



UNIVERSIDAD NACIONAL AUTÓNOMA DE MÉXICO

Maestría y Doctorado en Ciencias Bioquímicas

Evolución molecular de la proteína Spike del SARS-CoV2

TESIS

QUE PARA OPTAR POR EL GRADO DE:

Doctor en Ciencias

PRESENTA:

Georgina Ivette López Cortés

Dr. Enrique Ortega Soto
[Instituto de Investigaciones Biomédicas](#)

Dra. Gohar Gevorgyan Markosian
[Instituto de Investigaciones Biomédicas](#)

Dra. Adela Rodríguez Romero
[Instituto de Química](#)

Ciudad de México. Noviembre, 2021



Universidad Nacional
Autónoma de México

Dirección General de Bibliotecas de la UNAM

Biblioteca Central



UNAM – Dirección General de Bibliotecas
Tesis Digitales
Restricciones de uso

DERECHOS RESERVADOS ©
PROHIBIDA SU REPRODUCCIÓN TOTAL O PARCIAL

Todo el material contenido en esta tesis esta protegido por la Ley Federal del Derecho de Autor (LFDA) de los Estados Unidos Mexicanos (México).

El uso de imágenes, fragmentos de videos, y demás material que sea objeto de protección de los derechos de autor, será exclusivamente para fines educativos e informativos y deberá citar la fuente donde la obtuvo mencionando el autor o autores. Cualquier uso distinto como el lucro, reproducción, edición o modificación, será perseguido y sancionado por el respectivo titular de los Derechos de Autor.

Asignación de jurado



CGEP/PMDCB/1644/2021
Asunto: jurado de examen

López Cortés Georgina Ivette
Estudiante de Doctorado en Ciencias Bioquímicas
Presente

Los miembros del Subcomité Académico en reunión ordinaria del 20 de septiembre del presente año, conocieron su solicitud de asignación de **JURADO DE EXAMEN** para optar por el grado de **Doctorado en Ciencias**, con la réplica de la tesis "Evolución molecular de la proteína Spike del SARS-CoV2", dirigida por el Dr. Ortega Soto Enrique.

De su análisis se acordó nombrar el siguiente jurado integrado por los doctores:

| | |
|------------|------------------------------------|
| PRESIDENTE | Piñero Dalmau Daniel Ignacio |
| VOCAL | Becerra Bracho Arturo Carlos II |
| VOCAL | Fierro González Nora Alma |
| VOCAL | Fernández Velasco Daniel Alejandro |
| SECRETARIO | Segovia Forcella Lorenzo Patrick |

A LOS MIEMBROS DEL JURADO:

Es obligación de los tutores de este programa participar en éstas y otras actividades académicas encomendadas por nuestro Comité Académico. Sin embargo, en caso de que tenga un impedimento académico o de salud para cumplir con esta encomienda, es muy importante contar con su respuesta (Formato anexo) en un plazo no mayor a una semana.

Tome en cuenta que usted tiene **30 días hábiles** para emitir su voto con las rondas de revisión que considere necesarias.

Sin otro particular por el momento, aprovecho la ocasión para enviarle un cordial saludo.

Atentamente
"POR MI RAZA, HABLARÁ EL ESPÍRITU"
Cd. Universitaria, Cd. Mx., a 20 de septiembre de 2021
COORDINADORA

Dra. Claudia Lydia Treviño Santa Cruz

Contacto: mdcba@posgrado.unam.mx Tel. 55-5623-7006



Votos aprobatorios



DEPARTAMENTO DE
ECOLOGÍA EVOLUTIVA

Lic. Diana González Nieto
Directora de Certificación y Control Documental
Dirección General de Administración Escolar
P R E S E N T E

17 de noviembre de 2021

Me es grato comunicarle que después de revisar cuidadosamente y de discutir mis sugerencias y correcciones al documento de tesis titulado "Evolución molecular de la proteína Spike del SARS-CoV2", que para obtener el grado de Doctor presenta la alumna Georgina Ivette López Cortés con número de cuenta 516015338, inscrita en el **Doctorado en Ciencias Bioquímicas**, considero que la tesis reúne los requisitos establecidos y por ello emito mi **VOTO APROBATORIO** para que realice la réplica oral.

Agradezco de antemano la atención que se sirva prestar a la presente.

Atentamente

Dr. Daniel Ignacio Piñero Dalmau



Facultad de Ciencias, UNAM
Depto. De Biología Evolutiva
Email: abb@ciencias.unam.mx
21 de noviembre de 2021

Lic. Diana González Nieto
Directora de Certificación y Control Documental
Dirección General de Administración Escolar
UNAM
Presente

Me es grato comunicarle que después de revisar cuidadosamente y de discutir mis sugerencias y correcciones al documento de tesis titulado "Evolución molecular de la proteína Spike del SARS-CoV2", que para obtener el grado de Doctor presenta la alumna Georgina Ivette López Cortés con número de cuenta 516015338, inscrita en el **Doctorado en Ciencias Bioquímicas**, considero que la tesis reúne los requisitos establecidos y por ello emito mi **VOTO APROBATORIO** para que realice la réplica oral.

Agradezco de antemano la atención que se sirva prestar a la presente.

Atentamente,

Una firma manuscrita en tinta que parece ser "A. Becerra Bracho", escrita sobre una línea horizontal.

Dr. Arturo Carlos II Becerra Bracho

04/11/2021

Lic. Diana González Nieto
Directora de Certificación y Control Documental
Dirección General de Administración Escolar
UNAM
P r e s e n t e

Me es grato comunicarle que después de revisar cuidadosamente y de discutir mis sugerencias y correcciones al documento de tesis titulado " Evolución molecular de la proteína Spike del SARS-CoV2 ", que para obtener el grado de Doctor presenta la alumna Georgina Ivette López Cortés con número de cuenta 516015338, inscrita en el **Doctorado en Ciencias Bioquímicas**, considero que la tesis reúne los requisitos establecidos y por ello emito mi **VOTO APROBATORIO** para que realice la réplica oral.

Agradezco de antemano la atención que se sirva prestar a la presente.

Atentamente,



Dra. Nora Alma Fierro González

29/10/2021

Lic. Diana González Nieto
Directora de Certificación y Control Documental
Dirección General de Administración Escolar
UNAM
P r e s e n t e

Me es grato comunicarle que después de revisar cuidadosamente y de discutir mis sugerencias y correcciones al documento de tesis titulado " Evolución molecular de la proteína Spike del SARS-CoV2.", que para obtener el grado de Doctor presenta la alumna Georgina Ivette López Cortés con número de cuenta 516015338, inscrita en el **Doctorado en Ciencias Bioquímicas**, considero que la tesis reúne los requisitos establecidos y por ello emito mi **VOTO APROBATORIO** para que realice la réplica oral.

Agradezco de antemano la atención que se sirva prestar a la presente.

Atentamente,



Dr. Daniel Alejandro Fernández Velasco

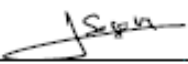
dd/mm/aaaa

Lic. Diana González Nieto
Directora de Certificación y Control Documental
Dirección General de Administración Escolar
UNAM
P r e s e n t e

Me es grato comunicarle que después de revisar cuidadosamente y de discutir mis sugerencias y correcciones al documento de tesis titulado "Evolución molecular de la proteína Spike del SARS-CoV2", que para obtener el grado de Doctor presenta la alumna Georgina Ivette López Cortés con número de cuenta 516015338, inscrita en el **Doctorado en Ciencias Bioquímicas**, considero que la tesis reúne los requisitos establecidos y por ello emito mi **VOTO APROBATORIO** para que realice la réplica oral.

Agradezco de antemano la atención que se sirva prestar a la presente.

Atentamente,



Dr. Lorenzo P. Segovia Forcella

Agradecimientos

Al programa de Maestría y Doctorado en Ciencias Bioquímicas de la UNAM por permitir la formación y exigencia a los alumnos, siempre con el apoyo para pulirnos en la excelencia.

Mis estudios de doctorado se lograron con la ayuda de la beca otorgada por CONACYT (No. CVU 699886), con el fin de dedicarme de tiempo completo a la realización de éstos.

El Programa de Apoyo a los Estudios de Posgrado (PAEP) de la UNAM aportó económicamente para la participación en el congreso de inmunología organizado por la Sociedad Mexicana de Inmunología, realizado en Cancún, México en Mayo del 2018.

La investigación en el Laboratorio del Dr. Enrique Ortega fue financiada con los donativos de DGAPA-PAPIIT-UNAM (IN208320) y de CONACyT-Ciencia de Frontera (87163).

Agradecimientos a título personal

Al Doctor Enrique Ortega Soto por fungir como tutor principal durante este periodo y orientarme a lo largo de este proceso. Por su rigurosidad, formalidad y ética ante la vida me han enseñado lo valioso de ser un científico integral. Lo respalda su trabajo y una sólida calidad como ser humano. Le agradezco haberme aceptado en su equipo de trabajo, por confiar en mis capacidades y en el desarrollo de las habilidades de sus alumnos. Por promover un ambiente de trabajo de compañerismo y superación.

Al Doctor Marco V. José del Instituto de Investigaciones Biomédicas por asesorar y guiar el trabajo presentado aquí y por apoyar mi proceso de titulación. Agradezco su compromiso en apoyar el potencial de los alumnos para que se desarrollen como investigadores de calidad. Su ejemplo inspira a dedicarse con esfuerzo y honestidad para que el trabajo hable por uno. Estoy muy agradecida por haberme integrado a su equipo para que se abriera una estrecha colaboración y que la suma de los esfuerzos individuales resultara en un equipo multidisciplinario basado en la excelencia.

A las Doctoras Adela Rodríguez Romero y Gohar Gevorkian Markosian por haber formado parte del comité tutorial, en el que sugirieron y criticaron constructivamente mi trabajo, con el fin de mejorarlo. Desde los estudios de Maestría se han tomado el tiempo para escuchar y alentarme a seguir en las mejores decisiones. Agradezco los ánimos y la seriedad de su trabajo que han servido de inspiración.

A la Doctora Claudia Garay Canales por su asistencia técnica mientras estuvo como técnico de laboratorio en el laboratorio del Doctor Enrique Ortega. Le agradezco las críticas y comentarios a mis proyectos e ideas, agradezco sus ánimos y la amistad que me ha brindado desde que llegué al laboratorio en la Maestría.

Al Biólogo Gonzalo Asariel Acero Galindo por la asesoría técnica en las primeras etapas del Doctorado, por su paciencia y dedicación para enseñarme y guiarme en su experiencia y conocimientos.

A todos los miembros del laboratorio del Dr. Enrique Ortega por sus aportaciones, críticas y comentarios a mis proyectos. Gracias por las muchas horas que pasamos juntos que fueron la causa de grandes amistades. A los que pasaron por el laboratorio o aún siguen porque tuve la oportunidad de aprender de ellos tanto en lo profesional como en lo personal.

A los miembros del grupo de Biología Teórica porque me extendieron el brazo para aprender de su área de conocimiento y conocer otras formas de trabajar. Me han apoyado y hemos ejecutado un gran trabajo, dirigido por la experiencia de cada quien.

A los co- autores de este trabajo, Miryam Palacios-Pérez, Gabriel S. Zamudio, Hannya F. Veledíaz, Enrique Ortega y Marco V. José, agradezco el apoyo y su participación en el desarrollo del trabajo principal que sustenta esta tesis.

Agradezco al Doctor Alfonso Islas Rodríguez que fue mi primer mentor y que con su ejemplo aprendí valores esenciales del qué hacer científico. Gracias por las lecciones de vida y las pláticas que me orientaron a tomar las decisiones que me han traído hasta ahora, de las cuales no me arrepiento y me siento muy orgullosa. Gracias por compartir su experiencia, pasiones y cultura, y por siempre le agradezco su apoyo en mi desarrollo y formación.

Agradezco a los profesores que me han inspirado y han contribuido en mi formación porque gracias a su trabajo, puedo sentir esta satisfacción por concluir esta etapa. Agradezco a mis formadores desde el CEDI, la Universidad de Guadalajara y la UNAM.

A mis amigos que en diferentes etapas han sido luz. Ahora varios en diferentes partes del mundo, realizándose y siguiendo sus sueños. Es un alivio contar con ustedes y les deseo siempre lo mejor.

A mi papá y a mi mamá porque su ejemplo ha sido inspiración para disfrutar del trabajo y de la superación personal. Porque gracias a ustedes he aprendido a encontrar satisfacción en los retos y me han acompañado en todos los procesos de mi vida. El consejo más sabio, cálido y honesto siempre va a ser el suyo. A mi familia por todo el cariño, por brindarme

su apoyo y esperanza, porque disfrutan verme sonreír. Y con gusto reconozco que todos han sido testigos del esfuerzo, satisfacción, de la pasión y de la alegría que me ha causado esta etapa.

A mi corasol, te agradezco los ánimos y el infinito apoyo. Has sido la templanza que necesito, el consejo y la dirección. Gracias por escucharme e interesarte en mis intereses, gracias por querer compartir tu vida conmigo.

... A todos les deseo el mayor de los éxitos en cada aspecto de sus vidas.

El trabajo principal presentado en esta tesis fue el resultado de una colaboración con el grupo de Biología Teórica del Instituto de Investigaciones Biomédicas de la UNAM.

Dedico esta tesis a quienes me han enseñado a ser como soy.

Índice

| | |
|---|-----------|
| ASIGNACIÓN DE JURADO | 2 |
| VOTOS APROBATORIOS | 3 |
| AGRADECIMIENTOS | 8 |
| AGRADECIMIENTOS A TÍTULO PERSONAL | 9 |
| ÍNDICE | 12 |
| GLOSARIO DE ABREVIATURAS | 14 |
| RESUMEN | 15 |
| ABSTRACT | 16 |
| INTRODUCCIÓN | 16 |
| MARCO TEÓRICO | 17 |
| SARS-CoV-2 | 17 |
| 1.1 Ciclo de replicación | 18 |
| 1.2 La proteína Espícula del SARS-CoV-2 | 20 |
| 1.3 Evolución del SARS-CoV-2 | 26 |
| COVID-19 | 29 |
| 2.1 Epidemia de la COVID-19 | 29 |
| 2.2 Fisiopatología | 30 |
| 2.3 Respuesta inmunológica del huésped ante la COVID-19 | 31 |
| Receptores de coronavirus humanos | 34 |
| 3.1 Aminopeptidasa N | 35 |
| 3.2 Dipeptidil peptidasa IV | 37 |
| 3.3 Enzima Convertidora de Angiotensina II | 38 |
| Teoría Neutral de la Evolución Molecular | 39 |
| 4.1 Prueba de neutralidad | 41 |
| JUSTIFICACIÓN | 44 |

| | |
|--|-----------|
| OBJETIVO | 44 |
| General | 44 |
| Particulares | 44 |
| HIPÓTESIS | 45 |
| MÉTODOS | 46 |
| Bases de datos | 46 |
| Modelo de evolución neutral | 46 |
| Análisis filogenético | 47 |
| Análisis estructural | 47 |
| FLUJO DE TRABAJO | 48 |
| RESULTADOS | 49 |
| DISCUSIÓN | 65 |
| CONCLUSIÓN | 75 |
| PERSPECTIVAS | 78 |
| REFERENCIAS | 79 |
| ANEXOS | 86 |
| Información suplementaria | 86 |
| Productos académicos generados durante los estudios de Doctorado | 87 |
| ARTÍCULO ORIGINAL | 88 |

Glosario de abreviaturas

- COVID-19 – Enfermedad causada por el coronavirus 2019
- SARS-CoV-2 – Coronavirus del Síndrome Respiratorio Agudo Severo 2
- SARS-CoV – Coronavirus del Síndrome Respiratorio Agudo Severo
- MERS-CoV – Coronavirus del síndrome Respiratorio del Medio Oriente
- OMS- Organización Mundial de la Salud
- CDC – Centros de Control y Prevención de Enfermedades
- nsp – proteína no estructural
- ORF – Marco de lectura abierto
- RTC - complejo de replicación y transcripción viral
- ERGIC - intermediario entre el retículo endoplásmico y el aparato de Golgi
- SARS2-S – proteína espícula del SARS-CoV-2
- RBD – Dominio de Reconocimiento del Receptor
- RBM - Motivo de Unión al receptor
- S1 - subdominio 1 de la proteína espícula
- S2 – subdominio 2 de la proteína espícula
- NTD - dominio Amino terminal de la subunidad 1 de la proteína espícula
- CTD - dominio Carboxilo terminal de la subunidad 1 de la proteína espícula
- ECA2 – Enzima Convertidora de Angiotensina II (ACE2 en inglés)
- TMPRSS2 – Serin Proteasa Transmembranal 2
- ANPEP- Aminopeptidasa N
- DPP IV- Dipeptidil peptidasa IV

Resumen

A medida que el SARS-CoV-2 se ha propagado y la pandemia se ha prolongado, el virus ha seguido evolucionando rápidamente, por lo que han surgido nuevas variantes altamente transmisibles que pueden ser motivo de preocupación para la salud pública. Los mecanismos evolutivos que impulsaron esta rápida diversidad no se han desentrañado, pero se parte de la teoría de la evolución molecular neutral como primer mecanismo que permite la evolución del virus. La teoría neutral de la evolución molecular establece que la mayoría de las mutaciones en las secuencias de ácidos nucleicos ocurren aleatoriamente. En seguida, estas pueden fijarse o desaparecer mediante selección purificadora o deriva génica. En este trabajo, se aplicó una prueba de sustitución de aminoácidos para dilucidar el tipo de presiones selectivas ejercidas sobre la proteína Espícula (proteína S) entre proteínas homólogas del género *Betacoronavirus* y se evaluaron las presiones ejercidas sobre los aminoácidos de la proteína obtenida de diferentes aislados clínicos de SARS-CoV-2. Encontramos que de manera particular, las mutaciones que dan origen a Tyr y Asn tienen mayor probabilidad de ser fijadas mientras que aminoácidos como la His y la Arg tienen menor probabilidad de fijarse tanto en la proteína completa como en el Dominio de Unión al receptor (RBD por sus siglas en inglés). En cuanto al fenómeno *in vivo* se demostró que mutaciones que dan origen a Glu, Lys, Phe, y Val tienen la mayor probabilidad de ocurrir en las nuevas partículas virales. En general, la aparición de los aminoácidos que tienen menor probabilidad de fijarse en relación con la del control neutral en sitios nuevos, son considerados como seleccionados negativamente, influyen en la estabilidad y/o funcionalidad de la proteína. Nuestros resultados muestran que la mayoría de las mutaciones únicas, ya sea para la proteína S del SARS-CoV-2 o de las variantes de preocupación para la salud, han sido seleccionadas por mecanismos relacionados con la evasión del sistema inmunológico, el aumento de la patogenicidad del virus o alteraciones en las interacciones proteína-proteína con proteínas del huésped. En conjunto, todas estas fuerzas han dado forma a la proteína S del coronavirus causante de la COVID-19.

Abstract

As the SARS-CoV-2 has spread and the pandemic has dragged on, the virus continued to evolve rapidly resulting in the emergence of new highly transmissible variants that can be of public health concern. The evolutionary mechanisms that drove this rapid diversity are not well understood but neutral evolution should open the first insight. The neutral theory of molecular evolution states that most mutations in the nucleic acid sequences are random and they can be fixed or disappeared by purifying selection. Herein, we performed a neutrality test to better understand the selective pressures exerted over neutral mutations in the SARS-CoV-2 Spike protein from homologue proteins of *Betacoronavirus*, as well as to the Spike proteins from human clinical isolates of the virus. Specifically, Tyr and Asn have the highest occurrence rates on the Receptor Binding Domain (RBD) and in the overall sequence of spike proteins of *Betacoronavirus*, whereas His and Arg have lower occurrence rates. The evolutionary phenomenon of SARS-CoV-2 occurring until the publication of this work shows that Glu, Lys, Phe, and Val have the highest probability of occurrence in the emergent viral particles. Amino acids that have higher occurrence than the expected by the neutral control, are favorable and are fixed in the sequence while the ones that have lower occurrence than expected, influence the stability and/or functionality of the protein. Our results show that most unique mutations either for SARS-CoV-2 or its variants of health concern are under selective pressures, which are related either to the evasion of the immune system, increasing the virus' fitness or altering protein – protein interactions with host's proteins. All these forces have shaped the Spike protein and the continually evolving variants.

Introducción

El SARS-CoV-2 ha causado innumerables contagios y muertes a lo largo de todo el mundo y ha polarizado la generación de conocimiento hacia el estudio de la biología del virus y el desarrollo de estrategias que pudieran frenar la propagación de este o bien reducir la severidad de los síntomas junto con la mortalidad. Este trabajo detalla un modelo evolutivo basado en mutaciones neutrales que pudiera servir para comprender el origen

del virus y los cambios que siguen ocurriendo y que han generado las variantes de interés. Importantemente para discutir sobre las mutaciones que han afectado la unión con el receptor, se definió el tipo de presión de selección en todas las mutaciones del RBD de SARS2-S. Así junto con otros parámetros se puede explicar el patrón evolutivo de la proteína. Se discutieron las mutaciones de las variantes de SARS-CoV-2 en la proteína S para observar el fenómeno evolutivo en tiempo real.

Marco teórico

SARS-CoV-2

El agente causal de la enfermedad COVID-19, el SARS-CoV-2, es un coronavirus (CoV) del orden Nidovirales, de la familia Coronaviridae y del género *Betacoronavirus*. Es un virus envuelto cuyo genoma es una molécula de ARN de cadena sencilla con polaridad positiva, de alrededor de 30,000 bases ^[1]. La estabilidad del genoma, es alcanzada gracias a que tiene una estructura cap en el extremo 5' y se encuentra poliadenilado en su extremo 3'. Además, el virus posee una exoribonucleasa (nsp14-ExoN) que le permite corregir errores de la transcripción, reduciendo así la tasa de mutación de la ARN polimerasa dependiente de ARN ^[2]. La cadena de ARN codifica para 25 proteínas no estructurales (nsp; non-structural protein) y cuatro proteínas estructurales: La de membrana (M; Membrane), la de la envoltura (E; Envelope), la nucleoproteína (N; Nucleoprotein) que tiene interacciones directas con el ARN y finalmente la proteína S (S; spike) (**Figura- 1.2**) ^[1,3,4]. Los primeros dos tercios del genoma codifican dos marcos de lectura abiertos (ORF; Open Reading Frame) que codifican por dos cadenas peptídicas que son proteolizadas para generar las proteínas involucradas en la transcripción y traducción de las proteínas virales, mientras que el último tercio comprende varios ORFs que codifican las diferentes proteínas estructurales y otras 9 proteínas accesorias ^[4,5].

Evolución molecular de la proteína Spike del SARS-CoV-2

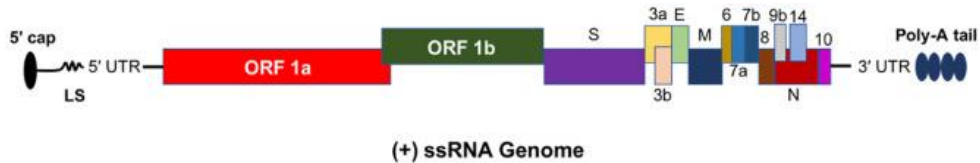


Figura 1.1- Estructura lineal del genoma del SARS-CoV-2. Imagen tomada de Naqvi, 2020.

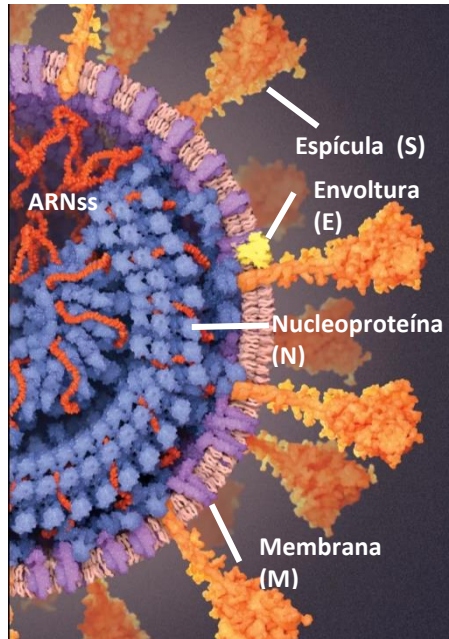


Figura 1.2- componentes estructurales del SARS-CoV-2. El virus envuelto por una bicapa lipídica contiene al genoma de ARN de una cadena que interacciona con Nucleoproteínas. Embebidas en la membrana se encuentran proteínas de Membrana, de Envoltura y las proteínas Espícula que hacen alusión a una corona, característica responsable del nombre de la familia. Imagen tomada de Falconeri V. 2020.

1.1 Ciclo de replicación

El mecanismo de entrada del virus comienza con la unión de la proteína S al receptor celular. Después, diferentes proteasas como la convertasa furina, las serin proteasas TMPRSS- 2 y 3 así como la catepsina- L, son necesarias para el procesamiento y activación del estado de fusión de SARS2-S ^[6]. En el caso de las proteasas furina y TMPRSS2, cortan a la proteína en la membrana dejando expuestos el péptido de fusión y las 2 secuencias de siete aminoácidos (HR), permitiendo que éstos aproximen más la partícula viral hacia la membrana celular hasta fusionar las membranas lipídicas ^[7]. También se ha reportado otra vía de entrada donde el virus es internalizado en vesículas, aquí las proteínas S se procesan por la catepsina- L, permitiendo así la liberación de la cápside al citoplasma ^[8]. Así el virus aprovecha las proteasas del huésped para favorecer la infección. Por lo tanto la interacción con estas proteasas genera una presión de selección

importante para la propagación del virus. La proteína S del SARS-CoV-2 presenta un motivo de corte de furina, el cual ha potenciado la propagación del virus ^[9,10]. Sin embargo, algunos estudios muestran que el sitio de corte de furina no es esencial para la internalización del SARS-CoV-2 ya que otras enzimas pueden llevar a cabo el procesamiento de la proteína S de manera inespecífica.

Una vez internalizada la partícula, los ribosomas del huésped comienzan a traducir el ORF1a y el ORF1b, de forma que rápidamente se sintetizan las proteínas pp1a y pp1ab respectivamente (**Figura 1.3**). Estas cadenas peptídicas son procesadas proteolíticamente por las proteasas virales PLpro y 3CLpro generando 16 proteínas no estructurales (nsp1-nsp16) que conforman al complejo de replicación y transcripción viral (RTC) ^[2]. Algunas de las proteínas nsp participan en la formación de vesículas de doble membrana, las cuales tienen su origen en el retículo endoplásmico ^[11]. Las vesículas de doble membrana sirven como nichos donde se concentran las moléculas necesarias para la replicación, traducción y empaquetamiento de las nuevas partículas virales. Cabe señalar que también evitan la activación del sistema inmunológico ya que al estar los componentes virales al interior de vesículas, éstos no se encuentran expuestos a los receptores del sistema inmune. Dentro de estas vesículas, se lleva a cabo la transcripción de aproximadamente nueve ARNs subgenómicos incluyendo los que se traducen en las proteínas estructurales E, M, S y N. Las vesículas de doble membrana dirigen proteínas virales recién sintetizadas al intermediario entre el retículo endoplásmico y el aparato de Golgi (ERGIC) ^[7,12]. Finalmente las partículas virales recién sintetizadas son transportadas a la membrana plasmática a través de vesículas que son secretadas por exocitosis. Particularmente los *Betacoronavirus* utilizan lisosomas del huésped como medio de transporte a la membrana, en vez de seguir a través del sistema excretor ^[13]. El mecanismo sugerido es la inhibición de las bombas de protones de los lisosomas, evitando su acidificación. De esta forma los viriones logran salir de la célula huésped para continuar con el ciclo infeccioso.

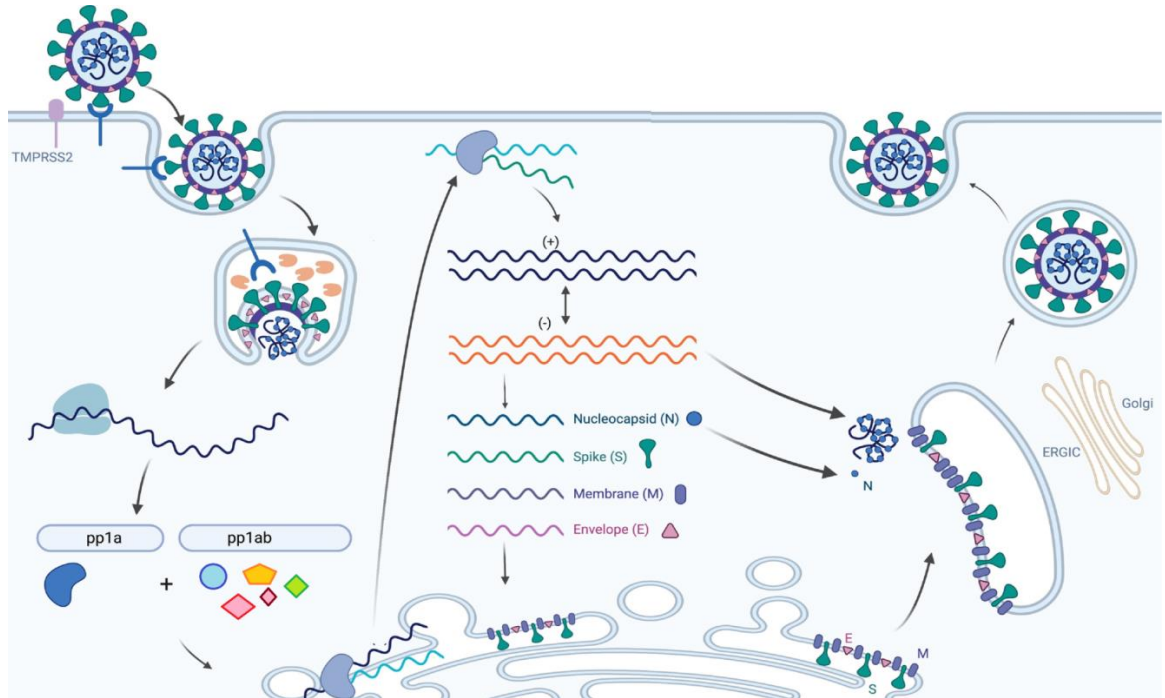


Figura 1.3- Ciclo replicativo del SARS-CoV-2. El virus puede fusionarse directamente en la membrana o introducirse en la célula por medio de vesículas. Las proteasas de membrana (TMPRSS) o en los endosomas (Catepsina L) promueven la fusión de membranas y liberación del genoma. El ARN de cadena sencilla es traducido en las poli-proteínas pp1a y pp1ab que inician la replicación y traducción del genoma en vesículas de doble membrana provenientes del retículo endoplásmico rugoso. Los ARN subgenómicos se traducen en las proteínas accesorias y estructurales del virus que pasan por el ERGIC para ser madurados. Finalmente las proteínas estructurales maduras son utilizadas para empaquetar las nuevas partículas virales que son transportadas a la membrana y exocitadas. Imagen modificada de Harrison, A., Lin, T., Wang, P. 2020.

