondiente al *dN, dS* y *ω*. Para las categorías inter e intragrupo se comparan las tasas estimadas de cada CoVs, al presentarse una comparación con un aislado del mismo tipo se considera intra, en caso contrario intergrupo.

Tabla 5. Estadísticas generales de tasas de selección.

Valores promedio, desviación estándar (SD) y mediana para las tasas *dN, dS y ω* para cada proteína estructural a partir de la comparación a pares de cada CoVs utilizando el paquete CODEML. En la tabla se muestra el promedio, la SD y la mediana.

Proteína Estructural	*dN*	*dS*	*ω*
S	0,70* ± 0,36× 0,84ł	1,47* ± 0,61 1,67ł	0,42* ± 0,31× 0,46ł
M	0,28* ± 0,16× 0,27ł	0,51* ± 0,33× 0,67ł	1,83* ± 1,09× 1,93ł
E	0,50* ± 0,27× 0,45ł	0,77* ± 0,40× 0,97ł	1,62* ± 0,97× 1,67ł

*: Promedio; ×: desviación estándar; ł: mediana.

4.3 Predicción, cuantificación y caracterización de *sequons* presentes en las proteínas estructurales de los CoVs

Utilizando el alineamiento de las proteínas estructurales se procedió a la predicción de los sitios de glicosilación, enlazados tanto en *N*- como de *O*-, a través de NetNGlyc y NetOGlyc. Para las proteínas S y M en los *Alfa-CoV* se predijeron un total de 3.413 y 208 *N*-glicosilaciones, respectivamente (con un promedio de 27,7 sitios predichos por secuencia para la proteína S y 1,7 sitios por secuencia de la proteína M). Para el caso de la proteína E, sólo se pudieron predecir sitios por debajo del valor umbral, por lo que no se consideraron predicciones válidas. Dentro de los *Alfa-CoV,* aquellas secuencias con mayor cantidad de glicosilaciones predichas corresponden a ortólogos provenientes de los tipos 229E-CoV y NL63-CoV, ambos capaces de causar un cuadro de resfriado común sólo en humanos. La cuantificación total de glicosilaciones predichas para los *Alfa-CoV* se puede visualizar en la Figura N°8, en donde se aprecian que las secuencias de 229E y NL63 poseen un promedio de 27,1 y 32 glicosilaciones por secuencia, respectivamente. Esto contrasta con otros tipos de *Alfa-CoV*, generalmente asociados a animales no humanos (tipos 229E-Camello, PEDV y HKU2-CoV), en donde

el promedio general de glicosilaciones es de 21,4 predicciones por secuencia. El promedio de predicciones por secuencia para ortólogos de los tipos 229E-Camello, PEDV y HKU2-CoV fue de 23, 22,1 y 19,9, respectivamente. En el caso de las secuencias de M, las glicosilaciones predichas variaron de 1 a 2 por secuencia, obteniendo un promedio de 1,6 predicciones por secuencia.

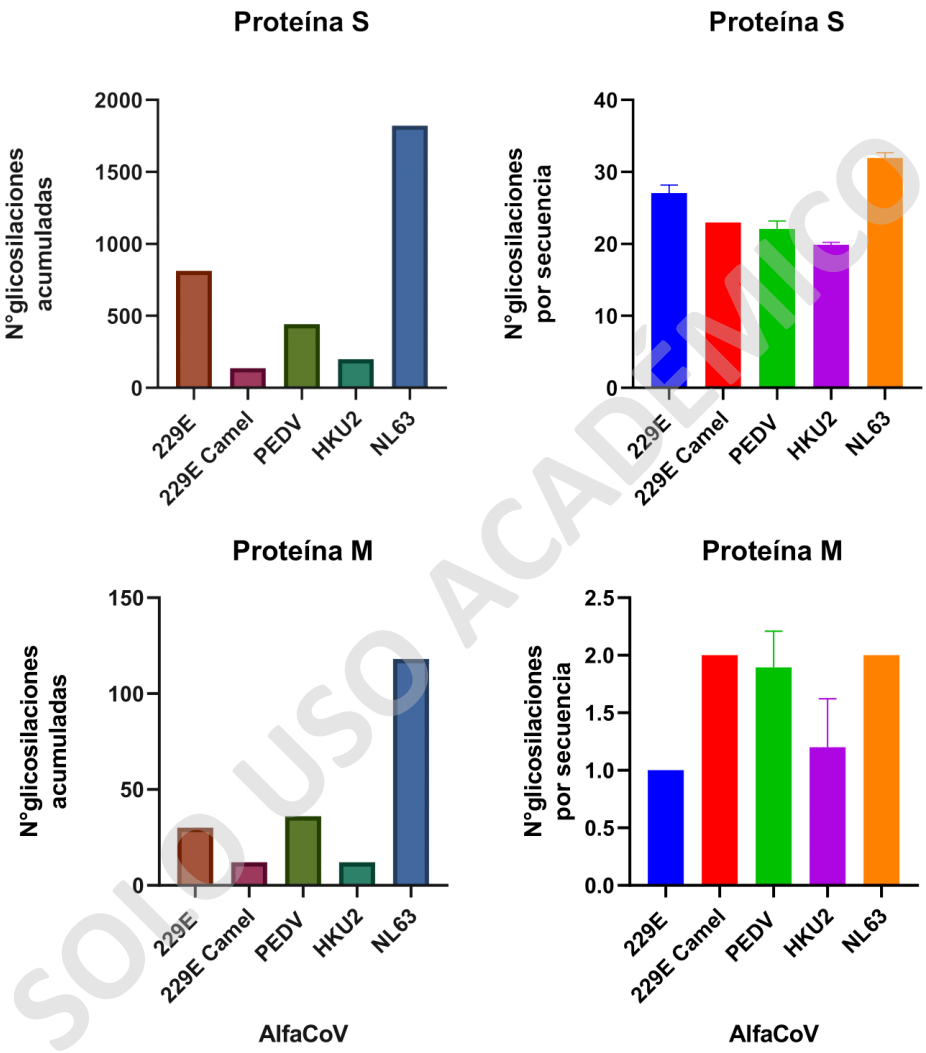

Figura 17. Cuantificación de *N*-glicosilaciones *Alfa-CoV*.

N-glicosilaciones cuantificadas en los *Alfa-CoV* para el total de secuencias y por cada secuencia para la proteína S y M, en donde 5 tipos de CoVs son aquellos que componen este género para este estudio. Siendo NL63 y 229E aislados pertenecientes a humanos, en contraste con 229E-Camel, PEDV y HKU2 que lo son para animales (Camello, cerdo y murciélago, respectivamente).

Para el caso de los *Beta-CoV,* se predijeron un total de 9.824 y 438 *N*-glicosilaciones para S y M (con un promedio de 17,1 sitios predichos por secuencia para la proteína S y 0,9 sitios por secuencia de

la proteína M). Al igual que los *Alfa-CoV,* las predicciones de glicosilación para la proteína E no alcanzaron valores confiables para predecir una glicosilación. La cuantificación de las predicciones para estos tipos de CoVs se puede visualizar en la Figura N°9. Para los tipos de CoV severos, se obtuvo un promedio de 17,8 *N*-glicosilaciones por secuencia. En contraste con S, la proteína M posee en general 1 glicosilación predicha por secuencia para los 16 tipos de *Beta-CoV.*

Para la búsqueda de diferencias significativas entre cada CoV y su número de glicosilaciones por secuencia, se dispuse del *Kruskal-Wallis test* (mencionada anteriormente en la sección metodología). Dado la extensión del número de comparaciones entre cada grupo, las diferencias significativas pueden ser observadas en las tablas resumen para cada tipo de glicosilación y la respectiva proteína estructural. En materiales suplementarios se pueden visualizar las Tablas SN°5-10.

En cuanto a las *O*-glicosilaciones predichas para los ortólogos de la proteína S, el promedio de detecciones en *Alfa-CoV* mostradas en la Figura N°10 (274 glicosilaciones predichas totales; 2,2 predicciones por secuencia) es menor al promedio de predicciones vistas en los *Beta-CoV* (2.243 predicciones totales y un promedio de 4,0 predicciones por secuencia, Figura N°10). Dentro de los *Alfa-CoV,* los tipos HKU2-CoV y NL63-CoV, poseen un número relativamente alto de glicosilaciones por secuencia (4 y 3, respectivamente). Por su parte, entre los *Beta-CoV*, algunos grupos alcanzaron valores inusualmente altos de predicciones por secuencia, por ejemplo, el grupo de los MHV (Figura N°10) alcanzaron un máximo promedio de 16 predicciones por secuencia.

Para los ortólogos de la proteína M en los *Alfa-CoV* se predijeron un total de 15, las cuales se presentaron en HKU2-CoV y NL63-CoV (Figura N°9), obteniendo un promedio por secuencia de 1,6 y 1, respectivamente. En cambio, para los *Beta-CoV* (Figura N°10) se obtuvo un total de 594, en donde se presentaron un promedio de 3,2 glicosilaciones por secuencia, siendo OC43-CoV, Bovino-CoV, HKU23-CoV, PHEV y MHV aquellos con el número más alto de glicosilaciones por monómero, es decir, 4.

Al realizar una comparación entre tipos de CoVs presentes en humanos y una segunda para animales, los resultados (Figura N°11) muestran que se presentan diferencias significativas para la cantidad de *N*-glicosilaciones entre los tipos humanos (*p*: 0,0325), en donde la cantidad de glicosilaciones varían en menor grado. En cambio, al comparar los tipos humanos se encontraron diferencias significativas entre los grupos de murciélago y rumiantes (*p*: 0,0001), al igual que entre murciélagos y murinos (*p*: 0,0081).

Ahora bien, en lo que respecta la comparación entre la cantidad de *N*-glicosilaciones entre tipos de CoVs en humanos y animales (Figura N°12) sugieren que la cantidad de estas MPTs en humanos

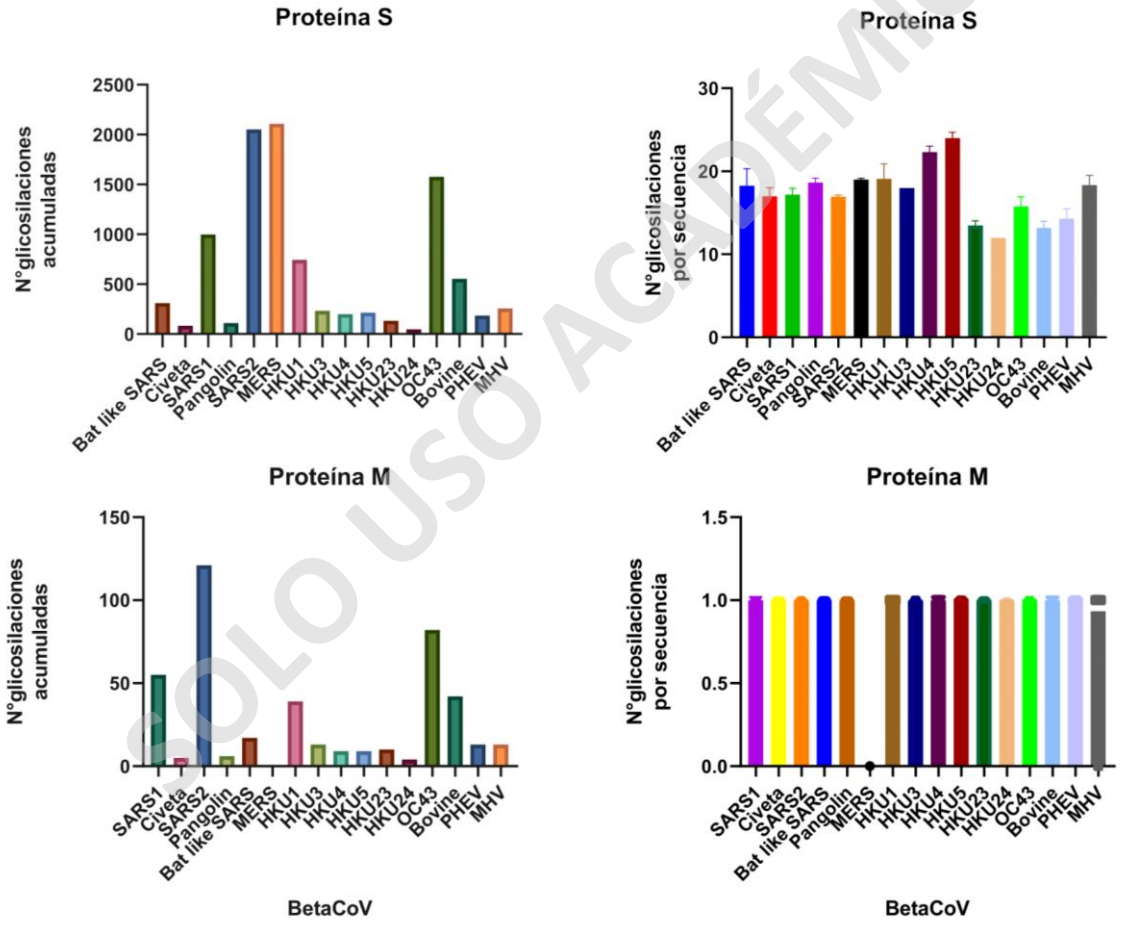

Figura 18. Cuantificación de *N*-glicosilaciones *Beta-CoV*.

N-glicosilaciones cuantificadas en los *Beta-CoV* para el total de secuencias y por cada secuencia para la proteína S y M, en donde son 14 tipos de CoVs que componen este género. Destacándose los severos (SARS-1, 2 y MERS) y de cuadros leves (OC43 y HKU1). Los demás corresponden a civeta, pangolín, murciélagos, camello, rata, bovino, cerdo y ratón (SARSr, HKU3, 4, 5, 23, 24, bovino, PHEV y MHV) todos aislados de animales.

poseen en promedio 19,6 versus las 17,5 en animales, encontrando diferencias significativas entre ambos grupos (*p*: 0,0016).

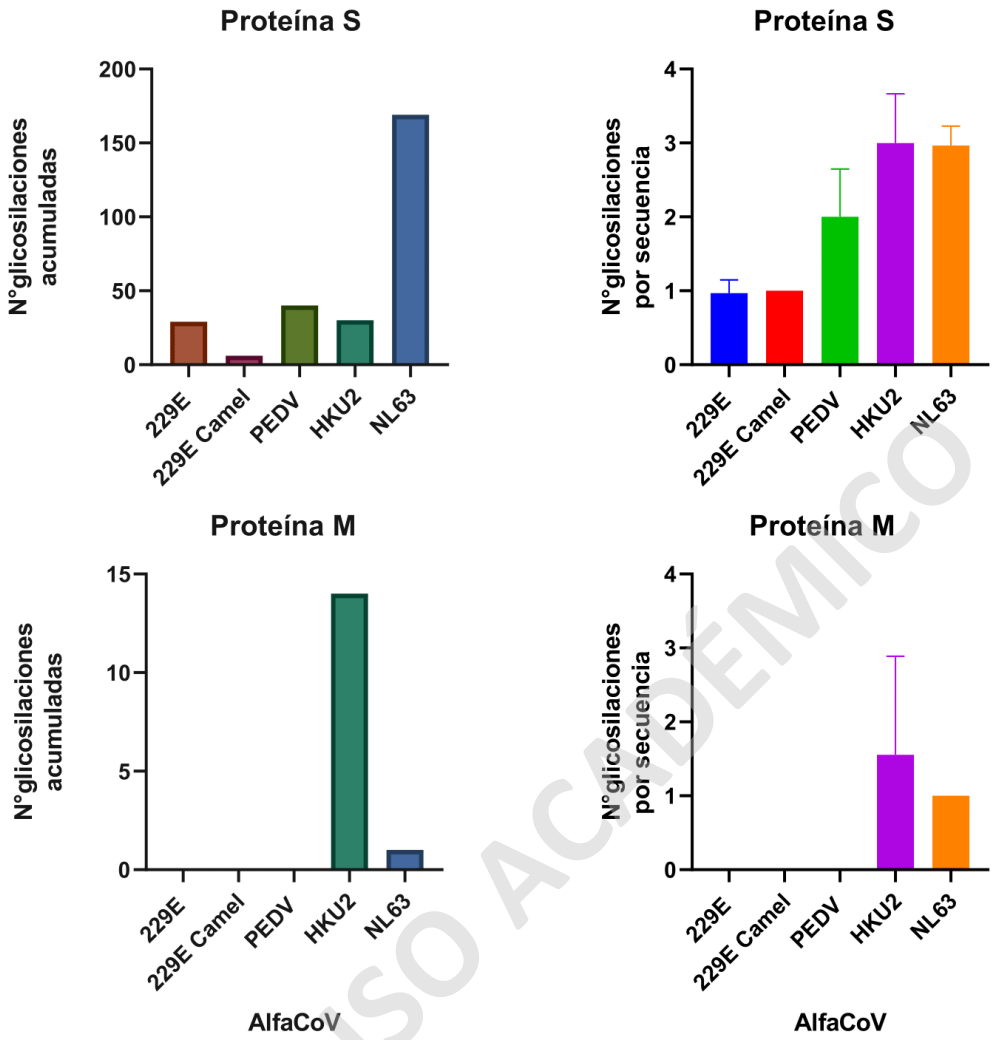

Figura 21. Cuantificación de *O*-glicosilaciones *Alfa-CoV*.

O-glicosilaciones cuantificadas en los *Alfa-CoV* para el total de secuencias y por cada una para las proteínas S y M, en donde 5 tipos de CoVs son aquellos que componen este género para este estudio. Siendo NL63 y 229E aislados pertenecientes a humanos, en contraste con 229E-Camel, PEDV y HKU2 que lo son para animales (camello, cerdo y murciélago, respectivamente).

Las Tablas N°6, 7, 8 y 9 resumen los valores estadísticos descriptivos obtenidos para cada tipo de CoV, según el tipo de glicosilación; mostrando la mediana, el promedio y a la desviación estándar.

En lo que respecta a la variación de las glicosilaciones, tanto *N*- como *O*-, por secuencia entre ambos géneros (Figura N°13) se puede visualizar que el número de *N*-glicosilaciones para las proteínas S y M, varían en mayor medida para el caso de *Alfa-CoV.* En contraste, las *O*-glicosilaciones varían en

mayor medida en los *Beta-CoV*, alcanzado cuantificaciones que llegan a 8 MPTs de este tipo en S y 4 en M, a diferencia de proteínas estructurales en *Alfa-CoV*, con un máximo de 4 en S y 3 en M.

Además, al comparar las *O*-glicosilaciones entre humanos y animales (Figura N°12), los resultados muestran diferencias significativas (*p*: 0,0001) y que este tipo de MPTs se presentan en mayor cantidad en aislados animales con un promedio de 4,1 versus 2,2 en humanos.

Tabla 6. Resumen estadístico de *N*-glicosilaciones de *Alfa-CoV*.

Resumen general de la cuantificación de *N*-glicosilaciones en proteínas estructurales presentes en aislados de *Alfa-CoV*.