1.2 La proteína Espícula del SARS-CoV-2

La proteína estructural S del SARS-CoV-2 (SARS2-S) ha sido el foco de atención de muchos estudios pues cumple varias funciones clave para el virus. En primer lugar es responsable de las interacciones iniciales con la célula huésped, en segundo lugar induce la fusión de las membranas, y tercero ha sido modelo de estudios inmunológicos ya que es una proteína antigénica. Es una proteína transmembranal tipo I, es decir su extremo amino terminal se encuentra hacia el exterior de la partícula viral y tiene un solo pase transmembranal. La cola que da hacia el interior es corta, de aproximadamente 40 amino ácidos ^[7]. En la superficie viral, la proteína S forma un complejo homo-trimérico (**Figura 1.4**) ^[14,15] donde los monómeros interaccionan entre sí. Cada monómero está formado por

una sola cadena peptídica que se divide en dos subdominios funcionales: el subdominio 1 (S1) que contiene el extremo Amino terminal (NTD) y el Dominio Carboxilo Terminal (CTD) de la subunidad; en el subdominio 2 (S2) se encuentra la maquinaria de fusión. La secuencia del CTD es la más variable ya que en éste se encuentra el Dominio de Unión al Receptor (RBD; Receptor Binding Domain) que a su vez contiene el Motivo de Unión al Receptor (RBM). En contraste, el S2 es altamente conservado pues se ha especializado en fusionarse a la membrana del huésped, lo que le permitiría internalizar el genoma y la posterior replicación del virus. El S2 contiene al péptido de fusión y dos estructuras con repeticiones de siete aminoácidos que en su conjunto forman la maquinaria de fusión

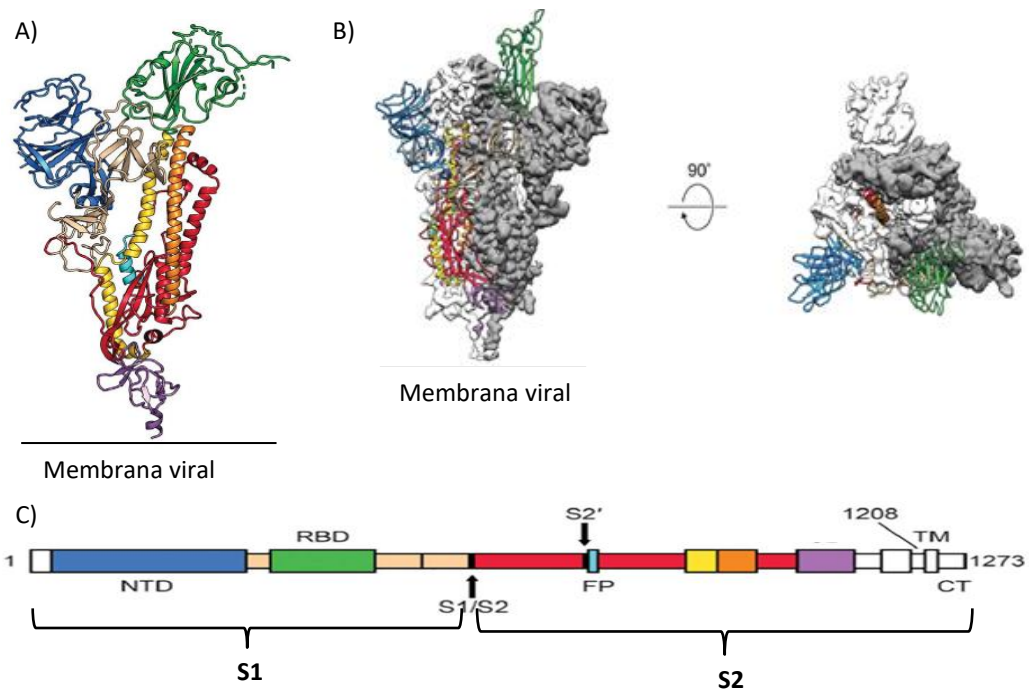


Figura 1.4- Estructura de la proteína S del SARS-CoV-2. A) Estructura tridimensional de la proteína S del causante de la COVID-19 por medio de CryEM con una resolución de 3.5Å. Se observan los motivos con el mismo código de colores que en C. En B) se muestra el trímero tanto en una vista lateral como horizontal por arriba del virus. Uno de los monómeros (en vista de listón y de colores) presenta el RBD hacia arriba, los otros dos monómeros (en blanco y gris con vista superficial) hacia abajo. En C) el NTD coloreado en azul y el RBD en verde forman parte del S1, mientras que el S2 en rojo, contiene al péptido de fusión en azul, las repeticiones de 7 aminoácidos en color amarillo y naranja, la región morada es un asa cercana a la membrana viral. La estructura correspondiente a las regiones blancas en la secuencia lineal (pase transmembranal (TM) y región citoplásmica (CM)), no se resolvieron en este experimento. Estructuras tomadas de Wrapp D, *et al.* 2020.

[15,16]. Además, la proteína SARS2-S cuenta con dos sitios susceptibles a cortes proteolíticos, que liberan la región S1 de la S2 y exponen el péptido de fusión para que se ancle en la membrana plasmática [17]. De tal forma, que cada uno de los monómeros es susceptible a cambios conformacionales donde el RBD se direcciona hacia “arriba” o hacia “abajo”, siendo la primera conformación la que interacciona con el receptor [18].

Diferentes estudios comprobaron que la proteína SARS2-S, se une principalmente a la proteasa Enzima Convertidora de Angiotensina 2 (ECA2) a través del RBD [6]. ECA2 es una proteína transmembranal implicada en la regulación de péptidos vaso-activos por lo que la desregulación de la proteasa altera el sistema Renina- Angiotensina- Aldosterona. Los sustratos más estudiados son la Angiotensina II y la Angiotensina I que son hidrolizados en Angiotensina 1-7 y angiotensina 1-9 respectivamente. En el humano ECA2 es expresada en células alveolares tipo II, células secretoras nasales, keratinocitos del esófago, células epiteliales del sistema gastrointestinal y cavidad oral, células pancreáticas β , endotelio venoso y arterial, células epiteliales del riñón y células del corazón [19]. Es por ello que el virus potencialmente puede infectar una variedad de tejidos y generar fallas sistémicas, mostrando tropismo por estas células.

Un estudio donde se expresó la enzima ECA2 de diferentes especies en células humanas, demostró que la proteína S de SARS-CoV-2 tiene la capacidad de interactuar también con la proteína ECA2 homóloga de murciélago, civeta y cerdo aunque no con la de ratón [3]. Por lo que se entiende que estas especies deben tener conservación en la región de la interface de ECA2. Se sabe que otros coronavirus tienen la capacidad de unirse a la proteasa ECA2, entre ellos, otros *Betacoronavirus* de murciélago similar al SARS (SL-CoV), el SARS-CoV, así como el *Alphacoronavirus* hCoV-NL63 [14,20]. De manera particular la afinidad de la proteína SARS2-S a ECA2 es mayor que la de SARS-CoV al mismo receptor [21,22]. Se han propuesto los aminoácidos responsables de la interacción de ambos virus y del receptor [16,23–26].

La proteína S es altamente conservada entre los CoVs, sin embargo existen diferencias importantes, principalmente en el RBD que reconoce al receptor de la célula

huésped y en la susceptibilidad al procesamiento proteolítico. Otra diferencia que desde un inicio llamó la atención en SARS2-S, radica en la inserción de una secuencia polibásica en la posición 682, correspondiente al sitio de unión de las subunidad 1 y 2 de la proteína, la cual es reconocido por la proteasa furina ^[7]. Este corte, permite la exposición de un motivo de unión a la proteína membranal Neuropilina, la cual también contribuye en la internalización del virus ^[27,28]. Existen otros virus que también presentan un motivo polibásico susceptible al corte de la proteasa furina, sin embargo, han llegado a presentarlo por sustituciones más que por inserción. Sin importar el origen, este motivo provee ventaja sobre las secuencias ancestrales ya que permite que aumente la patogenicidad de las partículas virales ^[10].

El procesamiento proteolítico de la proteína S es necesario para la fusión, pues gracias a que se separa la S1 de la S2, se permite la exposición del péptido de fusión, y así el acercamiento a la membrana plasmática (**Figura 1.5**) ^[18,29]. Las proteínas S presentan dos sitios de corte, en el primero se separan ambas subunidades, aunque se pueden mantener unidas por interacciones no covalentes ^[18]. La interacción entre ambas subunidades contribuye en mantener la estabilidad de la S2. Cuando el RBD interacciona con el receptor, ambas subunidades pueden separarse. Posteriormente la subunidad que se mantiene en contacto con la membrana viral, es nuevamente escindida de manera tal que expone el péptido de fusión ^[17]. Esto ocasiona el cambio conformacional de la S2 que implica la extensión de las repeticiones de siete amino ácidos en forma de hélices alfa que facilitan el anclaje del péptido de fusión en la membrana plasmática. En seguida las hélices alfa vuelven a su conformación original. Como resultado las membranas se acercan hasta fusionarse y formar un poro ^[7].

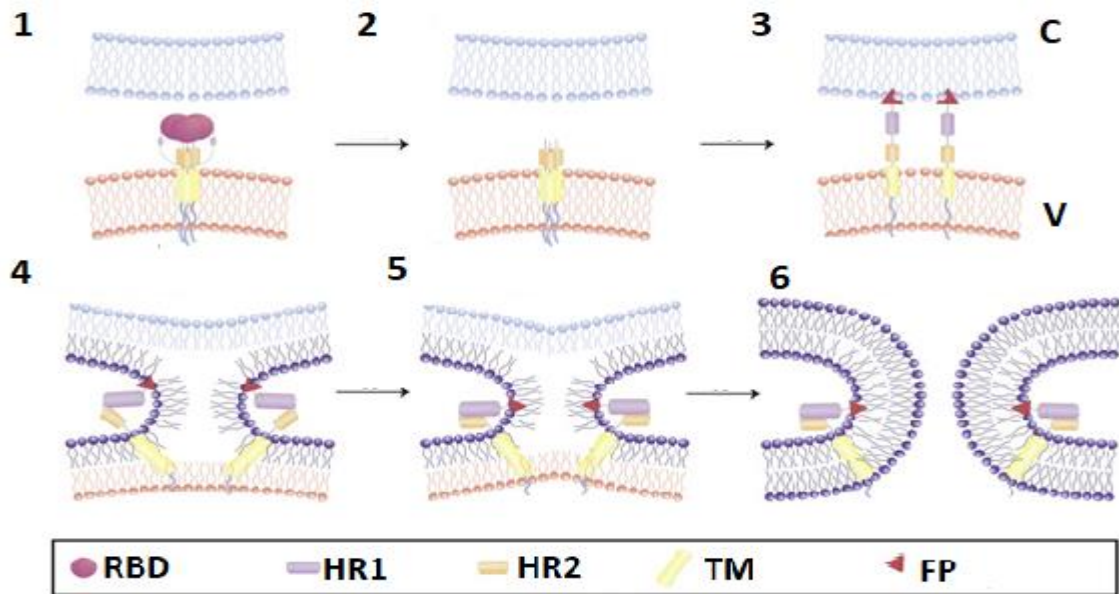


Figura 1.5- Mecanismo de fusión de las proteínas S de los coronavirus. La proteína S se preactiva cuando las proteasas hacen sus respectivos cortes, y se remueve la S1. Lo que permite que haya un cambio conformacional que aproxime a las membranas y se ancle el péptido de fusión (FP) en la membrana celular (C). En seguida la conformación original vuelve de forma que una de las capas de fosfolípidos es interrumpida para fusionarse con la membrana viral (V). Finalmente se forma un poro que permite el paso del genoma viral al citoplasma. Imagen modificada de Tang, et al. 2020.

Los *Betacoronavirus* aprovechan diferentes proteasas de los huéspedes que se expresan en diferentes compartimentos celulares para llevar a cabo el procesamiento de la proteína S. Consecuentemente, la proteólisis puede ocurrir en la membrana plasmática, durante el empaquetamiento de nuevas partículas virales, en vesículas o en el medio extracelular ^[29]. En el caso de la proteína SARS2-S, la furina escinde en S1/S2' a la proteína S recién sintetizada tanto en el ERGIC como en la membrana, de forma tal que la preactiva, posteriormente las serin proteasas TMPRSS- 2 y 3, o la enzima catepsina- L realizan el otro corte desde la membrana plasmática o desde los endosomas, respetivamente. Cabe señalar que el motivo de corte de furina incrementa la infectividad, sin embargo no es crucial para la infección ^[30].

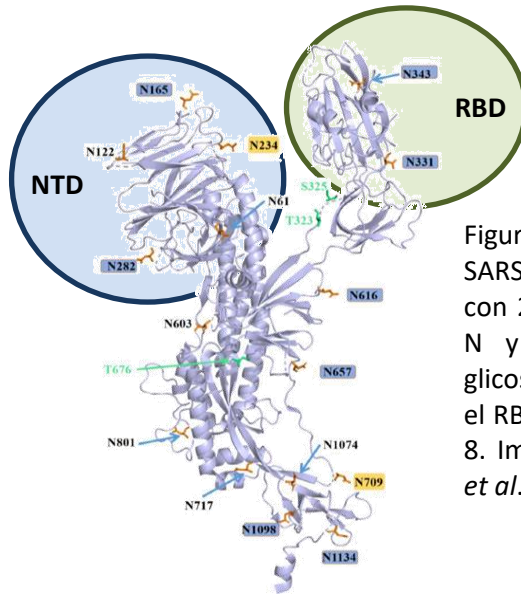


Figura 1.6- La proteína S del SARS-CoV-2 es una glicoproteína con 22 sitios de glicosilación tipo N y tres tipo O. Solo 2 N-glicosilaciones se encuentran en el RBD mientras que en el NTD se encuentran 8. Imagen tomada de Xiaohui, Z. *et al.* 2021.

Una característica más de la proteína SARS2-S es su alto grado de glicosilación (**Figura 1.6**). Hasta ahora se han descrito 22 N- glicosilaciones y 3 O- glicosilaciones y se ha descrito que las funciones principales de dichas glicosilaciones son contribuir a evadir al sistema inmunológico y al plegamiento adecuado de la proteína ^[31,32]. Cabe aclarar que la alta glicosilación es una característica conservada y que incluso hay coronavirus que se unen al receptor del huésped a través de las glicosilaciones del NTD de la S1, lo que implica que estos virus utilizan receptores de unión a carbohidratos ^[33]. Además de la ECA2, SARS2-S se une a otros receptores para mediar la infección a las células del huésped, entre ellos algunos que se unen a carbohidratos ^[34,35] (**Figura 1.7**).

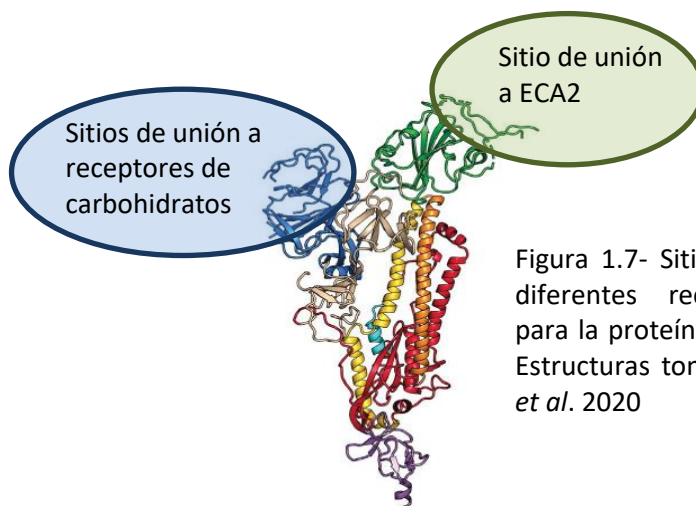


Figura 1.7- Sitios de unión de los diferentes receptores descritos para la proteína S del SARS-CoV-2. Estructuras tomadas de Wrapp D, *et al.* 2020

1.3 Evolución del SARS-CoV-2

Existen cuatro géneros de CoVs de los cuales solo algunos virus pertenecientes a los *Alphacoronavirus* y a los *Betacoronavirus* pueden infectar a humanos ^[36]. Estos dos géneros tienen un ancestro común que infectaba a los murciélagos, mientras que los *Gammacoronavirus* y *Deltacoronavirus* comparten un ancestro común que infectaba a aves ^[36]. La infección de humanos con los diferentes CoVs es el resultado de transmisiones zoonóticas ^[37,38]. Por lo tanto los CoVs de humano pueden estar directamente relacionados con un CoV de murciélago o con otros CoVs que infectan a otro mamífero. Hasta ahora se han identificado siete coronavirus que infectan a los humanos y causan principalmente enfermedades respiratorias e intestinales pero solo tres de ellos han representado amenazas epidémicas: el SARS-CoV, el MERS-CoV y el SARS-CoV-2 ^[26,39,40].

Desde la epidemia del SARS-CoV en el 2003, se secuenciaron una gran cantidad de muestras para identificar el posible ancestro del SARS-CoV. En ese ejercicio, se determinó que los murciélagos son el reservorio natural de muchos coronavirus. La invasión del humano a hábitats o bien la migración de animales infectados con los CoVs a sitios donde habita el humano, generan condiciones propicias para que ocurra una zoonosis. Sin embargo en cualquiera de los casos, se lleva a cabo un proceso de especiación que le permite al virus ancestral adaptarse a los nuevos huéspedes. Hasta ahora se desconoce el mecanismo por el cual el virus ancestral que era un tipo SARS de murciélago, llegó a especiarse dando origen a un virus distinto que infecta a los humanos ^[41].

La secuenciación del genoma de SARS-CoV-2 se hizo pública el 12 de enero del 2020, a sólo unos días de los primeros reportes de pacientes infectados. Este hecho permitió realizar análisis filogenéticos del virus para conocer su origen. El genoma del coronavirus de murciélago (*Rhinolophus affinis*) RaTG13 comparte el 96% con el genoma del SARS-CoV-2 ^[20]. La muestra de este coronavirus RaTG13, fue aislada en el 2013 a partir de heces de un individuo. Aparte de los murciélagos, se han identificado pocos huéspedes portando CoVs tipo SARS, por lo que se cree que el ancestro del SARS-CoV-2 debió de

haber infectado a estos murciélagos. Sin embargo, cabe señalar que Lam y colaboradores analizaron muestras congeladas de 18 pangolines malayo (*Manis javanica*) que se transportaban de contrabando en el 2018, y encontraron un coronavirus (Pn-CoV) que compartía el 92% de su genoma con el del SARS-CoV-2 ^[20,42,43]. Lo más interesante es que unas de las muestras compartían el 97% de identidad con el RBD de la proteína SARS2-S, mientras que el CoV RaTG13 comparte el 89% ^[20]. Estos hallazgos han permitido la hipótesis de que pudo haber ocurrido un evento de recombinación en algunos pangolines, los cuales transportaron el virus recombinante a la población humana y el virus se adaptó a esta nueva especie, dando origen al SARS-CoV-2. De hecho el S2 del SARS2-S es casi idéntico (100% y 99% similar al de RaTG13 y Pn-CoV respectivamente), mientras que las mayores diferencias se encuentran en el S1 y principalmente en el RBD de la secuencia del SARS-CoV-2 ^[20]. La conservación del S2 se entiende por la especialización de la maquinaria de fusión, y la variabilidad de S1 es en respuesta a la adaptación a los huéspedes.

Siguiendo el análisis filogenético, también se demostró que los otros coronavirus que han causado epidemias en los últimos 20 años, el SARS-CoV y el MERS-CoV, se encuentran más alejados filogenéticamente. A pesar de que el SARS-CoV pertenece al mismo linaje, comparte únicamente el 72% de identidad con el SARS-CoV-2 ^[20] sin embargo, es interesante ya que se unen al mismo receptor. Por lo tanto los estudios comparativos entre las proteínas de ambos virus han servido para identificar las características fundamentales que permitieron la adaptación al humano. En el caso de la proteína S de SARS-CoV, se comparte el 88% de identidad con la de SARS-CoV-2 y se desvía 3.8 Å de la estructura de la proteína, en RMSD y alineando 959 Cα ^[15]. Funcionalmente también tiene pocas diferencias ya que el SARS-S tiene una constante de afinidad con el receptor de 5nM, mientras que en el SARS2-S de 1.2 nM ^[22].

Dado el papel esencial de la proteína S en el ciclo de replicación del virus, se supone que ha sufrido fuertes presiones evolutivas para propiciar la internalización a células de una especie diferente, lo que conllevaría la adaptación del virus para infectar al huésped. Estas presiones deben de ser distintas en cada subunidad de la proteína según la especialización funcional de esta. Por ejemplo, el RBD debe una de las regiones de mayor

mutabilidad justamente porque la unión al receptor es crucial para adaptarse al receptor de las células del huésped. Incluso el trabajo de Lu y colaboradores proponen que las proteínas S de SARS-CoV-2, SARS-CoV y MERS-CoV permiten el reconocimiento de receptores homólogos de varias especies mientras ocurran solo algunas mutaciones ^[3,38]. Esto sugiere una gama de posibles huéspedes, así como la relativamente fácil adaptabilidad a otras especies.

Por otro lado, el surgimiento de las nuevas variantes de SARS-CoV-2 es resultado de la evolución continua del virus. A medida que se propaga el virus, la probabilidad de mutación y de selección de las variantes mejor adaptadas, se hace más evidente. Una variante es considerada como tal, cuando tiene al menos una mutación con respecto a la secuencia original; son consideradas de interés cuando epidemiológicamente se vuelven más representativas que otras variantes. Posteriormente esas variantes son evaluadas en cuanto a su transmisibilidad, la gravedad de la enfermedad que ocasionan y si son susceptibles a las vacunas desarrolladas, estos parámetros definen si una variante es de preocupación para la salud. Actualmente de todas las variantes, se consideran cuatro como variantes de preocupación para la salud: Alfa, Beta, Gamma y Delta. Desde sus inicios, diferentes organismos desarrollaron nomenclaturas basándose en las relaciones filogenéticas. Sin embargo, el número de variantes es cada vez mayor y no siempre es posible definir su origen monofilético. Es por ello que la OMS designó a las variantes de preocupación y a las de interés con letras griegas **(Tabla 1.1)** (<https://www.who.int/en/activities/tracking-SARS-CoV-2-variants>). Además ya se ha hecho una lista de mutaciones de preocupación ya que éstas están implicadas en el aumento de la patogenicidad del virus. Por lo tanto las variantes con esas mutaciones se han vuelto las de mayor representación en la población humana.

| Tabla 1.1- Nomenclaturas de las variantes de interés y de preocupación para la salud. | | | | |
|---|--------------------|---------------------|-------------------------|-------------------------|
| Nombre | Nomenclatura Pango | Nomenclatura GISAID | Nomenclatura Nextstrain | Aislada por primera vez |
| Alfa | B.1.1.7 | GRY (GR/501Y.V1) | 20I/S:501Y.V1 | Reino Unido, Sep-2020 |

| | | | | |
|---------|---------------------|------------|---------------|---------------------------------|
| Beta | B.1.351 | GH/501Y.V2 | 20H/S:501Y.V2 | Sudáfrica , May-2020 |
| Gamma | P.1 | GR/501Y.V3 | 20J/S:501Y.V3 | Brasil, Nov-2020 |
| Delta | B.1.617.2 | G/452R.V3 | 21A/S:478K | India, Oct-2020 |
| Epsilon | B.1.427/ B.1.429 | GH/452R.V1 | 20C/S.452R | USA, Mar-2020 |
| Zeta | P.2 | GR | 20B/S.484K | Brazil, Apr-2020 |
| Eta | B.1.525 | G/484K.V3 | 20A/S484K | Multiple countries, Dec-2020 |
| Theta | P.3 | GR | 20B/S:265C | Philippines, Jan-2021 |
| Iota | B.1.526 | GH | 20C/S:484K | USA, Nov-2020 |
| Kappa | B.1.617.1 | G/452R.V3 | 21A/S:154K | India, Oct-2020 |

COVID-19

2.1 Epidemia de la COVID-19

La epidemia de la COVID-19 comenzó oficialmente el 31 de diciembre de 2019 cuando la Organización Mundial de la Salud (OMS) fue notificada de la acumulación de casos de neumonía atípica, lo que ocasionó generar un estado de alerta de carácter sanitario. Unos días después, se dio a conocer que el agente etiológico era un nuevo coronavirus con características similares al coronavirus que ocasionó la epidemia del 2002-2003, el SARS-CoV. Fue el 12 de enero del 2020 que China hizo pública la secuencia del genoma del virus, lo que permitió la explosión en la investigación sobre el SARS-CoV-2 y la generación de vacunas contra este. El 30 de enero de 2020, la OMS declaró que era una Emergencia de Riesgo Internacional pues se reportaban pacientes en diferentes partes del mundo y finalmente el 11 de marzo del 2020 la OMS declaró que el acontecimiento tenía características de una pandemia. La gravedad de la pandemia ocasionó que durante los siguientes meses se destinaran grandes esfuerzos tanto económicos como de recursos humanos para desarrollar vacunas en tiempos récord.

Como era de esperarse, las pruebas clínicas fueron filtrando las vacunas adecuadas para su comercialización y aplicación. Actualmente (octubre del 2021) las cifras de contagios y muertes siguen aumentando aunque en un ritmo menos acelerado, ya que las campañas de vacunación en algunos países han resultado exitosas para frenar la propagación del virus. Sin embargo, no toda la población mundial ha recibido esta protección, por lo que la lucha continúa para evitar que el virus siga mutando y generando nuevas variantes más resistentes.

2.2 Fisiopatología

El SARS-CoV-2 infecta principalmente las vías respiratorias inferiores, al igual que el SARS-CoV y el MERS-CoV, ocasionando neumonía, fiebre, y otros síntomas. Según los reportes epidemiológicos, el periodo de incubación del virus es de entre 3 a 5 días, luego se van manifestando los primeros síntomas y no es sino hasta los 7 a 10 días posteriores a la infección que se manifiestan los síntomas de mayor intensidad. El progreso e intensidad de la enfermedad está sujeto a varios factores, entre ellos la respuesta inmunológica, el estado nutricional y la presencia de comorbilidades ^[19,44].

Los síntomas de la COVID-19 son diversos y van desde los leves a aquellos que requieren atención médica de urgencia. Los Centros de Control y Prevención de Enfermedades (CDC) de Estados Unidos han declarado que los principales son la tos seca, la dificultad para respirar, alta temperatura y escalofríos, dolor muscular, vómito y diarrea, así como la pérdida del olfato y del gusto (**Figura 1.8**). En un porcentaje de pacientes, la infección puede desatar una reacción inflamatoria intensa que lleva a la liberación excesiva de citocinas y de proteínas hepáticas que son marcadores de la inflamación. Así pues, el desarrollo de la enfermedad grave se sustenta en 4 eventos principales: la citotoxicidad propia por la infección del virus a las células, la desregulación del sistema Renina- Angiotensina- Aldosterona, las coagulopatías y la desregulación del sistema inmunológico ^[19]. Estos cuatro ejes ocasionan la falla multi-orgánica que puede ocasionar otras secuelas incluyendo la muerte ^[45].

| Leves | Medianos | Graves |
|--------------------|----------------------------|---------------------------------------|
| Fiebre | Nauseas | Dificultad respiratoria |
| Tos seca | Diarrea | Confusión |
| Fatiga | Conjuntivitis | Presión en el pecho |
| Dolores musculares | Pérdida del olfato y gusto | Coloración azul de cara |
| Dolor de garganta | | Incapacidad para mantenerse despierto |
| Dolor de cabeza | Pérdida del apetito | |

Figura 1.8- principales síntomas de la COVID-19. Los diferentes síntomas se pueden manifestar en distintas intensidades en cada individuo.

Actualmente no se han descifrado los factores que determinan la gravedad de la enfermedad, incluso hay estudios que estiman que la carga viral con la que algunos pacientes asintomáticos se infectaron, pudiera ser similar a la que infectó a algunos pacientes sintomáticos ^[46]. De hecho, se debate si todos los pacientes podrían tener la misma tasa de transmisibilidad del virus ^[46]. En ambos grupos de pacientes se han aislado células T de memoria específica a las proteínas virales de SARS-CoV-2 ^[47], lo que prueba que debió haber habido la activación del sistema inmune innato y adaptativo. Sin embargo se mantiene la pregunta de por qué no se manifestaron los síntomas típicos de una infección viral. Una de las hipótesis que pudiera explicar el fenómeno es la reactividad cruzada con otros coronavirus, sin embargo sigue siendo inconclusa. Se ha denominado que los pacientes asintomáticos son el talón de Aquiles en la lucha contra la transmisión del SARS-CoV-2 ^[48]. Esto justifica los estudios que describan la inmunología de los pacientes asintomáticos para determinar los factores que pueden influenciar en la gravedad de la enfermedad. Por ahora lo que es claro es que las comorbilidades como enfermedades metabólicas, hipertensión y daños al corazón influyen en este balance de la respuesta inmunológica.

2.3 Respuesta inmunológica del huésped ante la COVID-19

La infección por SARS-CoV-2 ocasiona un cuadro de respuesta inflamatoria. A inicios de la pandemia, la mayoría de evidencia de la fisiopatología y la respuesta inmunológica se recabó en pacientes graves o en tejidos *post-mortem*. Por ejemplo, se

encontraron Trampas Extracelulares de Neutrófilos (NETs), mediadores inflamatorios lipídicos, proteasas y Especies Reactivas de Oxígeno (ROS en inglés), la activación del inflamasoma NLRP3 en monocitos inflamatorios y los altos niveles séricos de IL-6, IL-18, lactato deshidrogenasa, la proteína C- reactiva y Dímero- D ^[49-51]. Con el paso del tiempo se ha determinado que inicialmente la infección provoca una inflamación local que recluta células inflamatorias además de otras que regulen la inflamación. Sin embargo en algunos pacientes esta no es regulada lo que promueve el aumento de la permeabilidad vascular causando edema que por un lado compromete el intercambio gaseoso y por otro facilita la migración de células inflamatorias al tejido pulmonar. Estas células ocasiona una liberación exacerbada de citocinas pro- inflamatorias que promueven la activación y reclutamiento de aún más células inmunológicas que dañan el tejido por sus diferentes funciones efectoras (ROS, enzimas proteolíticas, NETs). Ya en la circulación, las citocinas y las células inmunológicas activadas ocasionan diferentes fallas multi-orgánicas e incluso coagulopatías ^[52,53]. Aparte, la reparación del tejido suele estar mediado por fibroblastos más que por la población original, de manera que se reduce la capacidad pulmonar.

Las células infectadas por SARS-CoV-2 suelen morir por piroptosis, este tipo de muerte celular ocasiona la liberación de más partículas virales recién sintetizadas aparte de Patrones Moleculares Asociados a Daño (DAMPs en inglés) los cuales activan a los macrófagos alveolares y a otras células cercanas como neumocitos y células endoteliales. Todas estas células secretan una variedad de citocinas pro-inflamatorias, incluyendo MIP1 α , MIP1- β , IL-6, IL-8, IL-1 β , TGF- β , TNF y MCP1 que aumentan la permeabilidad vascular y reclutan otras células inmunológicas ^[52,54]. Los mecanismos propios de las células inmunológicas inflamatorias y el edema causan la hiperinflamación y el síndrome respiratorio agudo (ARDS, en inglés) que conlleva a la falla respiratoria, una mayor cantidad de muerte celular y la fibrosis ^[55].

Además, el genoma del SARS-CoV-2 de una cadena de ARN y el intermediario durante la replicación que es ARN de doble cadena, son detectados por los receptores tipo Toll (TLRs) 3, 7 y 8 en vesículas, así como por los receptores citoplasmáticos RIG-I y MDA-5. Estos últimos activan la vía que culmina en la fosforilación de IRF3, el cual induce la

transcripción de Interferon (IFN) tipo I y tipo III ^[56,57]. Los TLRs que son expresados principalmente por células mieloides, activan una cascada de señalización que induce la síntesis de los interferones antes mencionados y los genes regulados por NF-κB, que son citocinas pro-inflamatorias. Los IFN son un grupo de péptidos secretados por células que entre sus funciones principales, inhiben la replicación viral y originalmente se dividen en tres familias, los tipo I, II y III. Estos se unen a receptores membranales de manera autócrina y parácrina, y activan vías de señalización que promueven la expresión de genes estimulados por IFN (ISG en inglés). Consecuentemente disminuye la síntesis de nuevas proteínas y se liberan mediadores de la inflamación.

La respuesta inmunológica mediada por IFN es crucial ya que se ha observado un porcentaje de pacientes graves de COVID-19 con mutaciones en alguna de las proteínas de la vía del IFN (TLR3, UNC93B1, TICAM1, TBK1, IFR3, IFR7, IFNAR1 e IFNAR2) ^[56]. También se han reportado pacientes en condiciones graves con auto-anticuerpos contra alguno de los miembros de los IFN ^[58]. Aparte, se ha reportado que los coronavirus SARS-CoV, MERS-CoV y el SARS-CoV-2 tienen mecanismos para inhibir la respuesta mediada por IFN y para evitar la detección por parte de los receptores inmunológicos ^[56]. Proteínas tanto estructurales como accesorias del SARS-CoV-2 son capaces de inhibir diferentes puntos de la vía de señalización que induce la producción de IFN, la señalización mediada por IFN y las funciones efectoras de los ISGs ^[56,57]. Por lo que en muchos pacientes, la ausencia de algún mecanismo relacionado con el IFN o el retraso de las funciones efectoras mediadas por IFN, pueden determinar el progreso de la enfermedad.

Los anticuerpos (Ac) generados por linfocitos B también juegan un papel muy importante, pues los Acs opsonizan a las partículas virales, haciéndolas más susceptibles a su eliminación ya sea por el sistema de complemento o por células fagocíticas. De primer instancia los Acs son de clase IgM pero al cambiar de isotipo a IgG aumentan la afinidad por su antígeno. Los Acs neutralizantes sin importar el isotipo, evitan la internalización del virus a las células, reduciendo así el daño tisular causado por la infección. Los macrófagos alveolares y los neutrófilos reclutados son los que eliminan a las partículas opsonizadas en pacientes con enfermedad leve o asintomáticos. Un tratamiento que se ha utilizado con

pacientes graves es la transfusión pasiva de sueros de pacientes convalecientes, sin embargo esta no genera memoria inmunológica. En cambio las vacunas sí son una alternativa para prevenir síntomas graves ya que el individuo genera sus propias clonas de linfocitos B de memoria anti SARS-CoV-2 ^[59,60]. Se ha demostrado que tanto por infección como por vacunación, los anticuerpos anti SARS-CoV-2 permanecen en circulación por varios meses y que incluso las células B de memoria son detectables ^[61]. Además, un reporte demostró que las personas seleccionan linfocitos B de memoria que codifican para anticuerpos anti RBD, similares en sus secuencias génicas, utilizando dos vacunas distintas comparadas contra la infección natural en pacientes convalecientes ^[62]. Esto evidencia que la selección de linfocitos B es muy parecida entre individuos retados con antígenos similares y por lo tanto también se debe de entender que esta presión de selección promueve la selección de variantes de virus o al menos en su proteína S.

Receptores de coronavirus humanos

La proteína S de los coronavirus media la unión específica de los virus a los receptores celulares. Se han identificado los principales receptores que utiliza cada uno de los coronavirus que infectan al humano (**Tabla 1.2**). Estos receptores son proteínas que comparten características funcionales que podrían ser seleccionados por los virus. Cabe recalcar que esos receptores son expresados por diferentes tejidos, y están implicados en diferentes procesos fisiológicos. La aminopeptidasa N (también llamada CD13) y la dipeptidil peptidasa 4 (también CD26) son proteasas catalogadas como proteínas multifuncionales (moonlighting proteins en inglés), que juegan papeles importantes en la adhesión y señalización de leucocitos ^[44]. En la **Tabla 1.2** se mencionan algunas de las principales características de cada uno de los receptores virales y en la figura 1.6 se muestran características estructurales.

Tabla 1.2- Coronavirus humanos identificados. Los siete coronavirus humanos han surgido por transmisiones zoonóticas, de los cuales solo tres de ellos han causado epidemias.

| Coronavirus | Género | Identificado | Huésped natural | Receptor |
|-------------|-------------------------|--------------|---|--|
| HCoV- 229 E | <i>Alphacoronavirus</i> | 1965 | Murciélago <i>Hipposideros</i> y camélidos | Aminopeptidasa N (CD13) |
| HCoV- OC43 | <i>Betacoronavirus</i> | 1967 | Roedores y cerdos | Receptores de ácido siálico 9-O- acetilado |
| SARS-CoV | <i>Betacoronavirus</i> | 2002 | Murciélago <i>Rhinolophus</i> y civetas | ECA2 |
| HCoV- NL63 | <i>Alphacoronavirus</i> | 2004 | Murciélago <i>Triaenops afer</i> | ECA2 |
| HCoV- HKU1 | <i>Betacoronavirus</i> | 2005 | Roedores | Receptores de ácido siálico 9-O- acetilado |
| MERS-CoV | <i>Betacoronavirus</i> | 2012 | Murciélago y camello | Dipeptidil IV (CD26) |
| SARS-CoV-2 | <i>Betacoronavirus</i> | 2019 | Murciélago <i>Rhinolophus affini</i> Pangolín Malayo <i>Manis javanica</i> | ECA2 |

3.1 Aminopeptidasa N

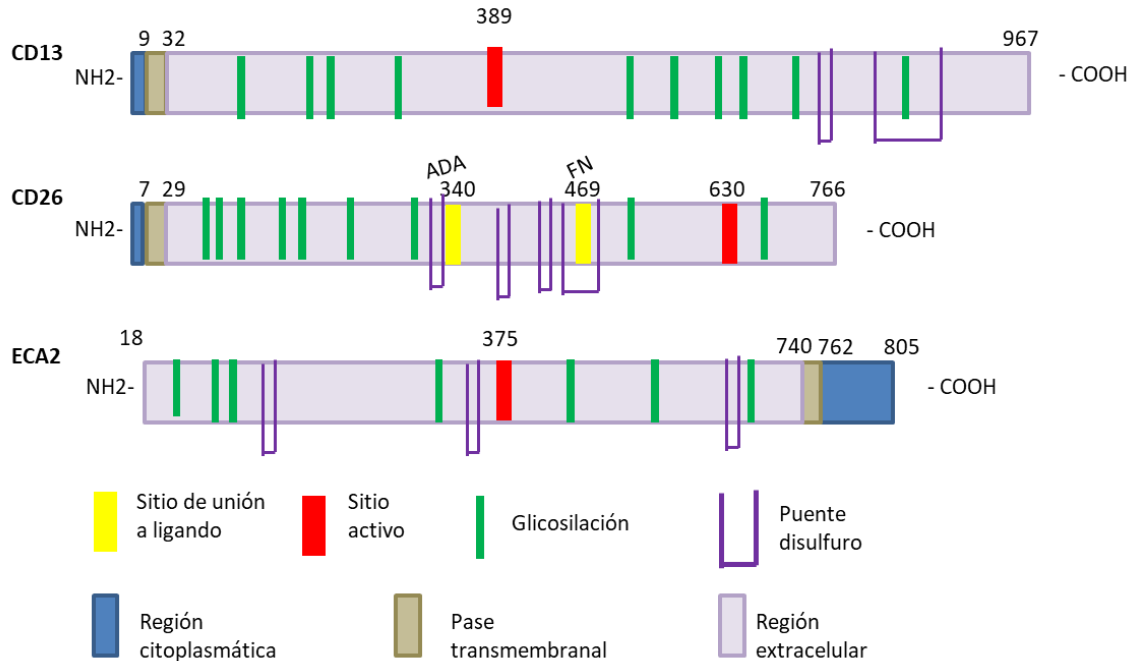
La Aminopeptidasa N (ANPEP, CD13, alanil peptidasa, gp150) (EC: 3.4.11.2)) es una glicoproteína de alrededor de 150 kDa que fue identificada inicialmente por su actividad catalítica y posteriormente se encontró que era la misma proteína que el marcador de leucemia mieloide CD13 ^[63]. A esta proteína se le había descrito con varios nombres, entre ellos MY7, SJ1D1 y L138 pero por consenso permanecieron Aminopeptidasa N (ANPEP) y CD13 ^[64]. ANPEP es una metaloproteína dependiente de Zn^{2+} que cataliza la hidrólisis de los enlaces peptídicos en residuos preferentemente neutros del extremo amino terminal de un péptido. Varios de sus sustratos son procesados secuencialmente; por ejemplo, la angiotensina III es hidrolizada en angiotensina IV la cual es nuevamente hidrolizada por ANPEP; o la encefalina que es escindida por ANPEP y posteriormente por CD10.

Estructuralmente, ANPEP es una proteína transmembranal de tipo II, es decir su extremo carboxilo es extracelular mientras que el extremo amino es citoplasmático (**Figura 1.9**). La ANPEP humana está constituido por 967 aminoácidos, con una gran porción extracelular, un solo cruce transmembranal y una pequeña porción citoplasmática de únicamente 8 aminoácidos ^[65]. Aparte de la gran masa proteica (109.5 kDa), alrededor del 40% de su peso (total 150 kDa) está conferido por glicosilaciones. Se expresa en la membrana celular en forma homo- dimerica; se puede encontrar intracelularmente en vesículas y también extracelularmente en forma soluble.

La expresión de esta peptidasa es constitutiva en células epiteliales de intestino y riñón, células endoteliales, fibroblastos, algunas neuronas y células mieloides, e interesantemente muchas células cancerosas sobre expresan dicha proteasa, por lo que se le considera un marcador en ciertos tumores ^[65]. La sobreexpresión de ANPEP en células cancerosas se ha relacionado con su potencial metastásico y esto ha estimulado la investigación sobre el uso de ANPEP como un posible blanco terapéutico ^[66].

ANPEP también es un marcador de diferenciación de células mieloides y participa en diferentes procesos celulares, ya sea gracias a su actividad enzimática (como en quimiotaxis, angiogénesis, degradación de péptidos durante la digestión) o independientemente de ella (como en migración, adhesión celular, agregación y fagocitosis) ^[65,67-70]. A pesar de ser una enzima de membrana con un segmento intracelular de solamente 9 a.a., algunas funciones las realiza actuando como receptor ya que es capaz de iniciar cascadas de señalización ^[68,71-74]. Sin embargo, aún no se conoce algún ligando natural que lo active en condiciones fisiológicas, por lo que es considerado un receptor huérfano.

Evolución molecular de la proteína Spike del SARS-CoV-2



1.9- Estructura lineal de los principales receptores de coronavirus de humanos. Los receptores identificados son proteasas de membrana: Aminopeptidasa N, Dipeptidil peptidasa IV y la Enzima Convertidora de Angiotensina 2.

3.2 Dipeptidil peptidasa IV

La enzima Dipeptidil peptidasa IV (DPP IV, CD26, Adenosin Deaminasa 2 (ADCP2), (EC: 3.4.14.5) tiene una actividad de peptidasa de serina, que escinde dipéptidos del extremo amino terminal de una cadena de aminoácidos ^[75]. Algunos de sus substratos son la proteína 1 tipo glucagón (GLP-1 en inglés) y la proteína inhibitoria gástrica (GIP en inglés), y la sustancia P, entre otros. Al igual que ANPEP, DPP IV es una proteína transmembranal tipo II con solo un pase transmembranal y 6 aminoácidos citoplasmáticos, es decir, la mayor parte de la proteína está expuesta extracelularmente (738 aminoácidos) ^[76]. En la región extracelular se han caracterizado dos dominios funcionales: el dominio homólogo a la α/β hidrolasa con la actividad de serin- proteasa (el cual se encuentra cercano al extremo C- terminal), y otro dominio que a su vez está dividido en dos subdominios: una región rica en cisteínas y otra región altamente glicosilada (**Figura 1.9**). Otra característica compartida con otras proteasas es que se encuentra formando homodímeros y precisamente las glicosilaciones facilitan la

formación del dímero y la interacción con otras proteínas ^[77]. En total posee nueve N-glicosilaciones y forma 5 puentes disulfuro que ayudan a mantener la estructura.

CD26 se expresa en células endoteliales, células epiteliales de riñón, hígado, pulmón, intestino, esófago, ovario, mama, próstata y en una variedad de poblaciones de leucocitos incluyendo a linfocitos T, B, células NK y células dendríticas ^[78]. Hay evidencias de que CD26 es sumamente importante en la activación y maduración de linfocitos T ya que interactúa con distintas proteínas involucradas en la vía de señalización de los linfocitos ^[79-83]. Además, DPP IV interactúa con la enzima deaminasa de adenosina (ADA en inglés) ^[84] que a su vez interactúa con el receptor 2 de adenosina expresado en células dendríticas ^[79]. Consecuentemente, forma un puente importante entre células. También CD26 puede unirse a fibronectina y colágeno tipo 1 de la matriz extracelular, habilidad que es aprovechada por células cancerosas metastásicas ^[75]. La sobre expresión de esta proteasa es inducida por IL-12 en linfocitos T activados; en contraste, la presencia de TNF regula negativamente la expresión de DPP IV.

3.3 Enzima Convertidora de Angiotensina II

ECA2 (EC: 3.4.17.23) es una carboxi-peptidasa de 805 aminoácidos que participa en el sistema renina- angiotensina- aldosterona, el cual regula la presión sanguínea. Esta regulación está fundamentada en regular el volumen sanguíneo así como la resistencia del sistema vascular. Debido a la participación en este proceso fisiológico tan importante, los sustratos más estudiados son la angiotensina I la cuál es hidrolizada en angiotensina 1-9 y la angiotensina II que es escindida a angiotensina 1-7. El primero evita la hipertrofia mientras que el segundo se opone a los efectos vasoconstrictores de la angiotensina II, ya que es un vasodilatador y evita la proliferación celular ^[85]. No obstante la ECA2 tiene otros sustratos como la neurotensina, kinetensina, bradikina, apelinas y casomorfina, entre otros ^[86].

ECA2 es una proteína transmembranal de tipo I (**Figura 1.9**), donde la mayoría de sus aminoácidos se encuentran en la cara extracelular, seguido por una región transmembranal y una región citoplasmática de 43 aminoácidos en el extremo carboxilo.

En la región extracelular tiene un sitio de unión para el ión Zn^{2+} y otro para Cl^- , siete glicosilaciones y tres puentes disulfuro y se identifican dos residuos fosforilables en la región citoplasmática: la Tyr781 y la Ser783. A diferencia de las proteasas anteriormente mencionadas, ECA2 sí tiene motivos canónicos de señalización, por ejemplo cuenta con un motivo de interacción con LC3 (LIR en inglés) (posiciones 778-786), un motivo de unión a SH2 (781-785), una señal de endocitosis (781-784), un dominio de unión a fosfotirosina (PTB en inglés) (791-795) y un motivo de unión a PDZ (803-805). De la cara extracelular, el dominio C- terminal de la proteasa es homóloga a colectrina.

La expresión de esta proteasa también es amplia, incluyendo células endoteliales, enterocitos, células de Leydig y de Sertoli, en el túbulo proximal renal, corazón, testículos, neumocitos tipo II, células epiteliales de los bronquios, nasales, de la córnea y vesícula. Se le encuentra en las membranas plasmáticas, en vesículas y en el citoplasma, así como en forma secretada. Su expresión es ubicua, y el nivel de expresión puede ser alterado a la alza por una falla cardíaca, la exposición al humo del tabaco, inflamación crónica de los bronquios, y por IFN- α , IFN- γ e IL-13^[87].

ECA2 interactúa con el transportador de aminoácidos neutros SL6A19. Dentro de sus funciones que dependen de la actividad enzimática se incluye la maduración de péptidos vasoactivos, homeostasis de la presión de fluidos, promueve la contracción de los miocitos, promueve el ensamblaje de las proteínas de uniones GAP, y promueve la proliferación celular.

Teoría Neutral de la Evolución Molecular

La teoría neutral de la evolución molecular propuesta por Motoo Kimura en 1968, establece que los cambios en las secuencias nucleotídicas ocurren de manera estocástica y pueden o no ser fijados por otros mecanismos evolutivos^[88]. Las mutaciones neutrales o al azar no influyen en la adaptación del organismo. La evolución de las especies es consecuencia de la variabilidad genética ocasionada por mutaciones que continuamente surgen y que pueden alterar las frecuencias alélicas en las poblaciones^[89,90]. De forma que la evolución adaptativa se debe a la selección natural de las mutaciones que ocurren

aleatoriamente. En consecuencia los cambios en aminoácidos también ocurren al azar. Kimura demuestra que la tasa de mutación se debe de evaluar en regiones del genoma donde no se ejerce una presión de selección; de ser así el número de mutaciones por generación sería igual a la tasa de mutación multiplicada por el tamaño poblacional. Estas variables en el tiempo permiten generar un “reloj molecular”, según el número de mutaciones ocurridas ^[91]. Kimura defendía que la tasa de evolución neutral depende solamente de la tasa de mutación y es independiente del tamaño poblacional. Sin embargo la mayoría de las mutaciones neutrales se eliminan por el mismo efecto estocástico y solo algunas de ellas serán fijadas entre los individuos de una población determinada. Por eso la probabilidad de que una nueva mutación se fije, sí dependerá del tamaño de la población, es decir, entre más pequeña sea la población, la probabilidad de fijación será más alta. Cabe recalcar que para una secuencia particular en una determinada población, inicialmente todas las mutaciones neutrales tienen la misma probabilidad de fijación.

La aparición de mutaciones neutrales genera una amplia variabilidad, sin embargo la mayoría van a ser eliminadas por selección purificadora. Por lo tanto las secuencias actuales reportadas en bases de datos solo son una proporción de las variantes originales de una secuencia nucleotídica. De inicio la ocurrencia de las mutaciones es estocástica permitiendo una gran cantidad de variantes que posteriormente son seleccionadas. Las mutaciones que afectan la adaptación son determinadas por la selección natural. La selección positiva se explica cuando una mutación aumenta la adecuación y la frecuencia alélica tenderá a aumentar hasta que se fije de manera tal que reemplace al alelo ancestral en la población. El caso contrario, la selección negativa tiende a reducir la frecuencia de una variante. Sin embargo, si la mutación no ejerce ningún tipo de ventaja o desventaja en la adecuación, la selección natural no opera. En este caso el cambio de frecuencias alélicas es ocasionado por la deriva génica (aquí llamado evolución neutral), es decir por las frecuencias iniciales de los alelos en una población.

4.1 Prueba de neutralidad

Bajo la visión de la teoría neutral de la evolución molecular se puede detectar si las presiones de selección alteran la probabilidad de fijación de las mutaciones neutrales. Generalmente los elementos funcionales de una proteína están sujetos a presiones de selección, por lo que su patrón de evolución neutral es distinto al esperado. Para detectar el tipo de selección ejercido sobre mutaciones se puede realizar una prueba de evolución neutral. Zamudio y colaboradores en el 2019 desarrollaron un método para evaluar el tipo de presión de selección o la neutralidad de las mutaciones puntuales en proteínas homólogas ^[92]. Este nuevo modelo calcula la probabilidad de ocurrencia de cada uno de los 20 aminoácidos tras aparecer por una mutación de sustitución en la secuencia de ácidos nucleicos. Este modelo incluso considera que la tasa de mutación por sustituciones de transición no es la misma a la tasa de mutación por transversiones (Figura 1.10).

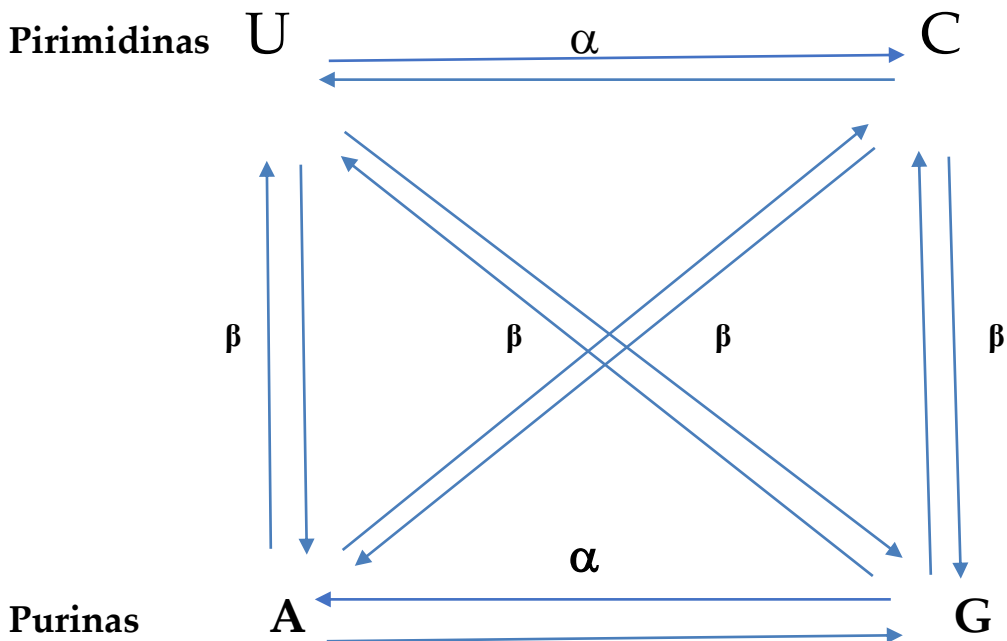


Figura 1.10 – Modelo evolutivo de sustitución de nucleótidos donde α es la tasa de sustitución por transición y β es la tasa de sustitución por transversión de los nucleótidos.

Cuando se observa que un aminoácido tiende a fijarse con una frecuencia menor a la esperada por la prueba de neutralidad, se entiende que ese aminoácido ha sufrido selección negativa. Sin embargo, esto no quiere decir que ese aminoácido no va a fijarse en ninguna posición, pues las posiciones donde se conserva ha sido porque fue seleccionado, a este fenómeno se le denomina selección negativa. En un segundo caso donde un aminoácido es favorable para una proteína, tenderá a ocurrir en mayor proporción a lo esperado y tenderá a fijarse en la secuencia. En este caso se considera que el aminoácido se encuentra bajo presión de selección positiva. El tercer caso es cuando un aminoácido se fija con la misma probabilidad a la predicha por mutaciones neutrales, en este caso el aminoácido se fija de manera neutra o por deriva génica. Estos aminoácidos no están bajo ninguna presión de selección por lo que su ocurrencia obedece al azar.

La prueba de neutralidad está relacionada con la degeneración del Código Genético Estándar de forma tal que la probabilidad de ocurrencia de cada uno de los aminoácidos está directamente relacionada a su codonicidad. Los datos de entrada son las secuencias de nucleótidos de la proteína a analizar y se alinean por codones de manera que toma en cuenta tanto mutaciones sinónimas como no sinónimas y los reporta en cambios de amino ácidos. Consecuentemente se genera un modelo de substitución que permite identificar aminoácidos que se desvían de las expectativas por mutaciones neutrales, es decir, cuáles aminoácidos se fijan por estar bajo selección positiva o negativa y cuáles aminoácidos se fijan por deriva génica o de manera neutral. Como Zamudio *et al.* demostraron (**Figura 1.11**), este modelo se puede aplicar entonces a proteínas con un origen monofilético que utilicen el código genético universal sin importar su grado de conservación.

Evolución molecular de la proteína Spike del SARS-CoV-2

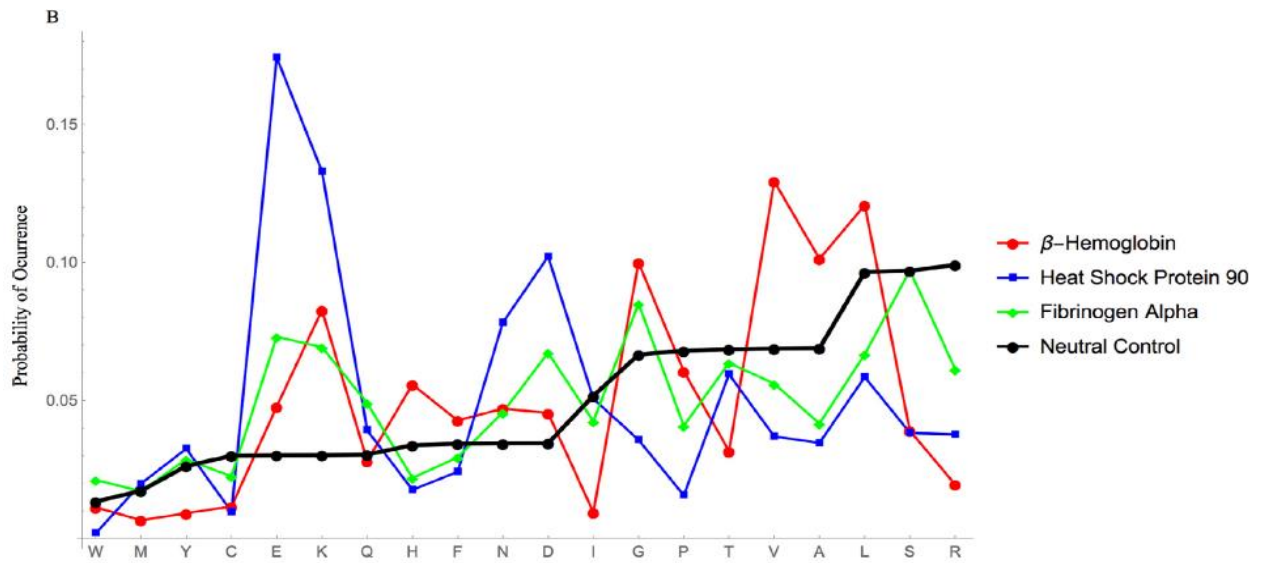


Figura 1.11- Prueba de neutralidad del modelo de Zamudio y col. Tres proteínas fueron probadas según la neutralidad de sus mutaciones puntuales. Las probabilidades superiores al control neutral (línea negra) están bajo presión de selección positiva. Las que están por debajo del control neutral están bajo presión de selección negativa y las probabilidades de ocurrencia que coinciden con el control no tienen presión de selección, por lo que se fijan de manera neutra o por derive génica. Gráfica tomada de Zamudio GS, et al. 2019.

Justificación

El SARS-CoV-2 ha impactado fuertemente a la humanidad y continúa evolucionando. Si bien es necesario comprender la biología del virus para desarrollar estrategias de prevención y de tratamiento que mitiguen su propagación, también es importante comprender los mecanismos evolutivos que operan sobre el SARS-CoV-2 y sus proteínas. A medida que la pandemia se siga prolongando, seguirán surgiendo variantes, entre ellas algunas de mayor preocupación para la salud y con potencial de que aumente su frecuencia en las poblaciones humanas. Los estudios evolutivos ayudan a comprender y calcular la tasa de mutación y la probabilidad de fijación de residuos que a su vez pueden influenciar la adaptación del virus. Por ello, se debe de calcular la probabilidad de ocurrencia de los aminoácidos para así descubrir si estos están sujetos a presiones de selección que influyan la afinidad por el receptor.

Objetivo

General

Evaluar la probabilidad de ocurrencia de los aminoácidos que han mutado en la proteína Espícula/Spike de los *Betacoronavirus* y de aislados clínicos de SARS-CoV2, para discernir los cambios de SARS2-S que han influenciado la unión al receptor y/o han conferido alguna ventaja que ocasione un incremento en la transmisibilidad del virus.

Particulares

- Describir las presiones selectivas ejercidas sobre las mutaciones neutrales tanto de la proteína completa como del RBD de la proteína S de los diferentes *Betacoronavirus*, y de aislados clínicos de SARS-CoV-2.
- Identificar los residuos únicos de la proteína S del SARS-CoV-2 que potencialmente hayan aumentado la adecuación del virus.
- Describir las mutaciones presentes en la proteína S de las variantes de preocupación de acuerdo a la presión de selección, y describir las consecuencias de dichas mutaciones en cuanto a la estructura y características químicas de la proteína.

- Evaluar si los potenciales sitios de glicosilación han sido afectados por las mutaciones neutrales de la proteína SARS2-S.

Hipótesis

La mayoría de las mutaciones observadas de la proteína S del SARS-CoV-2 y principalmente en el RBD, se han fijado en el genoma viral por presiones selectivas, ya que la proteína tiene funciones importantes donde la adaptación al receptor del huésped es crucial. Las mutaciones se seleccionaron y acumularon a través de varias generaciones para que la proteína tuviera afinidad al receptor ECA2 humano y así ocurriera la transmisión zoonótica. De igual manera, la mayoría de las mutaciones observadas en las variantes se encontrarán bajo presiones de selección que contribuyen a la adaptación del virus.

Métodos

Bases de datos

Las secuencias nucleotídicas y de aminoácidos de las proteínas S de diferentes coronavirus se obtuvieron del GenBank (<https://www.ncbi.nlm.nih.gov/>), incluyendo 6 proteínas S de *Alphacoronavirus*, 102 de *Betacoronavirus* y 561 secuencias del genoma completo de aislados clínicos de pacientes infectados con SARS-CoV-2. De estas se usaron 200 secuencias del gen de la proteína S para el análisis de neutralidad. Se descargaron del Protein Data Bank (<https://www.rcsb.org/>), las estructuras tridimensionales de los RBDs de SARS2-S así como de SARS-S unidos al receptor celular ECA2 (PDB: 6M0J y 2AJF respectivamente). Se obtuvo también la estructura de la proteína SARS2-S depositada en I-TASSER, el cual utiliza las estructuras determinadas experimentalmente para predecir una que incluya los extremos amino y carboxilo terminal que son difíciles de determinar experimentalmente. En esta plataforma se han depositado las estructuras de todas las proteínas de SARS-CoV-2. Finalmente, las mutaciones de la proteína S de las variantes de SARS-CoV-2 de interés y de preocupación para la salud, se recuperaron de los informes de situación reportados en www.outbreak.info.

Modelo de evolución neutral

Se alinearon 102 secuencias de aminoácidos de proteínas S del género *Betacoronavirus* utilizando el algoritmo MUSCLE^[93]. Esta alineación sirvió de molde para alinear correctamente la secuencia de nucleótidos, evitando los espacios que pudieran dividir los codones. A partir del alineamiento de nucleótidos, se calculó una tabla de mutaciones que representan el total de cambios en los codones. La tabla de mutaciones de codones se transformó en una matriz de mutación de aminoácidos sumando los valores de los codones para un aminoácido dado. Por lo tanto, esta matriz considera mutaciones sinónimas y no sinónimas. La matriz de mutación de aminoácidos se calculó para cada par de secuencias y se sumaron. Posteriormente, la matriz se normalizó por filas, de manera que cada fila sumara 1 y produjera una matriz de transición de probabilidad. La distribución estacionaria de la matriz de transición de probabilidad se derivó y se comparó con el control esperado tras mutaciones neutrales, como se describe

en ^[92]. Para evaluar la robustez estadística de la muestra de secuencias, se aplicó una prueba Jackknife, con un intervalo de confianza del 95%. Todo el proceso se le aplicó también a las secuencias de los RBDs de las mismas proteínas S analizadas así como de un conjunto de 200 secuencias de S de aislados clínicos de pacientes con SARS-CoV-2.

Análisis filogenético

Todos los análisis evolutivos se realizaron con el software MEGA X ^[94]. Los alineamientos múltiples de las secuencias de proteínas S se realizaron con el algoritmo MUSCLE ^[93]. Se construyó un árbol filogenético por máxima verosimilitud de las secuencias de S de los *Betacoronavirus*. También se hizo un segundo alineamiento múltiple donde se incluyeron las secuencias de aminoácidos de SARS-CoV-2, las proteínas S más proximales a SARS-CoV-2 y las proteínas S de los CoV de unión a ECA2. Esto permitió identificar las mutaciones únicas de la proteína SARS2-S.

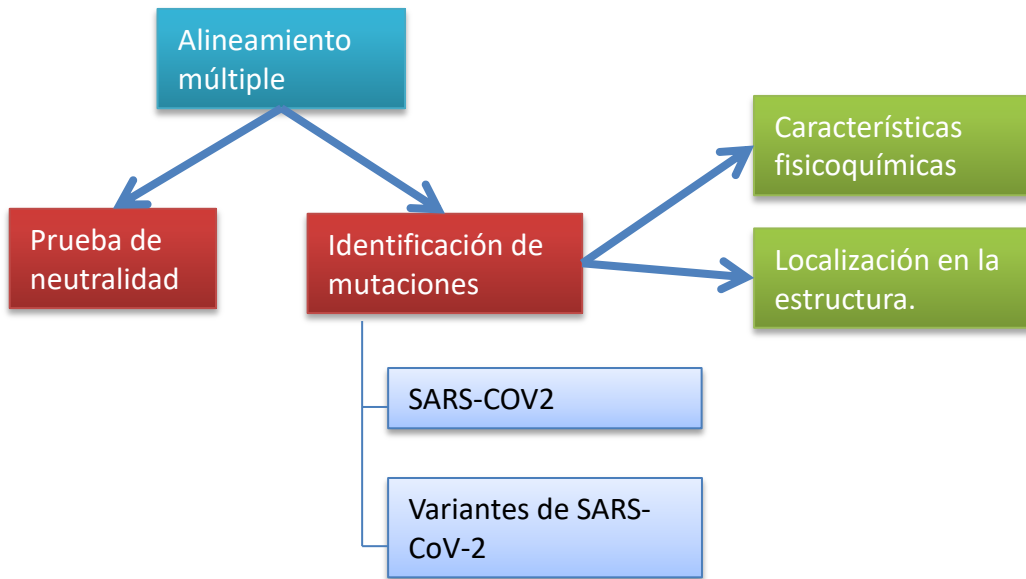
Finalmente, en la secuencia de la proteína S se identificaron los residuos glicosilados de SARS2-S según las reportadas por otros grupos ^[31,32]. Se tomaron en cuenta tanto N- glicosilaciones como O- glicosilaciones para compararles con el resto de secuencias de proteínas de unión a ECA2 con el fin verificar si los sitios de glicosilación son diferentes y proponer si estas pudieran afectar la unión con el receptor ECA2.

Análisis estructural

Después de descargar la estructura de la proteína SARS2-S, de los complejos de los RBDs de SARS2-S con ECA2 (6M0J) y de SARS-S unido a ECA2 (2AJF), se limpiaron de forma que se eliminaron cargas eléctricas fuera de la secuencia de aminoácidos y moléculas no proteicas. Posteriormente se procedió a realizar análisis estructurales visualizado y analizado con los programas Chimera ^[95] e I-TASSER ^[96-99]. Se identificaron en la estructura tridimensional las mutaciones únicas en la proteína SARS2-S que se habían obtenido en el análisis filogenético. Se hizo hincapié en los residuos involucrados en la interacción con el receptor, resaltando su conservación o si eran únicos. Se calcularon las distancias entre estos aminoácidos y los de ECA2 para establecer el número de contactos y número de puentes de hidrógeno que estabilizan la interacción. También se obtuvieron otros parámetros como el potencial hidrofóbico y el potencial electrostático para la superficie

de la interfaz. Los mismos parámetros se calcularon para el complejo del RBD del SARS-S con el mismo receptor con el fin de compararlos. Además de la estructura de SARS2-S, se resaltaron las mutaciones de las proteínas S de las variantes del SARS-CoV-2. En seguida, se hicieron las predicciones de las estructuras de las variantes a fin de compararlas para verificar si las mutaciones habían provocado cambios importantes en la estructura.

Flujo de trabajo



Resultados

La teoría neutral de la evolución molecular asume que la evolución es impulsada por mutaciones puntuales estocásticas, que eventualmente pueden ser fijadas por deriva genética o selección natural. Se aplicó un modelo de evolución neutral para describir el tipo de presión selectiva que podrían ejercerse sobre los aminoácidos tras surgir por una mutación en las proteínas S de los *Betacoronavirus*. El análisis de las proteínas S de este género reveló que Trp, Cys, His, Gly, Pro, Ser y Arg han sufrido presiones selectivas negativas, pues el número de cambios a esos aminoácidos es menor a lo esperado por evolución neutral. En contraste Tyr, Lys, Gln, Phe, Asn, Asp, Thr y Val han estado sustraron estar sujetas a presiones de selección positivas (**Figura 2.1 A**). Los aminoácidos Met, Glu, Ala y Leu se fijan de manera neutral.