PS	Estadísticas	229E	229E Camel	PEDV	HKU2	NL63
S	Mediana	27	23	22,5	20	32
	Promedio	27,1	23	22,1	19,9	32
	SD	1,1	0,0	1,1	0,3	0,7
M	Mediana	1,0	2,0	2,0	1,0	2,0
	Promedio	1,0	2,0	1,9	1,2	2,0
	SD	0,0	0,0	0,1	0,1	0,0

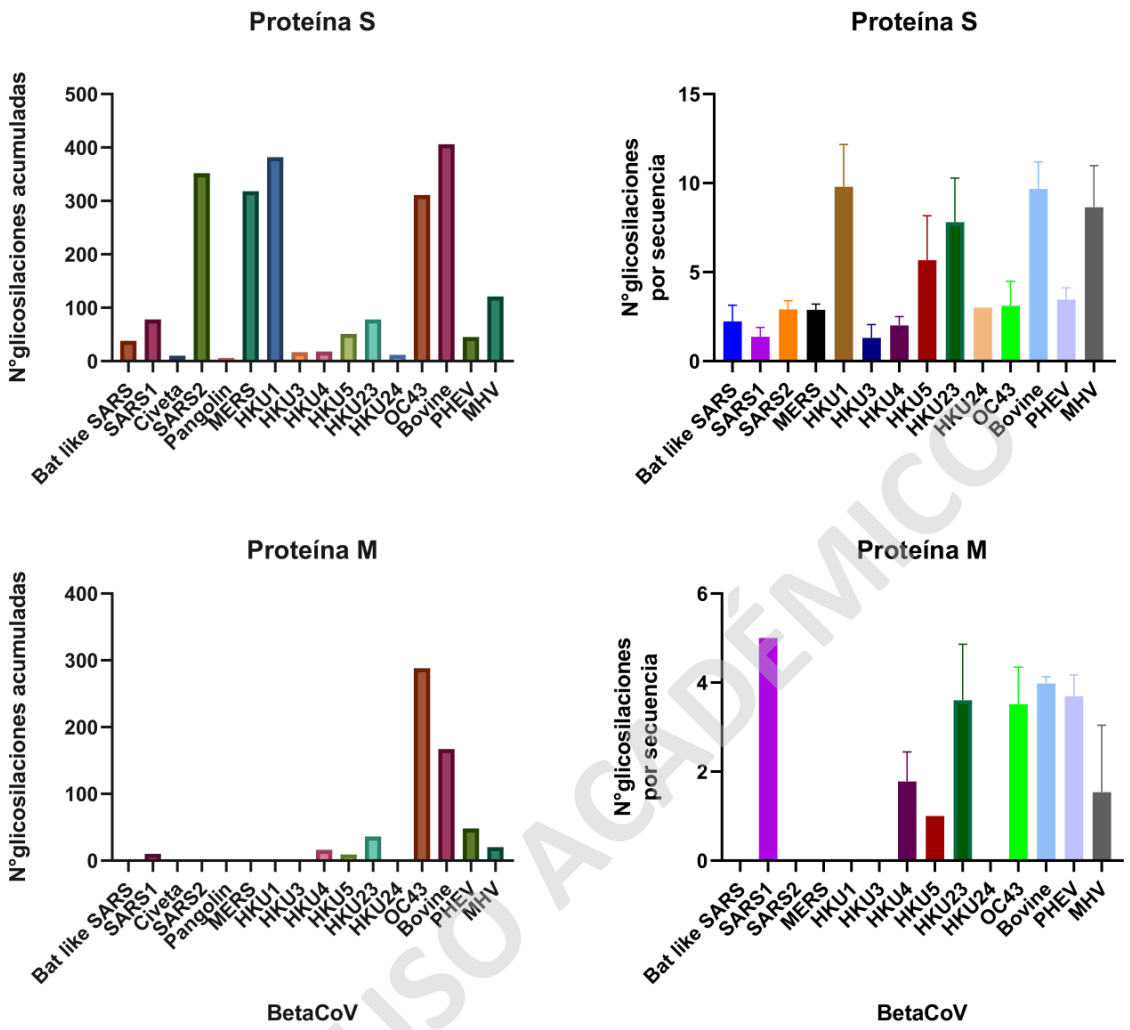

Figura 24. Cuantificación de *O*-glicosilaciones en *Beta-CoV*.

O-glicosilaciones cuantificadas de los *Beta-CoV* para el total de secuencias y por cada una para las proteínas S y M, en donde son 14 tipos de CoVs que componen este género. Destacándose los severos (SARS-1, 2 y MERS) y de cuadros leves (OC43 y HKU1). Los demás corresponden a civeta, pangolín, murciélagos, camello, rata, bovino, cerdo y ratón (SARSr, HKU3, 4, 5, 23, 24, bovino, PHEV y MHV) todos aislados de animales.

Tabla 7. Resumen estadístico de *N*-glicosilaciones en *Beta-CoV*.

Resumen general de la cuantificación de *N*-glicosilaciones en proteínas estructurales presentes en aislados de *Beta-CoV*.

PS	Estadísticas	SARSr-bat	Civeta	SARS 1	Pangolín	SARS 2	MERS	HKU1	HKU3	HKU4	HKU5
S	Mediana	19	17	17	19	17	19	18	18	22	24
	Promedio	18	17	17,2	18,7	17	19	19,1	18	22,3	24
	SD	2,0	1,0	0,8	0,5	0,2	0,2	2,0	0,0	0,7	0,7
M	Mediana	1,0	1,0	1,0	1,0	1,0	0,0	1,0	1,0	1,0	1,0
	Promedio	1,0	1,0	1,0	1,0	1,0	0,0	1,0	1,0	1,0	1,0
	SD	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0

PS	Estadísticas	HKU23	HKU24	OC43	Bovino	PHEV	MHV
S	Mediana	13,5	12	15,8	13,2	14,3	18,4
	Promedio	1,0	2,0	1,3	1,0	1,7	0,9
	SD	0,5	0,0	1,2	0,8	1,2	1,2
M	Mediana	1,0	1,0	1,0	1,0	1,0	1,0
	Promedio	1,0	1,0	1,0	1,0	1,0	0,9
	SD	0,0	0,0	0,0	0,0	0,0	0,3

A fin de poder indagar en la composición consenso de los sitios de *N*- y *O*-glicosilación, se procedió a la generación de logos de secuencias para visualizar el grado de preferencias de aminoácidos en los *sequons* y su entorno directo, reflejado en un *motif* (pequeña región aminoacídica compartida entre diferentes proteínas) de 4 residuos. Para las *N*-glicosilaciones de las proteínas S y M, en la Figura N°14 y 15, respectivamente, se detectó que para los grupos de los diferentes ortólogos de S los *sequons* tienden a variar en el *motif* alrededor de la glicosilación, en contraste con M donde tiende a ser conservado. Por ejemplo, para la segunda posición los ortólogos presentan preferencia por aminoácidos hidrofóbicos, siendo el grupo de los tipos severos aquel con mayor número de hidrofóbicos. De una manera más específica valina (*V*), isoleucina (*I*) y leucina (*L*) son aquellos con

mayor presencia en esta posición. La glicina (*G*) y la cisteína (*C*) son dos aminoácidos que igualmente se encuentran enriquecidos en la segunda posición, particularmente para los grupos de CoVs severos, de resfrío común y murinos (Figura N°15 a, c y e). Contrariamente, aminoácidos empobrecidos en esta posición son el ácido aspártico (*D*), asparagina (*N*) y glutamina (*Q*) aquellos con menor presencia. Siendo *D* el único con cadena cargada negativamente, para el caso de *N* y *Q*, ambos son polares sin carga en la cadena lateral.

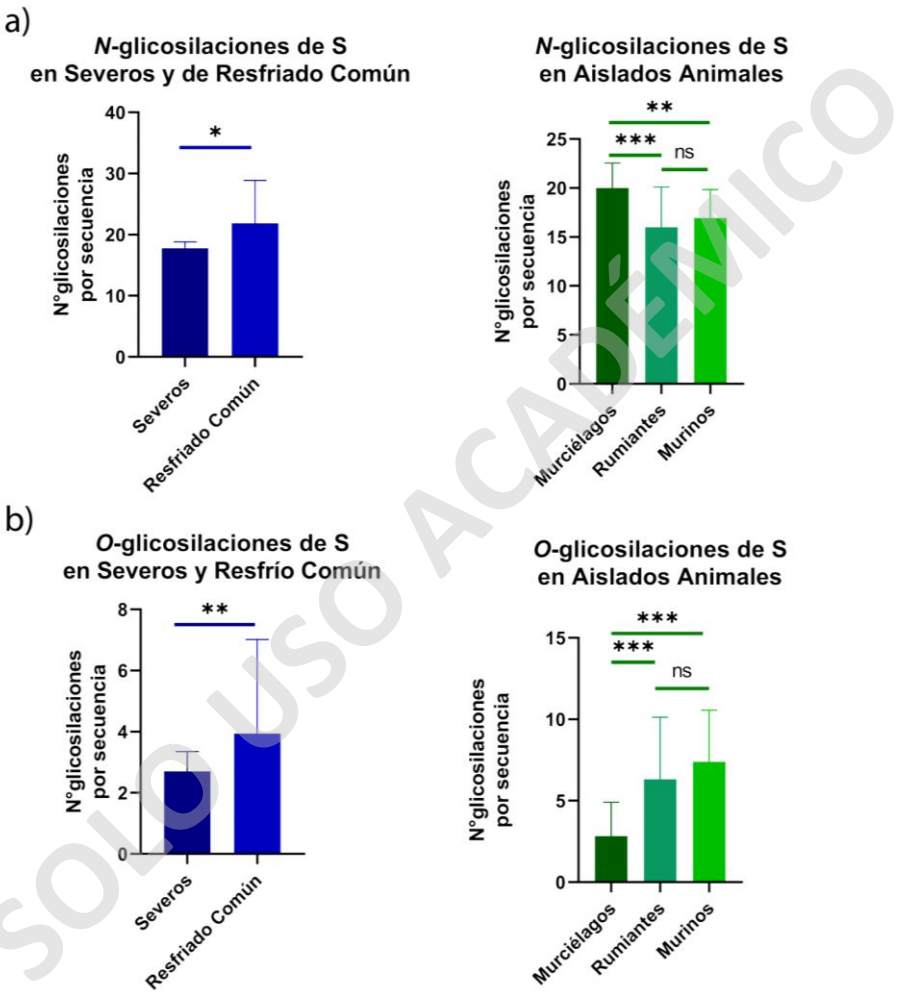

Figura N°11. Comparación en el número de glicosilaciones para diferentes tipos de asilados.

Cuantificación del número de glicosilaciones por secuencias para tipos de CoVs que circulan en humanos (severos y de resfrío común, colores azules) y en animales (murciélagos, rumiantes y murinos, colores verdes), donde en a) se muestran las comparaciones del tipo *N*-glicosilación y en b) las *O*-glicosilaciones. Los análisis estadísticos para identificar diferencias estadísticas entre grupos fueron *Mann-Whitney test* (humanos) y *Kruskal-Wallis test* (animales). Para identificar diferencias estadísticas significativas se fijó un valor $p < 0,05$ (*: significativo; ns: no significativo).

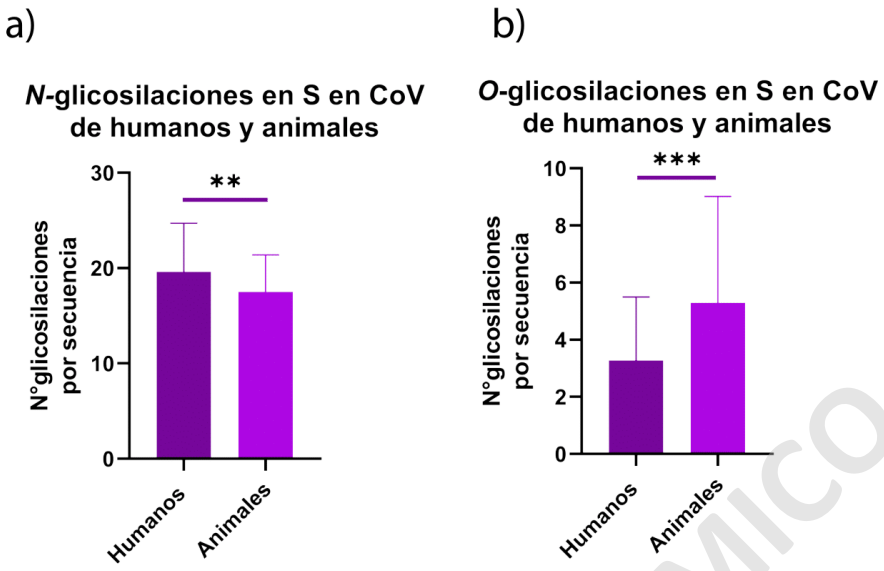

Figura N°12. Comparación en el número de glicosilaciones para diferentes tipos de CoVs según hospedero.

Cuantificación del número de glicosilaciones por secuencias para tipos de CoVs que circulan en humanos y en animales (murciélagos, rumiantes, civeta, pangolín y murinos), donde en a) se muestran las *N*-glicosilaciones y en b) las *O*-glicosilaciones. El análisis estadístico para identificar diferencias estadísticas entre grupos fue *Mann-Whitney test*. Para identificar diferencias estadísticas significativas se fijó un valor $p < 0,05$ (*: significativo; ns: no significativo).

A diferencia de la segunda posición en S, la tercera posee una fuerte preferencia por treonina (*T*) por sobre la serina (*S*), sin embargo, esto no se cumple para M (Figura N°15), donde aislados de CoVs de resfriado común, rumiantes y murciélagos, evidencian enriquecimiento en *Ser*. La cuarta posición muestra estar enriquecida principalmente por aminoácidos hidrofóbicos *I*, *V* y *L*. El aminoácido empobrecido en esta posición compartido por la mayoría de los grupos analizados es *C*.

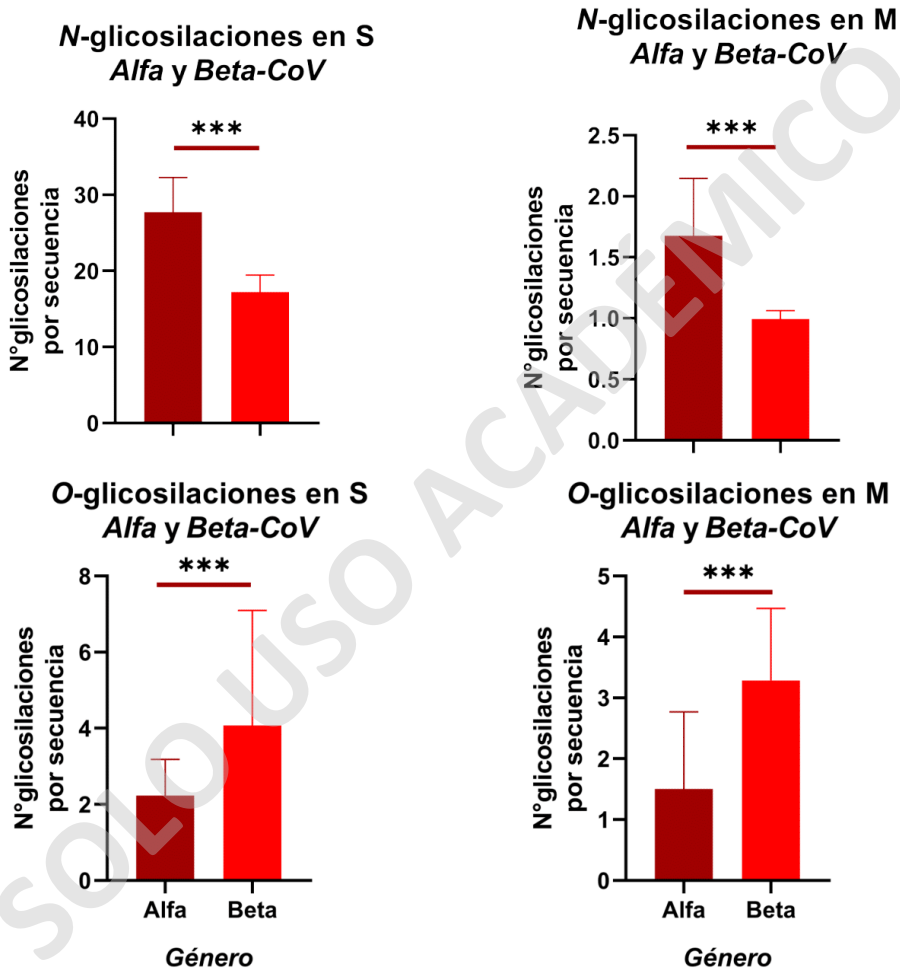

Figura N°13. Cuantificación de glicosilaciones basado en comparación entre *Alfa* y *Beta-CoV*.

Cuantificación del número de glicosilaciones por secuencia, para CoV provenientes de los géneros *Alfa* y *Beta-CoV*. Se utilizó el *Mann-Whitney test* para el análisis estadístico de los datos, considerando un valor $p < 0,05$ como límite de significancia estadística (*: significativo; ns: no significativo).

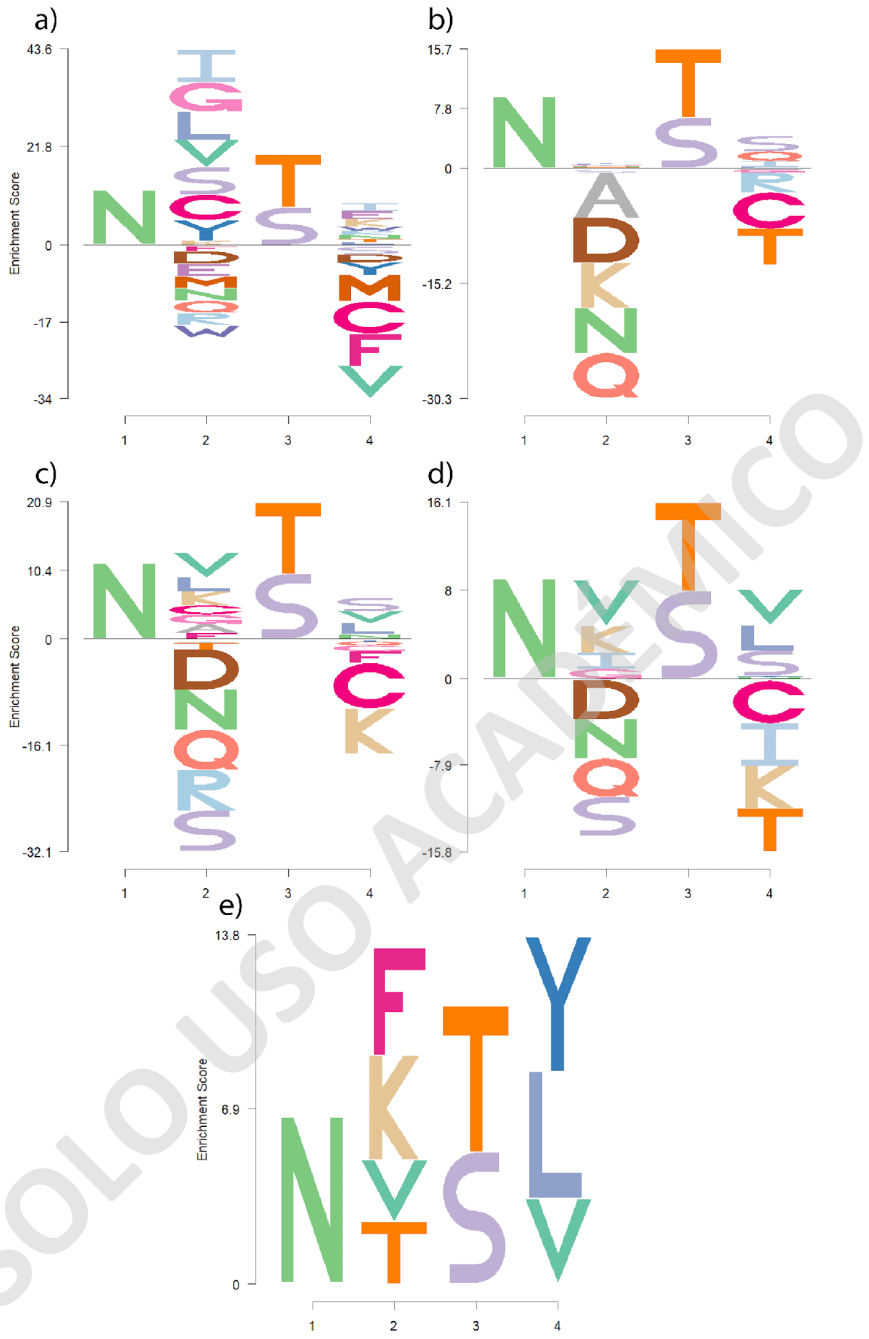

Figura N°14. EDlogos de *sequons* predichos para *N*-glicosilaciones en S.

EDlogo para los *sequons* de las *N*-glicosilaciones predichas para la proteína S de diferentes CoVs severos (a), asilados de murciélagos (b), aislados de cuadros "comunes" (c), aislados de rumiantes (d) y asilados de murinos (e).

Tabla 8. Resumen estadístico de *O*-glicosilaciones en *Alfa-CoV*.

Resumen general de la cuantificación de O-glicosilaciones en proteínas estructurales presentes en aislados de Alfa-CoV.

PS	Estadísticas	229E	229E Camel	PEDV	HKU2	NL63
S	Mediana	1,0	1,0	2,0	3,0	3,0
	Promedio	0,9	1,0	2,0	3,0	2,9
	SD	0,2	0,0	0,6	0,7	0,3
M	Mediana	0,0	0,0	0,0	2,0	1,0
	Promedio	0,0	0,0	0,0	1,6	1,0
	SD	0,0	0,0	0,0	0,4	0,0

Por otro lado, como bien se mencionó anteriormente para la proteína M el *sequon* de las *N*-glicosilaciones es bastante conservado entre grupos. La conservación de estos aminoácidos se puede evidenciar principalmente por la presencia de fenilalanina (*F*) y *G* presente en los 5 grupos analizados, en particular para la segunda posición. La tercera posición en el grupo severos y murinos (Figura N°15 a y e) son particularmente las únicas que presentan solamente un aminoácido para todo el grupo, es decir, *S* o *T*. Por último, la cuarta posición muestra un enriquecimiento de aminoácidos hidrofóbicos como, por ejemplo, *I, L* y *W.*

Tabla 9. Resumen estadístico de *O*-glicosilaciones en *Beta-CoV*.

Resumen general de la cuantificación de *O*-glicosilaciones en proteínas estructurales presentes en aislados de *Beta-CoV*.

PS	Estadísticas	SARSr-bat	Civeta	SARS1	Pangolín	SARS2	MERS	HKU1	HKU3	HKU4	HKU5
S	Mediana	2,0	2,0	1,0	1,0	3,0	3,0	9,0	1,0	2,0	5,0
	Promedio	2,2	2,0	1,4	1,0	2,9	2,9	9,8	1,3	2,0	5,7
	SD	0,9	0,0	0,5	0,0	0,5	0,3	2,4	0,8	0,5	2,5
M	Mediana	0,0	0,0	5,0	1,0	0,0	0,0	0,0	0,0	2,0	1,0
	Promedio	0,0	0,0	5,0	1,0	0,0	0,0	0,0	0,0	1,8	1,0
	SD	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,7	0,0

PS	Estadísticas	HKU23	HKU24	OC43	Bovino	PHEV	MHV
S	Mediana	9,0	3,0	3,0	9,5	3,0	8,0
	Promedio	7,8	3,0	3,1	9,7	3,5	8,6
	SD	2,5	0,0	1,4	1,5	0,7	2,3
M	Mediana	4,0	0,0	4,0	4,0	4,0	1,0
	Promedio	3,6	0,0	3,5	3,9	3,7	1,5
	SD	1,3	0,0	0,8	0,2	0,5	1,5

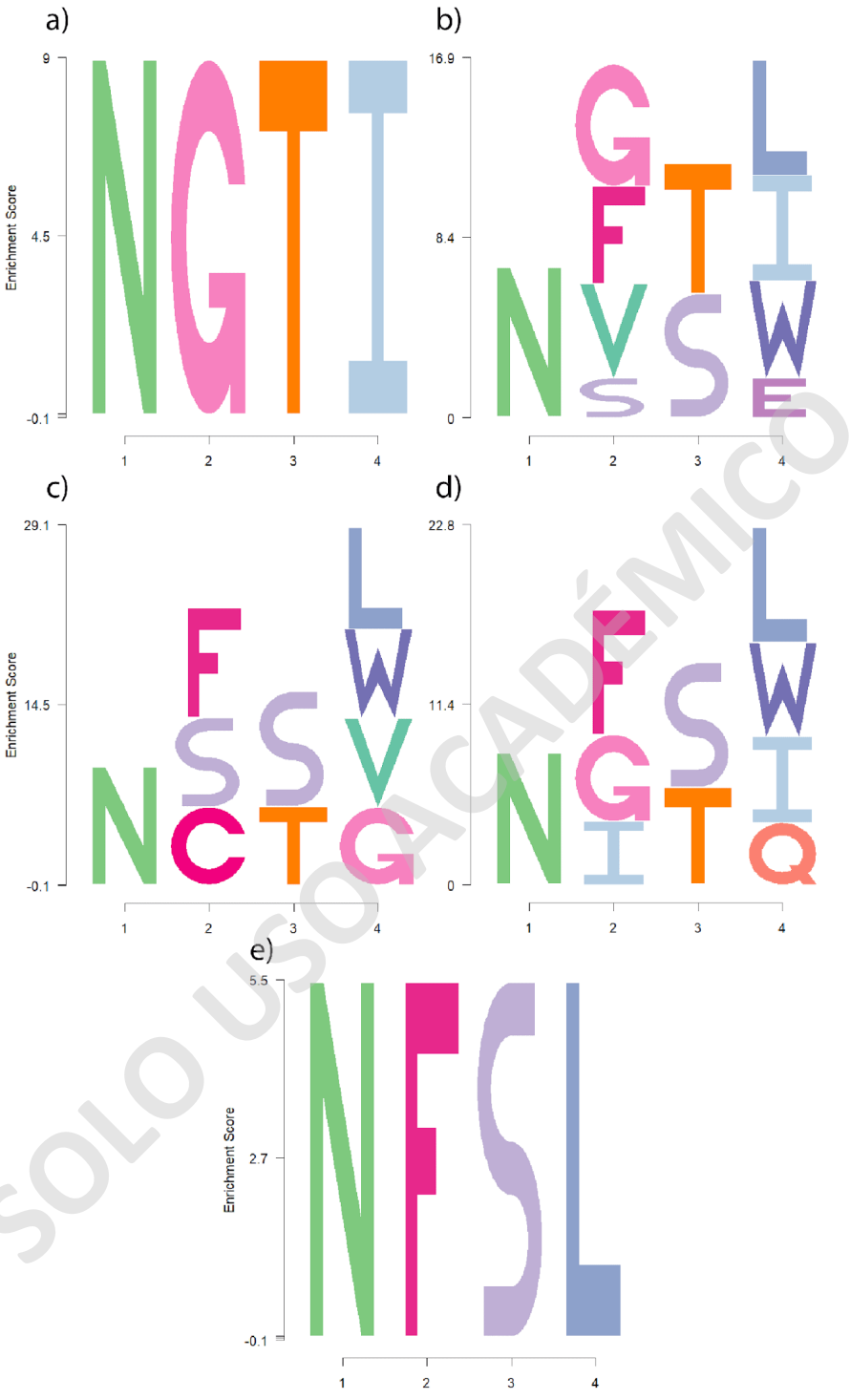

Figura N°15. EDlogos de *sequons* predichos para *N*-glicosilaciones en M.

EDlogo para los *sequons* de las *N*-glicosilaciones predichas para la proteína M de diferentes CoVs severos (a), aislados de murciélagos (b), aislados de cuadros "comunes" (c), aislados de rumiantes (d) y asilados de murinos (e).

Las *O*-glicosilaciones se pueden visualizar en la Figura N°16 principalmente para S, en cambio, el único grupo en el cual se predijeron *O*-glicosilaciones para la proteína M fue el de rumiantes, mostrada en la Figura SN°5. Las *O*-glicosilaciones al tener un *sequon* menos restrictivo, con una secuencia consenso de S/T, tienen un amplio rango de aminoácidos presentes en el *motif*, sin embargo, una característica particular es que hay una presencia enriquecida de *S* y *T* alrededor del mismo *sequon,* lo cual se cumple para todos los grupos analizados. De manera general los grupos de CoVs severos y de murciélagos sugieren no tener un *motif* particularmente conservado entre ellos, en cambio, al comparar los grupos de resfriado común, murinos y rumiantes se presentan aminoácidos enriquecidos comunes entre ellos, principalmente *G* y prolina (*P*), mostrado en la Figura N°16 d y e. Por otro lado, histidina (*H*) es el aminoácido más empobrecido en el *motif* alrededor del *sequon*, le sigue el *W* y *Q*, sin embargo, estos no son ubicuamente empobrecidos en todos los grupos, por ejemplo, en el grupo de severos uno de ellos (*Q*) se encuentran incluso enriquecidos en el *motif* (Figura N°16 a).

Como se mencionó anteriormente la proteína M en aislados de rumiantes (Figura SN°5) se encuentra particularmente enriquecida de *S*, *T* y en menor medida de *P*. En cambio, una particularidad del *motif* es que se encuentra empobrecidos de aminoácidos hidrofóbicos, en este caso de alanina (*A*) y metionina (*M*).

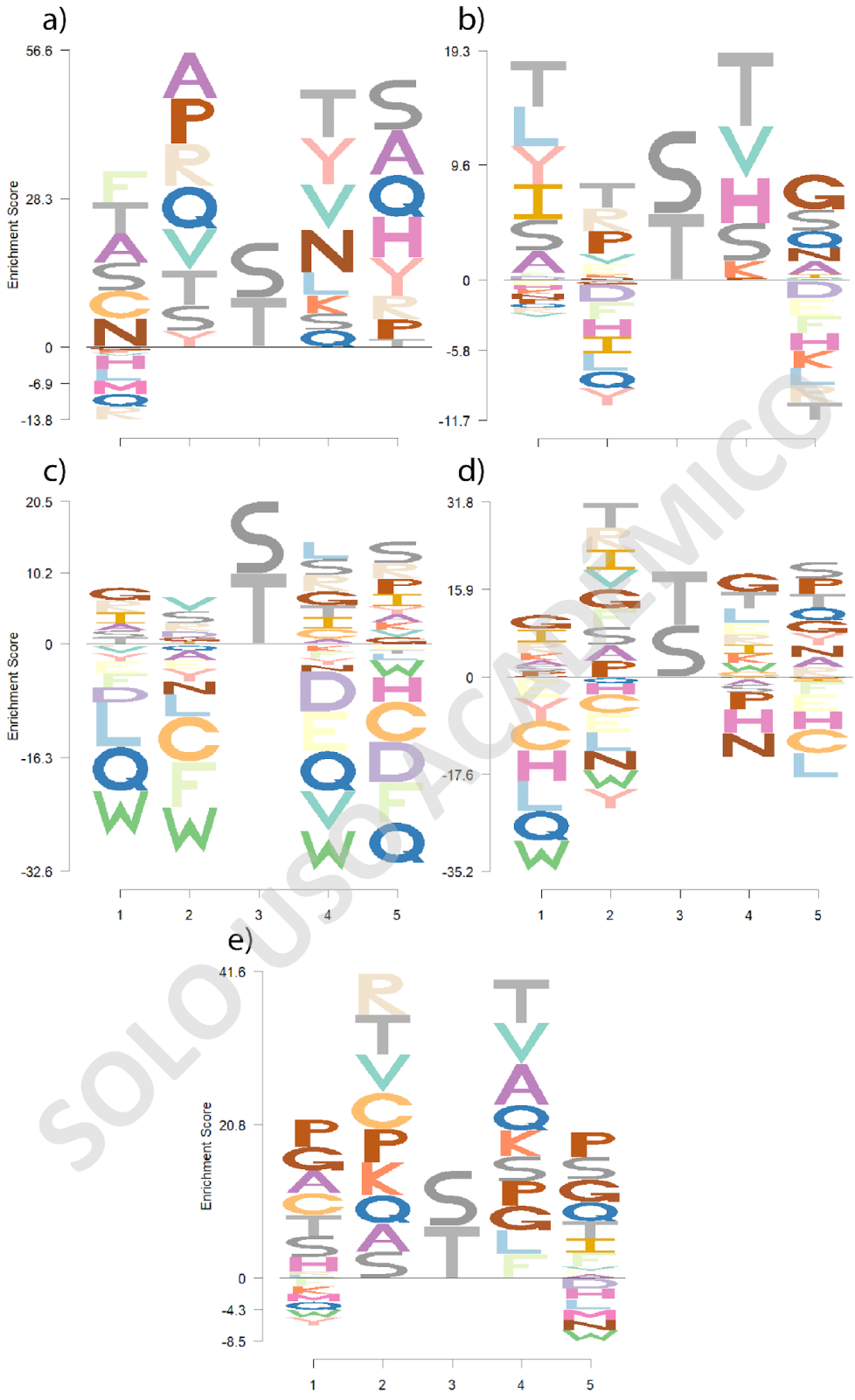

Figura N°16. EDlogo de *sequons* predichos para *O*-glicosilaciones en S.

EDlogo para los *sequons* de las *O*-glicosilaciones predichas para la proteína S de diferentes CoVs severos (a), asilados de murciélagos (b), aislados de cuadros "comunes o leves" (c), aislados de rumiantes (d) y asilados de murinos (e).

4.4 Selección a nivel de sitios a nivel de aminoácidos y su relación con sitios consenso de glicosilación en grupos relacionados de CoVs

Con el fin de evaluar el tipo de selección específica en los sitios asociados a glicosilación, se realizó el análisis de selección a lo largo de la secuencia de codones, para las proteínas S y M, dentro de diferentes grupos de CoVs, utilizando las herramientas FEL y FUBAR del servidor *Datamonkey*. Estas dos herramientas utilizan diferentes enfoques para detectar aquellos sitios dentro de las proteínas que posean selección positiva o negativa, utilizando un MSA dado. Además, la información de estos análisis fue comparados con el mapeo consenso de sitios de *N*- y *O*-glicosilación para el mismo grupo de datos utilizado.

En general, entre ambos métodos se identificaron un total de 11 sitios bajo selección positiva, siendo FUBAR aquel con más sitios identificados. La lista de la cantidad de sitios identificados como positivos y negativos es resumida en la Tabla N°10. Además, las glicosilaciones consenso reportadas se agrupan principalmente en zonas claves de las proteínas (Figura N°17), tales como el sitio de unión al receptor (RBD), o zonas colindantes al sitio de unión de la subunidad 1 y 2 de la proteína S (S1/S2), así como en el dominio N-terminal (NTD) de la proteína M (Figura N°18).

Respecto a la ubicación de los sitios seleccionados positivamente en la proteína S, para el caso del grupo de *Beta-CoV* severos, se reportaron 4 sitios con selección positiva cercanos al RBD y uno asociado a la *Heptad Repeat Regions* (HR), correspondiendo a las regiones entre los sitios 300 y 550, y entre los sitios 950 y 1000, respectivamente. Cercana a cada zona del RBD, se detectaron al menos dos *N*-glicosilaciones consenso y una *O*-glicosilación.

Para el caso de la proteína M, se reportaron dos sitios bajo selección positiva: uno cercano al dominio NTD, el cual se encuentra expuesto en la membrana del virión; y otro localizado en el domino C-terminal (CTD) tal como se observa en la Figura N°18.

Tabla 10. Resumen del tipo de selección predicha.

Resumen selección de sitios en las proteínas estructurales S y M utilizando las herramientas FEL y FUBAR.

Grupos	S		M	
	Positivos/Negativos/Totales		Positivos/Negativo/Totales	
	FEL	FUBAR	FEL	FUBAR
Severos	4/403/1262	1/620/1262	1/106/220	1/106/220
Murciélagos	0/626/1561	1/1100/1561	0/76/230	0/82/230
Comunes	1/402/1584	1/837/1584	1/75/233	0/83/233
Murinos	0/273/1404	0/543/1404	0/23/231	0/38/231
Rumiantes	0/399/1595	0/821/1595	0/62/233	0/66/233

El grupo de CoVs provenientes de murciélagos (Figura SN°6 y 7) corresponde a un grupo de aislados provenientes de diferentes especies de murciélagos, compuesto principalmente por dos subgrupos emparentados filogenéticamente (HKU2 y 3-CoV; HKU4 y 5-CoV) (Figura N°1). Estos aislados se seleccionaron y agruparon debido a su relación con aislados zoonóticos que dieron origen a los aislados severos ya conocidos. En este grupo, la ubicación del RBD presenta cierta variabilidad. La predicción de sitios en la proteína S para este grupo (Figura SN°6) mostró menos dos sitios bajo selección positiva: uno presente en el RBD y el otro en la porción endovirión. Al igual que el grupo de CoVs severos, se detectó un grupo de *N*-glicosilaciones consenso en zonas cercanas al RBD, e incluso se predicen glicosilaciones circundantes a la región endovirión. Para el caso de las *O*-glicosilaciones, se predijeron 3 de ellas cercanas al RBD y dos en el sitio S1/S2. La Figura SN°9 muestra el nivel de selección y glicosilaciones consenso para los sitios, a lo largo de la proteína M provenientes de aislados de murciélagos. Según los resultados de FEL y FUBAR, no se presentaron sitios positivamente seleccionados, detectando incluso una gran proporción de sitios bajo selección negativa. En lo que respecta a las glicosilaciones consenso predichas, es interesante de notar la

presencia de *O*-glicosilaciones en la región NTD de la proteína, la cual se encuentra expuesta al exterior de virión.