Para describir con mayor precisión las presiones que han actuado sobre los aminoácidos implicados en la unión al receptor, se realizó esta misma prueba únicamente con las secuencias correspondientes al RBD de las proteínas S de los mismos coronavirus. Se hizo un alineamiento múltiple para extraer la región que se alineaba con el RBD de SARS-CoV-2, las cuales fueron sometidas a la prueba de neutralidad. De manera similar a los resultados obtenidos al utilizar toda la proteína; Trp, Met, Glu, His, Pro, Ala, Leu y Arg mostraron estar sujetas a selecciones negativas, mientras que Tyr, Lys, Phe, Asn, Asp, y Thr manifestaron estar bajo selección positiva (**Figura 2.1 A**). Esto significa que es menos probable que se fije una Arg mediante una mutación aleatoria que una Tyr, porque la Arg está sometida a fuerte presión de selección negativa.

Evolución molecular de la proteína Spike del SARS-CoV-2

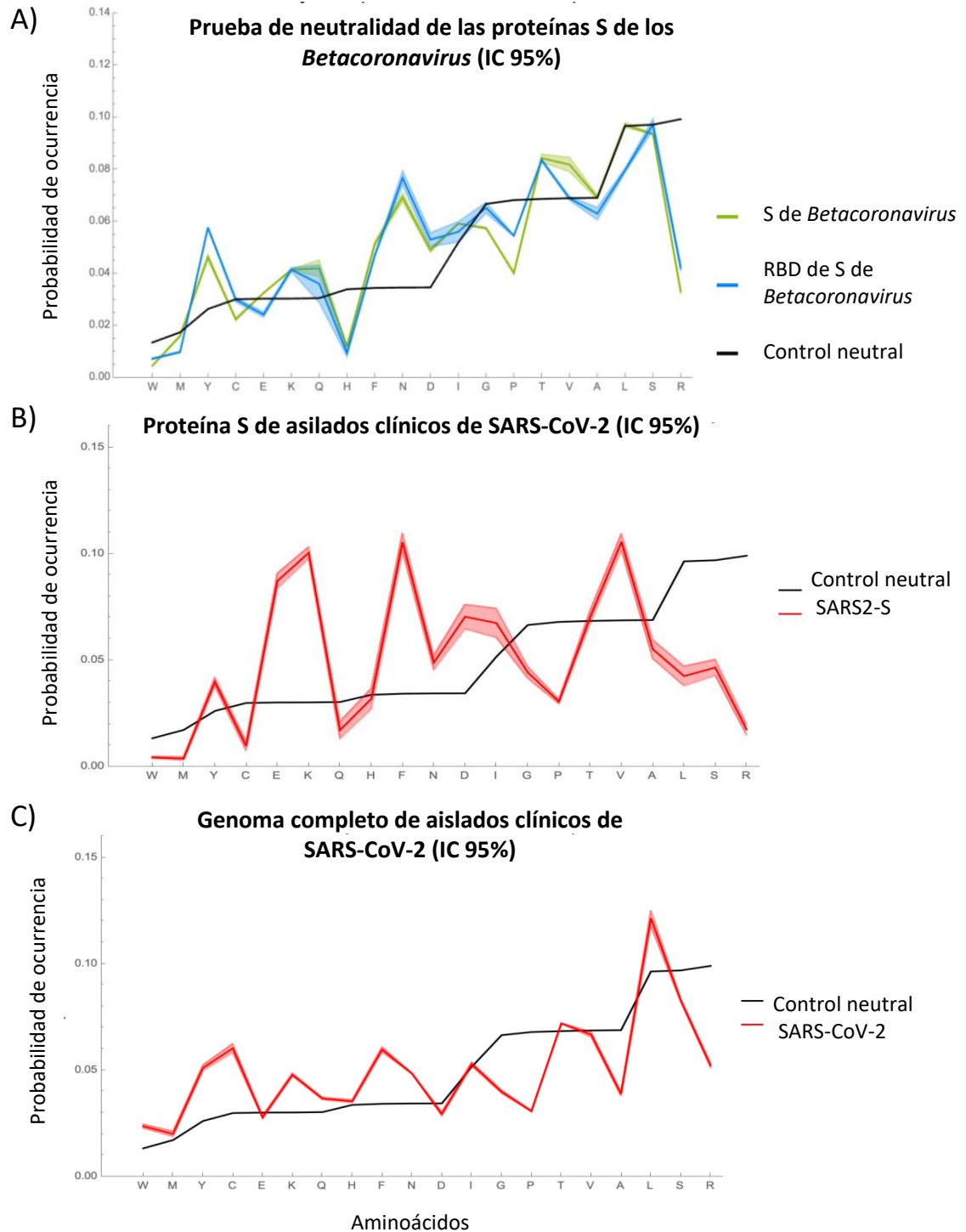


Figura 2.1- Prueba de neutralidad de las proteínas de los coronavirus y del genoma del SARS-CoV-2. A) Prueba de neutralidad de las proteínas S (verde) y de los RBDs (azul) de los *Betacoronavirus*. B) prueba de neutralidad de las proteínas S de aislados clínicos de SARS-CoV-2. en C) se hizo la prueba al genoma completo de los aislados clínicos. En todos los casos se compara contra su control por evolución neutral (líneas negras).

Para estudiar la evolución en curso del SARS-CoV-2, realizamos otra prueba de neutralidad de las proteínas S de SARS-CoV-2 de aislados humanos. La **Figura 2.1 B** muestra que Tyr, Glu, Lys, Phe, Asn, Asp, Ile y Val están bajo presión selectiva positiva, mientras que Trp, Met, Cys, Gln, Gly, Pro, Ala, Leu, Ser y Arg están bajo presión selectiva negativa. Las diferencias entre los gráficos de la **Figura 2.1 A y 2.1 B** resaltan la resolución del modelo para detectar las presiones selectivas en una proteína homóloga que pertenece a un taxón inferior. Además, realizamos la prueba de neutralidad a todo el genoma empleando 561 secuencias de SARS-CoV-2 de aislados humanos (**Figura 2.1 C**). Este análisis permitió retratar cómo todo el genoma del SARS-CoV-2 está sometido a presiones selectivas, que no son necesariamente las mismas a lo largo de todo el genoma viral. La evolución de otras proteínas del virus ha sido diferente a la de la proteína S. Diferentes proteínas virales están influenciadas por diferentes variables que modifican la probabilidad de aparición de sus propios aminoácidos.

Se realizaron relaciones filogenéticas de la proteína S de los *Betacoronavirus*, donde identificamos las proteínas S más cercanas al SARS-CoV-2. Como otros autores ya han reportado ^[20], el árbol filogenético resultante, mostró que la proteína S del coronavirus de murciélago RaTG13 exhibe la mayor similitud con la proteína SARS2-S, seguido por la proteína S del coronavirus PnCov. De hecho, se detectaron 79 mutaciones que se comparten entre estos tres coronavirus y los vuelven únicos. El árbol filogenético de la **Figura 2.2** muestra representantes de cada linaje para ubicar las proteínas S más cercanas a la del virus causante de la COVID-19 (se construyó con máxima verosimilitud). Aquellos coronavirus señalados con un círculo naranja, son aquellos que se ha demostrado tienen la capacidad de unirse al ECA2 humano. Algunos de estos coronavirus son de murciélago pero comparten algunas características químicas necesarias para la interacción con la proteína humana. Los virus con una marca verde, son otros coronavirus humanos que utilizan otros receptores diferentes a ECA2, como lo son el MERS-CoV a CD26 y el HCoV OC43 y HCoV HKU1 que se unen a los receptores de ácido siálico.

Evolución molecular de la proteína Spike del SARS-CoV-2

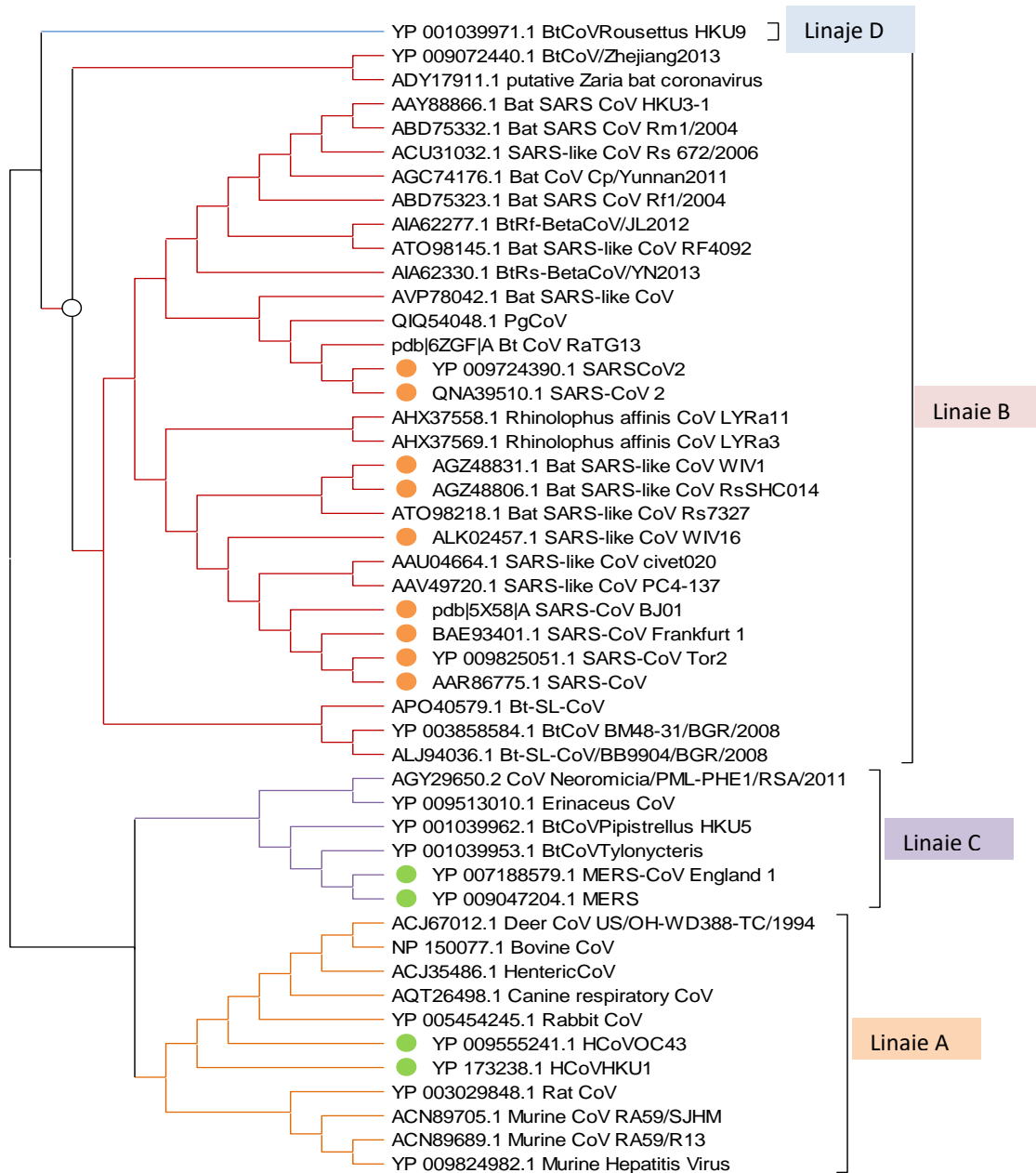


Figura 2.2.- Análisis evolutivo por método de máxima verosimilitud. Árbol filogenético de la proteína S del género *Betacoronavirus*. Se muestran representantes de los 4 linajes. Los CoV que se unen a hECA2 están marcados con un punto naranja, mientras que el marcador verde marca los CoV que infectan a los humanos y usan otros receptores.

Para identificar mutaciones que podrían ser ventajosas en la unión del SARS2-S a ECA2, se profundizó en la secuencia del RBD del subdominio S1. Por lo tanto, se realizó otro alineamiento múltiple que incluyó las secuencias de SARS2-S y de la proteína S de los CoV de unión de ECA2 humano ^[6]. Descartamos el *Alphacoronavirus* humano HCoV NL63

porque aunque se une a la proteasa ECA2, la orientación del RBD al unirse con dicho receptor es completamente diferente (**Figura S3**). El resultado fue la detección de todas las mutaciones únicas del RBD del SARS2-S, algunas de las cuales son compartidas con las secuencias de las proteínas S de CoV RaTG13 y PnCoV lo que sugiere que tal vez podrían unirse al ECA2 humano. La **Tabla 2.1** enlista todas las mutaciones puntuales en el RBD de SARS2- S comparado con el aminoácido correspondiente expresado en el resto de las proteínas S que se unen a ECA2, de las cuales la mayoría han sido seleccionadas. Los aminoácidos se resaltan según su naturaleza química, siendo amarillo los no polares, en verde los polares neutros, en azul los cargados positivamente y en rojo los cargados negativamente, además se resaltan en negrita los aminoácidos que interaccionan con el receptor.

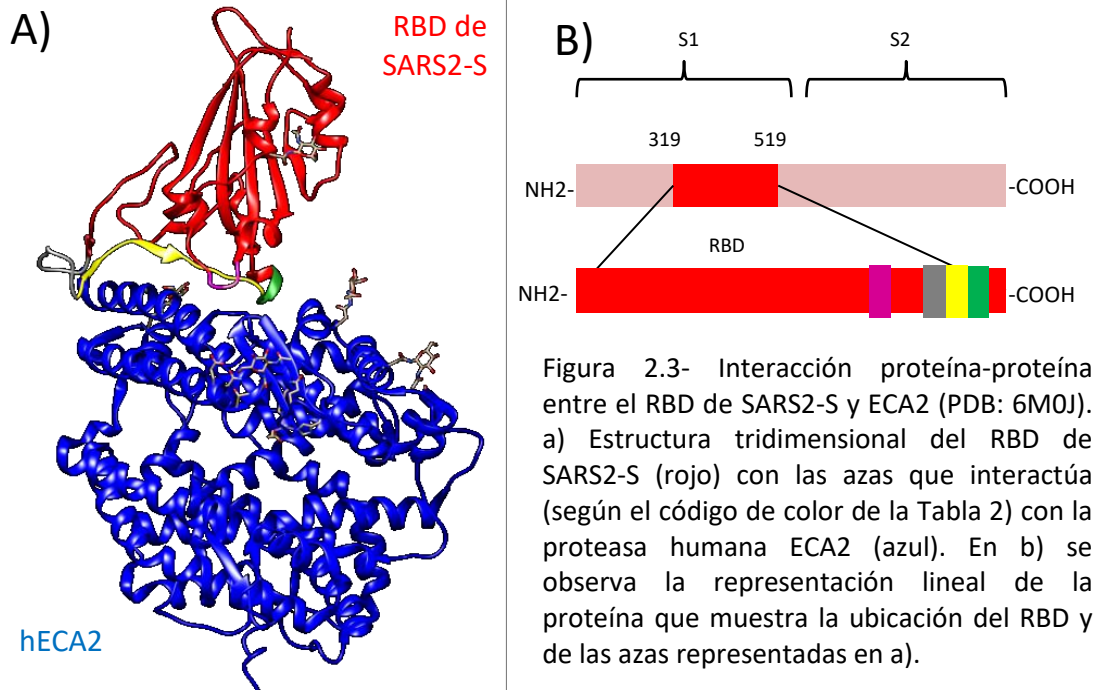
Se visualizó el RBD de SARS2-S interaccionando con ECA2 (PDB: 6M0J) y se determinó que la proteína viral interacciona con el receptor por medio de 4 asas (**Figura 2.3A**). Tres de ellas en realidad son secuenciales y forman una asa larga desordenada, y la otra se encuentra lejana en la secuencia lineal aunque cercana en un extremo de la interfaz (**Figura 2.3B**).

Tabla 2.1- Mutaciones únicas en el RBD de la proteína S de SARS-CoV-2.

| CoVs de unión a ECA2 | | SARS- CoV-2 | | | Características |
|----------------------|------|-------------|----------|-------------------|-------------------|
| Posición | a.a. | Posición | a.a. | Presión selectiva | |
| 335 | P | 348 | A | <i>Negativa</i> | Polaridad |
| 341 | E | 354 | N | <i>Positiva</i> | Carga |
| 359 | T | 372 | A | <i>Negativa</i> | Polaridad |
| 360 | F | 373 | S | <i>Neutral</i> | Polaridad |
| 371 | A | 384 | P | <i>Negativa</i> | Polaridad |
| 380 | S | 393 | T | <i>Positiva</i> | Polares |
| 389 | V | 402 | I | <i>Neutral</i> | No polares |
| 390 | K | 403 | R | <i>Negativa</i> | Misma carga (+) |
| 393 | D | 406 | E | <i>Negativa</i> | Misma carga (-) |
| 404 | V | 417 | K | <i>Positiva</i> | Polaridad y carga |
| 417 | M | 430 | T | <i>Positiva</i> | Polaridad |

Evolución molecular de la proteína Spike del SARS-CoV-2

| | | | | | | |
|----------------|------------|------------|------------|----------|----------|-------------------|
| Bucle 1 | 421 | L | 434 | I | Neutral | No polares |
| | 425 | T | 438 | S | Neutral | Polares |
| | 426 | R | 439 | N | Positiva | Carga |
| | 428 | I | 441 | L | Negativa | No polares |
| | 430 | A | 443 | S | Neutral | Polaridad |
| | 431 | T | 444 | K | Positiva | Carga |
| | 432 | S/Q | 445 | V | Neutral | Polaridad |
| | 433 | T | 446 | G | Neutral | Polares |
| | 439 | K | 452 | L | Negativa | Polaridad y carga |
| | 442 | Y/S | 455 | L | Negativa | No polares |
| | 443 | L | 456 | F | Positiva | No polares |
| | 445 | H | 458 | K | Positiva | Misma carga (+) |
| | 446 | G | 459 | S | Neutral | Polares |
| | 447 | K | 460 | N | Positiva | Carga |
| | 449 | R | 462 | K | Positiva | Misma carga (+) |
| | 457 | N | 470 | T | Positiva | Polares |
| | 458 | V | 471 | E | Negativa | Polaridad y carga |
| | 459 | P | 472 | I | Neutral | Polaridad |
| | 460 | F | 473 | Y | Positiva | Polaridad |
| | Bucle 2 | 461 | S | 474 | Q | Neutral |
| 462 | | P | 475 | A | Negativa | Polaridad |
| 463 | | D | 476 | G | Neutral | Carga |
| 464 | | G | 477 | S | Neutral | Polares |
| 465 | | K | 478 | T | Positiva | Carga |
| 467 | | T | 481 | N | Positiva | Polares |
| 469 | | P | 482 | G | Neutral | Polares |
| 470 | | P | 483 | V | Neutral | Polaridad |
| | | | 484 | E | Negativa | Inserción |
| 471 | | A | 485 | G | Neutral | Polaridad |
| 472 | | L | 486 | F | Positiva | No polares |
| Bucle 3 | | 476 | W | 490 | F | Positiva |
| | 479 | N | 493 | Q | Neutral | Polares |
| | 480 | D | 494 | S | Neutral | Carga |
| | 484 | Y | 498 | Q | Neutral | Polaridad |
| Bucle 4 | 485 | T | 499 | P | Negativa | Polares |
| | 487 | T | 501 | N | Positiva | Polares |
| | 489 | I | 503 | V | Neutral | No polares |
| | 505 | N | 519 | H | Negativa | Carga |
| Código de a.a. | | Polar | No polar | Polar + | Polar - | |



Ese mismo complejo sirvió para posteriormente identificar las mutaciones (rojo) en la estructura del RBD de SARS2-S (rosa pálido) e identificar los residuos responsables de la interacción con el receptor (azul) (**Figura 2.4**). Curiosamente, la mayoría de las mutaciones en el RBD de SARS2-S se encuentran cercanas a la interfaz con el receptor. En el acercamiento, se observan los aminoácidos involucrados en la interacción proteína-proteína, junto con sus cadenas laterales. Los que son comunes entre los coronavirus de unión a ECA2 permanecen en rosa pálido (Tyr 449, Tyr 453, Asn 487, Tyr 489, Thr 500, Gly 502, Tyr 505; **Tabla 2.2**) mientras que los aminoácidos únicos de SARS2-S se muestran en rojo (Lys 417, Gly 446, Leu 455, Phe 456, Ala 475, Phe 486, Gln 493, Gly 495, Gln 498 y Asn 501 **Tabla 2.3**). De los 17 aminoácidos del RBD que interactúan con aminoácidos de ECA2, 10 son únicos en SARS2-S, es decir 7 se comparten con otras proteínas S que tienen la capacidad de unirse al ECA2 humano. De esos 7, seis utilizan aminoácidos favorecidos en la evolución de la proteína, e interesantemente 4 son Tyr. De las 10 mutaciones únicas, 6 aminoácidos están bajo algún tipo de selección, y uno de ellos (K417) participa en la formación de un nuevo puente de hidrógeno con el receptor. En total, SARS2-S forma 11 puentes de hidrógeno con el receptor mientras que SARS-S forma 8.

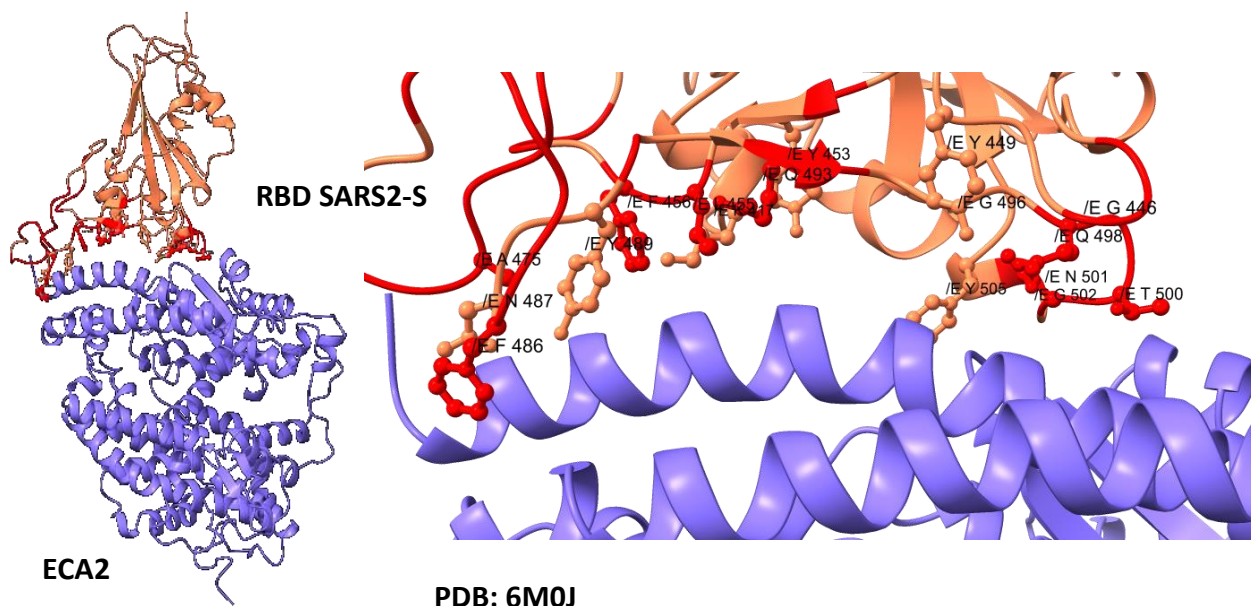


Figura 2.4- Interacción entre el RBD de SARS2-S con ECA2. La interacción entre el RBD del SARS2-S (rosa pálido) y el receptor humano ECA2 (azul) se aprecia mejor en el acercamiento donde se observan las cadenas laterales de los aminoácidos del RBD implicado en la unión con el receptor humano. Los aminoácidos únicos en SARS2-S se colorean en rojo, mientras que los conservados se mantienen en rosa pálido.

Tabla 2.2- Residuos conservados entre las proteínas S que se unen de ECA2 implicados en la interacción proteína-proteína con el receptor.

| a.a. conservados | Tipo de selección |
|------------------|-------------------|
| 449 | Y <i>Positiva</i> |
| 453 | Y <i>Positiva</i> |
| 487 | N <i>Positiva</i> |
| 489 | Y <i>Positiva</i> |
| 500 | T <i>Positiva</i> |
| 502 | G <i>Neutral</i> |
| 505 | Y <i>Positiva</i> |

Código de a.a. Polar

Tabla 2.3- Residuos únicos de SARS2-S implicados en la interacción proteína-proteína con ECA2.

| SARS2-S | | SARS-S | | Presión de selección |
|----------|------|----------|------|----------------------|
| Posición | a.a. | Posición | a.a. | |
| 417 | K | | | <i>Positiva</i> |
| 446 | G | 433 | T | <i>Neutral</i> |
| 455 | L | 442 | Y | <i>Negativa</i> |
| 456 | F | 443 | L | <i>Positiva</i> |
| 475 | A | | | <i>Negativa</i> |
| 486 | F | 472 | L | <i>Positiva</i> |
| 493 | Q | 479 | N | <i>Neutral</i> |
| 496 | G | | | <i>Neutral</i> |
| 498 | Q | 484 | Y | <i>Neutral</i> |
| 501 | N | 487 | T | <i>Positiva</i> |

Código de a.a. No polar Polar neutro Polar +

La distancia promedio entre los aminoácidos de la interfaz del RBD de SARS2-S con ECA2 comparadas con las distancias entre el RBD de SARS-S con ECA2, es similar por lo son necesarios otros parámetros para evaluar la estabilidad y afinidad de la interacción. Diecisiete aminoácidos del SARS2-S contactan con 17 aminoácidos del receptor, a una distancia media de 3.563 Å. En el caso de SARS-S, 15 aminoácidos tienen contacto con 18 aminoácidos del receptor, a una distancia media de 3.605 Å. Se compararon tanto el potencial hidrofóbico como el electrostático de la superficie de las interfaces de las proteínas S, y se observó que ambas tienen valores similares en cada parámetro. El potencial de hidrofobicidad de la interfaz de SARS2-S tiene un valor mínimo -27.2, media -4.743 y máximo de 22.83, de mientras que los valores para la interfaz de SARS-S son: mínimo -26.41, media -3.47 y máximo 23.19 (**Figura 2.5**). A pesar de ello, las mutaciones Phe 486 y la Phe 456 en el SARS2-S generan más contactos hidrófobos que las Leu en las mismas posiciones presentes en el SARS-S (**Figura 2.6**). Además, en el otro extremo de la interfaz de SARS2-S, la superficie es más hidrofílica gracias a la mutación Gln 498 que permite más contactos por fuerzas de van der Waals.

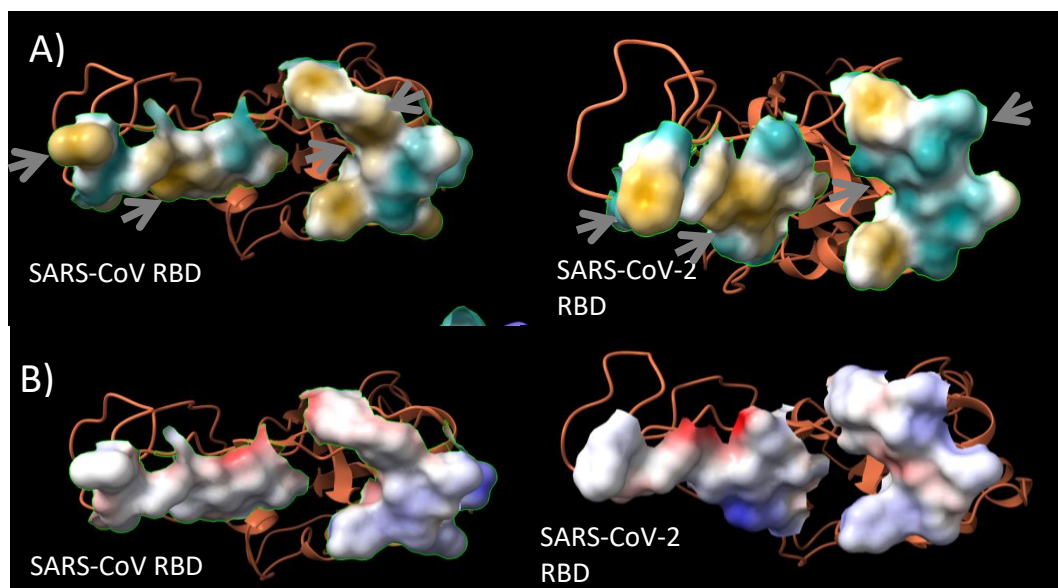


Figura 2.5 Características químicas de la superficie de contacto con el receptor del RBD de proteínas S. a) se comparan el potencial hidrófobo de los RBD del SARS-CoV (izquierda) y el SARS-CoV-2 (derecha) en degradado de azul (hidrófilo), a blanco (neutro) a dorado (hidrófobo). Las flechas apuntan hacia cambios importantes en los potenciales de hidrofobicidad. b) El potencial electrostático de la superficie de ambas interfaces muestra ligeras diferencias. La escala va de rojo (negativo), a blanco (neutro) a azul (positivo).

Para describir el comportamiento evolutivo continuo de la proteína, se analizaron las mutaciones de las proteínas S de algunas variantes de SARS-CoV-2. Primero se hizo una prueba de neutralidad con las proteínas S de 200 aislados clínicos reportados en todo el mundo (**Figura 2.1 B**). En seguida se predijeron las estructuras de las cuatro variantes de preocupación: Alfa, Beta, Gamma, Delta y tres de interés: Épsilon, Iota y Kappa (**Figura 2.7**) para definir si sus mutaciones habían cambiado sustancialmente la estructura de las proteínas S. Las proteínas S se desvían poco de la de referencia: la variante Alfa 3.32 Å, la Beta 3.14 Å, la Gamma 4.14 Å, la Delta 3.34 Å, la Épsilon 4.2 Å, la Iota 3.21Å y la Kappa 3.23Å (RMSD). Las variantes Gamma y Épsilon direccionan el RBD hacia otra posición, el resto de la estructura es muy parecida. Faltaría determinar por medio de cristalografía de rayos X o microscopía electrónica fría si efectivamente las estructuras cambian de orientación.

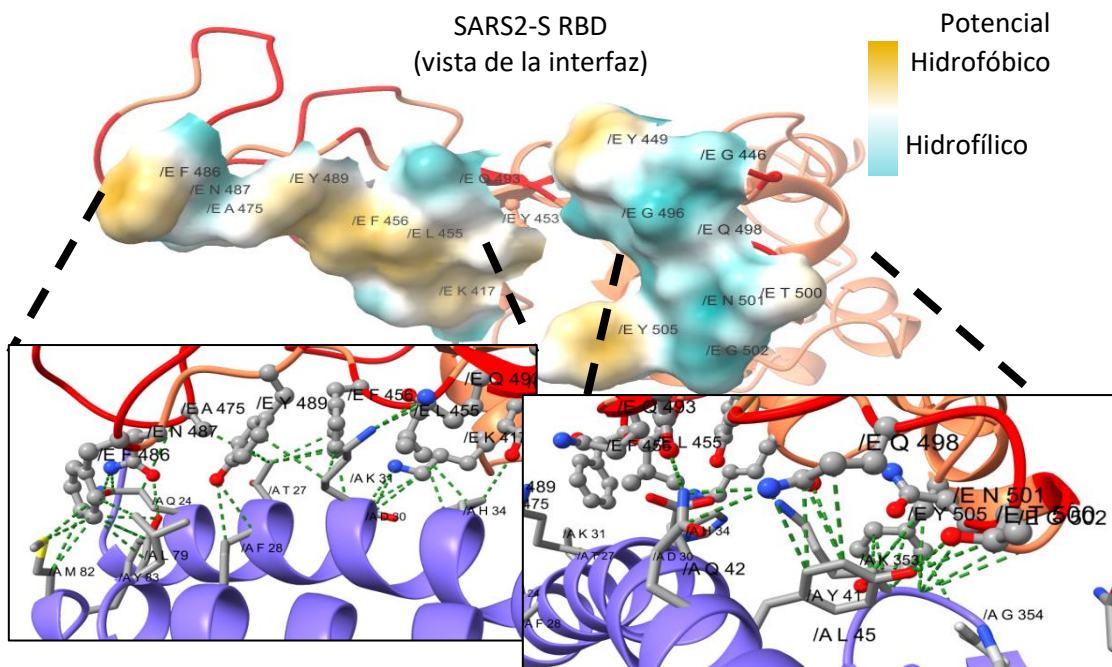


Figura 2.6- características de los aminoácidos del RBD de SARS2-S que se encuentran en la interfaz con ECA2. Con la estructura previamente presentada de la interacción entre el RBD del SARS2-S (rosa pálido) y el receptor humano ECA2 (azul) se resaltan los residuos Phe 456 y Phe 481 responsables del aumento de hidrofobicidad (recuadro izquierda) y la Gln 498 que aumenta la superficie hidrofílica (recuadro derecha).