En cuanto a los aislados de CoV que causan resfriado común (presentados en las Figuras SN°8 y 9), son el único grupo que presenta un sitio seleccionado predicho como positivo (642) por ambas herramientas (FEL y FUBAR); además, presenta *N*- y *O*-glicosilaciones consenso cercanos al sitio 642. Particularmente, esta posición se encuentra en las cercanías del S1/S2. En cuanto a la proteína M de este tipo de CoVs, se detectó un sitio seleccionado como positivo cercano a un sitio de glicosilación consenso, hallado en la posición 20, dentro del NTD (extravirión). Es interesante mencionar que posterior a este sitio, FEL predice 4 sitios bajo selección negativa (22, 23, 26 y 27).

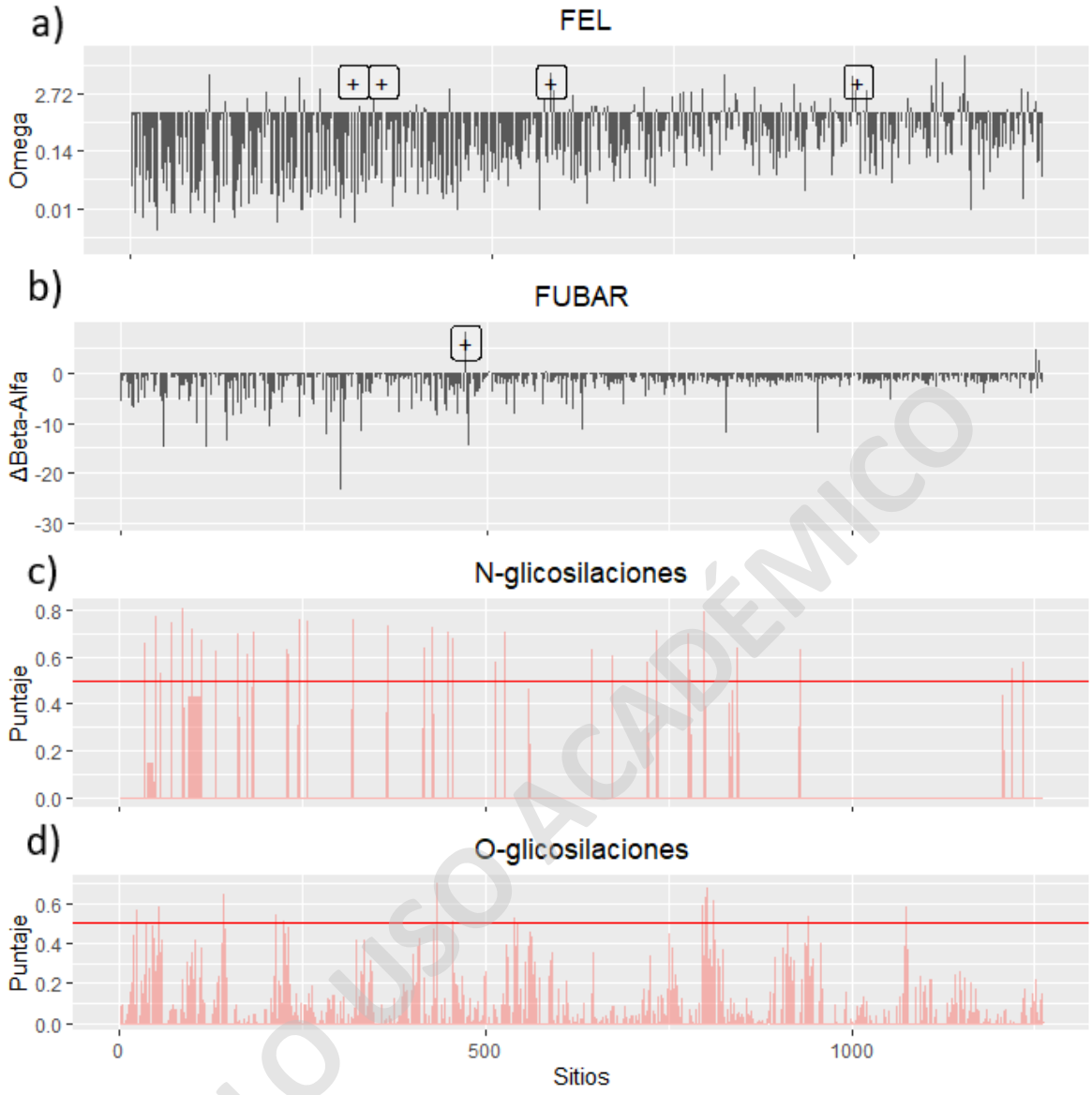

Figura N°17. Sitios seleccionados y glicosilaciones consenso para los CoVs severos en proteína S.

Nivel de selección a nivel de sitios en la proteína S, utilizando FEL y FUBAR para el grupo murciélagos: SARS-CoV-1, SARSr-CoV-1, MERS-CoV, MERSr-CoV, SARS-CoV-2, SARSr-CoV-2 y SARSr-pangolín-CoV. Los sitios reportados como positivos son marcados con signo (+), también se incluye el puntaje alcanzado por Omega en el caso de FEL (a) y la ΔBeta-Alfa reportada por FUBAR (b), para indicar sitios con algún grado de selección positiva o negativa. En la sección inferior se evidencia las glicosilaciones consenso predichas por NetGlyc, utilizando el puntaje mínimo para predecir dicha glicosilación, para las *N*-glicosilaciones (c) y *O*-glicosilaciones (d).

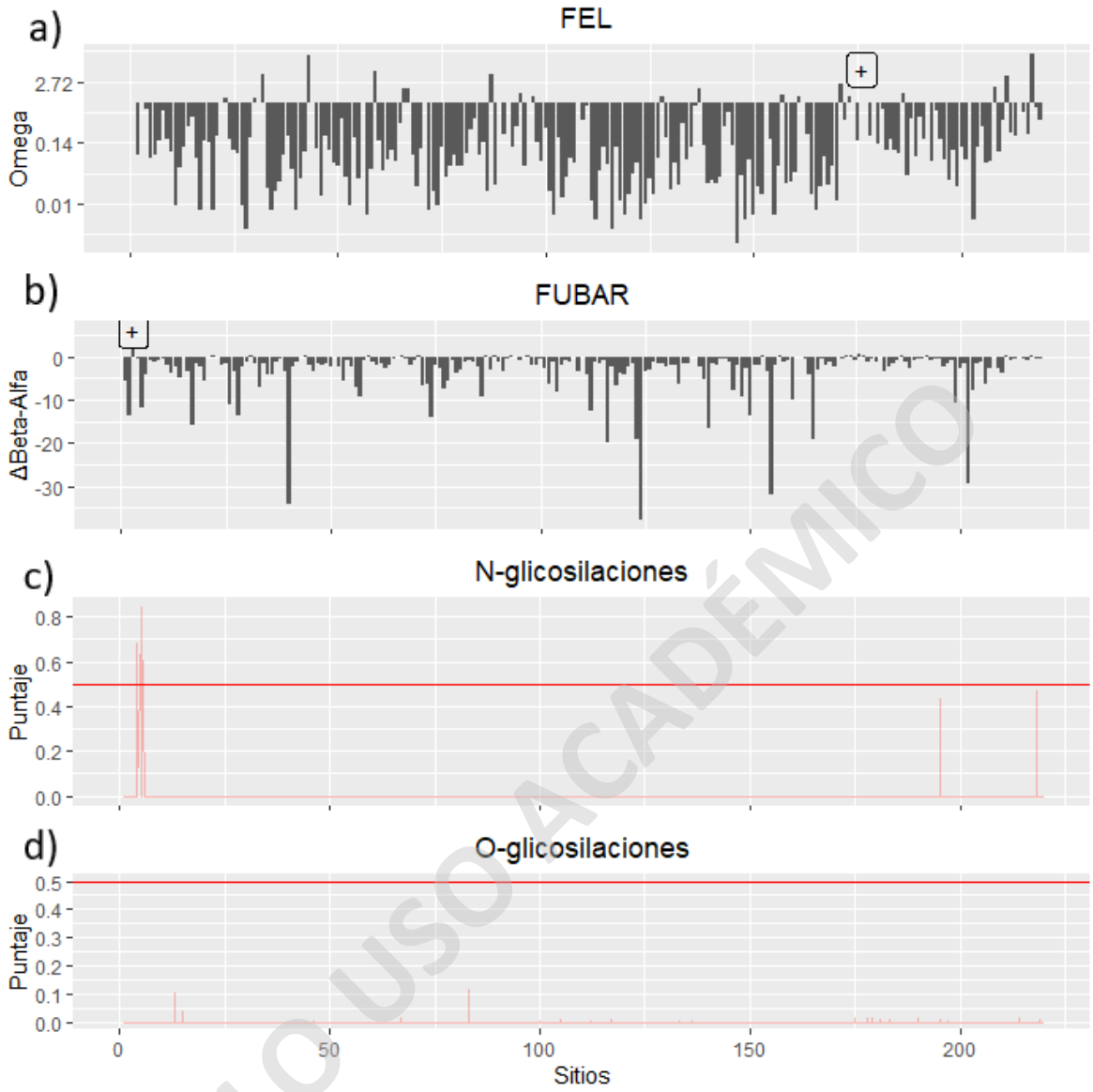

Figura N°18. Sitios seleccionados y glicosilaciones consenso para los CoVs severos en proteína M.

Nivel de selección a nivel de sitios en la proteína M, utilizando FEL y FUBAR para el grupo murciélagos: SARS-CoV-1, SARSr-CoV-1, MERS-CoV, MERSr-CoV, SARS-CoV-2, SARSr-CoV-2 y SARSr-pangolín-CoV. Los sitios reportados como positivos son marcados con signo (+), también se incluye el puntaje alcanzado por Omega en el caso de FEL (a) y la ΔBeta-Alfa reportada por FUBAR (b), para indicar sitios con algún grado de selección positiva o negativa. En la sección inferior se evidencia las glicosilaciones consenso predichas por NetGlyc, utilizando el puntaje mínimo para predecir dicha glicosilación, para las *N*-glicosilaciones (c) y *O*-glicosilaciones (d).

Al cuantificar y comparar las glicosilaciones entre CoVs severos y de resfrío común, evidencian que los causantes de cuadros leves son aquellos que poseen mayor cantidad de glicosilaciones (ambos

tipos). Por otro lado, la comparación entre animales sugiere que el grupo de CoVs pertenecientes a murciélagos son aquellos con mayor cantidad de *N*-glicosilaciones, en contraste con el tipo *O,* que lo son para murinos. Al comparar género, las *N*-glicosilaciones abundan en cantidad en lo que respecta *Alfa-CoV* en la proteína S y M, en contraste con las del tipo *O* para los *Beta-CoV*.

El análisis de preferencia de *sequons* sugieren que el tipo NXT es preferido en la proteína S para todos los tipos de CoVs analizados. Para la proteína M, lo es el tipo NXS, además de sugerir que los *sequons* en esta proteína tienden a ser conservados.

Finalmente, al evaluar el grado de selección a nivel de sitios, se detectaron un total de 11 bajo selección positiva, donde la proteína S en el grupo de CoVs severos es aquella que concentra la mayor cantidad, es decir, cinco. Ahora bien, no se encontró ningún *sequon* bajo selección positiva en ninguna de las dos proteínas en donde se predijeron glicosilaciones (S y M). Sin embargo, al evaluar la distribución de sitios de glicosilación en la proteína S, se pudo evidenciar que siguen la regla *O-follow-N*, es decir, que a las glicosilaciones del tipo *N*- le vienen acompañadas un conjunto de *O*-glicosilaciones, representación mostrada en la Figura N°19.

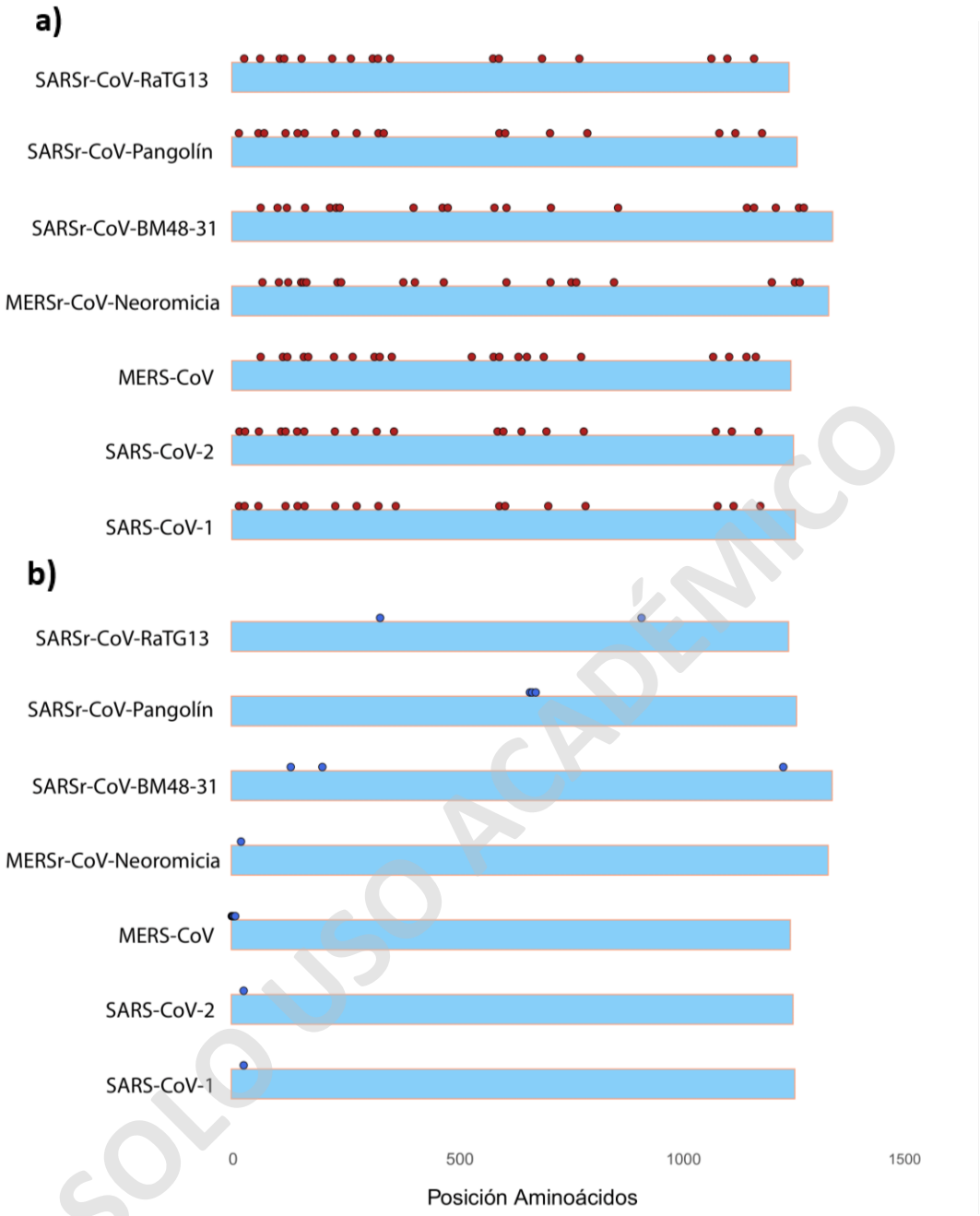

Figura N°19. Representación de proteína S y sus glicosilaciones en CoVs severos.

Representación esquemática de proteína estructural S en el conjunto de CoVs severos (SARS-CoV-1, SARS-CoV-2 y MERS-CoV) y sus relativos zoonóticos (SARSr-RaTG13, SARSr-CoV-BM48-31, SARSr-CoV-Pangolín y MERSr-CoV-Neoromicia). Cada círculo representa la posición del sitio de glicosilación predicho en donde se encuentran las a) *N*-glicosilaciones (círculos rojos) y b) *O*-glicosilaciones (círculos azules).

5. DISCUSIÓN DE RESULTADOS

Debido al potencial pandémico que han demostrado poseer los CoVs, es necesario entender la naturaleza de los cambios observables, a nivel genómico, en los diferentes tipos, así como comprender las dinámicas de zoonosis asociadas a su relación con el ser humano. La actual pandemia causada por el SARS-CoV-2 (4, 5), al igual que las epidemias de las pasadas décadas provocadas por el SARS-CoV-1 y MERS-CoV (6, 10), muestran que los eventos de zoonosis pueden ser aún más recurrentes y que los *Beta-CoV* pueden ser parte de potenciales futuros eventos (75). Este estudio apunta a un enfoque relacionado con el estudio del nivel de selección molecular y su relación con los sitios de glicosilaciones en las proteínas estructurales (S, M y E) en un amplio rango de CoVs.

El análisis filogenético realizado a partir de genomas completos (Figura N°1 y SN°1) sugiere que el alto número de tipos de *Beta-CoV* conocidos al inicio de la pandemia pudo haber sido aún mayor, dado a que se ha intensificado la búsqueda de estos patógenos pertenecientes a la familia *Coronaviridae*. Al momento, hay 6 tipos de CoVs presentes en murciélagos (SARSr-RaTG13, SARSr-BM4831, MERSr-Neoromicia, HKU-3, -4 y -5). El número de *Beta-CoV* descubiertos a lo largo de la pandemia ha aumentado, al igual que el entendimiento de los nichos ecológicos donde se desenvuelven y los procesos de transmisión (25), además del previamente conocido potencial zoonótico que poseen los virus que circulan en murciélagos (76). Los *Alfa-CoV* también tuvieron un aumento de nuevos tipos descubiertos, especialmente en murciélagos (77), pese a que en esta tesis solamente se presentaron 5 tipos diferentes, uno de los cuales es proveniente de murciélagos.

Los resultados de este estudio confirman que, en el grupo de los *Sarbecovirus* (78) (SARS-CoV-1, SARS-CoV-2, SARSr aislados de murciélago, pangolín y civeta) los tipos con sintomatología severa observados en humanos están más relacionadas con aislados virales provenientes de animales, por lo cual estaría asociado a eventos zoonóticos propuestos para los orígenes de cada tipo (sea SARS-CoV-1, SARS-CoV-2 o MERS-CoV) (10, 23, 25, 75). Por ejemplo, se ha reportado previamente que para el tipo SARS-CoV-2, los CoVs más cercanos filogenéticamente serían los tipos SARSr-RaTG13 y SARSr-pangolín (38). Mientras que en filogenias del genoma completo (Figura N°1), SARS-CoV-2 está más relacionado al grupo de SARSr-CoV que incluye a SARSr-RaTG13, que al grupo de SARSr-pangolín.