Posteriormente se identificaron las mutaciones en la estructura 3D (**Figura 2.8**) donde se colorearon todas las mutaciones reportadas en las variantes que se analizaron. Para ello se incluyeron otras variantes de interés: Zeta, Eta, Theta y Lambda. Se describió el tipo de presión de selección de cada mutación de las variantes. Curiosamente, todas las variantes comparten la mutación D614G y tres variantes de preocupación y una de interés presentan la sustitución N501Y. También las variantes Beta, Gamma, Zeta, Eta y Teta comparten E484K y la Delta, Épsilon y Kappa presentan L452R. Todos los aminoácidos con los que se sustituyeron están bajo un tipo de presión de selección. Por ejemplo la Tyr y la Lys son aminoácidos seleccionados positivamente tanto en las proteínas S de los *Betacoronavirus*, como en los RBD. La Gly y la Arg que están bajo selección negativa se han fijado en la posición 614 y 452 respectivamente por presiones compartidas. La variante Gamma, Zeta y Theta presentan una Phe en la posición 1176 aparte de que la Pro 681 es sustituida o por His o por Arg, dos aminoácidos con carga positiva y seleccionados negativamente.

Evolución molecular de la proteína Spike del SARS-CoV-2

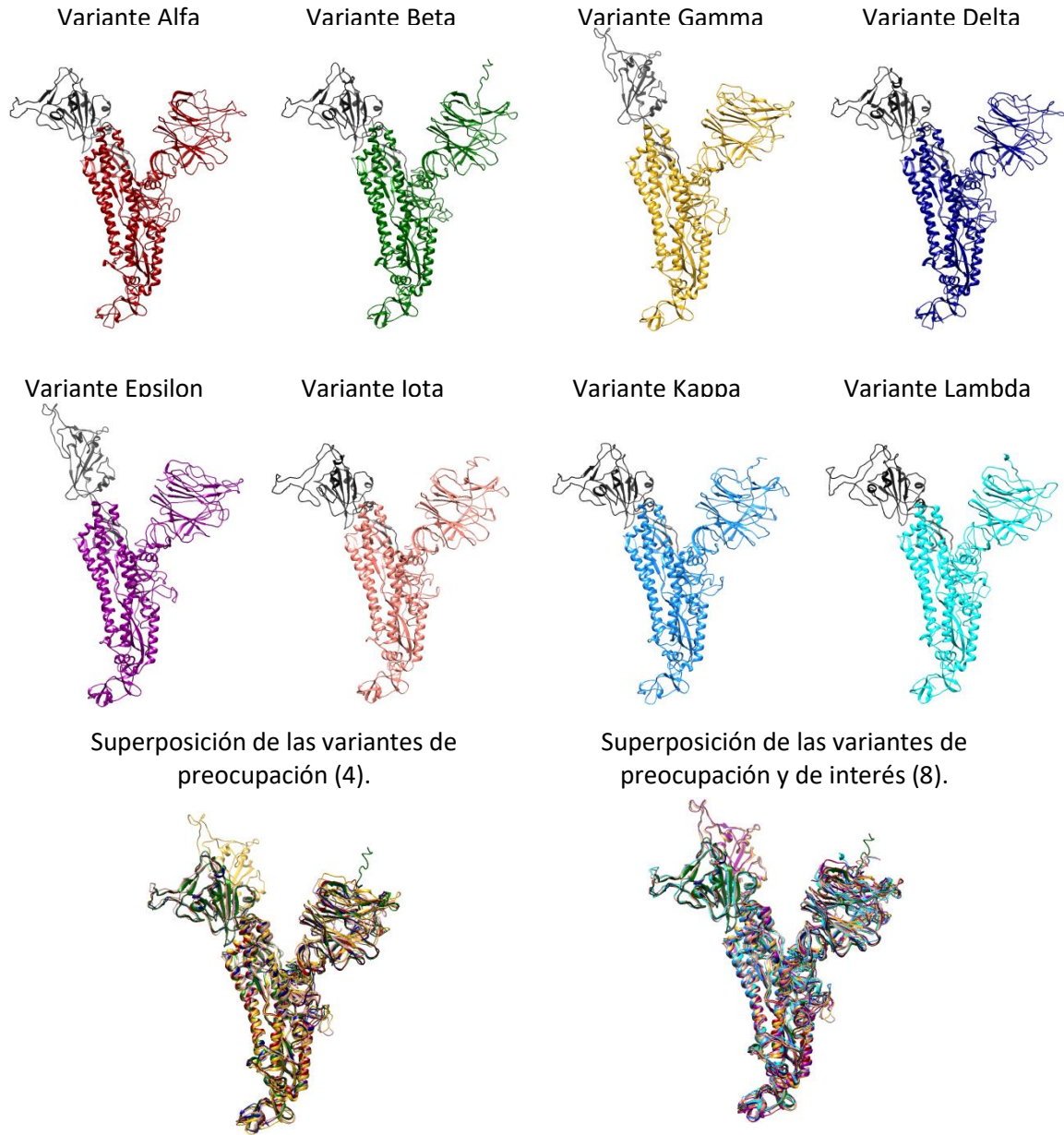


Figura 2.7- Predicción de las estructuras de las proteínas S de las variantes de SARS-CoV-2. Las primeras cuatro estructuras, en la primer fila, son las predicciones de las proteínas S de las variantes de preocupación (Alfa, Beta, Gamma, Delta), las siguientes cuatro son de las variantes de interés (Épsilon, Iota, Kappa y Lambda). La superposición de la izquierda compara las proteínas S de las cuatro variantes de preocupación mientras que en la de la derecha se añaden las variantes de interés.

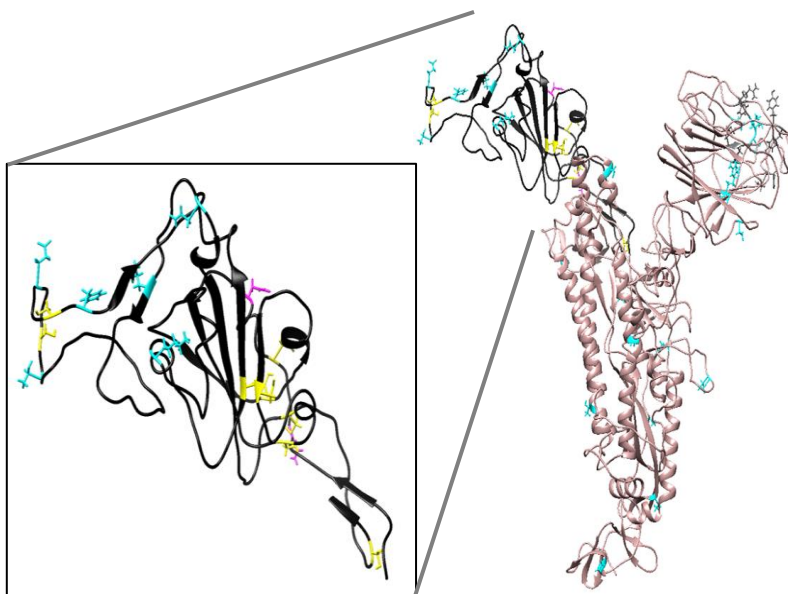


Figura 2.8- Estructura de referencia de la proteína SARS2S. Modificado de la estructura de referencia de I-TASSER La estructura del SARS2-S se muestra con el RBD resaltado en negro y el resto de la estructura en rosa claro. Las mutaciones puntuales de todas las variantes están en color cian y las deleciones en gris. El acercamiento del RBD muestra a las Cys de color amarillo y las dos Asn glicosiladas en magenta.

En total, la variante Alfa, aislada por primera vez en el Reino Unido, tiene 3 mutaciones seleccionadas positivamente, 3 seleccionadas negativamente y 1 mutación neutra, además de cuatro posiciones eliminadas (**Tabla 2.4**). La variante Beta identificada en Sudáfrica tiene 4 aminoácidos positivamente seleccionados, 2 aminoácidos seleccionados negativamente, 1 sustitución neutra y 3 aminoácidos eliminados. La variante Gamma identificada en Brasil, tiene la mayor cantidad de mutaciones: elevándose a 9 seleccionadas positivamente, 3 seleccionadas negativamente y ninguna neutra. La variante Delta que ha causado gran preocupación en los últimos meses, fue identificada por primera vez en la India; ésta tiene 2 mutaciones seleccionadas positivamente, 4 negativamente y 2 aminoácidos eliminados. Por último, en las variantes de interés se presentan varias mutaciones compartidas con las de preocupación, por ejemplo la Épsilon tiene solo una mutación seleccionada positivamente y 3 negativamente, la Zeta 2 positivas y una negativa, la Eta 2 mutaciones seleccionadas positivamente, 3 negativamente, 1 mutación sin presión de selección y 4 aminoácidos eliminados. La variante Theta expresa 5 mutaciones seleccionadas positivamente y 2 negativamente, la variante Iota únicamente presenta 2 mutaciones de interés (una seleccionada positivamente y la otra negativamente), la variante Kappa presenta 4 mutaciones seleccionadas negativamente y una de forma neutra y por último la Lambda

presenta 3 mutaciones seleccionadas de forma positiva, una negativa, 2 seleccionadas de forma neutra y 7 aminoácidos eliminados.

Tabla 2.4- Tipo de presión de selección en las mutaciones de las variantes de SARS-CoV-2.

| Posición en SARS2-S | | Variantes de preocupación | | | | Variantes de interés | | | | | | |
|---------------------|----------|---------------------------|------------------|--------------------|--------------------|------------------------|------------------|----------------|--------------------|------------------|--------------------|----------------------|
| | | Alfa (α) | Beta (β) | Gamma (γ) | Delta (δ) | Epsilon (ϵ) | Zeta (ζ) | Eta (η) | Theta (θ) | Iota (ι) | Kappa (κ) | Lambda (λ) |
| 13 | S | | | | | I + | | | | | | |
| 18 | L | | | F + | | | | | | | | |
| 19 | T | | | | R --- | | | | | | | |
| 20 | T | | | N ++ | | | | | | | | |
| 26 | P | | | S - | | | | | | | | |
| 52 | Q | | | | | | R --- | | | | | |
| 67 | A | | | | | | V + | | | | | |
| 69 | H | * | | | | | * | | | | | |
| 70 | V | * | | | | | * | | | | | |
| 75 | G | | | | | | | | | | | V + |
| 76 | T | | | | | | | | | | | I + |
| 80 | D | | A / | | | | | | | | | |
| 95 | T | | | | | | | | I + | | | |
| 138 | D | | | Y ++ | | | | | | | | |
| 144 | Y | * | | | | | * | | | | | |
| 145 | Y | * | | | | | * | | | | | |
| 152 | W | | | | | C - | | | | | | |
| 157 | F | | | | * | | | | | | | |
| 158 | R | | | | * | | | | | | | |
| 190 | R | | | S - | | | | | | | | |
| 215 | D | | G - | | | | | | | | | |
| 241 | L | * | | | | | | | | | | |
| 242 | L | * | | | | | | | | | | |
| 243 | A | * | | | | | | | | | | |
| 247 | S | | | | | | | | | | | * |
| 248 | Y | | | | | | | | | | | * |
| 249 | L | | | | | | | | | | | * |
| 250 | T | | | | | | | | | | | * |
| 251 | P | | | | | | | | | | | * |
| 252 | G | | | | | | | | | | | * |
| 253 | D | | | | | | | | | | | * |
| 417 | K | | N ++ | T + | | | | | | | | |
| 452 | L | | | | R --- | R --- | | | | R --- | | Q / |
| 478 | T | | | | K + | | | | | | | |

Evolución molecular de la proteína Spike del SARS-CoV-2

| | | | | | | | | | | | | |
|------|---|------|------|------|-----|-------|-----|------|------|-----|-------|------|
| 484 | E | | K + | K + | | | K + | K + | K + | | Q / | |
| 490 | F | | | | | | | | | | | S / |
| 501 | N | Y ++ | Y ++ | Y ++ | | | | | Y ++ | | | |
| 570 | A | D + | | | | | | | | | | |
| 614 | D | G - | G - | G - | G - | G - | G - | G - | G - | G - | G - | G - |
| 655 | H | | | Y ++ | | | | | | | | |
| 677 | Q | | | | | | | H -- | | | | |
| 681 | P | H -- | | | | R --- | | | H -- | | R --- | |
| 701 | A | | V + | | | | | | | | | |
| 716 | T | I + | | | | | | | | | | |
| 859 | T | | | | | | | | | | | N ++ |
| 888 | F | | | | | | | L / | | | | |
| 950 | D | | | | | N ++ | | | | | | |
| 982 | S | A / | | | | | | | | | | |
| 1027 | T | | | I + | | | | | | | | |
| 1071 | Q | | | | | | | | | | H -- | |
| 1092 | E | | | | | | | | K + | | | |
| 1101 | H | | | | | | | | Y + | | | |
| 1118 | D | H -- | | | | | | | | | | |
| 1176 | V | | | F + | | | F + | | F + | | | |

* indican aminoácidos eliminados.

Código de a.a. No polar Polar Polar + Polar -

Por último, dada la relevancia que se le ha encontrado recientemente a la capa protectora de carbohidratos ^[100], se analizaron los residuos glicosilados de la proteína SARS2-S. Se utilizó el alineamiento múltiple que se había generado con las secuencias de unión a ECA2, para identificar los residuos glicosilados reportados: 22 N-glicosilaciones y 3 O- glicosilaciones. Posteriormente se determinaron los residuos compartidos con dichas características, para proponerlos como potenciales formadores del mismo tipo de enlace glicosídico. La comparación reveló que SARS2-S expresa 4 Asn nuevas potencialmente glicosiladas (17, 74, 149 y 657) mientras que las otras 18 son compartidas entre todos los coronavirus de unión a ECA2 (**Tabla 2.5**). También se comparte solo un residuo que forma una O-glicosilación de tres que presenta la proteína SARS2-S. De hecho, observamos que las secuencias más cercanas, las proteínas S de PgCoV y CoV RaTG13, expresan exactamente las mismas Asn en esas posiciones que el agente causal de COVID-19. Estas

Asn únicas podrían haber sido importantes para la patogénesis de los virus, pero no para la unión al receptor ECA2 por que no se encuentran cercanas al RBM. Usando tanto el alineamiento múltiple como la estructura, se observó que ninguna mutación de las variantes sustituyen al residuo glicosilado pero si a aminoácidos cercanos. Esto podría alterar la estructura secundaria de la proteína alterando la superficie del escudo de carbohidratos y así los epítomos reconocidos por el sistema inmunológico.

Tabla 2.5- Sitios de glicosilación de SARS2-S comparados con las secuencias de las proteínas S que se unen a ECA2. Para cada sitio se indica si hay alguna variantes de SARS-CoV-2 presenta una mutación cercana.

| SARS2-S | Tipo de glicosilación | de Conservado | Variantes con mutaciones cercanas | SARS2-S | Tipo de glicosilación | de Conservado | Variantes con mutaciones cercanas |
|---------|-----------------------|---------------|-----------------------------------|---------|-----------------------|---------------|-----------------------------------|
| 17 | N- gli | No | γ, δ | 616 | N- gli | Si | α, β, γ, δ, ε, ζ, η, θ, ι, κ, λ |
| 61 | N- gli | Si | α, η | 657 | N- gli | No | γ |
| 74 | N- gli | No | α, η, λ | 676 | O-gli | Si | η |
| 122 | N- gli | Si | | 709 | N- gli | Si | β |
| 149 | N- gli | No | α, η | 717 | N- gli | Si | α |
| 165 | N- gli | Si | | 801 | N- gli | Si | |
| 234 | N- gli | Si | | 1074 | N- gli | Si | κ |
| 282 | N- gli | Si | | 1098 | N- gli | Si | |
| 323 | O-gli | No | | 1119 | N- gli | Si | α |
| 325 | O- gli | No | | 1134 | N- gli | Si | |
| 331 | N- gli | Si | | 1173 | N- gli | Si | γ, ζ,, θ |
| 343 | N- gli | Si | | 1194 | N- gli | Si | |
| 603 | N- gli | Si | | | | | |

Discusión

Los resultados más notables en este trabajo son: 1. Se estudió la evolución de la proteína S y de su Dominio de Unión al Receptor ECA2 humano. 2. Se utilizó una prueba de sustitución de aminoácidos basada tanto en la estructura del Código Genético Estándar como en su correspondiente gráfica fenotípica de aminoácidos, el cual nos permitió determinar, para cada aminoácido, su tipo de evolución: neutra, selección positiva o selección negativa. 3. Se analizaron las interacciones entre el RBD del SARS2-S con ECA2 con una distancia promedio de 3.2 Å. 4. Se predijeron las estructuras de las proteínas S de las principales variantes de preocupación/interés de SARS-CoV-2. 5. Combinando la prueba de neutralidad con los resultados de la unión entre la proteína S y ECA2 se observaron cambios muy importantes en la evolución del RBD que facilitaron la unión al receptor ECA2 en comparación con otras proteína S (de SARS-CoV-1 y otros coronavirus). 6. Se identificaron los potenciales sitios de glicosilación de la proteína S tanto del tipo N como del tipo O y se compararon con el resto de proteínas S de unión al receptor humano. 7. Se aplicó la prueba de neutralidad a 200 secuencias de proteínas S y del genoma completo de aislados clínicos humanos con SARS-CoV-2 y se observó que no sólo evolucionó la proteína S sino que también hubo varias mutaciones a lo largo del genoma viral en comparación con otros *Betacoronavirus*.

Basado en un modelo de evolución neutral ^[92], se calculó la probabilidad de ocurrencia de nuevos aminoácidos surgidos tras mutaciones de sustitución para las proteínas S de los *Betacoronavirus*, así como en el RBD de esas proteínas, y se hizo otra prueba de neutralidad a proteínas S de aislados clínicos de SARS-CoV-2. La selección de unas mutaciones sobre otras, sesga la evolución neutral, alterando la frecuencia de cada aminoácido con respecto a lo esperado por evolución neutral. De forma que se genera una presión de selección que suele ser similar entre proteínas homólogas, y es de las principales fuerzas impulsoras de la adaptación.

Las mutaciones que dan lugar a aminoácidos bajo selección negativa como Trp, Cys, His, Gly, Pro, Ser y Arg, suelen alterar la estabilidad termodinámica ya sea en

términos de la estructura, función o interacciones proteína-proteína. Por lo tanto la mayoría de estas mutaciones se eliminan y se conservan sólo ciertos residuos en determinadas posiciones. Consecuentemente, debido a la eliminación, la frecuencia de estos aminoácidos es menor a la esperada y por ello se le llama selección negativa. Por ejemplo, ocho Cys conservan la estructura del RBD formando enlaces disulfuro (C336-C361, C379-C432, C391-C525 y C480-C488) y 30 Cys en toda la proteína. Dada las características de las Cys, puede desestabilizar si se presentan en sitios nuevos, por lo tanto se eliminan o no se alcanzan a fijar salvo aquellas que favorezcan la estructura.

Desde un punto de vista funcional, las proteínas S necesitan ser procesadas por diferentes proteasas del huésped para separar la S1 de S2 y exponer el péptido de fusión, lo que es necesario para que la maquinaria de fusión pueda inducir la fusión de las membranas y así lograr la infección celular. Las proteínas S de otros coronavirus son escindidas por tripsina, convertasas tipo furina, catepsinas, serina proteasas, elastasas, plasmina, entre otros, para lograr la exposición del péptido de fusión ^[101]. Por tanto, las proteasas del huésped ejercen una presión selectiva importante sobre las proteínas de los picos y son responsables de los tropismos hacia las células que las expresan.

El hecho de que el motivo polibásico en SARS2-S (681-RRAR) sea el resultado de una inserción de 12 nucleótidos, ha generado una gran polémica acerca del origen del virus. La inserción se logra con los residuos (PRRA), de los cuales la Pro y Arg están bajo fuertes presiones negativas. Esto quiere decir que esos aminoácidos no suelen ser favorables en el resto de la proteína pero en ese sitio resultó ser muy favorecedor para aumentar la infectividad del virus ^[10]. Sin embargo, para la teoría neutral de la evolución, la probabilidad de que ocurran estos aminoácidos está directamente relacionada con la codonicidad de cada uno, y puesto que Pro es tetracodónico y las Arg y Ala son hexacodónicos, la probabilidad de aparición de dichos aminoácidos es alta. Una vez presentado el motivo, fue seleccionado ya que esta inserción ha sido un factor de virulencia crucial que permite la escisión por la proteasa furina que, en consecuencia, deja un motivo de unión a la proteína transmembranal neuropilina-1 que también participa en la internalización del virus ^[27,28].

Esta característica única entre los coronavirus más cercanos ha sorprendido y ha generado la polémica sobre el origen del SARS-CoV-2. Sin embargo otros investigadores han demostrado que la inserción del sitio de escisión de furina (RXXR) no es nueva entre los Coronavirus^[9] pues otros *Betacoronavirus* presentan motivos polibásicos similares. No obstante, hasta ahora nuestro análisis y las evidencias no permiten esclarecer si este motivo es favorecido por eventos epistáticos. Sin embargo, en todos los casos estudiados, el motivo incrementa la virulencia de esos virus. En nuestro alineamiento múltiple pudimos observar algunas de las proteínas S de Coronavirus con este motivo: el virus de la hepatitis murina (YP009824982.1), CoV murino RA59 / R13 (ACN89689.1), CoV murino RA59 / SJHM (ACN89705.1), CoV de rata (YP003029848.1), HCoV HKU1 (YP173238 .1), HCoV OC43 (YP009555241.1), Conejo CoV (YP005454245.1), Canino CoV (AQT26498.1), CoV entérico humano (ACJ35486.1), CoV bovino (NP150077.1) y ciervo Sambar CoV US / OH-WD388TC / 1994 (ACJ67012.1). De hecho la utilización de las diferentes proteasas del huésped permite un tropismo por las células blanco^[101], y se entiende que es el resultado de la presión de selección que se ejerce sobre los virus para procesar sus proteínas S.

La prueba de neutralidad mostró varios aminoácidos que están bajo presiones de selección parecidas tanto en el RBD como en la totalidad de las proteínas S de los *Betacoronavirus*. Por ejemplo Tyr y Asn están sujetas a altas presiones de selección positiva, las cuales favorecen que se fijen las mutaciones con estos aminoácidos. Por ejemplo son 4 Tyr las usadas en la interacción proteína – proteína, estas se presentan en todos los coronavirus de unión a ECA2, por lo que esa presión positiva hace difícil que sean sustituidas. Aparte, es interesante que algunas de las mutantes de interés y de preocupación presenten una Tyr más entre los aminoácidos que interaccionan con el receptor (N501Y). Más adelante se describirá esta mutación, sin embargo en este punto se quiere remarcar que las Tyr tienen características que favorecen la interacción con el receptor.

La presión positiva ejercida sobre la Asn permite que las mutaciones de Asn se fijen en frecuencias mayores que las esperadas por evolución neutral o deriva génica. Esta situación favorece que más Asn sean susceptibles a glicosilaciones, mientras se exprese el

motivo NXS/T (Asn, cualquier aminoácido excepto Pro, Ser o Thr). Las proteínas S de MERS-CoV, SARS-CoV y SARS-CoV-2 se han descrito como proteínas altamente glicosiladas, esta característica les permite evadir al sistema inmunológico, ya que varios de los potenciales epítomos neutralizantes se encuentran cubiertos por carbohidratos. También se usa la glicosilación como una protección para enmascarar regiones peptídicas que también podría ser reconocido por anticuerpos. Por último ya que las glicosilaciones son añadidas usando la maquinaria del huésped, el sistema inmunológico es engañado. Es decir las glicosilaciones son reconocidas como propias, y de hecho le ayudan a generar interacciones con el huésped. Por otro lado las glicosilaciones influyen en el plegamiento adecuado de la proteína, así como en las interacciones con el trímero de S. De las 22 N-glicosilaciones solo dos se encuentran en el RBD (N331, N343) pero se encuentran lejos del RBM, por lo que pudieran favorecer otras interacciones y no la interacción con el receptor. Aparte, la alineación con otras secuencias sugiere que probablemente en los otros coronavirus se formen enlaces similares. Cabe sugerir que las glicosilaciones únicas en SARS2-S hayan surgido antes de la zoonosis ya que los residuos glicosilados únicos en SARS2-S entre los de unión a ECA2, sí son compartidos con las secuencias de la proteína S de RaTG13 y PnCoV.

El reconocimiento por parte del sistema inmunológico provoca una fuerte presión selectiva en el virus, y las Asn se encuentran favorecidas ya que están bajo presión selectiva positiva. Esto favorece la formación del escudo protector de sacáridos que protegen a la proteína S del reconocimiento ^[32]. La alteración de los carbohidratos o de los aminoácidos cercanos podrían alterar tridimensionalmente a la proteína, exponiendo nuevos epítomos. Aunque ninguna de las sustituciones de las variantes reemplace a un residuo glicosilado, sí hay mutaciones cercanas, especialmente en los sitios superantigénicos del NTD: asas del residuo 14 al 20, del 140 a 158 y del 245 al 264 ^[102]. Según los reportes sobre las variantes de preocupación, estas sustituciones reducen el reconocimiento de sueros de pacientes ^[103]. De acuerdo a la prueba de neutralidad, hay más probabilidad de alterar residuos cercanos a la Asn involucrada en el enlace, que a la misma Asn. Entonces, posiblemente el carbohidrato o los carbohidratos van a permanecer

pero los residuos aledaños serán los responsables de alterar los epítomos. Queda por determinar experimentalmente si estas mutaciones evitan el enlace glucosídico o si la cadena de sacáridos se altera; y consecuentemente determinar los nuevos epítomos.

Por otro lado se ha demostrado que la proteína SARS2-S se une a receptores de carbohidratos de las células dendríticas como: CD206 (MR), CD209 (DC-SIGN), CD209L (L-SIGN/ CLEC4M), CD301 (MGL/CLEC10A) y Dectin-2/ CLEC6A ^[34,35]. La unión es a través de los carbohidratos del NTD. En otros coronavirus, el NTD igualmente glicosilado juega un papel importante en la unión a la célula huésped y en la evasión de la respuesta inmune. Para el virus de la influenza C y algunos coronavirus (HCoV HKU1 y HCoV OC43), la unión a través de los receptores de ácido siálico 9-O-acetilado es crucial y constituye otra barrera para la infección de distintas especies ^[33,104]. La interacción con proteínas específicas del huésped facilita el acercamiento de la maquinaria de fusión a la membrana celular. Por lo tanto, es importante conocer el impacto de la glicosilación de la proteína SARS2-S en el reconocimiento del receptor, la fusión a la membrana o la evasión del sistema inmunológico. Conocer el papel de los sacáridos unidos a la proteína S podría ayudar a comprender más la fisiopatología del virus y también a desarrollar otras estrategias profilácticas o terapéuticas, efectivas contra todas las variantes.

Al observar las **Tablas 2.2 y 2.3**, es evidente que la mayoría de los aminoácidos involucrados en la interacción con ECA2 están bajo presión de selección. Este resultado es de esperarse pues es un dominio que presenta gran adaptabilidad para mejorar su afinidad al receptor del huésped o para infectar a otras especies. Dentro de las mutaciones únicas en SARS2-S (**Tabla 3**) se identificaron mutaciones que han favorecido fuertemente la interacción con el receptor, por ejemplo las Phe 456 y Phe 486, que permiten mayor número de contactos hidrofóbicos con el receptor. Ambas sustituyen a una Leu, expresadas en otras secuencias de unión a ECA2. La presencia de las Phe en la interfaz, contribuye a generar interacciones más hidrofóbicas con átomos de ECA2: SARS2-S Phe456 con Asp 30, Thr 27 y Lys 31 de ECA2 y SARS2-S Phe486 con Leu 79, Met 82 y Tyr 83 de ECA2. Estas interacciones hidrofóbicas fueron probablemente la fuerza impulsora para mantener las mutaciones. Por otro lado, la Gln 498 permite mayor número de

contactos (7 en vez de 5) por fuerzas de van der Waals. La Lys 417 y Gly 426 permiten la formación de nuevos puentes de hidrógeno. Todos estos aminoácidos menos la Gly, se encuentran bajo presiones positivas.

Las variantes de SARS-CoV-2 se analizaron con la misma prueba de neutralidad para discutir el curso evolutivo de la proteína S en estos virus. Estudiar las mutaciones compartidas entre las variantes permitiría entender las presiones de selección ejercidas sobre el virus. Estas han tenido relevancia clínica porque sus frecuencias han aumentado rápidamente en las diferentes poblaciones. Se recuperaron todas las mutaciones en la proteína S que se han reportado en las variantes Alpha, Beta, Gamma, Delta, Epsilon, Iota, Kappa, Zeta, Eta, Theta y Lambda, y se localizaron en la estructura. La mayoría de las variantes presentan mutaciones en S1 e incluso varias supresiones o eliminaciones, por lo que se predijeron las estructuras de 7 variantes. Las estructuras predichas no muestran cambios sustanciales más que en la orientación del RBD. La habilidad para mantener la estructura y funcionalidad de los diferentes dominios a pesar de presentar algunos cambios, también debe ser motivo de selección.

La mayoría de las mutaciones compartidas entre las variantes, presentan aminoácidos bajo presiones de selección (positiva o negativa). En los casos particulares, la mutación D1118H de la variante Alfa introdujo otra His en S2, que puede alterar la estructura y función debido a la carga opuesta de los aminoácidos. La sustitución S982A presenta un aminoácido hexacodónico bajo neutralidad y en el mismo grupo según su requisito polar ^[105,106], por lo que la ocurrencia fue altamente probable. En la variante Beta, la mutación A701V se sustituye por un aminoácido con selección positiva a través de una sustitución en la segunda posición del triplete. La variante Gamma es la variante con más mutaciones a lo largo de la proteína S, y además la mayoría de estas mutaciones implican cambios hacia aminoácidos con selección positiva, como L18F, T20N, D138Y, H655Y, T1027I y V1176F. También P26S y R190S cambiaron a Ser (neutros o sin presión de selección). La variante de Épsilon tiene solo una mutación impulsada por selección positiva S13I y dos mutaciones seleccionadas negativamente W152C y L452R, la primera

de las cuales resultó en el cambio de aminoácidos con requisitos polares similares, y la segunda mutación que involucra a.a. con diferentes requisitos polares.

Resaltan las mutaciones compartidas entre las variantes. Estas se deben por presiones selectivas en común, por ejemplo Pro 681 que se encuentra justo antes del sitio de corte de la proteasa furina es substituida por His o por Arg, los cuales son aminoácidos con carga positiva seleccionados negativamente. La consecuencia directa es que puede causar una ligera diferencia en la estructura secundaria inmediatamente antes del sitio de escisión de la furina. Esto basado en que Pro introduce cambios de direccionalidad en el esqueleto de Carbonos alfa ($C\alpha$) ya que el grupo amino se une tanto al $C\alpha$ como a la cadena lateral en forma cíclica. La ausencia de la Pro podría alterar la susceptibilidad a la escisión por la furina, sin embargo se necesitaría confirmarlo experimentalmente. Con análisis computacional se confirmó que la mutación P681R aumenta las propiedades básicas del motivo, lo que aumenta la susceptibilidad al corte por la furina ^[107]. Y de manera experimental se comprobó que las propiedades fusogénicas de la proteína aumentan, ya que forma más sincitios. Aparte de que la substitución permite una región de baja complejidad la cual contribuye a la evasión del sistema inmunológico ^[108].

La secuencia de referencia de SARS2-S tiene una Lys en la posición 417 que permite la formación de un puente de hidrógeno, sin embargo las variantes de Beta y Gamma lo sustituyen por aminoácidos polares sin carga (K417N/T). Queda por confirmar si estas proteínas S podrían mantener el puente de hidrógeno en la interacción proteína-proteína con el receptor, ya que el átomo de Nitrógeno de la cadena lateral que estaba donando el electrón, ya no se encuentra. De manera similar, la mutación E484K, también compartida, ocurre hacia un aminoácido con casi el mismo requisito polar a pesar de que la carga cambia, conduce a un aminoácido con selección positiva. Se ha descrito que esta substitución es resistente a algunos anticuerpos neutralizantes ^[109]. Otra mutación compartida es la substitución D614G que está presente en todas las variantes analizadas. La mutación se traduce en un aminoácido pequeño sin carga que se selecciona negativamente. Se ha comprobado que esta substitución aumenta la unión a ECA2 en comparación con el virus ancestral, por lo que también ha aumentado la infectividad y

transmisión de las variantes ^[61,110,111]. De hecho, los pacientes infectados por virus expresando esta mutación, presentan mayor carga viral ^[112]. D614G altera la afinidad por el receptor debido a un cambio conformacional que provoca que los RBD se dispongan en la posición “arriba”, lo que es necesario para el reconocimiento del receptor ^[113]. Además, la sustitución compartida N501Y que agrega otra Tyr a la interfaz con el receptor, aumenta la unión al receptor ^[114,115]. Esta sustitución permite la infección a células de ratones a través de la interacción con ECA2 de ratón ^[116] generando así la posibilidad de implementar otros modelos *in vivo* para la investigación.

Es probable que estas mutaciones compartidas entre diferentes variantes ya se hayan fijado en las partículas virales de SARS-CoV-2 que actualmente están infectando humanos. Algunas de las mutaciones presentes en variantes de orígenes distintos refleja una convergencia evolutiva proveniente de una fuerte selección en común. Se ha reportado que las mutaciones compartidas entre las variantes de preocupación aumentan la transmisibilidad del virus ^[109], de forma que ciertas mutaciones se vuelven mayormente representadas. Las presiones de selección en este coronavirus, están relacionadas con la eficiencia para internalizarse a las células, las interacciones con proteínas del huésped, la susceptibilidad a los cortes proteolíticos, la eficiencia de replicación y la evasión al sistema inmunológico ^[117], y por supuesto la competencia con otras partículas virales. Por ejemplo se sabe que la variante Delta rápidamente desplazó en menos de un mes a la variante Alfa en Reino Unido y Estados Unidos en mayo del 2021 ^[118].