Por otro lado, la filogenia utilizando la secuencia de S (Figura N°2a) muestra una mayor cercanía entre SARS-CoV-2 y los del SARSr-pangolín (80). Esta incongruencia ha sido reportada por Boni y colaboradores (79), y se ha propuesto que esto es particularmente un producto de eventos de recombinación entre diferentes CoVs (77, 81).

Los *Merbecovirus* (78) (MERS-CoV, MERSr-CoV, HKU-4 y -5), muestran agrupamientos esperados, entre las variantes del tipo severo MERS-CoV, con tipos HKU-4 y -5 (provenientes de murciélago). Esto es esperado, considerando diferentes reportes que sugieren que el virus epidémico del 2012-2013 necesitó de *Quirópteros* como un potencial reservorio intermediario (82). No obstante, los resultados de este estudio contrastan con lo reportado previamente sobre el rol del camello en la aparición del MERS-CoV, dado a que las filogenias de genoma completo y de las proteínas estructurales presentadas acá muestran que el HKU-23-CoV (reportado como cercano al MERS-CoV producto de potenciales eventos de recombinación) (83), no se encuentra directamente emparentado con el MERS-CoV. En contraste, el grupo de los *Embecovirus* (78) (MHV-CoV, HKU1 y 24-CoV), poseen una mayor cercanía filogenética con el HKU23-CoV. Este subgénero es aquel que posee menor grado de incongruencia filogenética a lo largo de las diferentes filogenias presentadas con la única excepción de la proteína S. Esto se puede explicar producto de la gran diversidad de CoVs presente en África (84), en donde el 70% de especies de dromedarios se encuentran en el Este de África (85), propiciando que diferentes géneros de CoVs circulen en estos mamíferos. En cuyo caso, la cercanía filogenética encontrada con *Embecovirus* sugiere ser producto de varios eventos de recombinación, los cuales incluyen CoVs presentes en roedores (85).

Pedacovirus (PEDV-CoV), *Setracovirus* (NL63-CoV), *Duvinacovirus* (229E-CoV) y *Rinacovirus* (HKU2-CoV), componen principalmente a los *Alfa-CoV* analizados en este estudio. Se pudo observar que, entre las filogenias para proteínas estructurales, el árbol de la proteína estructural M presenta incongruencias filogenéticas respecto a los árboles hechos a partir de la información deducida de las proteínas S y E. Esto se evidencia a partir de los agrupamientos de los tipos PEDV-CoV y 2293-CoV, donde dicha incongruencia se podría atribuir a la evolución convergente reportada en la historia evolutiva de estos tipos de CoVs (NL-63-CoV, PEDV-CoV y 229E-CoV) (86).

Los agrupamientos estimados para la filogenia de los tipos HKU1-CoV, MHV y HKU24-CoV muestran resultados esperados según previos reportes, que relacionan a estos *Beta-CoV* con tipos provenientes de roedores. Particularmente, HKU1-CoV no posee un origen zoonótico confirmado;

sin embargo, diferentes estudios han estimado que la cercanía filogenética compartida con MHV y HKU24-CoV serían producto de la circulación de estos últimos en roedores. No obstante, su cercanía con OC43-CoV, BCoVs y PHEV (87), también refuerza la idea de que estos tipos de CoVs han evolucionado en los mamíferos rumiantes y que el proceso zoonótico pudo haber ocurrido aproximadamente hace un siglo atrás (86).

Una vez ya conocidas las relaciones filogenéticas, se realizaron análisis en comparación de secuencias para calcular el grado de selección molecular, en las familias de ortólogos de las proteínas estructurales. El análisis sugirió que la gran mayoría de las comparaciones mostraron bajos valores de radio ω (23, 88). La comparación entre secuencias de los mismos grupos (Figura N°3 y 4) mostró que gran parte de las comparaciones presentan estos bajos valores, algo totalmente esperado debido a que los cambios esperados para secuencias del mismo tipo poseen menor grado de divergencia (89).

En contraste con las comparaciones intragrupos, las comparaciones entre diferentes tipos de CoVs mostró un alto número de valores $\omega > 1$ (Figura N°6 y SN°4); en la práctica esto puede ser un resultado esperado para secuencias divergentes, sin embargo, debe ser tomado con precaución cuando se trata de secuencias virales (89).

Estas potenciales sobreestimaciones de selección positiva se basan, por ejemplo, en que, en algunos casos, aislados virales provenientes de diferentes hospederos poseen distintos efectos del nivel de selección. Como previamente se ha observado en las tasas de sustitución de sitios no sinónimos, lo cuales son significativamente más bajas en arbovirus que en otros virus de ARN (90); también, el uso de muestras no contemporáneas en la comparación podría forzar incorrectamente el supuesto de que todas las secuencias codificantes están bajo las mismas limitaciones de presión selectiva (91). Por lo que sólo deberían ser tomadas en cuenta aquellas secuencias que provengan de poblaciones que coexisten a un mismo tiempo dado (92) como, por ejemplo, aislados obtenidos en una misma población de murciélagos que potencialmente contengan tipos de *Alfa-CoV* y/o *Beta-CoV* en una región en particular. Sin embargo, esto puede ser un problema difícil de resolver para muestras virales, tal como se ha reportado por Kryazhimskiy y colaboradores:

> "Es difícil determinar la escala de tiempo apropiada asociada con un conjunto de datos de secuencias microbianas muestreadas, particularmente para un virus muestreado en diferentes momentos" (89).

Para corregir y/o estabilizar dichas sobreestimaciones es necesario utilizar modelos de sustitución que impongan restricciones al tipo de selección en secuencias codificantes, determinar las escalas de tiempo ideales para comparar muestras virales relacionadas (46, 89) o estimar selección a nivel de secuencias codificantes virales pertenecientes a tipos virales del mismo género (89).

En cuanto a los valores *ω* estimados para las diferentes proteínas estructurales (Figura N°5 y SN°6) los resultados sugieren que la proteína S es aquella con mayor cantidad de estimaciones con un $\omega > 1$. Sin embargo, es interesante notar que M y E poseen estimaciones con un $\omega < 2$, o inclusive, $\omega < 3$. Esto puede ser una señal de selección positiva para estas proteínas estructurales, no obstante, también puede ser resultado de una sobreestimación producto de los fenómenos anteriormente mencionados. Ahora bien, enfocándose en las tasas no sinónimas (*dN*) los resultados (Tabla N°5) sugieren que la mediana en S (0,85) es mayor que la de M y E (0,28 y 0,46, respectivamente). La naturaleza de las mutaciones no sinónimas contribuye a que se produzcan cambios a nivel de aminoácidos, estos cambios contribuyen al constante cambio de residuos en la proteína estructural S, en contraste con M y E. Estos cambios se pueden esperar para la proteína S, dado que es aquella que se encuentra bajo mayor presión selectiva producto del reconocimiento inmune (94). Además, los resultados de análisis para ortólogos de S en los dos géneros analizados en este estudio sugieren que esta tasa alta de *dN*, en comparación con las de M y E, potenciarían los cambios en ciertos residuos aminoacídicos (89). Otra potencial explicación puede ser que la adaptación molecular en ciertos tipos de CoVs severos (SARS-CoV-1, SARS-CoV-2 y MERS-CoV) o aquellos que fueron aislados de hospederos intermediarios (murciélagos, civeta, pangolín y/o camello), dado que es reportado que los virus que se encuentran en procesos de adaptación a nuevos hospederos tienen a sufrir mayores cambios para su adaptación, conduciendo a mayor cantidad de sustituciones (95, 96).

Para el caso de las tasas *dS* en la proteína estructural S es mayor que las de las proteínas M y E (Tabla N°5), particularmente interesante dado que estas mutaciones tienden a mantener el aminoácido. Una potencial explicación para la alta tasa *dS* en S corresponde al hecho de que ciertos residuos en la estructura contribuyen a otorgar cierta ventaja, por lo que a nivel global contribuyen a que el ciclo viral sea estable (93). Independiente de las tasas sinónimas en M y E sean menor con respecto a S, siguen siendo mayores que sus contrapartes no sinónimas (M y E) lo que sugiere que estas proteínas se mantienen sin mayores cambios. Esto podría ser producto de que se ha reportado que los virus de ARN de hebra simple (*single-stranded RNA*), los cuales tienden a poseer mayores tasas de

sustituciones sinónimas que contribuyen a compactar el genoma, principalmente en estructuras genómicas que contribuyen al ensamblaje del virión (98), como M y E.

Es interesante notar que ambas proteínas, M y E, son conservadas a lo largo del género *Beta-CoV*, inclusive comparte más de un 90% de identidad de secuencia con homólogos del SARS-CoV (99). Por otro lado, para el caso de los *Alfa-CoV* este grado de conservación es menor, dado el largo variable de estas proteínas dentro de este género y la presencia de la glicoproteína asociada a la envoltura, la hemaglutinina-esterasa (HE) (100).

Este estudio se centró en predecir sitios de glicosilación en diferentes ortólogos de las tres proteínas estructurales presentes al exterior del virión. Una vez determinado el grado de selección molecular a nivel de secuencias para cada tipo de CoV, se dio paso a la predicción de los tipos de MPT relacionadas a glicanos. La *N*-glicosilación se predice según la aparición de un *sequon* (S-X-S/T, donde X no puede ser P), en cambio, la secuencia consenso necesario para las *O*-glicosilación corresponde a la presencia de S o T en la estructura proteica (35).

La proteína S se encuentran altamente glicosilada en los diferentes tipos de CoVs (géneros *Alfa-CoV* y *Beta-CoV*); en contraste, las proteínas M y E se han reportado con sólo una o ninguna glicosilación, respectivamente (30). En los *Alfa-CoV* se predijeron en promedio 7,36 *N*-glicosilaciones en S y 1,67 en M. En los *Beta-CoV,* el promedio de *N*-glicosilaciones para S es de 3,59 y de 0,99 para M.

Es importante mencionar que para la proteína E no se predijo ningún sitio como potencialmente glicosilado, para ambos tipos de glicosilación. Sin embargo, se ha reportado que dicha proteína sí posee potenciales *N*-glicosilaciones, como se ha observado, por ejemplo, en el SARS-CoV-1. Donde la proteína E posee dos glicosilaciones cuando la proteína adopta una topología particular en la membrana del virión (102). Tomando en cuenta esto, no se tomaron en cuenta dichas potenciales glicosilaciones debido a que su aparición está condicionada por el cambio de topología e interacción con otras proteínas para formar complejos proteicos durante el proceso de replicación viral (en conjunto con S y M) (30, 102).

A nivel del tipo de huésped al cual infecta cada tipo de CoV se evaluó el número de glicosilaciones según huésped, en donde se compara a los CoVs presentes en humanos, tanto severos como de resfrío común, y de aquellos que se encuentran principalmente en animales no humanos, los cuales corresponden a civetas, pangolines, rumiantes, murciélagos y murinos (Figura N°12). Al comparar

estos grupos se encontró que ambos tipos de glicosilaciones poseen diferencias significativas entre CoVs presentes en humanos y animales no humanos. En donde para el caso del *N*-glicosilaciones de la proteína S de aislados virales de humanos poseen un número mayor de glicosilaciones por monómero, en contraste con los de animales. Por otro lado, el caso contrario se presenta para las *O*-glicosilaciones, siendo los aislados virales de animales aquellos con mayor cantidad. Este fenómeno se ha reportado principalmente para la Influenza A, específicamente en la HA. Este fenómeno se encuentra presente específicamente en aves (la diferencia en el número de glicosilaciones es aún mayor), donde porciones de la HA poseen zonas mínimamente glicosiladas, en contraste con sus contrapartes humanas que acumulan entre 7 a 9 veces más sitios de glicosilación (104). Vale recalcar que no se ha reportado dicho fenómeno en los géneros de CoVs abordado en el estudio, independiente de que la diferencia en el número sea menor. Sin embargo, los resultados pueden dar indicios que estas diferencias en número de glicosilaciones sí están presentes en los géneros *Alfa-CoV* y *Beta-CoV*.

Es importante hacer notar que el número de glicosilaciones predichas en este estudio es menor que la reportada en otros estudios, esto se debe a que los estudios llevan a cabo tanto la cuantificación, como la caracterización de éstas. Ahora bien, otro fenómeno que vale la pena mencionar es el hecho de que muchas de las glicosilaciones predichas no siempre se encontrarán glicosiladas. Para ejemplificar este fenómeno se utilizará al SARS-CoV-2 y MERS-CoV. Para el caso de SARS-CoV-2 se predijeron en promedio 17 glicosilaciones por monómero, es decir, cuando se forma el trímero se tiene un total de 51 potenciales glicosilaciones predichas a lo largo de la proteína S. Sin embargo, se han reportado que cada monómero posee 22 a 23 *N*-glicosilaciones (37, 38, 46), por lo que se tendrían 69 glicosilaciones para la estructura proteica completa. Para el caso de MERS-CoV se predijeron un total de 17 para cada trímero, en contraste con las 21 a 23 reportadas por otros estudios (46, 106). Esta discrepancia se puede explicar producto del hecho que no todas las glicosilaciones predichas se encontrarán ocupadas por glicanos una vez expresada la proteína viral en la superficie del virión (30); esto también depende del sistema de expresión que utilizaron, por ejemplo, células de mamíferos o insecto (105), o de la disponibilidad de enzimas encargadas de los procesos de glicosilación en el Compartimento Intermedio Endoplásmico Retículo-Golgi (*Endoplasmic Reticulum-Golgi Intermediate Compartment*, ERGIC) (30, 38).

Las predicciones para *O*-glicosilaciones en *Alfa-CoV* y *Beta-CoV* mostraron que para la proteína S se estimaron un promedio de 2,2 y 4,1 glicosilaciones, respectivamente. Cabe destacar que este tipo

de glicosilaciones son aún más difíciles de predecir y caracterizar (101), principalmente porque para que se produzca una *O*-glicosilación se deben reunir ciertos requisitos que propicien su aparición; por ejemplo, los aminoácidos adyacentes que dan origen al *sequon* (101), especialmente la posición preferente del residuo *P* cercano al *sequon* (102). El reducido número de *O*-glicanos en la proteína S de los CoVs, puede contrastarse con otros patógenos virales que infectan humanos como el citomegalovirus y/o virus de Epstein Barr, los cuales están extensamente *O*-glicosilados. Particularmente interesante es que algunos virus utilizan un extenso repertorio de *N*-glicosilaciones para proteger proteínas del reconocimiento del sistema inmune (VIH), en cambio, el rol de sugerido para el amplio número de *O*-glicosilaciones se relaciona principalmente con enmascarar epítopos inmunodominantes de anticuerpos y células citotóxicas (virus de la varicela-zóster) (103).

Para la proteína M, se tiene un promedio de 1,5 y 3,3 predicciones en *Alfa-CoV* y *Beta-CoV*, respectivamente; es interesante notar que las *O*-glicosilaciones consenso solamente se presentan en 2 tipos de *Alfa-CoV,* y están presentes en la mayoría de los *Beta-CoV*, con 6 excepciones. Se ha reportado que las *O*-glicosilaciones en proteínas virales juegan un papel biológico, por ejemplo, contribuyendo a la inducción de interferón del tipo I por parte del MHV-CoV (110), o mediando la unión al receptor del huésped en SARS-CoV-2 (112), entre otros roles. Por ejemplo, dada la evolución que ha tenido este patógenos en la población humana, la aparición de nuevas variantes ha ocasionado cambios en residuos de la proteína S, los cuales potencian la unión al receptor ACE2 (alterando la infectividad y severidad de la infección). Sin embargo, a pesar del alto número de nuevas variantes, aún no se entiende del todo si altera el perfil de glicosilación del SARS-CoV-2 (111). Esto impulsa la oportunidad de investigar si dicho perfil se mantiene o cambia constantemente, por ejemplo, como lo hace la Influenza A (104).

Con el objetivo de investigar los *sequons,* tanto para las *N*- como *O*-glicosilaciones, el uso de EDlogo permite la observación de los consensos de los residuos aledaños al sitio de glicosilación en los diferentes tipos de CoVs, enfocándose en la relación observable entre aminoácidos preferidos y no preferidos. El proceso de glicosilación depende de una maquinaría enzimática del huésped, el cual corresponde a un grupo de enzimas (presentes a lo largo del Compartimiento Retículo Endoplasmático Aparato de Golgi, ERGIC), que reconocen el *sequon* correspondiente para generar una glicosilación (38); sin embargo, ciertas configuraciones del *sequon* pueden potenciar o atenuar la glicosilación de un sitio dado, una vez que la proteína se encuentre completa para realizar sus funciones (114). Por ejemplo, para las *N*-glicosilaciones, residuos como *G* y *A* son preferidos, dado

su tamaño reducido. Otros grupos preferidos en el reconocimiento de la zona de glicosilación corresponden a los residuos hidrofóbicos (*C*, *I* y *V*) y aromáticos (*F* y *Y*).

Dentro del *motif* de glicosilación, la segunda posición cumple con gran parte de estas preferencias; por ejemplo, para la proteína S (Figura N°14), en la segunda posición se presentan *G*, *A*, *F*, *V* e *I*, las cuales potenciarían la aparición de una glicosilación. Sin embargo, a veces también se presentan aminoácidos que reducen la posibilidad de que el residuo de *N* sea modificado como, por ejemplo, *K.* Éste es un aminoácido básico, que se encuentra con un estado no-preferido, que disminuye la posibilidad de glicosilación en la región cercana (115). Considerando las tres primeras posiciones, la tercera posición (*S* o *T* en el consenso clásico) presenta una preferencia por *T*, contrastando con el *sequon* con mayor densidad reportado para la proteína gp120 del Virus de la Inmunodeficiencia Humana (VIH), la cual presenta mayor preferencia por *S* en la tercera posición, de acuerdo con un estudio de aislados recolectados por un periodo de 29 años (114). Teniendo en cuenta esto, no se ha reportado si un género o subgénero de CoVs poseen un *sequon* con algún patrón de cambio particular con respecto al tiempo, lo cual deja abierta la posibilidad de investigar *sequons* para proteínas virales de CoVs.

Para el caso de la proteína M (Figura N°15), la segunda posición del *motif* de glicosilación también presenta residuos de tamaño menor como la *G,* o residuos aromáticos, como la *F*. Es interesante mencionar que se ha reportado que la presencia de *S* o *T* en la segunda posición potenciaría la aparición de una glicosilación, basado en proteínas humanas (116) y S del SARS-CoV-2 (117). En el caso de la proteína M, se observó que *S* o *T* estarían presentes en el 40% de los *sequons* de la proteína estructural M. En la tercera posición del *motif*, posee una preferencia por *S*, en contraste con la proteína S.