A pesar de que todas las regiones de una proteína tienen la misma probabilidad de mutar, las mutaciones en el subdominio S2 son menos evidentes. Esto es porque la mayoría de las mutaciones que caen en esta región, se eliminan antes de que se detecten. Por consiguiente, resulta un subdominio conservado. Esto a su vez está impulsado por la alta especialización del dominio: la fusión con la membrana de la célula huésped. Este subdominio está involucrado en la fusión con la membrana plasmática, por lo que se entiende que requiere de características fisicoquímicas específicas. Contiene tanto al péptido de fusión como a las repeticiones de siete aminoácidos que forman la maquinaria de fusión ^[7]. Por el contrario, S1 es muy importante para el reconocimiento del receptor,

específicamente el CTD donde está contenido el RBD, por lo que la secuencia de este dominio es crucial para adaptarse al hospedador actual o para adaptarse a otras especies. Interesantemente, algunas de las variantes de SARS-CoV-2 presentan proporcionalmente mayor número de mutaciones en S2, que otras variantes. El significado biológico exacto de esta observación no es claro. Es probable que las variantes con menor número de mutaciones: i) habían estado mutando igual en ambos subdominios pero para cuando se aislaron las muestras, las mutaciones del S2 ya habían sido eliminadas, o ii) la mayoría de las mutaciones ocurrieron en S1 desde el principio. Para apoyar alguna de las hipótesis deberían ser evaluadas mayor número de variantes y considerar tanto la tasa de mutación del virus, como de transmisión en los diferentes países.

Cada una de las proteínas virales tiene distintas presiones de selección que las moldean y por lo tanto su patrón de evolución es distinto a la de cualquier otra. De igual forma tomar diferentes genes o el genoma completo del virus, muestra la probabilidad de ocurrencia de los aminoácidos en toda esa secuencia. Los análisis de neutralidad con el genoma completo con la proteína S y con el RBD de las secuencias de unión a ECA2 humano, demuestra justamente las diferencias según el tamaño de secuencia a analizar. Aquí se calcula la probabilidad de ocurrencia de los aminoácidos según el tipo de presiones de selección. Las presiones de selección ejercidas sobre cada proteína, tanto estructurales como no estructurales, están directamente relacionadas con las interacciones con otras proteínas, funcionalidad e incluso dónde se encuentran dentro de la partícula. Por ende es de esperarse que el análisis del genoma completo arroje probabilidades distintas para cada aminoácido a las probabilidades detectadas en una de las proteínas virales.

La acumulación de mutaciones está relacionada con la capacidad de corregir errores. Los coronavirus codifican una ARN polimerasas dependiente de ARN, que son propensas a equivocarse, pero a diferencia de otros virus de ARN, también tienen una exoribonucleasa de 3' a 5' (nsp14-ExoN) que corrige la nueva secuencia ^[2]. La nsp14-ExoN contribuye a mantener el genoma largo y estable. Por lo tanto, la acumulación de errores se disminuye. La prueba de neutralidad usa el principio de evolución neutral molecular

Evolución molecular de la proteína Spike del SARS-CoV-2

donde se consideran mutaciones sinónimas y no sinónimas; lo que permite obtener información de los tipos de presiones selectivas que influyeron en las mutaciones neutrales.

Conclusión

La larga duración de la pandemia, la amplia distribución geográfica y el rápido contagio, principalmente durante las olas epidémicas, han influenciado la generación de variantes del SARS-CoV-2. A escala global y local, se puede apreciar la evolución de este virus. Las vacunas han cumplido con el objetivo de detener la transmisión en algunas ciudades, sin embargo todavía falta una gran cobertura.

Los estudios evolutivos juegan un papel importante en la prevención de catástrofes epidemiológicas y en el desarrollo de mejores tratamientos que cubran la mayoría de las variantes virales. En este caso se debe partir de la concepción de que el primer mecanismo evolutivo se basa en el hecho de que la mayoría de las mutaciones ocurren estocásticamente. No son el resultado de una presión selectiva que lleve a conferir alguna ventaja o desventaja, sino que suceden y posteriormente la selección o la deriva génica pueden operar para fijar dichas mutaciones. Es decir, algunas mutaciones se mantendrán hasta que se produzca otra sustitución o bien hasta que se fije por algún otro mecanismo evolutivo. Se calcula que si una mutación no es sustituida en un lapso de cuatro generaciones, esta será fijada de manera neutral, es decir sin estar sujeta a presiones de selección. Es así como la deriva génica toma importancia para explicar mutaciones que no están directamente relacionadas con la adecuación.

Otras de las mutaciones ocurridas estocásticamente, serán seleccionadas por poseer características favorables para la proteína (presión positiva) mientras que otras mutaciones se eliminarán por no ser favorables (presión negativa); por lo que consecuentemente, en ambos casos se fijarán en una proporción distinta a la esperada. Las presiones selectivas son similares entre proteínas homólogas, más aún, se mantienen a largo plazo. Es por ello que es de esperarse encontrar mutaciones compartidas entre variantes de SARS-CoV-2 con diferentes orígenes geográficos. Las presiones selectivas importantes para la evolución del SARS-CoV-2 están relacionadas con la patogénesis, como mutaciones que mejoran la adecuación, alteran las interacciones con las proteínas del huésped o evaden al sistema inmunológico del huésped. Las principales mutaciones

compartidas tanto en las variantes de interés como en las de preocupación son E484K, N501Y, D614G, así como los aminoácidos eliminados cercanos a los sitios de glicosilación. Como consecuencia, estas se seleccionaron por conducir a la adaptación y el éxito del virus.

La prueba de neutralidad nos permitió calcular el tipo de presión selectiva para cada aminoácido en la proteína S de los *Betacoronavirus*. Este estudio evolutivo permite comprender y describir los cambios en la secuencia del SARS2-S que afectan su estabilidad, estructura o función. Pudimos discutir sobre la inserción de un motivo que permite la escisión de una proteasa, que a pesar de que individualmente los aminoácidos no están favorecidos, el motivo fue seleccionado porque proveyó una clara ventaja en la transmisibilidad del virus. Es decir, la selección del motivo polibásico se impuso al mecanismo de eliminación de dichas inserciones que era lo esperado por la prueba. Además, se discutieron las presiones selectivas en el RBD que favorecieron la capacidad del SARS-CoV-2 para infectar a los seres humanos. Y se discutió que estas presiones están relacionadas con las propiedades químicas de los aminoácidos que interactúan con el receptor, es decir, algunos cambios favorecieron aumentando el número de puentes de hidrógeno o permitiendo un mayor número de contactos con el receptor. Como era de esperarse para el RBD, aquí se ha demostrado que la mayoría de las mutaciones han sido fijadas por selección más que por neutralidad. Y hemos ido más lejos determinando cuáles aminoácidos son favorecidos por estas presiones. La mayoría de las mutaciones en la interfaz y las variantes de interés se vieron favorecidas con la presencia de Tyr y Asn.

La prueba de neutralidad realizada a la proteína SARS2-S de aislados humanos muestra el comportamiento evolutivo actual del virus. Pudimos aplicar la prueba principalmente porque la rápida propagación del virus ha permitido la fijación de mutaciones. Aquí demostramos de manera cuantitativa las presiones selectivas reales ejercidas sobre el virus. La Figura 2.1B muestra que algunos de los aminoácidos seleccionados para las proteínas S en los *Betacoronavirus* todavía están bajo selección positiva en las proteínas de SARS2-S (es decir, Tyr, Lys, Phe, Asn, Asp).

De los 17 aminoácidos que interaccionan con el receptor, solo 2 de ellos han mutado en algunas de las variantes, y éstas son mutaciones compartidas y catalogadas de interés. En perspectiva, sería interesante comprobar si el cambio de K417N/T afecta en la formación del puente de hidrógeno con el receptor, el cual favorece la estabilidad de la unión. Se sabe que N501Y, otro aminoácido que interacciona, aumenta la afinidad por el receptor ^[114]. Sin embargo las mutaciones compartidas en el RBD se encuentran muy cercanas a los residuos que interaccionan con el receptor, por lo que se sugiere que haya cambios en la interacción, los cuales se podrían estudiar experimentalmente o *in silico* por dinámicas moleculares.

Por lo tanto, la hipótesis planteada inicialmente se acepta, ya que la mayoría de las mutaciones neutrales, sobre todo en el RBD, han estado sujetas a presiones que han aumentado la transmisión del virus. Después de hacer la prueba de neutralidad, se identificaron las mutaciones únicas en la proteína SARS2-S tanto en el RBD como en la proteína completa y también de las variantes de preocupación. Para cada una se evaluó el tipo de presión ejercida y posteriormente se tomaron en cuenta sus propiedades químicas así como la posición dentro de la estructura. Estos análisis, junto con la revisión de la literatura, sugirieron que el impacto de algunas mutaciones fue en el aumento de la afinidad por el receptor (Phe456, Phe 486, Q498), en la interacción con otras proteínas del huésped o en la evasión del reconocimiento del sistema inmunológico.

La pandemia causada por el SARS-CoV-2 ha impactado a nivel mundial en diferentes aspectos de la vida moderna; desde la salud física y mental hasta el desarrollo económico. A más de 18 meses se cuenta ya con varias vacunas que están siendo aplicadas a millones de personas, y otras tantas aún en desarrollo. El ritmo al que se ha obtenido el conocimiento del SARS-CoV-2 ha sido impresionante gracias al desarrollo tecnológico de los últimos años, permitiendo la secuenciación del genoma y el desarrollo de modelos *in vitro* e *in vivo* que a su vez han generado un mundo de conocimiento alrededor de la biología del virus así como de su tratamiento. Hasta ahora se siguen descubriendo características de su ciclo de vida así como de su posible origen y evolución.

Perspectivas

Este trabajo permitió generar nuevas preguntas sobre la evolución del SARS-CoV-2 que permiten ampliar esta línea de investigación. Por ejemplo, la mayoría de los sitios susceptibles a glicosilación se han conservado en los *Betacoronavirus* del linaje B, sin embargo no sabemos si los patrones de glicosilación se mantienen. Ya que los virus utilizan las enzimas glicosil- transferasas del huésped, la única información para determinar las cadenas glicosídicas en los sitios correctos, es la misma secuencia proteica y la expresión de las distintas glicosil- transferasas en las células blanco. Según la literatura, las glicosilaciones juegan un papel importante en el encubrimiento de determinantes antigénicos. Sin embargo, si la mayoría de los patrones de glicosilación no cambian, los diferentes receptores de carbohidratos del sistema inmunológico e incluso anticuerpos serían capaces de distinguir específicamente a la proteína S mediante el reconocimiento de los carbohidratos conservados. Es por ello que se deberían estudiar los patrones de glicosilación a lo largo de varias generaciones y en distintas células huésped para poder determinar su verdadero papel en la evasión del sistema inmunológico.

El estudio de los receptores virales podría ayudar a entender la co-evolución de los coronavirus humanos con los receptores. La tasa de mutación de los virus, así como su replicación es mucho mayor a la del huésped por lo que se podría considerar que las proteínas del huésped son casi constantes. Consecuentemente se parte de la idea de que la adaptación del virus en torno al receptor es determinante para su éxito. Entonces ¿será posible que estos receptores hayan sido seleccionadas por la importancia sistémica que representan, se expresan en una variedad de tejidos, etc. o simplemente los coronavirus se encuentran fácilmente con células que los expresan?, ¿habrá más que descubrir de los receptores virales, tienen otro papel aparte de ser ancla para el virus?

Referencias

1. Khailany, R. A., Safdar, M. & Ozaslan, M. Genomic characterization of a novel SARS-CoV-2. *Gene Reports* **19**, 100682 (2020).
2. Smith, E. C. & Denison, M. R. Coronaviruses as DNA Wannabes: A New Model for the Regulation of RNA Virus Replication Fidelity. *PLoS Pathog.* **9**, e1003760 (2013).
3. Zhou, P. *et al.* A pneumonia outbreak associated with a new coronavirus of probable bat origin. *Nature* **579**, 270–273 (2020).
4. Naqvi, A. A. T. *et al.* Insights into SARS-CoV-2 genome, structure, evolution, pathogenesis and therapies: Structural genomics approach. *Biochim. Biophys. Acta. Mol. Basis Dis.* **1866**, 165878 (2020).
5. Parlikar, A. *et al.* Understanding genomic diversity, pan-genome, and evolution of SARS-CoV-2. *PeerJ* **2020**, (2020).
6. Hoffmann, M. *et al.* SARS-CoV-2 Cell Entry Depends on ACE2 and TMPRSS2 and Is Blocked by a Clinically Proven Protease Inhibitor. *Cell* **181**, 271-280.e8 (2020).
7. Tang, T., Bidon, M., Jaimes, J. A., Whittaker, G. R. & Daniel, S. Coronavirus membrane fusion mechanism offers a potential target for antiviral development. *Antiviral Research* vol. 178 104792 (2020).
8. Zhao, M.-M. *et al.* Cathepsin L plays a key role in SARS-CoV-2 infection in humans and humanized mice and is a promising target for new drug development. *Signal Transduct. Target. Ther.* **2021 61 6**, 1–12 (2021).
9. Wu, Y. & Zhao, S. Furin cleavage sites naturally occur in coronaviruses. *Stem Cell Res.* **50**, 102115 (2021).
10. Johnson, B. A. *et al.* Loss of furin cleavage site attenuates SARS-CoV-2 pathogenesis. *Nature* **591**, 293–299 (2021).
11. Snijder, E. J. *et al.* A unifying structural and functional model of the coronavirus replication organelle: Tracking down RNA synthesis. *PLoS Biol.* **18**, 1–25 (2020).
12. Mandala, V. S. *et al.* Structure and drug binding of the SARS-CoV-2 envelope protein transmembrane domain in lipid bilayers. *Nat. Struct. Mol. Biol.* **27**, 1202–1208 (2020).
13. Ghosh, S. *et al.* β -Coronaviruses Use Lysosomes for Egress Instead of the Biosynthetic Secretory Pathway. *Cell* 1–16 (2020) doi:10.1016/j.cell.2020.10.039.
14. Letko, M., Marzi, A. & Munster, V. Functional assessment of cell entry and receptor usage for SARS-CoV-2 and other lineage B betacoronaviruses. *Nat. Microbiol.* **5**, 562–569 (2020).
15. Wrapp, D. *et al.* Cryo-EM Structure of the 2019-nCoV Spike in the Prefusion Conformation. *Science* (80-.). (2020) doi:10.1101/2020.02.11.944462.
16. Huang, Y., Yang, C., Xu, X. feng, Xu, W. & Liu, S. wen. Structural and functional properties of SARS-CoV-2 spike protein: potential antivirus drug development for COVID-19. *Acta Pharmacol. Sin.* **41**, 1141–1149 (2020).
17. Madu, I. G., Roth, S. L., Belouzard, S. & Whittaker, G. R. Characterization of a Highly Conserved Domain within the Severe Acute Respiratory Syndrome Coronavirus Spike Protein S2 Domain with Characteristics of a Viral Fusion Peptide. *J. Virol.* **83**, 7411–7421 (2009).

18. Walls, A. C. *et al.* Structure, Function, and Antigenicity of the SARS-CoV-2 Spike Glycoprotein. *Cell* **181**, 281-292.e6 (2020).
19. Gupta, A. *et al.* Extrapulmonary manifestations of COVID-19. *Nat. Med.* **26**, 1017–1032 (2020).
20. Jaimes, J. A., André, N. M., Chappie, J. S., Millet, J. K. & Whittaker, G. R. Phylogenetic Analysis and Structural Modeling of SARS-CoV-2 Spike Protein Reveals an Evolutionary Distinct and Proteolytically Sensitive Activation Loop. *J. Mol. Biol.* **432**, 3309–3325 (2020).
21. Wang, Q. *et al.* Structural and Functional Basis of SARS-CoV-2 Entry by Using Human ACE2. *Cell* **181**, 894-904.e9 (2020).
22. Yan, R. *et al.* Structural basis for the recognition of SARS-CoV-2 by full-length human ACE2. *Science (80-.)*. **367**, 1444–1448 (2020).
23. Benton, D. J. *et al.* Receptor binding and priming of the spike protein of SARS-CoV-2 for membrane fusion. *Nature* (2020) doi:10.1038/s41586-020-2772-0.
24. Lan, J. *et al.* Structure of the SARS-CoV-2 spike receptor-binding domain bound to the ACE2 receptor. *Nature* **581**, 215–220 (2020).
25. Li, F., Li, W., Farzan, M. & Harrison, S. C. Structural biology: Structure of SARS coronavirus spike receptor-binding domain complexed with receptor. *Science (80-.)*. **309**, 1864–1868 (2005).
26. Wang, C., Horby, P. W., Hayden, F. G. & Gao, G. F. A novel coronavirus outbreak of global health concern. *Lancet* **395**, 470–473 (2020).
27. Cantuti-Castelvetri, L. *et al.* Neuropilin-1 facilitates SARS-CoV-2 cell entry and infectivity. *Science (80-.)*. **2985**, (2020).
28. Daly, J. L. *et al.* Neuropilin-1 is a host factor for SARS-CoV-2 infection. *Science (80-.)*. **3072**, (2020).
29. Li, F. Structure, Function, and Evolution of Coronavirus Spike Proteins. *Annual Review of Virology* vol. 3 237–261 (2016).
30. Papa, G. *et al.* Furin cleavage of SARS-CoV-2 Spike promotes but is not essential for infection and cell-cell fusion. *PLoS Pathog.* **17**, (2021).
31. Zhao, X., Chen, H. & Wang, H. Glycans of SARS-CoV-2 Spike Protein in Virus Infection and Antibody Production. *Front. Mol. Biosci.* **8**, 53 (2021).
32. Sikora, M. *et al.* Computational epitope map of SARS-CoV-2 spike protein. *PLoS Comput. Biol.* **17**, e1008790 (2021).
33. Tortorici, M. A. *et al.* Structural basis for human coronavirus attachment to sialic acid receptors. *Nat. Struct. Mol. Biol.* **26**, 481–489 (2019).
34. Hoffmann, D. *et al.* Identification of lectin receptors for conserved SARS-CoV-2 glycosylation sites. *bioRxiv* 2021.04.01.438087 (2021) doi:10.1101/2021.04.01.438087.
35. Gao, C. *et al.* SARS-CoV-2 Spike Protein Interacts with Multiple Innate Immune Receptors. *bioRxiv* (2020) doi:10.1101/2020.07.29.227462.
36. Woo, P. C. Y. *et al.* Discovery of Seven Novel Mammalian and Avian Coronaviruses in the Genus Deltacoronavirus Supports Bat Coronaviruses as the Gene Source of Alphacoronavirus and Betacoronavirus and Avian Coronaviruses as the Gene Source of Gammacoronavirus and Deltacoronavirus. *J. Virol.* **86**, 3995–4008 (2012).

37. Zhao, X., Ding, Y., Du, J. & Fan, Y. 2020 update on human coronaviruses: One health, one world. *Med. Nov. Technol. Devices* **8**, 100043 (2020).
38. Lu, G., Wang, Q. & Gao, G. F. Bat-to-human: Spike features determining 'host jump' of coronaviruses SARS-CoV, MERS-CoV, and beyond. *Trends Microbiol.* **23**, 468–478 (2015).
39. Peiris, J. S. M. *et al.* Coronavirus as a possible cause of severe acute respiratory syndrome. *Lancet* **361**, 1319–1325 (2003).
40. Zaki, A. M., van Boheemen, S., Bestebroer, T. M., Osterhaus, A. D. M. E. & Fouchier, R. A. M. Isolation of a Novel Coronavirus from a Man with Pneumonia in Saudi Arabia. *N. Engl. J. Med.* **367**, 1814–1820 (2012).
41. Zhao, J., Cui, W. & Tian, B. P. The Potential Intermediate Hosts for SARS-CoV-2. *Front. Microbiol.* **11**, (2020).
42. Lam, T. T. Y. *et al.* Identifying SARS-CoV-2-related coronaviruses in Malayan pangolins. *Nature* **583**, 282–285 (2020).
43. Han, G.-Z. Pangolins Harbor SARS-CoV-2-Related Coronaviruses. *Trends Microbiol.* **28**, 515–517 (2020).
44. López- Cortés, G. I., Díaz- Alvarez, L. & Ortega, E. Membrane enzymes play the cell adhesion game. *Front. Immunol.*
45. Ayres, J. S. A metabolic handbook for the COVID-19 pandemic. *Nat. Metab.* **2**, 572–585 (2020).
46. Boyton, R. J. & Altmann, D. M. The immunology of asymptomatic SARS-CoV-2 infection: what are the key questions? *Nat. Rev. Immunol.* 2021 1–7 (2021) doi:10.1038/s41577-021-00631-x.
47. T, S. *et al.* Robust T Cell Immunity in Convalescent Individuals with Asymptomatic or Mild COVID-19. *Cell* **183**, 158-168.e14 (2020).
48. M, G., DS, Y. & DV, H. Asymptomatic Transmission, the Achilles' Heel of Current Strategies to Control Covid-19. *N. Engl. J. Med.* **382**, 2158–2160 (2020).
49. Rodrigues, T. S. *et al.* Inflammasomes are activated in response to SARS-cov-2 infection and are associated with COVID-19 severity in patients. *J. Exp. Med.* **218**, (2020).
50. Barnes, B. J. *et al.* Targeting potential drivers of COVID-19: Neutrophil extracellular traps. *J. Exp. Med.* **217**, 1–7 (2020).
51. Ramos-Martínez, E. *et al.* Multiple Origins of Extracellular DNA Traps. *Front. Immunol.* **12**, 1–12 (2021).
52. Tay, M. Z., Poh, C. M., Rénia, L., MacAry, P. A. & Ng, L. F. P. The trinity of COVID-19: immunity, inflammation and intervention. *Nat. Rev. Immunol.* **20**, 363–374 (2020).
53. Felsenstein, S., Herbert, J. A., McNamara, P. S. & Hedrich, C. M. COVID-19: Immunology and treatment options. *Clin. Immunol.* **215**, 108448 (2020).
54. Polidoro, R. B., Hagan, R. S., de Santis Santiago, R. & Schmidt, N. W. Overview: Systemic Inflammatory Response Derived From Lung Injury Caused by SARS-CoV-2 Infection Explains Severe Outcomes in COVID-19. *Front. Immunol.* **11**, (2020).
55. Matthay, M. A. *et al.* Acute respiratory distress syndrome. *Nat. Rev. Dis. Prim.* 2019 51 **5**, 1–22 (2019).
56. Kim, Y.-M. & Shin, E.-C. Type I and III interferon responses in SARS-CoV-2 infection.

- Exp. Mol. Med.* 2021 535 **53**, 750–760 (2021).
57. Park, A. & Iwasaki, A. Type I and Type III Interferons – Induction, Signaling, Evasion, and Application to Combat COVID-19. *Cell Host Microbe* **27**, 870–878 (2020).
 58. P, B. *et al.* Autoantibodies against type I IFNs in patients with life-threatening COVID-19. *Science* **370**, (2020).
 59. Wang, Z. *et al.* Exposure to SARS-CoV-2 generates T-cell memory in the absence of a detectable viral infection. *Nat. Commun.* 2021 121 **12**, 1–8 (2021).
 60. Chi, X. *et al.* A neutralizing human antibody binds to the N-terminal domain of the spike protein of SARS-CoV2. *Science (80-)*. **369**, 650–655 (2020).
 61. Zhou, B. *et al.* SARS-CoV-2 spike D614G variant confers enhanced replication and transmissibility. *bioRxiv* (2020) doi:10.1101/2020.10.27.357558.
 62. Gaebler, C. *et al.* Evolution of Antibody Immunity to SARS-CoV-2. *bioRxiv* (2020) doi:https://doi.org/10.1101/2020.11.03.367391.
 63. Look, A. T., Ashmun, R. A., Shapiro, L. H. & Peiper, S. C. Human myeloid plasma membrane glycoprotein CD13 (gp150) is identical to aminopeptidase N. *J. Clin. Invest.* **83**, 1299–1307 (1989).
 64. Griffin, J. D., Ritz, J., Nadler, L. M. & Schlossman, S. F. Expression of myeloid differentiation antigens on normal and malignant myeloid cells. *J Clin Invest* **68**, 932–941 (1981).
 65. Mina-Osorio, P. The moonlighting enzyme CD13: old and new functions to target. *Trends Mol. Med.* **14**, 361–371 (2008).
 66. Garay-Canales, C. A., Díaz-Alvarez, L. & Lopez-Cortes, G. I. Novel immunotherapy strategies involving matrix metalloproteinase (MMP) family. in *Immunotherapy in resistant cancer: from the lab bench work to its clinical perspectives* (eds. MORALES-MONTOR, J. & SEGOVIA-MENDOZA, M.) vol. 2 227–247 (Academic Press Inc., 2020).
 67. Mina-Osorio, P. *et al.* CD13 is a novel mediator of monocytic/endothelial cell adhesion. *J. Leukoc. Biol.* **84**, 448–459 (2008).
 68. Licona-Limon, I., Garay-Canales, C. A., Munoz-Paletta, O. & Ortega, E. CD13 mediates phagocytosis in human monocytic cells. *J. Leukoc. Biol.* **98**, 85–98 (2015).
 69. Mina-Osorio, P., Shapiro, L. H. & Ortega, E. CD13 in cell adhesion: aminopeptidase N (CD13) mediates homotypic aggregation of monocytic cells. *J. Leukoc. Biol.* **79**, 719–730 (2006).
 70. Lu, C., Amin, M. A. & Fox, D. A. CD13/Aminopeptidase N Is a Potential Therapeutic Target for Inflammatory Disorders. *J. Immunol.* **204**, 3–11 (2020).
 71. Gabrilovac, J., Breljak, D. & Čupić, B. Regulation of aminopeptidase N (EC 3.4.11.2; APN; CD13) on the HL-60 cell line by TGF- β 1. *Int. Immunopharmacol.* **8**, 613–623 (2008).
 72. Santos, N., Langner, J., Herrmann, M. & Riemann, D. Aminopeptidase N/CD13 is directly linked to signal transduction pathways in monocytes. *Eur. PMC* **201**, 22–32 (2000).
 73. Subramani, J. *et al.* Tyrosine Phosphorylation of CD13 Regulates Inflammatory Cell-Cell Adhesion and Monocyte Trafficking. *J. Immunol.* **191**, 3905–3912 (2013).
 74. Ghosh, M. *et al.* Molecular mechanisms regulating CD13-mediated adhesion. *Immunology* **142**, 636–647 (2014).

75. Hung-Chi Cheng, Mossaad Abdel-Ghany, and B. U. P. A Novel Consensus Motif in Fibronectin Mediates Dipeptidyl Peptidase IV Adhesion and Metastasis. *J. Biol. Chem.* **278**, 24600–24607 (2003).
76. Klemann, C., Wagner, L., Stephan, M. & von Hörsten, S. Cut to the chase: a review of CD26/dipeptidyl peptidase-4's (DPP4) entanglement in the immune system. *Clinical and Experimental Immunology* vol. 185 1–21 (2016).
77. Engel, M. *et al.* The crystal structure of dipeptidyl peptidase IV (CD26) reveals its functional regulation and enzymatic mechanism. *Proc. Natl. Acad. Sci.* **100**, 5063–5068 (2003).
78. Abbott, C. A., McCaughan, G. W., Baker, E. & Sutherland, G. R. Genomic organization, exact localization, and tissue expression of the human CD26 (dipeptidyl peptidase IV) gene. *Immunogenetics* **40**, 331–338 (1994).
79. Fan, H. *et al.* Molecular mechanism and structural basis of interactions of dipeptidyl peptidase IV with adenosine deaminase and human immunodeficiency virus type-1 transcription transactivator. *European Journal of Cell Biology* vol. 91 265–273 (2012).
80. Morimoto, C. & Schlossman, S. F. The structure and function of CD26 in the T-cell immune response. *Immunol. Rev.* **161**, 55–70 (1998).
81. Kameoka, J., Tanaka, T., Nojima, Y., Schlossman, S. F. & Morimoto, C. Direct association of adenosine deaminase with a T cell activation antigen, CD26. *Science (80-)*. **261**, 466–469 (1993).
82. Lambeir, A. M., Durinx, C., Scharpé, S. & De Meester, I. Dipeptidyl-peptidase IV from bench to bedside: An update on structural properties, functions, and clinical aspects of the enzyme DPP IV. *Critical Reviews in Clinical Laboratory Sciences* vol. 40 209–294 (2003).
83. Havre, P. A. *et al.* CD26 Expression on T-Anaplastic Large Cell Lymphoma (ALCL) Line Karpas 299 is associated with increased expression of Versican and MT1-MMP and enhanced adhesion. *BMC Cancer* **13**, (2013).
84. Gine, S. *et al.* Regulation of epithelial and lymphocyte cell adhesion by adenosine deaminase-CD26 interaction. *Biochem. J* vol. 361 (2002).
85. Samavati, L. & Uhal, B. D. ACE2, Much More Than Just a Receptor for SARS-COV-2. *Front. Cell. Infect. Microbiol.* **0**, 317 (2020).
86. Gheblawi, M. *et al.* Angiotensin-Converting Enzyme 2: SARS-CoV-2 Receptor and Regulator of the Renin-Angiotensin System: Celebrating the 20th Anniversary of the Discovery of ACE2. *Circ. Res.* 1456–1474 (2020) doi:10.1161/CIRCRESAHA.120.317015.
87. Purkayastha, A. *et al.* Direct Exposure to SARS-CoV-2 and Cigarette Smoke Increases Infection Severity and Alters the Stem Cell-Derived Airway Repair Response. *Cell Stem Cell* **27**, 869-875.e4 (2020).
88. Kimura, M. The neutral theory of molecular evolution and the world view of the neutralists. *Genome* **31**, 24–31 (1989).
89. Jensen, J. D. *et al.* The importance of the Neutral Theory in 1968 and 50 years on: Aresponse to Kern and Hahn 2018. *Evolution* **73**, 111 (2019).
90. Kimura, M. The neutral theory of molecular evolution: a review of recent evidence.

- Japanese J. Genet.* **66**, 367–386 (1991).
91. Zuckerkandl, E. & Pauling, L. B. Molecular disease, evolution, and genic heterogeneity. in *Horizons in Biochemistry* (eds. In Kasha, M. & Pullman, B.) 189–225 (1962).
 92. Zamudio, G. S., Prosdocimi, F., de Farias, S. T. & José, M. V. A neutral evolution test derived from a theoretical amino acid substitution model. *J. Theor. Biol.* **467**, 31–38 (2019).
 93. Edgar, R. C. MUSCLE: Multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.* **32**, 1792–1797 (2004).
 94. Kumar, Stecher, Li, Knyaz & Tamura. MEGA version X. (2018).
 95. Pettersen, E. F. *et al.* UCSF Chimera - A visualization system for exploratory research and analysis. *J. Comput. Chem.* **25**, 1605–1612 (2004).
 96. Roy, A., Kucukural, A. & Zhang, Y. I-TASSER: A unified platform for automated protein structure and function prediction. *Nat. Protoc.* **5**, 725–738 (2010).
 97. Yang, J. *et al.* The I-TASSER suite: Protein structure and function prediction. *Nature Methods* vol. 12 7–8 (2015).
 98. Yang, J. & Zhang, Y. I-TASSER server: New development for protein structure and function predictions. *Nucleic Acids Res.* **43**, W174–W181 (2015).
 99. Zhang, C. *et al.* I-TASSER Genome wide structure and function modeling of SARS-CoV2. vol. 19 <https://zhanglab.ccmb.med.umich.edu/COVID-19/> (2020).
 100. Zhao, X., Chen, H. & Wang, H. Glycans of SARS-CoV-2 Spike Protein in Virus Infection and Antibody Production. *Front. Mol. Biosci.* **8**, 53 (2021).
 101. Millet, J. K. & Whittaker, G. R. Host cell proteases: Critical determinants of coronavirus tropism and pathogenesis. *Virus Res.* **202**, 120–134 (2015).
 102. McCallum, M. *et al.* N-terminal domain antigenic mapping reveals a site of vulnerability for SARS-CoV-2. *Cell* **0**, (2021).
 103. Plante, J. A. *et al.* The Variant Gambit: COVID’s Next Move. *Cell Host Microbe* 104743 (2021) doi:10.1016/j.chom.2021.02.020.
 104. Hulswit, R. J. G. *et al.* Human coronaviruses OC43 and HKU1 bind to 9-O-acetylated sialic acids via a conserved receptor-binding site in spike protein domain A. *Proc. Natl. Acad. Sci. U. S. A.* **116**, 2681–2690 (2019).
 105. Mathew, D. C. & Luthey-Schulten, Z. On the physical basis of the amino acid polar requirement. *J. Mol. Evol.* **66**, 519–528 (2008).
 106. Woese, C. R., Dugre, D. H., Saxinger, W. C. & Dugre, S. A. The molecular basis for the genetic code. *Proc. Natl. Acad. Sci. U. S. A.* **55**, 966–974 (1966).
 107. Saito, A. *et al.* SARS-CoV-2 spike P681R mutation enhances and accelerates viral fusion. *bioRxiv* **4**, 2021.06.17.448820 (2021).
 108. Becerra, A., Muñoz-velasco, I., Cottom-salas, W., Vázquez-salazar, A. & Hernández-morales, R. Two Short Low Complexity Regions (LCRs) are Hallmark Sequences of the Delta SARS-CoV- 2 Variant Spike Protein. 1–14.
 109. Gobeil, S. M.-C. *et al.* Effect of natural mutations of SARS-CoV-2 on spike structure, conformation and antigenicity. *bioRxiv Prepr. Serv. Biol.* **6226**, 1–24 (2021).
 110. Hou, Y. J. *et al.* SARS-CoV-2 D614G variant exhibits efficient replication ex vivo and transmission in vivo. *Science (80-.).* **370**, 1464–1468 (2021).