Los *sequons* correspondientes a las *O*-glicosilaciones (Figura N°16) contribuyen a la aparición de esta modificación postraduccional en un residuo de *S* o *T* (30). Se sabe que la proteína S posee un grupo de *O*-glicosilaciones en su estructura; sin embargo, se han hecho todavía pocos estudios para caracterizar y cuantificar estas modificaciones de manera efectiva (35, 112, 117). Por ejemplo, para el caso del SARS-CoV-2 se han reportado 3 sitios de *O*-glicosilación conservados (*S*673, *T*678 y *S*686) (37), mientras que otros estudios reportan una mayor cantidad de este tipo de glicosilaciones (117). La función de dichas glicosilaciones no posee un consenso, sin embargo, para el actual virus pandémico se ha reportado que ciertas mutaciones anulan la aparición de ciertas *O*-glicosilaciones

(variantes Alfa y Delta), potenciando el corte de sitio de furina. Esto afecta directamente el tropismo e infectividad viral producto de la modulación que tienen estas modificaciones postraduccionales (111).

La Figura N°16 muestra una predicción de los sitios de 4 residuos, presentes en el *motif* de *O*-glicosilación para los 5 grupos de CoVs analizados. En la tercera posición, donde se localiza la glicosilación, se observa una preferencia de *S* por sobre *T* en el 80% de los grupos, siendo la excepción solamente el grupo de rumiantes (Figura N°16, panel d). Otra característica destacada es la presencia de *P* en las posiciones 1, 2, 3 y 4; se sabe que *P* contribuye a que la probabilidad de que se produzca una *O*-glicosilación aumente (102). Adicionalmente, la presencia de *S* y *T* aledañas al residuo receptor de la glicosilación contribuye a la aparición de múltiples sitios, dado a que propician el ambiente necesario para formar un conglomerado de *O*-glicosilaciones (111). Este rasgo, observado en los grupos de CoVs analizados (Figura N°16) sugeriría entonces la posibilidad de zonas con sitios con múltiples modificaciones en estas proteínas.

El empobrecimiento de *W* en el 60% de los grupos concuerda con el hecho de que la presencia de aminoácidos aromáticos disminuye la probabilidad de la aparición de *O*-glicosilaciones. Sin embargo, se puede apreciar que en el 60% se presenta *Y* en los sitios 1, 4 y 5 (Figura N°16), lo cual reduciría la eficiencia con la que estos sitios puedan ser glicosilados, al igual que la presencia de *F* en el grupo de CoVs severos (Figura N°16a). Este fenómeno se da principalmente por el impedimento estérico que genera el tamaño de la cadena lateral de estos aminoácidos aromáticos (109).

Finalmente, una vez definidos los sitios de glicosilación consenso para los diferentes tipos de CoVs analizados en este estudio, se llevó a cabo la evaluación del grado de selección molecular a nivel de sitios, para las proteínas estructurales S y M, a fin de contrastar esta información con la de sitios de glicosilaciones predichas a lo largo de la estructura (Figura N°17, 18 y SN°6-13). Para el grupo de CoVs severos, (SARS-CoV-1, SARS-CoV-2, SARSr-CoV y MERS-CoV), las proteínas S y M poseen 7 sitios predichos bajo selección positiva; tres de estas posiciones (209, 250 y 582 dentro del alineamiento consenso), poseen *N*- u *O*-glicosilaciones aledañas. En el caso particular de M, en donde uno de los sitios seleccionados positivamente está ubicado en la porción exovirión, posee *N*-glicosilaciones consenso, lo cual sugiere que estas modificaciones podrían estar bajo algún grado de selección diversificadora.

En el caso del conjunto de CoVs asociados a resfrío común (Figura SN°8 y 9) (NL-63-CoV, OC43-CoV, HKU1-CoV y 229E-CoV), el sitio 642 de la proteína S se encuentra bajo selección positiva, el cual está cercano a *O*-glicosilaciones, aunque no se encuentra directamente en uno de los *sequons*. A pesar de ello, este sitio está rodeado de sitios de *N*-glicosilación. El sitio 20 presente en la estructura de M corresponde a otro sitio bajo selección positiva, en donde particularmente se concentran *N*- y *O*-glicosilaciones. En el resto de los grupos de CoVs (Figuras SN°6, 7, 10, 11, 12 y 13) (murciélagos, rumiantes y murinos), se concentra solamente 3 sitios reportados bajo selección positiva; en el caso de la proteína S del grupo de CoVs de murciélagos, uno de los sitios bajo selección positiva (sitio 34), también se encuentra rodeado de *N*-glicosilaciones. Estos agrupamientos de glicosilaciones alrededor de sitios bajo selección positiva, tanto de *N*- como *O*-glicosilaciones, han sido reportados para coronavirus como el SARS-CoV-1 y el MERS-CoV. En donde los análisis de radios ω en sitios expuestos, es decir, sin glicanos que los cubran, son aquellos que poseen mayores valores de ω, en contraste con aquellos residuos con glicanos (46). Esto concuerda con nuestros resultados, dado que los sitios reportados como positivos, tanto en S como en M, se encuentran expuestos y con glicosilaciones aledañas.

La ausencia de *sequons* bajo selección positiva no necesariamente significa ausencia de este tipo de selección natural, en cuyo caso reevaluar nuestra hipótesis utilizando diferentes metodologías *in silico* puede ser una alternativa para aceptar o refutar nuestra hipótesis definitivamente. Por ejemplo, evaluar la probabilidad de que se produzcan ciertas sustituciones de aminoácidos que culminen en la aparición de sitios de glicosilación, además de evaluar la tasa ω en los residuos de las proteínas estructurales en ambos géneros de CoVs presentes en este estudio. Esta metodología alternativa ha sido utilizada para evaluar la aparición *sequons* en la proteína HA de la Influenza A H3N2 (40). Además, evaluar el grado ganancia a lo largo de la historia evolutiva de cada CoV también puede ser un factor a tomar en cuenta. Por ejemplo, en la Influenza A H1N1 se ha reportado la ganancia de diversos sitios de glicosilación en un periodo de tiempo de 3 décadas en la proteína HA (104) o la variabilidad de sitios glicosilados en el VIH en *gp120* a lo largo de su historia evolutiva (114).

Entre los grupos virales mencionados anteriormente, se puede observar un fenómeno particular en cuanto a las glicosilaciones consenso: la mayoría de los sitios predichos se encuentran en agrupamientos que rodean sitios bajo selección positivas. Este fenómeno se puede presentar producto de que las glicosilaciones enmascaran los residuos en donde se encuentran presentes, por

lo que los sitios aledaños presentan un mayor grado de selección positiva, debido a que se encuentran expuestos al reconocimiento humoral del sistema inmune (106), fenómeno mencionado anteriormente. El sitio 3 en la proteína M del grupo de los CoVs severos (Figura N°16), por ejemplo, presenta un grado de selección que sugiere ser la causante de la ganancia del sitio de *N*-glicosilación, de una forma parecida a una observación descrita previamente en análisis evolutivos del H3N2 de la Influenza A (40). El sitio 642 en la proteína S del grupo de CoVs de tipo resfrío común, posee un *sequon* de *Ser* que genera la aparición de *O*-glicosilación.

La distribución de las glicosilaciones consenso para S y M en los diferentes grupos analizados siguen un patrón particular, reportado previamente para el SARS-CoV-2 (109). Tal patrón se refiere a la regla *O-Follow-N*, que consiste en que el conglomerado de *N*-glicosilaciones viene acompañado principalmente un conjunto de *O*-glicosilaciones cercanas a los sitios consenso predichos (117). Para una revisión más específica y sin tomar en cuenta las glicosilaciones consenso, la Figura N°19 permite visualizar en un grupo de CoVs severos y sus relativos zoonóticos, las porciones de la proteína S en donde se presentan aglomeraciones de *N*-glicosilaciones (círculos rojos, Figura N°19a), y en donde se puede ver la presencia de *O*-glicosilaciones cercanos a ellos (círculos azules, Figura N°19b). La potencial explicación de esto es que las transferasas GalNAc (GalNAc-tS), que catalizan la aparición de una *O*-glicosilación, poseen un dominio similar a lectina (*Lectin-like Domain)* que se une a glicanos (123). En el caso de la proteína estructural S, aquellos agrupamientos se observan principalmente en la S1; en cambio, para M, esto ocurre en la porción exovirión de la proteína. Este fenómeno ha sido reportado para otro conjunto de proteínas (117, 124, 125); sin embargo, aún no es claro sí el *sequon* S/T después de una *N* glicosilada posee la tendencia a sufrir una *O*-glicosilación.

La selección natural podría estar causando que se produzcan ciertas sustituciones que contribuyan a la aparición de *N*, *S* o *T* en formas de conglomerados en ciertas porciones de la proteína estructural S en ambos géneros de CoVs, lo cual permitiría la aparición de *N*- y *O*-glicosilaciones en diferentes porciones de la proteína.

Según los resultados obtenidos y las comparaciones realizadas entre CoVs y otros virus, en especial el Influenza A, proponemos un posible modelo para ganancia y/o pérdida de glicosilaciones en los CoVs (*Alfa-CoV* y *Beta-CoV*) mostrado en la Figura N°20. Esto modelo solamente toma en cuenta mutaciones producto de la selección natural (omite eventos de recombinación entre segmentos de virus influenza y eventos de recombinación de porciones de genomas de CoVs, los cuales pueden

tener efectos aún mayores en cambios en la composición del genoma). Según lo discutido, los CoVs analizados poseen sequons relativamente estables, en contraste con Influenza (la ganancia y pérdida en periodos de tiempo es evidente en la HA). Por lo tanto, el modelo propuesto para los CoVs es que estos poseen *sequons* conservados (cambian relativamente poco), los cuales al sufrir una mutación producto de la selección natural poseen la maquinaria necesaria para corregirla, producto del mecanismo de corrección de errores (11), en contraste con el virus Influenza, el cual por su naturaleza de ssARN (-) depende de la polimerasa dependiente de ARN (*RdRp*) para replicar su genoma, la cual posee una alta tasa de mutaciones (118). Por otro lado, enfocado en los cambios de residuos en la proteína S de los CoVs, un ejemplo, es el SARS-CoV-2, donde según lo reportado las variantes que han hecho aparición poseen mutaciones que le permiten modificar la unión al receptor ACE2 y/o modificar el reconocimiento inmune, sin embargo, estos cambios no se han traducido en cambio de posición de *sequons* (ganancia o pérdida de estos), dado que la mayoría de las mutaciones están adyacentes a glicosilaciones (7). Por ende, proponemos que los CoVs no utilizan una estrategia de cambios continuos de glicosilaciones, en contrate al virus Influenza, independiente que la proteína S se encuentre extensamente glicosilada por *N*- y un número menor de *O*-glicosilaciones, por lo que los cambios mutacionales producto de la selección natural no se ven reflejados en la modificación de los *sequons*.

Teniendo en cuenta los resultados obtenidos, ampliar el número de CoVs de ambos géneros sería el principal objetivo para futuras investigaciones dado que los resultados de la vigilancia genómica han culminado en el descubrimiento de nuevos *Alfa-CoV* y *Beta-CoV* (77). Además de utilizar análisis *in silico* alternativos que permitirían definir el patrón de *sequons* que poseen ambos subgéneros, definir si hay algún fenómeno evolutivo actuando sobre los sitios de glicosilación y/o qué ventajas evolutivas les otorgan a las glicosilaciones estos fenómenos a lo largo de su historia evolutiva.

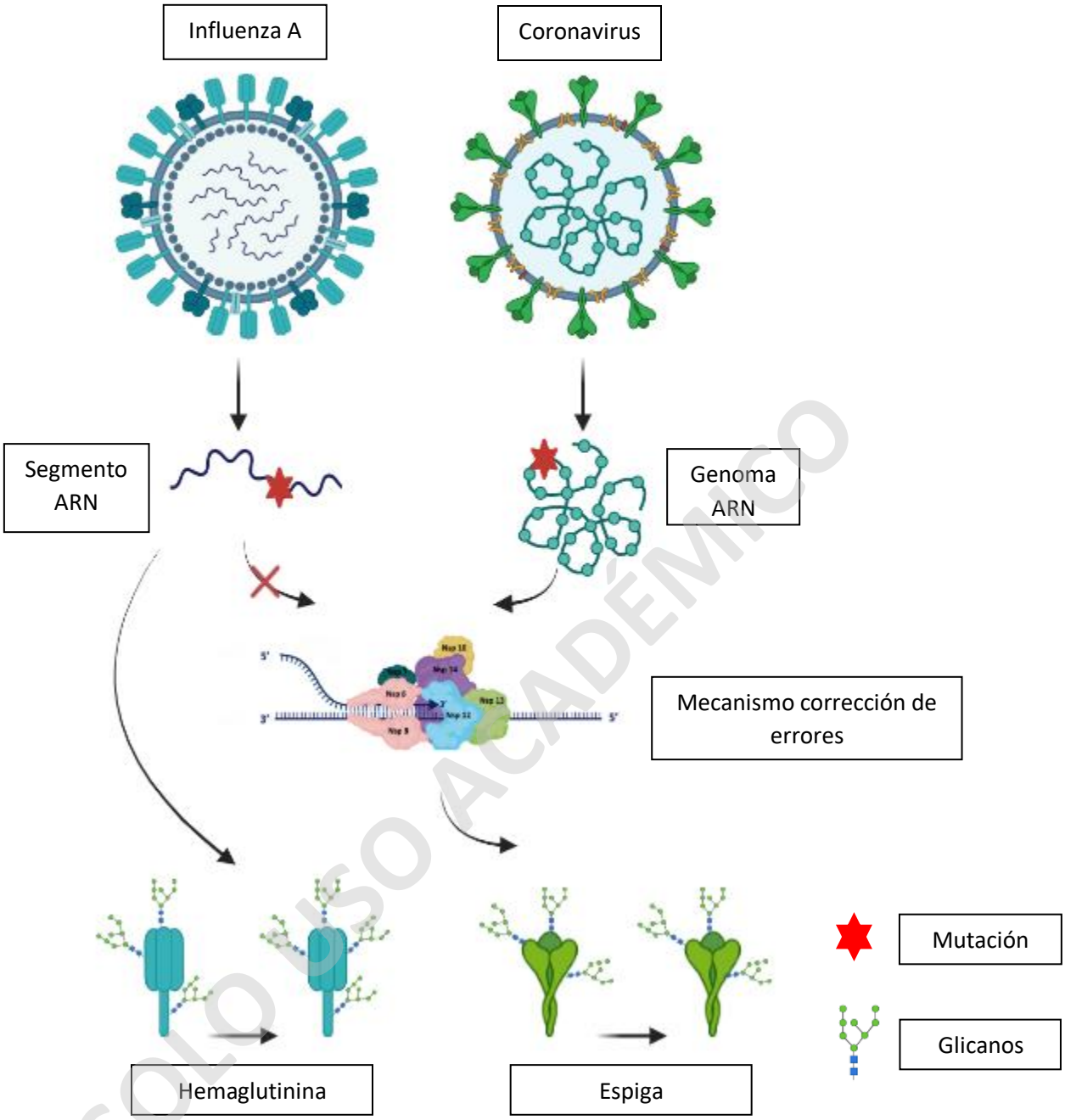

Figura N°20 Modelo propuesto para ganancia y/o pérdida de *sequons* por selección molecular en CoVs.

Comparación entre el virus Influenza y CoVs, donde se muestra el genoma de ARN segmentado (Influenza) y el genoma de ARN continuo (CoVs). La naturaleza del genoma puede influir en la tasa de cambios producto de la selección natural. La presencia del mecanismo de corrección de errores en los CoVs es un factor importante a tener en cuenta, dado que la ausencia de este en Influenza se traduce en tasas de errores mayores en comparación con los CoVs. Por lo que proteínas de superficie como la HA tienen mayor oportunidad de sufrir eventos de ganancia o pérdida de glicosilaciones (lo que se traduce en cambios de sequons), en contraste con los CoVs (donde se mantienen relativamente estable los sequons). La estrella roja representa las mutaciones producto de la selección natural.

6. CONCLUSIONES

Los géneros *Alfa-CoV* y *Beta-CoV* son un grupo diverso de CoVs que infectan a un amplio rango de hospederos, incluido el ser humano. Determinar el grado y tipo de selección al cual están sujetos estos patógenos intracelulares es un desafío importante dado las dinámicas evolutivas que poseen. Sin embargo, este estudio se centra en evaluar el grado de selección natural sobre los residuos, relacionados directa o indirectamente, a las modificaciones postraduccionales del tipo glicosilación.

El conjunto de CoVs analizados en esta tesis dan indicios de cierto patrón que comparten los géneros *Alfa-CoV* y *Beta-CoV* conocidos al inicio de la pandemia del SARS-CoV-2. Al analizar las filogenias presentadas en nuestro objetivo inicial, logramos verificar relaciones filogenéticas esperadas entre los tipos de CoVs, incluyendo agrupamientos entre CoVs severos y sus contrapartes de origen animal más cercanos. Una excepción a este patrón se observó para el caso de los agrupamientos correspondientes a CoVs de rumiantes y murinos cuyo perfil filogenético podría atribuirse a fenómenos de recombinación, como ya se ha visto en las variantes del tipo HKU-23, Bovino-CoV y/o HKU-24. El efecto de la recombinación entre diferentes tipos de CoVs se ha visto sólo de forma parcial en grupos por separado, por lo que quedaría evaluar su efecto en la filogenia utilizando un conjunto más completo de la familia *Coronaviridae*.

A nivel de secuencias de las proteínas estructurales, las tasas ω sugieren que las proteínas M y E poseen un grado mayor de selección positiva, sin embargo, las tasas *dN* y *dS* son más altas en la proteína S, lo cual indica que esta proteína es aquella que sufre más cambios mutacionales en su secuencia aminoacídica en ambos géneros de CoVs. Es necesario destacar que la evaluación del nivel de selección natural en proteínas puede estar sujeta a diferentes factores y es un proceso dinámico, por lo que su evaluación y análisis debe ser constante para poder evaluar cambios de importancia en estas proteínas.

Los sitios de glicosilación predichos en este estudio se concentran principalmente en las proteínas estructurales S y M. Las comparaciones entre *Alfa-CoV* y *Beta-CoV* sugieren una mayor abundancia de *N*-glicosilaciones detectadas en los ortólogos de *Alfa-CoV*, y una mayor abundancia de *O*-glicosilaciones en los *Beta-CoV*. Las proteínas provenientes de diferentes tipos de CoVs que infectan humanos poseen ligeramente mayor cantidad de sitios de glicosilación predichos en contraste con aquellas proteínas provenientes de virus que infectan otros animales no humanos. Este rasgo ha sido observado en otros virus, como la influenza A. Dentro de las restricciones que se tienen para la

cuantificación de los sitios de glicosilación en este estudio, se puede mencionar la falta de análisis de la biología estructural de los *sequons* y/o la evaluación de ganancia o pérdida a nivel del historial evolutivo para cada tipo de CoVs, principalmente en aquellos tipos pobremente estudiados que no se encuentran relacionados con sus contrapartes pandémicas.

El análisis de *sequons* sugiere diferentes preferencias en los grupos de CoVs analizados: para las *N*-glicosilaciones, se presenta una preferencia del *sequon* NXT por sobre el NXS, algo previamente observado en otros tipos de virus. La importancia evolutiva de esta preferencia en CoVs es un aspecto que merece consideración en estudios posteriores.

También se puede concluir que las glicosilaciones consenso reportadas para la proteína estructural S siguen un patrón *O-follow-N*, en el cual el conglomerado de *N*-glicosilaciones presentes a lo largo de la proteína S se ven acompañadas por un número menor de *O*-glicosilaciones, patrón que se ha visto previamente en el SARS-CoV-2 y que merece ser explorado en otros tipos dentro de la familia *Coronaviridae*.

A pesar de no haber descubierto residuos bajo selección positiva directamente relacionados a los sitios de glicosilación, se pudo observar que los residuos bajo este tipo de presión evolutiva se encuentran en los residuos aledaños a los sitios de glicosilación.

Finalmente, podemos concluir que no se encontró ningún *sequon* en las proteínas estructurales S, M y E de los géneros *Alfa-CoV* y *Beta-CoV* que tenga alguna relación con residuos aminoacídicos que se encuentren bajo selección natural del tipo positiva.

7. IMPLICANCIAS, RECOMENDACIONES Y/O PROYECCIONES FUTURAS

El enfoque evolutivo de esta investigación permitió indagar sobre la relación que existe entre la presencia de glicosilaciones y la selección natural asociada con esos *motifs* o los sitios aledaños, en diferentes tipos de coronavirus. Este enfoque es útil dado el limitado número de estudios que se refieren a la cantidad de glicosilaciones en el conglomerado de CoVs conocidos, así como de sitios seleccionados en las proteínas estructurales.

La cuantificación y caracterización de glicosilaciones en proteínas estructurales de tipos emergentes de CoVs contribuye a formar una idea acerca de los patrones de origen y conservación de estos MPT. La comprensión de los procesos que están involucrados en MPT en proteínas de la familia *Coronaviridae* puede posibilitar mejores estrategias para la generación y/o modificación de vacunas contra esta clase de patógenos emergentes.

El estudio reportado en esta tesis aborda 19 tipos de CoVs, en donde se extiende de forma novedosa la relación entre selección molecular y modificaciones postraduccionales en proteínas estructurales. El estudio del rol de los sitios de glicosilación en las proteínas de diferentes CoVs es hasta la fecha un aspecto que todavía requiere una investigación más profunda, considerando el rol de estas modificaciones en el reconocimiento de estos patógenos.

Junto con el amplio rastreo y caracterización de nuevos tipos o variantes de CoVs, especialmente provenientes de *Quirópteros* (el orden de mamíferos que contiene a los murciélagos), se recomienda ampliar dicho rastreo a más tipos de animales, así como implementar una vigilancia genómica y clínica continua; merecería especial atención, por ejemplo, la vigilancia genómica para animales que compartan nichos ecológicos con murciélagos y humanos, debido a su potencial rol como intermediarios en procesos de zoonosis. Adicionalmente, el análisis de los *sequons* y *motifs* de glicosilación, provenientes de proteínas de estos nuevos tipos y/o variantes, podría ayudar a predecir potenciales epítopos con cierta relevancia terapéutica.

Junto con lo anterior, dentro del marco de la vigilancia genómica, el uso de la información de selección natural para la elección de potenciales antígenos también puede ser relevante. En un estudio anterior (126) se propuso un método *in silico* que contempla el uso de péptidos o regiones bajo selección positiva como candidatos para generación de vacunas. En ese estudio, se desarrolló la búsqueda de proteínas provenientes de *Toxoplasma gondii* que contuviesen sitios bajo selección

positiva, debido a que este tipo de selección podría evidenciar su relación con los mecanismos de defensa del hospedero. Se propone que las proteínas sometidas a selección positiva por el sistema inmune son candidatos idóneos para la generación de vacunas, en donde los cambios producidos por esta selección natural generan patrones que son rastreables a lo largo de la proteína.

El siguiente paso corresponde a seleccionar y evaluar los candidatos epítopos. Por ejemplo, para la proteína S apuntamos a zonas conservadas como la subunidad 2, donde hay un rango de aminoácidos conservado entre CoVs (945-1100) (127), lo cual puede ser evaluado para un grupo más amplio de proteínas S, inclusive en nuestro estudio es la zona con menor glicosilaciones consenso. La proteína M, contiene dominios transmembrana que poseen epítopos para células T, lo cual contribuye a inmunidad celular proteína-específica (128), los cuales no poseen glicosilaciones consenso. Finalmente, para la proteína E, se ha reportado que todos los CoVs comparten la misma arquitectura general para esta proteína, por lo que proponemos que la elección del epítopo se centre principalmente en el dominio de unión a PDZ (P para densidad postsináptica, la proteína de unión septada de Drosophila Discs-largo y la proteína de unión estrecha epitelial ZO-1). Algunos CoVs, incluido el SARS-CoV-1, emplean este dominio para modular procesos celulares que repercuten directamente en la patogénesis viral (129), para esta proteína estructural es necesario evaluar la presencia de glicosilaciones según la disposición tridimensional.

El enfoque que proponemos es el uso de nanotecnología para la generación de estructuras mosaicos que contengan antígenos, por ejemplo, los evaluados como candidatos a través de análisis *in silico*. Esto sería algo similar a lo realizado por Cohen y colaboradores (130), quienes llevaron a cabo el ensamblaje de complejos de antígenos en nanopartículas, los que fueron evaluados según su capacidad para generar respuesta inmune frente a un espectro de CoVs en murinos. Este enfoque basado en la generación de un pan-vacuna, evidenció que la respuesta inmune de los murinos frente al espectro de *Sarbecovirus* utilizados en el estudio, dio indicios positivos para estimular una respuesta inmune eficiente, lo cual abre la posibilidad de la potencial generación de pan-vacunas frente a patógenos con relaciones zoonóticas, como los *Beta-CoV.* Para el desarrollo de esta vacuna proponemos un enfoque diferente al de Cohen y colaboradores (130), basado en el uso de las tres proteínas estructurales de este estudio (S, M y E), según la evaluación de selección molecular y epítopos mencionados que se encuentran aledaños a sitios de glicosilación consenso, por ende, expuestos. Dicha estrategia de utilizar proteínas estructurales ha sido diseñada *in silico* para las proteínas S-N y E-M por separado, centrada solamente en CoVs que infectan humanos (131).

La integración de la vigilancia genómica, el testeo activo de candidatos a epítopos bajo selección positiva y la caracterización de sitios de glicosilación en dichas proteínas, puede contribuir a desarrollar un amplio espectro de proteínas de CoVs para ser evaluadas como potenciales candidatos a pan-vacunas. Esta investigación puede contribuir al desarrollo de estas áreas, para una posterior exploración de potenciales innovaciones biotecnológicas, cubriendo alternativas de tratamiento frente a patógenos de carácter zoonótico.

8. REFERENCIAS

1. Grubaugh ND, Ladner JT, Lemey P, Pybus OG, Rambaut A, Holmes EC, Andersen KG. Tracking virus outbreaks in the twenty-first century. Nature microbiology. 2019 Jan; 4(1): 10-9.
2. Dennehy JJ. Evolutionary ecology of virus emergence. Annals of the New York Academy of Sciences. 2017 Feb;1389(1):124.
3. Centers for Disease Control and Prevention (CDC). Principles of Epidemiology in Public Health Practice. 2012 Oct;1(3)-511.
4. Cucinotta D, Vanelli M. WHO declares COVID-19 a pandemic. Acta Bio Medica: Atenei Parmensis. 2020;91(1):157.
5. Andersen KG, Rambaut A, Lipkin WI, Holmes EC, Garry RF. The proximal origin of SARS-CoV-2. Nature medicine. 2020 Apr;26(4):450-2.
6. Su S, Wong G, Shi W, Liu J, Lai AC, Zhou J, Liu W, Bi Y, Gao GF. Epidemiology, genetic recombination, and pathogenesis of coronaviruses. Trends in microbiology. 2016 Jun 1;24(6):490-502.
7. Plante JA, Mitchell BM, Plante KS, Debbink K, Weaver SC, Menachery VD. The variant Gambit: COVID's next move. Cell host & microbe. 2021 Mar 1.
8. WHO. WHO Coronavirus (COVID-19) Dashboard; 2022. World Health Organization. [Cited 24 Jan]. Available from: https://covid19.who.int/?gclid=CjwKCAjwtNf6BRAwEiwAkt6UQg9OUFk6pIC07i6Vxe3Ar8fH1PDMYp2_1c42YJ4ZevNdM_nqFA_GmBoCcIsQAvD_BwE
9. Narh CA. Genomic cues from beta-coronaviruses and mammalian hosts sheds light on probable origins and infectivity of SARS-CoV-2 causing COVID-19. Frontiers in genetics. 2020;11.
10. Cleaveland S, Laurenson MK, Taylor LH. Diseases of humans and their domestic mammals: pathogen characteristics, host range and the risk of emergence. Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences. 2001 Jul 29;356(1411):991-9.
11. Robson F, Khan KS, Le TK, Paris C, Demirbag S, Barfuss P, Rocchi P, Ng WL. Coronavirus RNA proofreading: molecular basis and therapeutic targeting. Molecular cell. 2020 Aug 4.
12. Ye ZW, Yuan S, Yuen KS, Fung SY, Chan CP, Jin DY. Zoonotic origins of human coronaviruses. International journal of biological sciences. 2020;16(10):1686.

13. Martirosyan L, Paget WJ, Jorgensen P, Brown CS, Meerhoff TJ, Pereyaslov D, Mott JA. The community impact of the 2009 influenza pandemic in the WHO European region: a comparison with historical seasonal data from 28 countries. BMC infectious diseases. 2012 Dec;12(1):1-1.

14. Van Hoek AJ, Underwood A, Jit M, Miller E, Edmunds WJ. The impact of pandemic influenza H1N1 on health-related quality of life: a prospective population-based study. PloS one. 2011 Mar 2;6(3):e17030.

15. Wang C, Zhang H, Gao Y, Deng Q. Comparative Study of Government Response Measures and Epidemic Trends for COVID-19 Global Pandemic. Risk Analysis. 2021 Sep 5.

16. Li X, Geng W, Tian H, Lai D. Was mandatory quarantine necessary in China for controlling the 2009 H1N1 pandemic?. International journal of environmental research and public health. 2013 Oct;10(10):4690-700.

17. Nicola M, Alsafi Z, Sohrabi C, Kerwan A, Al-Jabir A, Iosifidis C, Agha M, Agha R. The socio-economic implications of the coronavirus pandemic (COVID-19): A review. International journal of surgery. 2020 Jun 1;78:185-93.

18. Sallusto F, Lanzavecchia A, Araki K, Ahmed R. From vaccines to memory and back. Immunity. 2010 Oct 29;33(4):451-63.

19. Han S. Clinical vaccine development. Clinical and experimental vaccine research. 2015 Jan 1;4(1):46-53.

20. Chang D, Zaia J. Why glycosylation matters in building a better flu vaccine. Molecular & Cellular Proteomics. 2019 Dec 1;18(12):2348-58.

21. María RR, Arturo CJ, Alicia JA, Paulina MG, Gerardo AO. The impact of bioinformatics on vaccine design and development. Vaccines. 2017 Sep 6;2:3-6.

22. Bagdonaite I, Wandall HH. Global aspects of viral glycosylation. Glycobiology. 2018 Jul;28(7):443-67.

23. Forni D, Cagliani R, Clerici M, Sironi M. Molecular evolution of human coronavirus genomes. Trends in microbiology. 2017 Jan 1;25(1):35-48.

24. Graham RL, Donaldson EF, Baric RS. A decade after SARS: strategies for controlling emerging coronaviruses. Nature Reviews Microbiology. 2013 Dec;11(12):836-48.

25. Vakulenko Y, Deviatkin A, Drexler JF, Lukashev A. Modular Evolution of Coronavirus Genomes. Viruses. 2021 Jul;13(7):1270.

26. Wang Q, Vlasova AN, Kenney SP, Saif LJ. Emerging and re-emerging coronaviruses in pigs. Current opinion in virology. 2019 Feb 1;34:39-49.

27. Cui J, Li F, Shi ZL. Origin and evolution of pathogenic coronaviruses. Nature Reviews Microbiology. 2019 Mar;17(3):181-92.

28. Drexler JF, Corman VM, Drosten C. Ecology, evolution and classification of bat coronaviruses in the aftermath of SARS. Antiviral research. 2014 Jan 1;101:45-56.

29. McBride R, Van Zyl M, Fielding BC. The coronavirus nucleocapsid is a multifunctional protein. Viruses. 2014 Aug;6(8):2991-3018.

30. Fung TS, Liu DX. Post-translational modifications of coronavirus proteins: roles and function. Future virology. 2018 May 21;13(6):405-30.

31. Schoeman D, Fielding BC. Coronavirus envelope protein: current knowledge. Virology journal. 2019 Dec;16(1):1-22.

32. Song W, Gui M, Wang X, Xiang Y. Cryo-EM structure of the SARS coronavirus spike glycoprotein in complex with its host cell receptor ACE2. PLoS pathogens. 2018 Aug 13;14(8):e1007236.

33. Watanabe Y, Allen JD, Wrapp D, McLellan JS, Crispin M. Site-specific glycan analysis of the SARS-CoV-2 spike. Science. 2020 Jul 17;369(6501):330-3.

34. Graham RL, Baric RS. Recombination, reservoirs, and the modular spike: mechanisms of coronavirus cross-species transmission. Journal of virology. 2010 Apr 1;84(7):3134-46.

35. Varki A, Cummings RD, Esko JD, Stanley P, Hart GW, Aebi M, Darvill AG, Kinoshita T, Packer NH, Prestegard JH, Schnaar RL. Essentials of Glycobiology [internet].

36. Thompson AJ, de Vries RP, Paulson JC. Virus recognition of glycan receptors. Current opinion in virology. 2019 Feb 1;34:117-29.

37. Shajahan A, Supekar NT, Gleinich AS, Azadi P. Deducing the N-and O-glycosylation profile of the spike protein of novel coronavirus SARS-CoV-2. Glycobiology. 2020 Dec;30(12):981-8.

38. Watanabe Y, Bowden TA, Wilson IA, Crispin M. Exploitation of glycosylation in enveloped virus pathobiology. Biochimica et Biophysica Acta (BBA)-General Subjects. 2019 Oct 1;1863(10):1480-97.

39. da Silva AG. Measuring natural selection. InBioinformatics 2017 (pp. 315-347). Humana Press, New York, NY.

40. Suzuki Y. Positive selection for gains of N-linked glycosylation sites in hemagglutinin during evolution of H3N2 human influenza A virus. Genes & genetic systems. 2011;86(5):287-94.

41. Lau SK, Feng Y, Chen H, Luk HK, Yang WH, Li KS, Zhang YZ, Huang Y, Song ZZ, Chow WN, Fan RY. Severe acute respiratory syndrome (SARS) coronavirus ORF8 protein is acquired from SARS-related coronavirus from greater horseshoe bats through recombination. Journal of virology. 2015 Oct 15;89(20):10532-47.

42. Sabir JS, Lam TT, Ahmed MM, Li L, Shen Y, Abo-Aba SE, Qureshi MI, Abu-Zeid M, Zhang Y, Khiyami MA, Alharbi NS. Co-circulation of three camel coronavirus species and recombination of MERS-CoVs in Saudi Arabia. Science. 2016 Jan 1;351(6268):81-4.

43. Tang X, Wu C, Li X, Song Y, Yao X, Wu X, Duan Y, Zhang H, Wang Y, Qian Z, Cui J. On the origin and continuing evolution of SARS-CoV-2. National Science Review. 2020 Jun 1;7(6):1012-23.

44. Lv L, Li G, Chen J, Liang X, Li Y. Comparative genomic analysis revealed specific mutation pattern between human coronavirus SARS-CoV-2 and Bat-SARSr-CoV RaTG13. BioRxiv. 2020 Jan 1.

45. Pyrc K, Dijkman R, Deng L, Jebbink MF, Ross HA, Berkhout B, Van der Hoek L. Mosaic structure of human coronavirus NL63, one thousand years of evolution. Journal of molecular biology. 2006 Dec 15;364(5):964-73.

46. Watanabe Y, Berndsen ZT, Raghwani J, Seabright GE, Allen JD, Pybus OG, McLellan JS, Wilson IA, Bowden TA, Ward AB, Crispin M. Vulnerabilities in coronavirus glycan shields despite extensive glycosylation. Nature communications. 2020 May 27;11(1):1-0.

47. Altman MO, Angel M, Košík I, Trovão NS, Zost SJ, Gibbs JS, Casalino L, Amaro RE, Hensley SE, Nelson MI, Yewdell JW. Human influenza A virus hemagglutinin glycan evolution follows a temporal pattern to a glycan limit. MBio. 2019 Apr 2;10(2):e00204-19.

48. Watanabe Y, Raghwani J, Allen JD, Seabright GE, Li S, Moser F, Huiskonen JT, Strecker T, Bowden TA, Crispin M. Structure of the Lassa virus glycan shield provides a model for immunological resistance. Proceedings of the National Academy of Sciences. 2018 Jul 10;115(28):7320-5.

49. Hatcher EL, Zhdanov SA, Bao Y, Blinkova O, Nawrocki EP, Ostapchuck Y, Schäffer AA, Brister JR. Virus Variation Resource–improved response to emergent viral outbreaks. Nucleic acids research. 2017 Jan 4;45(D1):D482-90.

50. Brister, R.; Ako-Adjei, D.; Bao, Y.; Blinkova, O. NCBI viral genomes resource. Nucleic Acids Res. 2014, 43, D571–D577.

51. Weaver S, Shank SD, Spielman SJ, Li M, Muse SV, Kosakovsky Pond SL. Datamonkey 2.0: a modern web application for characterizing selective and other evolutionary processes. Molecular biology and evolution. 2018 Mar 1;35(3):773-7.

52. Cock PJ, Antao T, Chang JT, Chapman BA, Cox CJ, Dalke A, Friedberg I, Hamelryck T, Kauff F, Wilczynski B, De Hoon MJ. Biopython: freely available Python tools for computational molecular biology and bioinformatics. Bioinformatics. 2009 Jun 1;25(11):1422-3.

53. Stajich JE, Block D, Boulez K, Brenner SE, Chervitz SA, Dagdigian C, Fuellen G, Gilbert JG, Korf I, Lapp H, Lehväslaiho H. The Bioperl toolkit: Perl modules for the life sciences. Genome research. 2002 Oct 1;12(10):1611-8.

54. Nakamura T, Yamada KD, Tomii K, Katoh K. Parallelization of MAFFT for large-scale multiple sequence alignments. Bioinformatics. 2018 Jul 15;34(14):2490-2.

55. Larsson A. AliView: a fast and lightweight alignment viewer and editor for large datasets. Bioinformatics. 2014 Nov 15;30(22):3276-8.