111. Plante, J. A. *et al.* Spike mutation D614G alters SARS-CoV-2 fitness. *Nature* 1–6 (2020) doi:10.1038/s41586-020-2895-3.
112. Korber, B. *et al.* Tracking Changes in SARS-CoV-2 Spike: Evidence that D614G Increases Infectivity of the COVID-19 Virus. *Cell* **182**, 812–827.e19 (2020).
113. Yurkovetskiy, L. *et al.* Structural and Functional Analysis of the D614G SARS-CoV-2 Spike Protein Variant. *Cell* **183**, 739–751.e8 (2020).
114. Luan, B., Wang, H. & Huynh, T. Enhanced binding of the N501Y-mutated SARS-CoV-2 spike protein to the human ACE2 receptor: insights from molecular dynamics simulations. *FEBS Lett.* **595**, 1454–1461 (2021).
115. Liu, Y. *et al.* The N501Y spike substitution enhances SARS-CoV-2 transmission. *bioRxiv Prepr. Serv. Biol.* (2021) doi:10.1101/2021.03.08.434499.
116. Gu, H. *et al.* Adaptation of SARS-CoV-2 in BALB/c mice for testing vaccine efficacy. *Science (80-.).* **369**, 1603–1607 (2020).
117. Cai, Y. *et al.* Structural basis for enhanced infectivity and immune evasion of SARS-CoV-2 variants. *bioRxiv Prepr. Serv. Biol.* **9745**, 1–14 (2021).
118. Bolze, A. *et al.* Rapid displacement of SARS-CoV-2 variant B.1.1.7 by B.1.617.2 and P.1 in the United States. *medRxiv* 2021.06.20.21259195 (2021).

Anexos

Información suplementaria

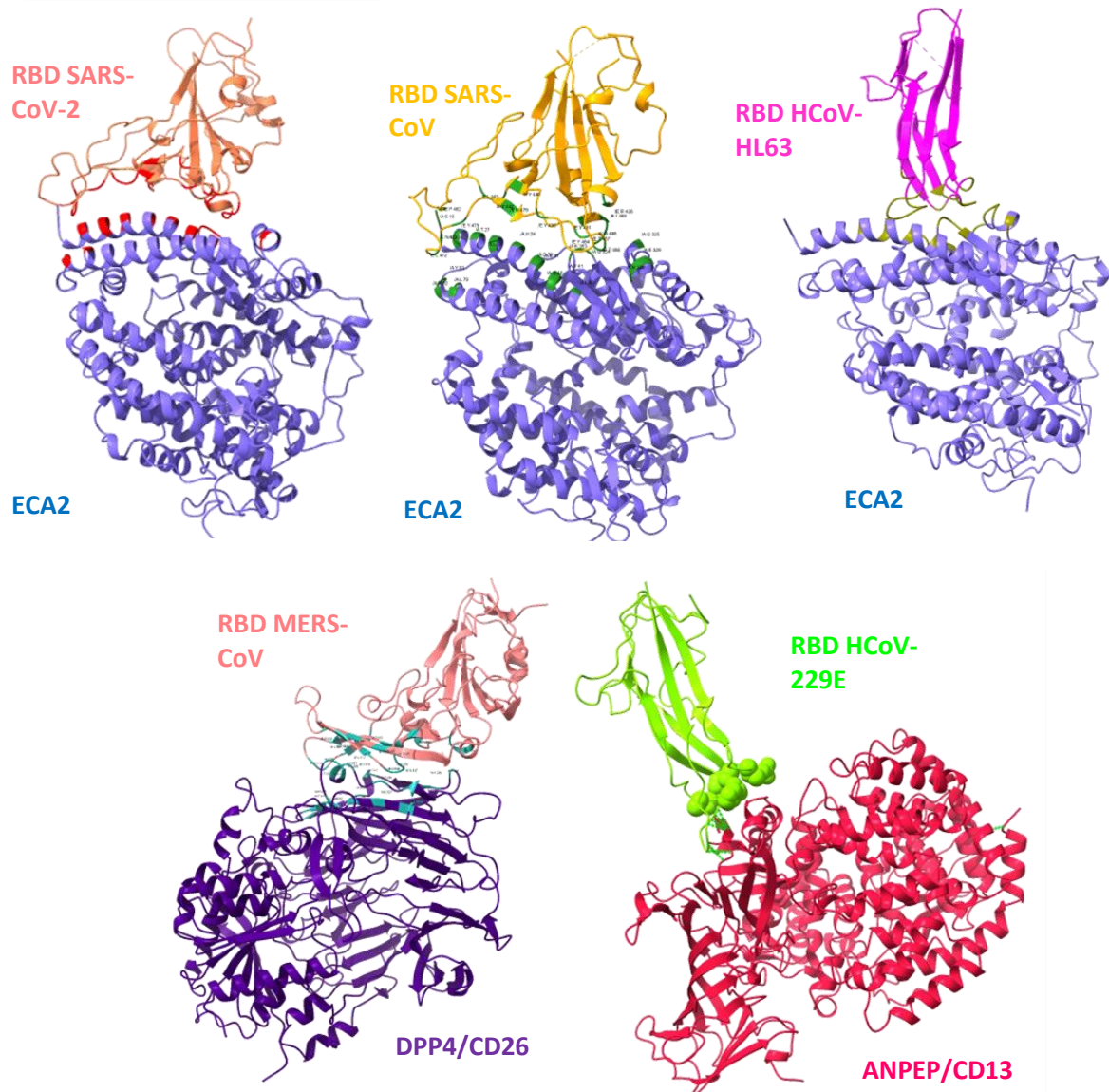


Figura S1- Interacción de los RBDs de diferentes coronavirus humanos con su receptor celular. Estructuras de las interacciones de diferentes RBDs de proteínas S de coronavirus con su respectivo receptor, en la parte superior de izquierda a derecha el RBD del SARS2-S (PDB:6M0J), SARS-S (PDB: 2AJF) y del HCoV-HL63- (PDB:3KBH) los tres con ECA2 y en la parte inferior izquierda la interacción de DPP4 con el RBD del MERS-CoV (PDB: EKRO) y la interacción de ANPEP con el RBD de HCoV-229E (PDB: 6ATK). Los residuos de contacto se encuentran resaltados en cada interacción.

Productos académicos generados durante los estudios de Doctorado

Capítulos de libro

- Garay- Canales, C.A.; Díaz, L. y **López- Cortés, G.I.** (2020). Novel immunotherapy strategies involving matrix metalloproteinase (MMP) family. En: Morales-Montor, J. y Segovia- Mendoza, M. Immunotherapy in resistant cancer: from the lab bench work to its clinical perspectives, Volumen 2. *Elsevier*, Inglaterra.
- **López- Cortés, G.I.**; Magallanes, A. y Ortega, E. Vitamin D as a modulator of the immune response during viral infections. En: Morales-Montor, J. The innate immune system in health and disease: from the lab bench work to its clinical implications (*aceptado*)

Artículos de revisión

- Ramos-Martínez E, Hernández-González L, Ramos-Martínez I, Pérez-Campos Mayoral L, **López-Cortés GI**, Pérez-Campos E, Mayoral Andrade G, Hernández-Huerta MT, José MV. (2021). Multiple Origins of Extracellular DNA Traps. *Front Immunol.* 12:621311. doi: 10.3389/fimmu.2021.621311. PMID: 33717121; PMCID: PMC7943724.
- **López-Cortés, G.I.**; Díaz-Alvarez, L. y Ortega, E. 2021. Membrane enzymes play the cell adhesion game. *Frontiers Immunology* 10.3389/fimmu.2021.742292 (*aceptado*)

Artículo original

- **López-Cortés, G.I.**; M. Palacios-Pérez, G. Zamudio, H. Velez, Ortega, E. y M. V. José. 2021. Neutral evolution test of the spike protein of SARS-CoV-2 and its implications in the binding to ACE2. *Sci Rep* **11**, 18847. <https://doi.org/10.1038/s41598-021-96950-z>



OPEN

Neutral evolution test of the spike protein of SARS-CoV-2 and its implications in the binding to ACE2

Georgina I. López-Cortés^{1,2}, Miryam Palacios-Pérez², Gabriel S. Zamudio²,
Hannya F. Veledíaz^{2,3}, Enrique Ortega^{1,2,3} & Marco V. José^{2,3,4}

As the SARS-CoV-2 has spread and the pandemic has dragged on, the virus continued to evolve rapidly resulting in the emergence of new highly transmissible variants that can be of public health concern. The evolutionary mechanisms that drove this rapid diversity are not well understood but neutral evolution should open the first insight. The neutral theory of evolution states that most mutations in the nucleic acid sequences are random and they can be fixed or disappear by purifying selection. Herein, we performed a neutrality test to better understand the selective pressures exerted over SARS-CoV-2 spike protein from homologue proteins of *Betacoronavirus*, as well as to the spikes from human clinical isolates of the virus. Specifically, Tyr and Asn have higher occurrence rates on the Receptor Binding Domain (RBD) and in the overall sequence of spike proteins of *Betacoronavirus*, whereas His and Arg have lower occurrence rates. The *in vivo* evolutionary phenomenon of SARS-CoV-2 shows that Glu, Lys, Phe, and Val have the highest probability of occurrence in the emergent viral particles. Amino acids that have higher occurrence than the expected by the neutral control, are favorable and are fixed in the sequence while the ones that have lower occurrence than expected, influence the stability and/or functionality of the protein. Our results show that most unique mutations either for SARS-CoV-2 or its variants of health concern are under selective pressures, which could be related either to the evasion of the immune system, increasing the virus' fitness or altering protein – protein interactions with host proteins. We explored the consequences of those selected mutations in the structure and protein – protein interaction with the receptor. Altogether all these forces have shaped the spike protein and the continually evolving variants.

The ongoing COVID-19 pandemic caused by the rapid global transmission of SARS-CoV-2¹ illustrates the planetary consequences of recurrent episodes of zoonotic transmission from animals to human populations. At least seven coronaviruses have been identified to infect humans causing principally respiratory difficulties but only three of them pose potential pandemic threats^{2–4}. Among the 4 genera of coronavirus, only some virus of the *Alphacoronavirus* and the *Betacoronavirus* can infect humans⁵. These two genera have a common ancestor that infects bats while the *Gammacoronavirus* and *Deltacoronavirus* have a bird coronavirus origin⁶. This means that human coronavirus may be directly related to bat coronavirus or to other mammals as intermediate hosts. Phylogenetic analyses have revealed that SARS-CoV-2 is a *Betacoronavirus* related to the bat *Rhinolophus affinis* coronavirus Bat-SL-RaTG13 and the Malayan pangolin (*Manis javanica*) coronavirus, and that SARS-CoV-2 and SARS-CoV belong to the same B lineage, whereas MERS-CoV belongs to the C lineage^{1,6}.

Structural and genomic analysis of viral components are key for understanding the evolution of the virus and developing vaccines and therapeutic strategies both to combat the pandemic and to prevent further spread. As all the coronavirus, SARS-CoV-2 recognizes and fuses into the host cells membranes through the spike glycoprotein⁷. The SARS-CoV-2 spike glycoprotein (SARS2-S) attaches to the human Angiotensin Converting Enzyme 2 (ACE2) expressed on the cell membrane and is further processed by host's proteases⁸ which are necessary for fusion. SARS2-S is made up of the Subunit 1 (S1) that contains the Receptor Binding Domain (RBD) and Subunit 2 (S2), responsible for fusion with the cell membrane^{9,10}. Two conformations are described for the SARS2-S where the RBD is in a down position, or in an up position, this last one interacts with the receptor¹¹. In the viral membrane, the spike protein interacts with other two spike proteins forming an homotrimer^{10,11}. Given

¹Department of Immunology, Instituto de Investigaciones Biomédicas, Universidad Nacional Autónoma de México, 04510 CDMX, México. ²Theoretical Biology Group, Instituto de Investigaciones Biomédicas, Universidad Nacional Autónoma de México, 04510 CDMX, México. ³Universidad Latinoamericana, Nutrición, Campus Cuernavaca, Morelos, México. [✉]email: ortoto@unam.mx; marcojose@biomedicas.unam.mx

the essential role of this protein in the viral life cycle, it is assumed that it has undergone strong evolutionary pressures to ensure the propagation of the virus. The human membrane protease ACE2 has been identified as the viral receptor for several coronavirus, including other bat SARS-like coronavirus, the SARS-CoV and SARS-CoV-2, as well as the *Alphacoronavirus* hCoV-NL63³². Different analyses revealed that the RBD of spike proteins of SARS-CoV-2, SARS-CoV and MERS-CoV allow binding to the receptor from various species while staying within a range of possible mutations^{33,31}, although neither the binding affinities nor the effect of such mutations on the affinity have been quantified. What has been certainly demonstrated through structural analysis and surface plasmon resonance, is that the binding affinity of SARS2-S protein to its receptor ACE2 is greater than the one of SARS-CoV to the same receptor^{13,13}. The most probable amino acids responsible for the increase in affinity that could have resulted in enhancing the spread of the virus SARS-CoV-2 have already been proposed^{9,14–18}.

Herein, we analyze the reference sequence of the spike protein of the SARS-CoV-2 and compare it to the sequences of spike proteins from other coronavirus, to best fit an evolution model that explains the amino acids preferences that have been selected for a higher affinity binding to the host's receptor. Furthermore, we demonstrated the differences of the selective pressures exerted over the ongoing evolving virus. To understand the evolution of the spike protein, we applied an amino acid substitution test (neutrality test) to identify the amino acids that deviate from neutral selection. This test is directly related to the degeneracy of the Standard Genetic Code (SGC)¹⁷ and was applied to both the whole sequence of the spike protein and to the sequence of the RBD as well as to the spikes from human clinical isolates. Then we identified unique mutations in SARS2-S that increased the protein's binding to its receptor, in terms of their physicochemical properties. Finally, we discuss how the mutations of variants could affect both the protein structure and the success of the virus.

Materials and methods

Data sources. The nucleotide and amino acid sequences of the spike proteins of 102 *Betacoronavirus* reported in the GenBank (<https://www.ncbi.nlm.nih.gov/>) were obtained. In addition, we downloaded 561 whole genome sequences of SARS-CoV-2 from human isolates reported in the GenBank. From the whole genomes, 200 spike sequences were obtained. The structures of the RBDs of SARS2-S in complex with ACE2 (6M0J), SARS-S RBD bound to ACE2 (2AJF), were downloaded from the Protein Data Bank (<https://www.rcsb.org/>). Also, the reference structure of SARS2-S was downloaded from the SARS-CoV-2-dedicated ZhangLab webpage (<https://zhanglab.ccmb.med.umich.edu/COVID-19/>). SARS-CoV-2 variants' spike sequences were retrieved from Situation Reports deposited in the site outbreak.info.

Neutral evolution model. The SGC has been mathematically modeled into a 6-dimensional (6D) hypercube using group theory¹⁸. The 6D representation of the SGC means that any codon has 6 other codons to which it is adjacent. The 6D-model has been further transformed into its amino acid phenotypic graph representation¹⁹. In this graph of the 20 canonical amino acids, the vertices represent the amino acids; and two amino acids (aa1 and aa2) are joined by an edge; at least one codon that encodes aa1 is at one edge distance to at least one codon that encodes aa2 in the 6D-model of the SGC. For a given amino acid, the set of adjacencies of its codons represent the possible changes the amino acid can undergo with a single nucleotide mutation. The normalization by rows of this adjacency matrix leads to a probability transition matrix of a stochastic process whose states are the amino acids. Adding the Markov property to the stochastic process results in a discrete time stochastic process with no memory. The limiting distribution (or stationary probability distribution), of the average transition matrix is a neutral control of the changes present in a protein history at amino acid level. The substitution matrices for each set of protein sequences are a heuristic evolutionary model for each protein. The stationary distributions for each protein reflect the probability of occurrence of each amino acid in the proteins if they continue evolving with their current evolutionary model. Thus, our amino acid neutral test considers all positions as equally likely to change, with the sole constraint that they obey the degeneracy of the SGC. If a component from the stationary distribution, obtained from protein sequences, has a greater value than its corresponding component at the neutral control, it will be interpreted as positive selection, and lower values will be considered as negative selection. The values that lie at or close to the neutral control will be interpreted as consistent with a neutral mutation according to the random drift hypothesis of molecular evolution. This test measures unambiguously the levels of positive or negatively selected amino acids as well as those that are neutral or close to neutrality, without the need of a phylogenetic tree²⁰. Our results cannot be directly visualized or measured from a phylogenetic tree.

Data. One hundred two spike sequences of *Betacoronavirus* were pairwise aligned. Each pair of protein sequences was aligned using MUSCLE²¹ with default parameters. The protein alignment was used as template to derive a nucleotide alignment that would not have gaps that could split codons. From the nucleotide alignment a table of mutations was computed that account for the total of changes in codons. The table of codon mutations was transformed into an amino acid mutation matrix by adding up the values of the codons for a given amino acid. Hence, this matrix considers synonymous and non-synonymous mutations. The amino acid mutation matrix was computed for every pair of sequences and added up. Then, the matrix was normalized by rows, so that each row adds up to 1, and yields a probability transition matrix. The stationary distribution of the probability transition matrix was derived and compared to the control of neutral evolution as described elsewhere¹⁷. To assess the statistical robustness of the sample of sequences, a jackknife procedure was applied. The procedure of deriving the stationary distribution from the probability transition matrix of a sample of sequences was repeated to all possible subsets of 50 sequences. A confidence interval of 95% was computed around the stationary distribution derived from the set of sequences. The whole procedure was also applied to the RBD sequences of the same spikes of *Betacoronavirus* and to 200 sequences of the spike protein of SARS-CoV-2.

Phylogenetic analysis. All the evolutionary analyses were conducted with MEGA X software^{21,22}. The multiple alignments of the spike sequences were performed with MUSCLE algorithm. The reference sequence of SARS2-S was compared to both groups i) the most proximal amino acid (a.a) sequences and ii) the ACE2 binding CoVs. Consequently, unique mutations for SARS2-S and conserved residues were identified manually. Potential glycosylation sites were identified in both the linear and structural models.

Structural analysis. The structures were cleaned to have the most accurate protein and complexes of the RBDs to its receptor. The structural analysis was visualized and analyzed with Chimera^{23,24}, and I-TASSER^{25–29}. The complex of SARS2-S RBD with the receptor was used to point out unique and conserved residues. Distances between the amino acids involved in the protein–protein interaction were computed. Other parameters like the hydrophobicity and electrostatic potential were calculated for the a.a. in the interface. The same was calculated for the complex SARS2-S RBD with the receptor. The number of contacts, number of hydrogen bonds, hydrophobicity and mean distances were compared. Besides, spike mutations of the most prominent variants of SARS-CoV-2 were identified in the three-dimensional model of SARS2-S and the structural models were predicted for each of them with I-TASSER platform. The reference structure was downloaded from the SARS-CoV-2-dedicated ZhangLab webpage (<https://zhanglab.ccmb.med.umich.edu/COVID-19/>).

Results

The neutral theory of molecular evolution assumes that evolution is driven by random point mutations that eventually may be fixed by genetic drift or selection. From this point of view, we applied an amino acid neutrality test¹⁷ to better understand the type of selective pressure, if any, acting upon a.a. present in the spike protein. This analysis revealed that in the *Betacoronavirus* genus, Trp, Cys, His, Gly, Pro, Ser, and Arg underwent negative selective pressures, as the number of changes in these a.a. are lower than the expected by neutral evolution. In contrast, Tyr, Lys, Gln, Phe, Asn, Asp, Thr, and Val, displayed positive selection and Met, Glu, Ala, and Leu are amino acids fixed by genetic drift or neutral forces (Fig. 1A). To accurately analyze the adaptation of SARS-CoV-2 to its receptor, we applied the neutrality test to the RBD of the spike sequences. Importantly, Trp, Met, Glu, His, Pro, Ala, Leu, and Arg showed negative selection, whereas Tyr, Lys, Phe, Asn, Asp, and Thr manifested positive selection in the domain (Fig. 1A). This means that it is less likely to find an Arg that appeared by random mutation than a Tyr or an Asn, because the hexa-codon Arg is under high negative selective pressure.

To better understand the ongoing evolution of SARS-CoV-2, we performed another neutrality test to the spike proteins of SARS-CoV-2 from human isolates. Figure 1B shows that Tyr, Glu, Lys, Phe, Asn, Asp, Ile, and Val are under positive selective pressure whereas Trp, Met, Cys, Gln, Gly, Pro, Ala, Leu, Ser, and Arg are under negative selective pressures. The differences between graphs of Fig. 1A, B highlights the sharpness of the model to detect the selective pressures in a protein belonging to a lower taxon. Moreover, we performed the neutrality test to the whole genome of 561 sequences of SARS-CoV-2 from human isolates (Figure S1). This analysis permits to portray how the whole genome of SARS-CoV-2 is subjected to selective pressures, which are not necessarily the same along the whole viral genome. Different viral proteins are influenced by different variables that modify the probability of occurrence of their own amino acids.

We carried out a multiple alignment of the a.a. sequences of spike proteins of coronavirus including *Alphacoronavirus* and *Betacoronavirus*. Consequently, we identified the most related spike sequences to SARS-CoV-2 (not shown). We focused in the RBD sequence of the S1 subdomain to identify mutations that could be advantageous for SARS2-S binding to ACE2. Therefore, another multiple alignment was performed including the spikes of SARS-CoV-2, RaTG13, PnCoV, and several spikes known to bind to human ACE2⁸. We discarded the *Alphacoronavirus* HCoV NL63 because even though it binds to the human ACE2, the orientation of the RBD in the protein–protein interaction is different. We noticed 49 unique mutations in SARS2-S compared to the others ACE2 binding sequences (Table S1). Some of these mutations are present in CoV RaTG13 and PnCoV spikes, what would suggest that these spike proteins could bind to the human ACE2 as well. Those mutations were localized in the structure of the complex with ACE2 (PDB: 6M0J). Interestingly, most mutations are located at the interface with ACE2 (Figure S2). Figure 2A shows the interaction between SARS2-S RBD and ACE2 (modified from PDB: 6M0J), all the unique amino acids are colored in red. Figure 2B shows a close-up where the side chains of the amino acids involved in the protein–protein interaction are shown with sticks. We compared the conservation of the residues among all the ACE2 binding sequences. The conserved residues are shown in pale pink as the rest of the structure (i.e., Tyr 449, Tyr 453, Asn 487, Tyr 489, Thr 500, Gly 502, Tyr 505) (Table 1) while the unique a.a. for SARS2-S, are shown in red (Fig. 2B). Among the 17 a.a. involved in the interaction, 10 are unique for SARS2-S, including Lys 417, Gly 446, Leu 455, Phe 456, Ala 475, Phe 486, Gln 493, Gly 495, Gln 498 and Asn 501 (Table 2). We analyzed unique a.a. that raised affinity to the receptor, therefore were selected during the evolutionary process. The mutations Gly 446 and Lys 417 enable the formation of new hydrogen bonds with the receptor, SARS2-S forms 8 hydrogen bonds with the receptor while SARS2-S forms 11 (Supplementary Material). Specifically, Phe 456 and Phe 486 that are under positive selection were selected because they generate more hydrophobic contacts with the receptor than the Leu's at the same positions present in SARS2-S. At the other edge of the interface, Gln 498 which was selected neutrally in the RBD enables a hydrophilic surface that permits interactions with the receptor by van der Waals forces (Fig. 3).

Then we concentrated in the mutations of spikes proteins of SARS-CoV-2 variants to unravel the evolutionary behavior of the virus. The variants that have received most attention due to their importance for public health are those identified as variants of concern and variants of interest. To know if the mutations have modified the structure of spike of some variants, we downloaded the reference structure of SARS2-S from I-TASSER and modified to point all mutations in the variants in cyan and deletions in grey (Fig. 4A); the RBD is magnified in the inset (black) (Fig. 4B). Also, the Cys of the RBD are shadowed in yellow and the two glycosylated Asn are in

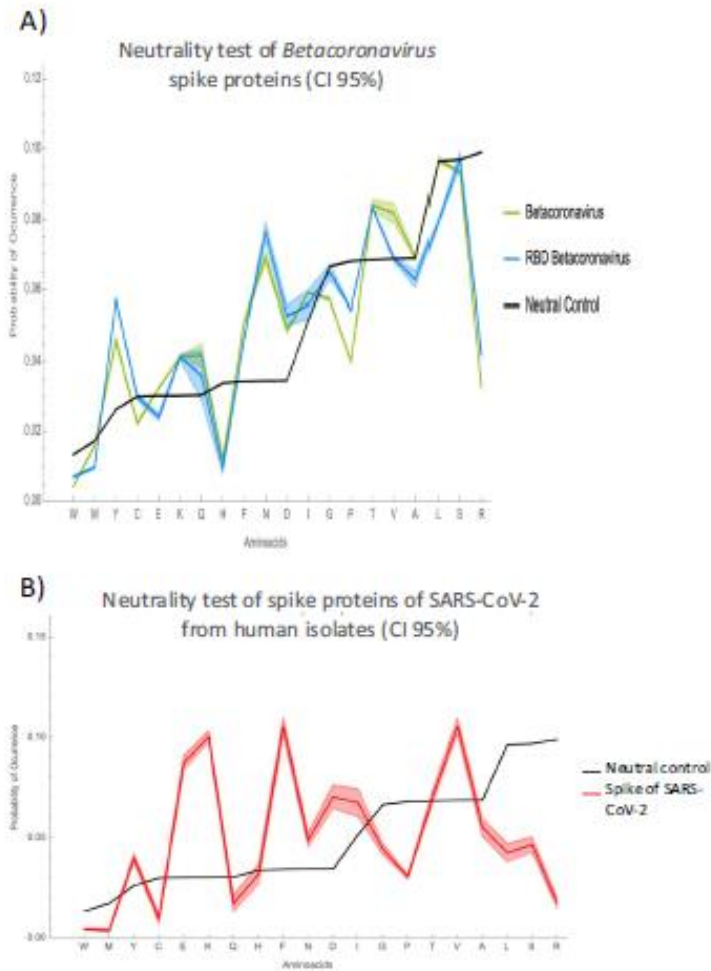


Figure 1. Neutrality evolution test of the a.a. of the spike protein and the RBD. The probability of occurrence of each a.a. in a hypothetical protein that has nonselective pressures over its mutations (black line), compared to (A) the probability of occurrence for a.a. of the spike protein of *Betacoronavirus* (green) and of the a.a. of the RBD (blue). (B) the neutrality test to the spike protein of SARS-CoV-2 from human isolates (red). A.a. with higher occurrence than that predicted by purely stochastic changes are under positive selection pressure, while frequencies lower than the neutral prediction are amino acids that underwent negative selection pressures. A Jackknife procedure was performed to obtain confidence interval of 95%. All calculations were made with Wolfram Mathematica 12.3.

magenta (Fig. 4B). Figure S2 shows the predicted structures of the four variants of health concern: Alpha, Beta, Gamma and Delta as well as three of interest: Epsilon, Iota, and Kappa; they are all merged with the reference structure. To note, the predicted structure of Gamma and Epsilon variants protrude the RBD at a different position than that in the reference structure and the rest of the variants. From the reference structure of SARS2-S downloaded, Alpha deviates only 2.52 Å, Beta 1.97 Å, Gamma 4.14 Å, Delta 3.34 Å, Epsilon 4.2 Å, Iota 3.21 Å and Kappa 3.23 Å (Root Mean Square Deviation). Then, we identified the selective pressures over each mutation in the spike proteins of the variants mentioned above, and four more variants of interest (Zeta, Eta, Theta and Lambda). Interestingly, all variants share the mutation D614G, three of the variants of concern and one of interest presents N501Y substitution (Table 3). Also, the Beta, Gamma, Zeta, Eta and Theta share E484K, and the Delta, Epsilon and Kappa variants share L452R. Tyr and Lys are positively selected a.a. both for the complete spike

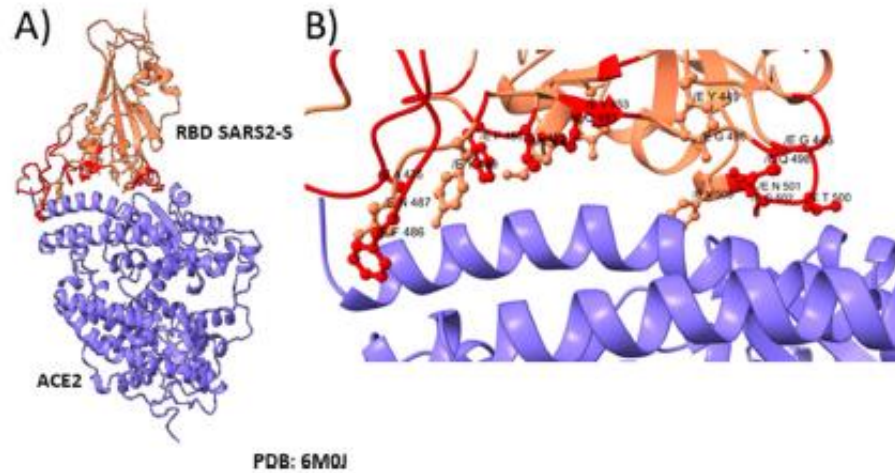


Figure 2. Interaction between the RBD of the spike protein of SARS-CoV-2 with ACE2. (A) Interaction between the RBD of SARS2-S (pale pink) and the human receptor ACE2 (blue). (B) A close-up of the interface shows the R side chains of the a.a. of the RBD involved in the binding with the human receptor. Unique a.a. for SARS-CoV-2 are colored in red.

| Conserved a.a. | | Selection according to neutrality |
|----------------|---|-----------------------------------|
| 449 | Y | Positive |
| 453 | Y | Positive |
| 487 | N | Positive |
| 489 | Y | Positive |
| 500 | T | Positive |
| 502 | G | Neutral |
| 505 | Y | Positive |

Table 1. Conserved residues involved in protein-protein interaction with ACE2 among the ACE2 binding spikes. The selection type according to the neutrality test are indicated. a.a. code Polar

proteins as well as for the RBDs of *Betacoronavirus* and SARS-CoV-2 human isolates. In contrast, Gly and Arg were clearly negatively selected at position 614 and 452 respectively (Fig. 1). Gamma, Zeta, and Theta present a Phe in position 1176 and the Pro 681 is often substituted either by His or Arg.

Finally, as shown recently, SARS2-S is highly glycosylated which forms a protective shield against recognition of the immune system. Hence, we considered the conservation of the glycosylation sites important for evolution of the virus. The glycosylation sites reported for SARS2-S²⁹ were identified and compared to the other ACE2 binding sequences (Table 4). Four Asn (17, 74, 149 and 657) are unique to SARS2-S, while the rest (18) are shared and from the three a.a. that forms O-glycosylations, only one is conserved. To note, PgCoV and bat CoV RaTG13 express all potential glycosylation sites identified in the sequence of the causal agent of COVID-19, meaning that those a.a. were selected even before SARS-CoV-2 infected humans. Nevertheless, they all contribute to avoid the recognition of the immune system without affecting the binding to the receptor ACE2. We also identified mutations close to glycosidic sites in the SARS-CoV-2 variants of concern. All of them possess mutations near to these sites.

| SARS2-S | | SARS-S | | Selection according to neutrality |
|---------|---|--------|---|-----------------------------------|
| 417 | K | | | Positive |
| 446 | G | 433 | T | Neutral |
| 455 | L | 442 | Y | Negative |
| 456 | F | 443 | L | Positive |
| 475 | A | | | Negative |
| 486 | F | 472 | L | Positive |
| 493 | Q | 479 | N | Neutral |
| 496 | G | | | Neutral |
| 498 | Q | 484 | Y | Neutral |
| 501 | N | 487 | T | Positive |

Table 2. Unique residues for SARS2-S involved in protein–protein interaction with ACE2. The selection type according to the neutrality test are mentioned specifically for SARS2-S. Other ACE2 binding spike proteins expressed different a.a. Here SARS-S is shown as an example. The empty spaces in SARS-S are a.a. that do not

| interact with ACE2. a.a. code | Non-polar | Polar | Polar positive |
|-------------------------------|-----------|-------|----------------|
| | | | |

Discussion

The neutral theory of evolution states that most mutations in nucleic acid sequences are random, and these can be fixed by genetic drift and/or natural selection. We exploited an amino acid neutrality test to interrogate the molecular evolution of the spike protein and to consider the possible implications for binding to its receptor that is why we focused on the RBD sequences. Positive selective pressures cause a.a. to be fixed in higher frequencies than by neutral evolution or genetic drift, while negative selective pressures cause a.a. to appear in lower frequencies than the neutral control. Fixed mutations under negative selective pressures may remain because they are advantageous for the virus by contributing to the protein stability or increasing the affinity for its receptor. These selective pressures are similar among homologous proteins and importantly, they are the major driving forces for adaptation.