56. Procter JB, Carstairs GM, Soares B, Mourão K, Ofoegbu TC, Barton D, Lui L, Menard A, Sherstnev N, Roldan-Martinez D, Duce S. Alignment of biological sequences with Jalview. InMultiple Sequence Alignment 2021 (pp. 203-224). Humana, New York, NY.

57. Minh BQ, Schmidt HA, Chernomor O, Schrempf D, Woodhams MD, Von Haeseler A, Lanfear R. IQ-TREE 2: new models and efficient methods for phylogenetic inference in the genomic era. Molecular biology and evolution. 2020 May 1;37(5):1530-4.

58. Kalyaanamoorthy S, Minh BQ, Wong TK, Von Haeseler A, Jermiin LS. ModelFinder: fast model selection for accurate phylogenetic estimates. Nature methods. 2017 Jun;14(6):587-9.

59. Hoang DT, Chernomor O, Von Haeseler A, Minh BQ, Vinh LS. UFBoot2: improving the ultrafast bootstrap approximation. Molecular biology and evolution. 2018 Feb 1;35(2):518-22.

60. Letunic I, Bork P. Interactive Tree Of Life (iTOL) v4: recent updates and new developments. Nucleic acids research. 2019 Jul 2;47(W1):W256-9.

61. Rombel IT, Sykes KF, Rayner S, Johnston SA. ORF-FINDER: a vector for high-throughput gene identification. Gene. 2002 Jan 9;282(1-2):33-41.

62. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. Journal of molecular biology. 1990 Oct 5;215(3):403-10.

63. Suyama M, Torrents D, Bork P. PAL2NAL: robust conversion of protein sequence alignments into the corresponding codon alignments. Nucleic acids research. 2006 Jul 1;34(suppl_2):W609-12.

64. Yang Z. PAML: a program package for phylogenetic analysis by maximum likelihood. Computer applications in the biosciences. 1997 Oct 1;13(5):555-6.

65. Mazola Y, Chinea G, Musacchio A. Integrating bioinformatics tools to handle glycosylation. PLoS computational biology. 2011 Dec 29;7(12):e1002285.

66. Gupta R, Brunak S. Prediction of glycosylation across the human proteome and the correlation to protein function. InPac Symp Biocomput 2001 Dec 12 (Vol. 7, pp. 310-22).

67. Steentoft C, Vakhrushev SY, Joshi HJ, Kong Y, Vester-Christensen MB, Schjoldager KT, Lavrsen K, Dabelsteen S, Pedersen NB, Marcos-Silva L, Gupta R. Precision mapping of the human O-GalNAc glycoproteome through SimpleCell technology. The EMBO journal. 2013 May 15;32(10):1478-88.

68. "One-way ANOVA followed by Dunnett's multiple comparisons test was performed using GraphPad Prism version 8.0.0 for Windows, GraphPad Software, San Diego, California USA, www.graphpad.com".

69. Kosakovsky Pond SL, Frost SD. Not so different after all: a comparison of methods for detecting amino acid sites under selection. Molecular biology and evolution. 2005 May 1;22(5):1208-22.

70. Murrell B, Moola S, Mabona A, Weighill T, Sheward D, Kosakovsky Pond SL, Scheffler K. FUBAR: a fast, unconstrained bayesian approximation for inferring selection. Molecular biology and evolution. 2013 Feb 18;30(5):1196-205.

71. Dey KK, Xie D, Stephens M. A new sequence logo plot to highlight enrichment and depletion. BMC bioinformatics. 2018 Dec;19(1):1-9.

72. Brennan P. drawProteins: a Bioconductor/R package for reproducible and programmatic generation of protein schematics. F1000Research. 2018;7.

73. Crooks GE, Hon G, Chandonia JM, Brenner SE. WebLogo: a sequence logo generator. Genome research. 2004 Jun 1;14(6):1188-90.

74. Tajima F, Nei M. Estimation of evolutionary distance between nucleotide sequences. Molecular biology and evolution. 1984 Apr 1;1(3):269-85.

75. Frutos R, Serra-Cobo J, Pinault L, Lopez Roig M, Devaux CA. Emergence of bat-related betacoronaviruses: hazard and risks. Frontiers in microbiology. 2021 Mar 15;12:437.

76. Mollentze N, Streicker DG. Viral zoonotic risk is homogenous among taxonomic orders of mammalian and avian reservoir hosts. Proceedings of the National Academy of Sciences. 2020 Apr 28;117(17):9423-30.

77. Zhou H, Ji J, Chen X, Bi Y, Li J, Wang Q, Hu T, Song H, Zhao R, Chen Y, Cui M. Identification of novel bat coronaviruses sheds light on the evolutionary origins of SARS-CoV-2 and related viruses. Cell. 2021 Jun 9.

78. Mavrodiev EV, Tursky ML, Mavrodiev NE, Ebach MC, Williams DM. On Classification and Taxonomy of Coronaviruses (Riboviria, Nidovirales, Coronaviridae) with special focus on severe acute respiratory syndrome-related coronavirus 2 (SARS-Cov-2). bioRxiv. 2020 Jan 1.

79. Boni MF, Lemey P, Jiang X, Lam TT, Perry BW, Castoe TA, Rambaut A, Robertson DL. Evolutionary origins of the SARS-CoV-2 sarbecovirus lineage responsible for the COVID-19 pandemic. Nature microbiology. 2020 Nov;5(11):1408-17.

80. Lam TT, Jia N, Zhang YW, Shum MH, Jiang JF, Zhu HC, Tong YG, Shi YX, Ni XB, Liao YS, Li WJ. Identifying SARS-CoV-2-related coronaviruses in Malayan pangolins. Nature. 2020 Jul;583(7815):282-5.

81. Li X, Giorgi EE, Marichannegowda MH, Foley B, Xiao C, Kong XP, Chen Y, Gnanakaran S, Korber B, Gao F. Emergence of SARS-CoV-2 through recombination and strong purifying selection. Science Advances. 2020 Jul 1;6(27):eabb9153.

82. Bobay LM, O'Donnell AC, Ochman H. Recombination events are concentrated in the spike protein region of Betacoronaviruses. PLoS genetics. 2020 Dec 17;16(12):e1009272.

83. Han HJ, Yu H, Yu XJ. Evidence for zoonotic origins of Middle East respiratory syndrome coronavirus. The Journal of general virology. 2016 Feb;97(Pt 2):274.

84. So RT, Chu DK, Miguel E, Perera RA, Oladipo JO, Fassi-Fihri O, Aylet G, Ko RL, Zhou Z, Cheng MS, Kuranga SA. Diversity of dromedary camel coronavirus HKU23 in African camels revealed multiple recombination events among closely related betacoronaviruses of the subgenus Embecovirus. Journal of virology. 2019 Dec 1;93(23):e01236-19.

85. Chu DK, Hui KP, Perera RA, Miguel E, Niemeyer D, Zhao J, Channappanavar R, Dudas G, Oladipo JO, Traoré A, Fassi-Fihri O. MERS coronaviruses from camels in Africa exhibit region-dependent genetic diversity. Proceedings of the National Academy of Sciences. 2018 Mar 20;115(12):3144-9.

86. Vijgen L, Keyaerts E, Moës E, Thoelen I, Wollants E, Lemey P, Vandamme AM, Van Ranst M. Complete genomic sequence of human coronavirus OC43: molecular clock analysis suggests

a relatively recent zoonotic coronavirus transmission event. Journal of virology. 2005 Feb 1;79(3):1595-604.

87. Lau SK, Woo PC, Li KS, Tsang AK, Fan RY, Luk HK, Cai JP, Chan KH, Zheng BJ, Wang M, Yuen KY. Discovery of a novel coronavirus, China Rattus coronavirus HKU24, from Norway rats supports the murine origin of Betacoronavirus 1 and has implications for the ancestor of Betacoronavirus lineage A. Journal of virology. 2015 Mar 15;89(6):3076-92.

88. Jeffares DC, Tomiczek B, Sojo V, dos Reis M. A beginners guide to estimating the non-synonymous to synonymous rate ratio of all protein-coding genes in a genome. InParasite genomics protocols 2015 (pp. 65-90). Humana Press, New York, NY.

89. Kryazhimskiy S, Plotkin JB. The population genetics of dN/dS. PLoS genetics. 2008 Dec 12;4(12):e1000304.

90. Holmes EC. Patterns of intra-and interhost nonsynonymous variation reveal strong purifying selection in dengue virus. Journal of virology. 2003 Oct 15;77(20):11296-8.

91. Sharp PM, Bailes E, Chaudhuri RR, Rodenburg CM, Santiago MO, Hahn BH. The origins of acquired immune deficiency syndrome viruses: where and when?. Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences. 2001 Jun 29;356(1410):867-76.

92. Viboud C, Bjørnstad ON, Smith DL, Simonsen L, Miller MA, Grenfell BT. Synchrony, waves, and spatial hierarchies in the spread of influenza. science. 2006 Apr 21;312(5772):447-51.

93. Lin JJ, Bhattacharjee MJ, Yu CP, Tseng YY, Li WH. Many human RNA viruses show extraordinarily stringent selective constraints on protein evolution. Proceedings of the National Academy of Sciences. 2019 Sep 17;116(38):19009-18.

94. Van Egeren D, Novokhodko A, Stoddard M, Tran U, Zetter B, Rogers M, Pentelute BL, Carlson JM, Hixon M, Joseph-McCarthy D, Chakravarty A. Risk of rapid evolutionary escape from biomedical interventions targeting SARS-CoV-2 spike protein. PloS one. 2021 Apr 28;16(4):e0250780.

95. Bhatt S, Holmes EC, Pybus OG. The genomic rate of molecular adaptation of the human influenza A virus. Molecular biology and evolution. 2011 Sep 1;28(9):2443-51.

96. Geoghegan JL, Holmes EC. The phylogenomics of evolving virus virulence. Nature Reviews Genetics. 2018 Dec;19(12):756-69.

97. Nguyen TT, Pathirana PN, Nguyen T, Nguyen QV, Bhatti A, Nguyen DC, Nguyen DT, Nguyen ND, Creighton D, Abdelrazek M. Genomic mutations and changes in protein secondary

structure and solvent accessibility of SARS-CoV-2 (COVID-19 virus). Scientific Reports. 2021 Feb 10;11(1):1-6.

98. Tubiana L, Božič AL, Micheletti C, Podgornik R. Synonymous mutations reduce genome compactness in icosahedral ssRNA viruses. Biophysical journal. 2015 Jan 6;108(1):194-202.

99. Arya R, Kumari S, Pandey B, Mistry H, Bihani SC, Das A, Prashar V, Gupta GD, Panicker L, Kumar M. Structural insights into SARS-CoV-2 proteins. Journal of molecular biology. 2021 Jan 22;433(2):166725.

100. Turlewicz-Podbielska H, Pomorska-Mól M. Porcine coronaviruses: overview of the state of the art. Virologica Sinica. 2021 Mar 15:1-9.

101. Dong X, Chen C, Yan J, Zhang X, Li X, Liang X. Comprehensive O-glycosylation analysis of the SARS-CoV-2 spike protein with biomimetic Trp-Arg materials. Analytical Chemistry. 2021 Jul 20;93(30):10444-52.

102. Yuan Q, Liao Y, Torres J, Tam JP, Liu DX. Biochemical evidence for the presence of mixed membrane topologies of the severe acute respiratory syndrome coronavirus envelope protein expressed in mammalian cells. FEBS letters. 2006 May 29;580(13):3192-200.

103. Bagdonaite I, Nordén R, Joshi HJ, King SL, Vakhrushev SY, Olofsson S, Wandall HH. Global mapping of O-glycosylation of varicella zoster virus, human cytomegalovirus, and Epstein-Barr virus. Journal of Biological Chemistry. 2016 Jun 3;291(23):12014-28.

104. Altman MO, Angel M, Košík I, Trovão NS, Zost SJ, Gibbs JS, Casalino L, Amaro RE, Hensley SE, Nelson MI, Yewdell JW. Human influenza A virus hemagglutinin glycan evolution follows a temporal pattern to a glycan limit. MBio. 2019 Apr 2;10(2):e00204-19.

105. Gong Y, Qin S, Dai L, Tian Z. The glycosylation in SARS-CoV-2 and its receptor ACE2. Signal transduction and targeted therapy. 2021 Nov 15;6(1):1-24.

106. Walls AC, Xiong X, Park YJ, Tortorici MA, Snijder J, Quispe J, Cameroni E, Gopal R, Dai M, Lanzavecchia A, Zambon M. Unexpected receptor functional mimicry elucidates activation of coronavirus fusion. Cell. 2020 Dec 10;183(6):1732.

107. Brooks SA. Appropriate glycosylation of recombinant proteins for human use. Molecular biotechnology. 2004 Nov;28(3):241-55.

108. Wandall HH, Nielsen MA, King-Smith S, de Haan N, Bagdonaite I. Global functions of O-glycosylation: promises and challenges in O-glycobiology. The FEBS Journal. 2021 Aug 4.

109. Christlet TH, Veluraja K. Database analysis of O-glycosylation sites in proteins. Biophysical journal. 2001 Feb 1;80(2):952-60.

110. De Haan CA, De Wit M, Kuo L, Montalto-Morrison C, Haagmans BL, Weiss SR, Masters PS, Rottier PJ. The glycosylation status of the murine hepatitis coronavirus M protein affects the interferogenic capacity of the virus in vitro and its ability to replicate in the liver but not the brain. Virology. 2003 Aug 1;312(2):395-406.

111. Zhang L, Mann M, Syed Z, Reynolds HM, Tian E, Samara NL, Zeldin DC, Tabak LA, Ten Hagen KG. Furin cleavage of the SARS-CoV-2 spike is modulated by O-glycosylation. bioRxiv. 2021 Jan 1.

112. Zhang Y, Zhao W, Mao Y, Chen Y, Zheng S, Cao W, Zhu J, Hu L, Gong M, Cheng J, Yang H. O-glycosylation landscapes of SARS-CoV-2 spike proteins. Frontiers in Chemistry. 2021:729.

113. Rao RS, Bernd W. Do N-glycoproteins have preference for specific sequons?. Bioinformation. 2010;5(5):208.

114. Rao RS, Wollenweber B. Subtle evolutionary changes in the distribution of N-glycosylation sequons in the HIV-1 envelope glycoprotein 120. International journal of biological sciences. 2010;6(5):407.

115. Manwar Hussain MR, Iqbal Z, Qazi WM, Hoessli DC. Charge and Polarity Preferences for N-glycosylation: a genome-Wide In Silico study and its implications regarding constitutive Proliferation and adhesion of carcinoma cells. Frontiers in oncology. 2018 Feb 28;8:29.

116. Huang YW, Yang HI, Wu YT, Hsu TL, Lin TW, Kelly JW, Wong CH. Residues comprising the enhanced aromatic sequon influence protein N-glycosylation efficiency. Journal of the American Chemical Society. 2017 Sep 20;139(37):12947-55.

117. Tian W, Li D, Zhang N, Bai G, Yuan K, Xiao H, Gao F, Chen Y, Wong CC, Gao GF. O-glycosylation pattern of the SARS-CoV-2 spike protein reveals an “O-Follow-N” rule. Cell research. 2021 Oct;31(10):1123-5.

118. Peck KM, Lauring AS. Complexities of viral mutation rates. Journal of virology. 2018 Jul 15;92(14):e01031-17.

119. Sanda M, Morrison L, Goldman R. N-and O-glycosylation of the SARS-CoV-2 spike protein. Analytical chemistry. 2021 Jan 7;93(4):2003-9.

120. Elhammer AP, Poorman RA, Brown E, Maggiora LL, Hoogerheide JG, Kezdy FJ. The specificity of UDP-GalNAc: polypeptide N-acetylgalactosaminyltransferase as inferred from a database of in vivo substrates and from the in vitro glycosylation of proteins and peptides. Journal of Biological Chemistry. 1993 May 15;268(14):10029-38.

121. Zhang M, Gaschen B, Blay W, Foley B, Haigwood N, Kuiken C, Korber B. Tracking global patterns of N-linked glycosylation site variation in highly variable viral glycoproteins: HIV, SIV, and HCV envelopes and influenza hemagglutinin. Glycobiology. 2004 Dec 1;14(12):1229-46.

122. Petrescu AJ, Milac AL, Petrescu SM, Dwek RA, Wormald MR. Statistical analysis of the protein environment of N-glycosylation sites: implications for occupancy, structure, and folding. Glycobiology. 2004 Feb 1;14(2):103-14.

123. Imberty A, Piller V, Piller F, Breton C. Fold recognition and molecular modeling of a lectin-like domain in UDP-GalNac: polypeptide N-acetylgalactosaminyltransferases. Protein engineering. 1997 Dec 1;10(12):1353-6.

124. Chandrasekhar KD, Lvov A, Terrenoire C, Gao GY, Kass RS, Kobertz WR. O-glycosylation of the cardiac IKs complex. The Journal of physiology. 2011 Aug 1;589(15):3721-30.

125. Riethmueller S, Somasundaram P, Ehlers JC, Hung CW, Flynn CM, Lokau J, Agthe M, Düsterhöft S, Zhu Y, Grötzinger J, Lorenzen I. Proteolytic origin of the soluble human IL-6R in vivo and a decisive role of N-glycosylation. PLoS biology. 2017 Jan 6;15(1):e2000080.

126. Goodswen SJ, Kennedy PJ, Ellis JT. A gene-based positive selection detection approach to identify vaccine candidates using Toxoplasma gondii as a test case protozoan pathogen. Frontiers in genetics. 2018 Aug 20;9:332.Yurina V. Coronavirus epitope prediction from highly conserved region of spike protein. Clinical and experimental vaccine research. 2020 Jul;9(2):169.

127. Yurina V. Coronavirus epitope prediction from highly conserved region of spike protein. Clinical and experimental vaccine research. 2020 Jul;9(2):169.

128. Liu J, Sun Y, Qi J, Chu F, Wu H, Gao F, Li T, Yan J, Gao GF. The membrane protein of severe acute respiratory syndrome coronavirus acts as a dominant immunogen revealed by a clustering region of novel functionally and structurally defined cytotoxic T-lymphocyte epitopes. The Journal of infectious diseases. 2010 Oct 15;202(8):1171-80.

129. Schoeman D, Fielding BC. Is there a link between the pathogenic human coronavirus envelope protein and immunopathology? A review of the literature. Frontiers in microbiology. 2020 Sep 3;11:2086.

130. Cohen AA, Gnanapragasam PN, Lee YE, Hoffman PR, Ou S, Kakutani LM, Keeffe JR, Wu HJ, Howarth M, West AP, Barnes CO. Mosaic nanoparticles elicit cross-reactive immune responses to zoonotic coronaviruses in mice. Science. 2021 Feb 12;371(6530):735-41.

131. Li M, Zeng J, Li R, Wen Z, Cai Y, Wallin J, Shu Y, Du X, Sun C. Rational Design of a Pan-Coronavirus Vaccine Based on Conserved CTL Epitopes. Viruses. 2021 Feb;13(2):333.