The neutrality test applied to the RBD sequences, shows that there are amino acids that have similar selective pressures as the whole protein, with few notable exceptions. For example, Tyr is subject to a higher positive selection in the RBD than in the whole protein; actually, four Tyr are involved in protein – protein interaction with the receptor. In fact, most common residues between the ACE2 binding proteins have positive selection. However, unique mutations in SARS2-S RBD were fixed by selective pressures rather than neutral forces. Hence, the adaptability of this domain is explained by quantitative means. Specifically, the mutations that contributed to a higher affinity to the receptor, compared to SARS-S, were also identified. Phe 456 and Phe 486 generate more hydrophobic interactions with atoms from ACE2, Gln 498 enhances interaction with the receptor by van der Waals forces, and Lys 417 enabled the formation of new hydrogen bonds with the receptor. These mutations certainly were selected for increasing the affinity with the receptor.

Besides spike proteins need to be processed by different host proteases to disassemble S1 from S2 and to expose the fusion peptide, consequently the fusion machinery could induce the membranes fusion and thus to achieve the cell infection. Other coronavirus are cleaved by trypsin, furin-like convertases, cathepsins, serin proteases, elastases, plasmin, among others, in order to achieve the fusion peptide exposure³⁰. Thus, host proteases exert an important selective pressure to the spikes proteins and are responsible for tropisms toward cells who express them. The fact that the polybasic motif (682-RRAR) in SARS2-S resulted from an insertion of 12 nucleotides (681-PRRA), has generated the COVID lab-leak hypothesis, a claim that a coronavirus was manually manipulated in a laboratory and sparked doubts about the origin of SARS-CoV-2. However, others have discussed that this motif has appeared multiple times and independently in spike proteins of others coronavirus³¹. It is not clear whether there was an epistasis event that enables the 4 amino acid insertion but certainly the motif was selected because it increases the pathogenicity of the virus³². The cleavage not only disrupts the binding between S1 and S2 and approaches the fusion machinery, but also generates a neuropilin-1 binding motif^{33,34} which contributes to the internalization of the virus³⁵. Within the context of the neutral theory of evolution, the probability of occurrence of each amino acid is related to their codonicity; in other words, since Arg is hexa-codonic, and Pro and Ala are tetra-codonic, the probability of appearance of these specific amino acids turns to be high. According to our results, the probability of occurrence of Pro and Arg is lower than the expected in the whole protein. Nevertheless, once presented, they were selected because they were part of a motif that gave a specific advantage to the virus. So, the expression of a polybasic motif at S1/S2¹ which is susceptible to a specific protease

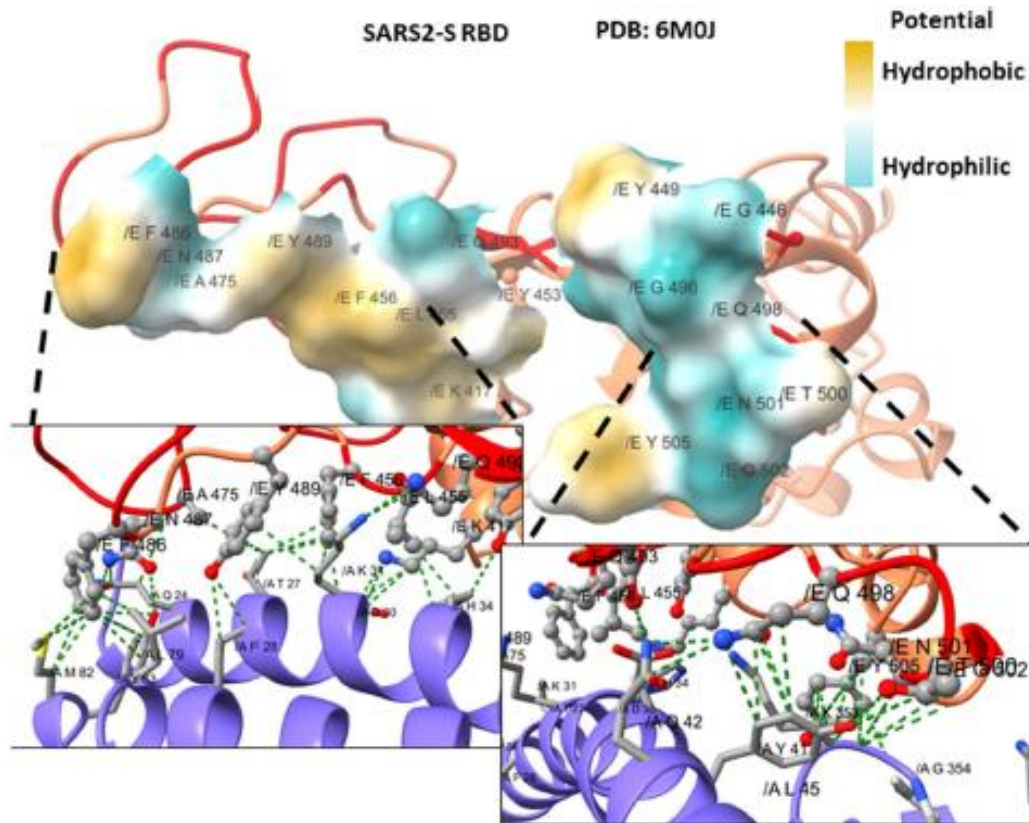


Figure 3. Unique residues of SARS2-S contact more the receptor ACE2. The surface of the amino acids involved in protein–protein interaction with the receptor is colored by its hydrophobic potential. Specific pockets are zoom in to show the responsible residues. Compared to other ACE2 binding spikes, SARS2-S has Phe 456 and Phe 486 in the interface which enables more hydrophobic area (left down) and Gln 498 enables a hydrophilic area, all three establishes more contacts with the receptor compared to other ACE2 binding proteins.

should not be regarded as a surprise. Other coronavirus that present similar polybasic motifs in the S1'/S2' site, where shown in our multiple alignment of spikes of *Betacoronavirus*: the spikes of the Murine Hepatitis Virus (YP009824982.1), Murine CoV RA59/R13 (ACN89689.1), Murine CoV RA59/SJHM (ACN89705.1), Rat CoV (YP003029848.1), HCoV HKU1 (YP173238.1), HCoV OC43 (YP009555241.1), Rabbit CoV (YP005454245.1), Canine CoV (AQT26498.1), Human enteric CoV (ACJ35486.1), Bovine CoV (NP150077.1) and the Sambar deer CoV US/OH-WD388TC/1994 (ACJ67012.1).

Herein, we present a model that analyzes the preference of amino acids by means of the selective pressures that enhances transmissibility or increase the stability of the interaction with the receptor. The substitutions at position 681 may cause slight differences in the secondary structure immediately before the furin cleavage site since Pro introduces slight bends to the protein structure. We certainly calculated the predicted structure of the protein which is very similar to the known structure. Only the Gamma and Epsilon variants had a predicted structure slightly different from the reference. Further predictions should be tested to know if there are greater differences in the trimer and if the interaction with the receptor is enhanced in any spike variant. Regarding this site, it is predicted by computational analysis that the mutation P681R increased the basic properties of the motif, which in turn increased the affinity to the furin enzymatic site³⁶. In fact, the mutation does augment the fusogenic properties and enhances syncytia formation. It seems that this substitution permits a Low Complexity Region that contributes to immune system evasion by increasing antigenic variability³⁷.

The neutrality test performed to the spike protein of SARS-CoV-2 from human isolates shows the current evolutionary behavior of the virus. We could apply the test mainly because the fast spread of the virus increased its population size that allowed the appearance of mutations. Here we demonstrated in a quantitative manner the actual selective pressures exerted over the virus. Figure 1B shows that some of the positive selected a.a. for

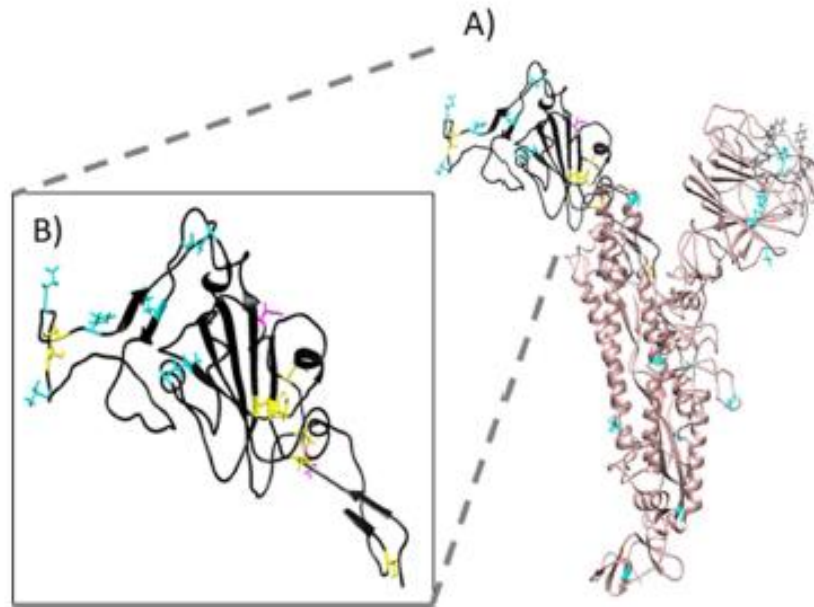


Figure 4. Structure of the spike protein of SARS-CoV-2 and the variants of concern. (A) Structure of the spike protein of SARS-CoV-2 in rosy-brown color. (B) Zoom of the RBD structure of SARS2-S colored in black. Cys of the RBD are shadowed in yellow and the two glycosylated Asn are magenta. All sites of point mutations in the variants (from Alpha to Lambda) are shadowed in cyan and deletions in grey. Molecular graphics and analyses performed with UCSF Chimera, developed by the Resource for Biocomputing, Visualization, and Informatics at the University of California, San Francisco, with support from NIH P41-GM103311.

spike proteins in *Betacoronavirus* are still under positive selection in the spike proteins of SARS-CoV-2 (i.e. Tyr, Lys, Phe, Asn, Asp), although the probability could be altered.

By the neutrality test we observed that the common mutations present in different variants of concern were fixed by common selective pressures. For example, the Beta and Gamma variants substitute a Lys at position 417. It remains to confirm if these spikes maintain the hydrogen bond formed in the ancestral sequence as Asp has no electron donor. The mutation E484K, present in the same variants and others three, occurs towards an a.a. with almost the same polar requirement^{45,46} and it is strongly positive selected. This shared mutation leads to neutralizing antibodies resistance³⁸. Other shared mutation is D614G substitution which is present in all variants analyzed. The mutation translates to a small a.a. without charge that is negatively selected. It has been proven that this substitution enhances the binding to ACE2 in comparison with the ancestral virus, therefore infectivity and transmission of the variant also has raised^{47–50}. In fact, patients carrying virus with this mutation, have a significantly higher viral load³⁹. D614G also alters the conformation of the RBDs towards the up position, which is necessary for receptor recognition⁴¹. Besides, N501Y substitution another shared mutation increases the binding to the receptor^{41,42}. Actually this substitution enables infection to mice cells through interaction with mouse ACE2⁵¹.

Shared mutations of the variants of concern are reported to increase transmissibility and/or favor a conformational state change⁴⁶. The strong selective pressure exerted on these substitutions has already driven them to fixation in SARS2-S. Shared mutations among different health concern variants arose by convergent evolution coming from selective pressures regarding the viral fitness by increasing the pathogenicity, protein-protein interaction, and the evasion of the immune system⁴³. However, another selective pressure comes from other viral particles if the viral particles are constantly competing between them. By now Delta variant has displaced rapidly the Alpha variant, which was the predominant variant in UK and USA⁴⁴.

The accumulation of mutations is linked to the capability to correct errors. Coronavirus have RNA-dependent RNA polymerases which are prone to mistake, unlike other RNA virus, they have also a 3' to 5' exoribonuclease (nsp14-ExoN) that proofreads the new sequence⁴¹. Nsp14-ExoN is one of the major factors enabling long and stable RNA genomes, nevertheless some mutations escape from the proofreading. Mutations in the S2 subdomain would probably be less frequent to reach fixation because the sequence is crucial for the function. This subdomain is involved in fusing with the membrane, so the domains require specific physicochemical characteristics. Therefore, purifying selection deletes most neutral mutations, and consequently, there is a high degree of conservation. On the contrary, S1 is important for receptor recognition, specifically the Carboxyl Terminal Domain (CTD) where the RBD is contained, so the sequence of this domain is crucial for adapting to the actual

| Position | SARS2-S | Variants of concern | | | | Variants of interest | | | | | | |
|----------|---------|---------------------|----------|-----------|-----------|----------------------|----------|---------|-----------|----------|-----------|------------|
| | | Alpha (α) | Beta (β) | Gamma (γ) | Delta (δ) | Epsilon (ε) | Zeta (ζ) | Eta (η) | Theta (θ) | Iota (ι) | Kappa (κ) | Lambda (λ) |
| 13 | S | | | | | I + | | | | | | |
| 18 | L | | | F + | R - | | | | | | | |
| 19 | T | | | | | | | | | | | |
| 20 | T | | | N + | | | | | | | | |
| 26 | P | | | S - | | | | | | | | |
| 52 | Q | | | | | | R - | | | | | |
| 67 | A | | | | | | V + | | | | | |
| 69 | H | * | | | | | * | | | | | |
| 70 | V | * | | | | | * | | | | | |
| 75 | G | | | | | | | | | | | |
| 76 | T | | | | | | | | | | | V + |
| 80 | D | | A - | | | | | | | | | I + |
| 95 | T | | | | | | | | I + | | | |
| 138 | D | | | Y + | | | | | | | | |
| 144 | Y | * | | | | | * | | | | | |
| 145 | Y | * | | | | | * | | | | | |
| 152 | W | | | | | C - | | | | | | |
| 157 | F | | | | * | | | | | | | |
| 158 | R | | | | * | | | | | | | |
| 190 | R | | | S - | * | | | | | | | |
| 215 | D | | G - | | | | | | | | | |
| 241 | L | | * | | | | | | | | | |
| 242 | L | | * | | | | | | | | | |
| 243 | A | | * | | | | | | | | | |
| 247 | S | | | | | | | | | | | * |
| 248 | Y | | | | | | | | | | | * |

Table 3. (continued)

| | | | | | | | | | | | |
|------|---|-----|-----|-----|-----|-----|-----|-----|-----|-----|-------|
| 249 | L | | | | | | | | | | * |
| 250 | T | | | | | | | | | | * |
| 251 | P | | | | | | | | | | * |
| 252 | G | | | | | | | | | | * |
| 253 | D | | | | | | | | | | * |
| 417 | K | | N + | T / | | | | | | R - | Q - |
| 432 | L | | | | | | | | | K + | |
| 478 | T | | | | | | | | | K + | |
| 484 | E | | K + | K + | | | | | | K + | Q - |
| 490 | F | | | | | | | | | K + | S - |
| 501 | N | Y + | Y + | Y + | | | | | | Y + | |
| 570 | A | D + | | | | | | | | | |
| 614 | D | G - | G - | G - | G - | G - | G - | G - | G - | G - | G - |
| 655 | H | | | Y + | | | | | | | |
| 677 | Q | | | | | | | H / | | | |
| 681 | P | H / | | | | | | R - | | | R --- |
| 701 | A | | V + | | | | | | | H / | |
| 716 | T | I + | | | | | | | | | |
| 859 | T | | | | | | | | | | N + |
| 888 | F | | | | | | | L - | | | |
| 950 | D | | | | | N + | | | | | |
| 982 | S | A - | | | | | | | | | |
| 1027 | T | | | I + | | | | | | | |
| 1071 | Q | | | | | | | | | | H -- |
| 1092 | S | | | | | | | | K + | | |
| 1101 | H | | | | | | | | Y + | | |
| 1118 | D | H / | | | | | | | | | |
| 1176 | V | | | F + | | | | | F + | | |

Table 3. Type of selective pressure for mutations in health concern variants. The mutations colored by physicochemical properties are enlisted for each variant and the type of selective pressure applied for each a.a. is shown. The type of pressure is specified either for the whole protein or for amino acids positioned in the RBD. "Position" column is according to the RefSeq. The substituting amino acid on each variant is indicated in one-letter code next and the selective pressure exerted over the a.a. according to the neutrality test. Positive (+), negative (-), neutral (/). Asterisks indicate deletion.

a.a. code Non-polar Polar Polar positive Polar negative

| SARS2-S | Glycosylation | Conserved | Variants with mutations near by | SARS2-S | Glycosylation | Conserved | Variants with mutations near by |
|---------|---------------|-----------|---------------------------------|---------|---------------|-----------|---------------------------------|
| 17 | N-gly | No | γ, δ | 616 | N-gly | Yes | α, β, γ, δ, ε, ζ, η, θ, ι, κ, λ |
| 61 | N-gly | Yes | α, η | 657 | N-gly | No | γ |
| 74 | N-gly | No | α, η, λ | 676 | O-gly | Yes | η |
| 122 | N-gly | Yes | | 709 | N-gly | Yes | β |
| 149 | N-gly | No | α, η | 717 | N-gly | Yes | α |
| 165 | N-gly | Yes | | 801 | N-gly | Yes | |
| 234 | N-gly | Yes | | 1074 | N-gly | Yes | κ |
| 282 | N-gly | Yes | | 1098 | N-gly | Yes | |
| 323 | O-gly | No | | 1119 | N-gly | Yes | α |
| 325 | O-gly | No | | 1134 | N-gly | Yes | |
| 331 | N-gly | Yes | | 1173 | N-gly | Yes | γ, ζ, θ |
| 343 | N-gly | Yes | | 1194 | N-gly | Yes | |
| 603 | N-gly | Yes | | | | | |

Table 4. Conservation of the glycosylation sites of SARS2-S. Most of the glycosylation sites reported from SARS2-S are shared with all the ACE2 binding sequences except for 6 residues 4 Arg, 1 Thr and 1 Ser. All variants of concern present mutations near at least one unique glycosylated residue.

host or for the ability to infect other species. Interestingly, some variants have more mutations in S2 whereas others present a few. The exact biological significance of this observation is not clear. It is probable that the variants with few mutations could either i) have been mutating before and by the time they were reported they may have had enough time to fix or disappear neutral mutations at S2, or ii) they may have appeared recently and have mutated only in S1. Now we cannot sustain any hypothesis because we discussed only eleven variants and we lack the accurate mutation rate of each variant and the transmission rate at different places.

Evidence shows the influence of glycosylation in the success of SARS-CoV-2^{28,42}, so we included this characteristic in the study. In other coronavirus the highly glycosylated Amino Terminal Domain (NTD), plays an important role in attachment to the host cell and immune response evasion. For influenza C virus and some coronavirus (HCoV HKU1, and HCoV OC43), attachment through 9-O-acetylated sialic acid receptors is crucial and constitute another species barrier^{52–55,56}. The interaction with specific hosts' proteins facilitates the approaching of the fusion machinery to the cell membrane. Furthermore, saccharides mask potential epitopes recognized by antibodies, making it difficult to the immune system to eliminate the virus. It has been observed that the spike glycoprotein must be shielded by the protective glycans from the immune system^{28,50}. For the spike protein, Asn mutations are subjected to positive selection both in the RBD and in the whole protein of *Betacoronavirus*. This favors the fixation of new Asn susceptible to glycosylation, only if the sequon NXT/S is expressed too (N stands for Asn, X is any a.a. except Pro, and the third position stands for Thr or Ser). Therefore, the four new glycosylation sites were enhanced by positive selection. It remains to be determined whether the mutations near glycosylation sites interfere with the formation of glycosidic bonds in the variants of health concern. We found that all variants present mutations near glycosylation sites, which directly alter the epitopes recognized by antibodies they have been reported reduce recognition by antibodies²⁹, by creating new epitopes. Hence, evasion to the immune system recognition is an important selective pressure that has influenced the fixation of neutral mutations. It is important to undercover to what extent the glycosylation of SARS-CoV-2 spike protein could help us understand the physiopathology of the virus and to develop better prophylactic or therapeutic strategies, effective against all variants.

The neutrality test constructed here considers both synonymous and non-synonymous mutations, which allowed us to obtain information of the types of selective pressures that influenced deviation in fixation from neutral mutations. The limitation of our study is that we cannot detect mutations fixed by epistasis or originated by recombination or concomitant mutations. Therefore, the common mutations are considered only as the result of natural selection or genetic drift. Another consideration is that the neutrality test showed that the amino acids with the capability to be glycosylated are favored but we cannot certainly conclude that specific residues are bonded to a carbohydrate.

Conclusion

The long-lasting pandemic, the wide geographic distribution, and the rapid contagiousness mainly during epidemic waves, have influenced the generation of variants of SARS-CoV-2. At global or local scale, the evolution of this virus can be appreciated. Vaccines and drugs have been developed and tested aiming to stop transmission which would also result in preventing the virus from mutating and developing new variants. Therefore, evolutionary studies play an important role in the prevention of epidemiological catastrophes and in the development of better treatments that covers most viral variants. The neutrality test computed shows the type of selective pressure for each amino acid in the spike protein of *Betacoronavirus*. This evolutionary study enables to understand and describe changes in SARS2-S sequence that affects its stability, structure, or function. The neutrality test could be exploited more even in SARS2-S sequences to better appreciate the evolutionary behavior of SARS-CoV-2. Further analysis should be done to know whether if the predicted structures of the variants enhance the binding with the receptor and if the interactions in the trimer are not disturbed. In the end mutations have allowed SARS-CoV-2 to become a persistent threat to mankind, on the scale of a pandemic.

Received: 23 April 2021; Accepted: 18 August 2021

Published online: 22 September 2021

References

- Zhao, X., Ding, Y., Du, J. & Fan, Y. update on human coronaviruses: One health, one world. *Med. Nov. Technol. Devices* **8**, 100043 (2020).
- Petrik, J. S. M. et al. Coronavirus as a possible cause of severe acute respiratory syndrome. *Lancet* **361**, 1319–1325 (2003).
- Zaki, A. M., van Boheemen, S., Bestebroer, T. M., Osterhaus, A. D. M. E. & Fouchier, R. A. M. Isolation of a Novel Coronavirus from a Man with Pneumonia in Saudi Arabia. *N. Engl. J. Med.* **367**, 1814–1820 (2012).
- Wang, C., Horby, P. W., Hayden, F. G. & Gao, G. F. A novel coronavirus outbreak of global health concern. *Lancet* **395**, 470–473 (2020).
- Woo, P. C. Y. et al. Discovery of seven novel mammalian and avian coronaviruses in the genus deltacoronavirus supports bat coronaviruses as the gene source of alphacoronavirus and betacoronavirus and avian coronaviruses as the gene source of gammacoronavirus and deltacoronavirus. *J. Virol.* **86**, 3995–4008 (2012).
- Jalmes, J. A., André, N. M., Chapple, J. S., Millet, J. K. & Whitaker, G. R. Phylogenetic analysis and structural modeling of SARS-CoV-2 spike protein reveals an evolutionary distinct and proteolytically sensitive activation loop. *J. Mol. Biol.* **432**, 3309–3325 (2020).
- Letko, M., Marzi, A. & Munster, V. Functional assessment of cell entry and receptor usage for SARS-CoV-2 and other lineage B betacoronaviruses. *Nat. Microbiol.* **5**, 562–569 (2020).
- Hoffmann, M. et al. SARS-CoV-2 cell entry depends on ACE2 and TMPRSS2 and is blocked by a clinically proven protease inhibitor. *Cell* **181**, 271–280.e8 (2020).
- Huang, Y., Yang, C., Xu, X., Feng, X., Wu, W. & Liu, S. W. Structural and functional properties of SARS-CoV-2 spike protein: potential antiviral drug development for COVID-19. *Acta Pharmacol. Sin.* **41**, 1141–1149 (2020).

10. Wrapp, D. et al. Cryo-EM structure of the 2019-nCoV spike in the prefusion conformation. *Science* <https://doi.org/10.1101/2020.02.11.944462> (2020).
11. Walls, A. C. et al. Structure, function, and antigenicity of the SARS-CoV-2 spike glycoprotein. *Cell* **181**, 281–292.e6 (2020).
12. Zhou, P. et al. A pneumonia outbreak associated with a new coronavirus of probable bat origin. *Nature* **579**, 270–273 (2020).
13. Wang, Q. et al. Structural and functional basis of SARS-CoV-2 entry by using human ACE2. *Cell* **181**, 894–904.e9 (2020).
14. Benion, D. J. et al. Receptor binding and priming of the spike protein of SARS-CoV-2 for membrane fusion. *Nature* <https://doi.org/10.1038/s41586-020-2772-0> (2020).
15. Lan, J. et al. Structure of the SARS-CoV-2 spike receptor-binding domain bound to the ACE2 receptor. *Nature* **581**, 215–220 (2020).
16. Li, F., Li, W., Farzan, M. & Harrison, S. C. Structural biology: Structure of SARS coronavirus spike receptor-binding domain complexed with receptor. *Science* **309**, 1864–1868 (2005).
17. Zamudio, G. S., Prosdoci, F., de Farias, S. T. & José, M. V. A neutral evolution test derived from a theoretical amino acid substitution model. *J. Theor. Biol.* **467**, 31–38 (2019).
18. José, M. V., Zamudio, G. S. & Morgado, E. R. A unified model of the standard genetic code. *R. Soc. Open Sci.* **4**, 66 (2017).
19. José, M. V. & Zamudio, G. S. On the importance of asymmetry in the phenotypic expression of the genetic code upon the molecular evolution of proteins. *Symmetry* **12**, 997 (2020).
20. Baker, J., Meade, A., Pagel, M. & Venditti, C. Positive phenotypic selection inferred from phylogenies. *Biol. J. Linn. Soc.* **118**, 95–115 (2016).
21. Edgar, R. C. MUSCLE: Multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.* **32**, 1792–1797 (2004).
22. Kumar, S., Stecher, G., Li, M., Niyazov, C. & Tamura, K. (2018). MEGA X: Molecular Evolutionary Genetics Analysis across Computing Platforms. *Mol. Biol. Evol.* **35**(6):1547–1549 <https://doi.org/10.1093/molbev/msy096>
23. Petersen, E. F. et al. UCSF Chimera—A visualization system for exploratory research and analysis. *J. Comput. Chem.* **25**, 1605–1612 (2004).
24. Goddard, T. D. et al. UCSF ChimeraX: Meeting modern challenges in visualization and analysis. *Protein Sci.* **27**, 14–25 (2018).
25. Roy, A., Kucukural, A. & Zhang, Y. I-TASSER: A unified platform for automated protein structure and function prediction. *Nat. Protoc.* **5**, 725–738 (2010).
26. Yang, J. et al. The I-TASSER suite: Protein structure and function prediction. *Nat. Methods* **12**, 7–8 (2015).
27. Yang, J. & Zhang, Y. I-TASSER server: New development for protein structure and function predictions. *Nucleic Acids Res.* **43**, W174–W181 (2015).
28. Zhang, C. et al. I-TASSER Genome Wide Structure and Function Modeling of SARS-CoV-2, vol 19 (2020).
29. Zhao, X., Chen, H. & Wang, H. Glycans of SARS-CoV-2 spike protein in virus infection and antibody production. *Front. Mol. Biosci.* **8**, 53 (2021).
30. Millet, J. K. & Whittaker, G. R. Host cell proteases: Critical determinants of coronavirus tropism and pathogenesis. *Virus Res.* **202**, 120–134 (2015).
31. Wu, Y. & Zhao, S. Furin cleavage sites naturally occur in coronaviruses. *Stem Cell Res.* **50**, 102115 (2021).
32. Papa, G. et al. Furin cleavage of SARS-CoV-2 Spike promotes but is not essential for infection and cell–cell fusion. *PLoS Pathog.* **17**, 66 (2021).
33. Daly, J. L. et al. Neuropilin-1 is a host factor for SARS-CoV-2 infection. *Science* **3072**, eabd3072 (2020).
34. Cantuti-Castelvetri, L. et al. Neuropilin-1 facilitates SARS-CoV-2 cell entry and infectivity. *Science* **2985**, eabd2985 (2020).
35. Tang, T., Bidon, M., James, J. A., Whittaker, G. R. & Daniel, S. Coronavirus membrane fusion mechanism offers a potential target for antiviral development. *Antiviral Research* **178**, 104792 (2020).
36. Saito, A. et al. SARS-CoV-2 spike P681R mutation enhances and accelerates viral fusion. *bioRxiv* **4**, 448820 (2021).
37. Becerra, A., Muñoz-Velasco, I., Collom-Salas, W., Vázquez-Salazar, A. & Hernández-Morales, R. Two Short Low Complexity Regions (LCRs) are Hallmark Sequences of the Delta SARS-CoV-2 Variant Spike Protein. *Research Square* 1–14.
38. Gobell, S.M.-C. et al. Effect of natural mutations of SARS-CoV-2 on spike structure, conformation and antigenicity. *bioRxiv Prepr. Serv. Biol.* **6226**, 1–24 (2021).
39. Korber, B. et al. Tracking changes in SARS-CoV-2 spike: Evidence that D614G increases infectivity of the COVID-19 virus. *Cell* **182**, 812–827.e19 (2020).
40. Yurkovetskiy, L. et al. Structural and functional analysis of the D614G SARS-CoV-2 spike protein variant. *Cell* **183**, 739–751.e8 (2020).
41. Luan, B., Wang, H. & Huynh, T. Enhanced binding of the N501Y mutated SARS-CoV-2 spike protein to the human ACE2 receptor: Insights from molecular dynamics simulations. *FEBS Lett.* **595**, 1454–1461 (2021).
42. Liu, Y. et al. The N501Y spike substitution enhances SARS-CoV-2 transmission. *bioRxiv Prepr. Serv. Biol.* <https://doi.org/10.1101/2021.03.08.434499> (2021).
43. Cai, Y. et al. Structural basis for enhanced infectivity and immune evasion of SARS-CoV-2 variants. *bioRxiv Prepr. Serv. Biol.* **9745**, 1–14 (2021).
44. Bolze, A. et al. Rapid displacement of SARS-CoV-2 variant B.1.1.7 by B.1.617.2 and P.1 in the United States. *medRxiv* 2021.06.20.21259195 (2021).
45. Mathew, D. C. & Luthy-Schulten, Z. On the physical basis of the amino acid polar requirement. *J. Mol. Evol.* **66**, 519–528 (2008).
46. Woese, C. R., Dugre, D. H., Saxinger, W. C. & Dugre, S. A. The molecular basis for the genetic code. *Proc. Natl. Acad. Sci. USA* **55**, 966–974 (1966).
47. Hou, Y. J. et al. SARS-CoV-2 D614G variant exhibits efficient replication ex vivo and transmission in vivo. *Science* **370**, 1464–1468 (2021).
48. Planke, J. A. et al. Spike mutation D614G alters SARS-CoV-2 fitness. *Nature* <https://doi.org/10.1038/s41586-020-2895-3> (2020).
49. Zhou, B. et al. SARS-CoV-2 spike D614G variant confers enhanced replication and transmissibility. *bioRxiv* <https://doi.org/10.1101/2020.10.27.357558> (2020).
50. Sikora, M. et al. Computational epitope map of SARS-CoV-2 spike protein. *PLoS Comput. Biol.* **17**, e1008790 (2021).
51. Gu, H. et al. Adaptation of SARS-CoV-2 in BALB/c mice for testing vaccine efficacy. *Science* **369**, 1603–1607 (2020).
52. Smith, E. C. & Denison, M. R. Coronaviruses as DNA Wannabes: A new model for the regulation of RNA virus replication fidelity. *PLoS Pathog.* **9**, e1003760 (2013).
53. Hulsweil, R. J. G. et al. Human coronaviruses OC43 and HKU1 bind to 9-O-acetylated sialic acids via a conserved receptor-binding site in spike protein domain A. *Proc. Natl. Acad. Sci. U. S. A.* **116**, 2681–2690 (2019).
54. Alejandra Torricci, M. et al. Structural basis for human coronavirus attachment to sialic acid receptors. *Nat. Struct. Mol. Biol.* **26**, 481–489 (2019).
55. Gao, C. et al. SARS-CoV-2 Spike Protein Interacts with Multiple Innate Immune Receptors. *bioRxiv* (2020). doi:<https://doi.org/10.1101/2020.07.29.227462>.

Author contributions

Conceptualization: G.L.-C., M.V.J.; methodology: G.L.-C., G.S.Z., M.P.-P., M.V.J.; software: G.S.Z., M.P.-P.; validation: G.L.-C., G.S.Z., M.P.-P.; formal analysis: G.L.-C., G.S.Z., M.P.-P., M.V.J.; investigation: G.L.-C., G.S.Z.,