PONTIFICIA UNIVERSIDAD CATÓLICA DEL ECUADOR

FACULTA DE HÁBITAT, INFRAESTRUCTURA Y CREATIVIDAD

PROYECTO DE GRADO:

CARACTERIZACIÓN DE LA DIVERSIDAD GENÉTICA DE DELECIONES E INSERCIONES EN LOS GENOMAS DE SARS CoV-2 EN ECUADOR Y SU RELACIÓN CON EL FITNESS VIRAL ENTRE OCTUBRE DE 2023 Y SEPTIEMBRE DE 2024

PROYECTO DE GRADO PREVIO A LA OBTENCIÓN DEL TÍTULO DE:

MÁSTER EN BIOLOGÍA COMPUTACIONAL

AUTORA: GARCÍA CANDO DANIELA SOFÍA

TUTOR: PhD SERGIO ALAN CERVANTES PÉREZ

Dedicatoria

Todo el esfuerzo y trabajo que puse en este proyecto no hubiera sido posible sin las personas que estuvieron junto a mi nutriéndome con su amor y brindándome apoyo en cada paso. Por eso, se lo dedico a mi hermana, Cristina, a mamá, a papá y a Dasha, quienes han sido mi soporte, mi inspiración y motivación para no rendirme.

También, me lo dedico, porque pese a las adversidades que se presentaron en el camino, no me rendí, y creí en mi misma y en lo que puedo conseguir.

En memoria de: Angie y Jacinto

Agradecimientos

Considero que hoy en día, en una sociedad que cada vez nos exige ser más individualistas, es bueno recordar que cada meta alcanzada es un logro colectivo.

Por ello quiero aprovechar para agradecer a mi familia, por su profundo amor, a mis amigos por su aliento constante y a mis compañeros de trabajo por su respaldo.

También, expresar gratitud a los investigadores y colegas que, con su trabajo y pasión por realizar ciencia en Latinoamérica, me han inspirado.

Y a la guía y consejos de muchos de ellos, incluyendo Alan, mi tutor, quienes no me dejaron rendir.

Gracias por ser mi red de apoyo

Índice de Contenidos

Índice de Tablas

Índice de Imágenes

Índice de Figuras

Resumen

El monitoreo epidemiológico de SARS CoV-2 con secuenciación es indispensable actualmente, ya que permite identificar de manera oportuna variantes que pueden ser de interés por su impacto epidemiológico. Como consecuencia de este cambio de enfoque actualmente se cuenta con una gran cantidad de información genómica cuya interpretación es fundamental para que las instituciones de salud y gobiernos puedan tomar decisiones informadas sobre el control y contención del virus. En este contexto, el análisis secundario, es decir el llamado de variantes genómicas es importante pues implica hacer que esta información sea accionable al identificar eventos genéticos que puedan impactar el fitness viral aumentando la transmisibilidad, patogenicidad y disminuyendo la eficiencia de los tratamientos disponibles. Si bien dentro de estos eventos genéticos, las sustituciones de nucleótidos han sido muy estudiadas, este no ha sido el caso para inserciones y deleciones. Razón por la que en este estudio se pretende identificar eventos de deleción e inserción que tengan un impacto positivo en el fitness viral. Esto dentro de cerca de 800 secuencias del virus del SARS CoV-2 en Ecuador que se descargaron del GISAID. También se incluyó más de 1000 secuencias de Colombia, México y Brasil, para contar con un panorama más amplio sobre el comportamiento del virus en la región y con diferentes presiones ambientales dentro del periodo de octubre de 2023 a septiembre de 2024. Esto mediante el uso de herramientas bioinformáticas como NextClade de NextStrain, ancladas en la nube y realizando gráficos y clasificación de datos en Excel, y pruebas estadísticas en R.

Abstract

Since the beginning of the 2020 COVID-19 epidemic, sequencing has become a major tool for epidemiological monitoring, because it enables an early identification of variants of concern or interest due to their epidemiological impact. Genomic information has increased exponentially, but interpreting it remains a bottleneck in this workflow. This abundance of information, which could help build strategies to combat potential contagious variants, becomes useless.

Secondary analysis of this information is key to making it actionable, through identifying genetic events as substitutions, deletions, or insertions that may impact viral fitness by increasing its transmissibility and pathogenicity or decreasing the effectiveness of available treatments. Despite extensive research on substitutions, insertions and deletions remained unexplored. Consequently, this study sought to identify deletion and insertion events that could potentially enhance viral fitness by analyzing nearly 2000 SARS CoV-2 sequences from four countries: Ecuador, Colombia, Brazil, and Mexico. This work was carried out using bioinformatics tools such as the cloud-based platform: NextClade from NextStrain. Microsoft Excel for data classification and R Studio for statistical analysis.

1 Introducción

El virus del SARS CoV-2 y la pandemia que ocasionó, cambiaron totalmente la forma cómo se realizaba vigilancia genómica a nivel mundial, ocasionando una revolución que permitió un mayor acceso a la secuenciación de fragmentos cortos en países del sur global, como Ecuador. Esta tecnología demostró ser fundamental para monitorear casi en tiempo real la evolución y propagación del virus (Farkas, Mella, Turgeon, & Haigh , 2021). Un virus que desde los inicios de la pandemia ha demostrado, por la naturaleza de su genoma de ARN en sentido positivo, ser altamente variable y adaptable a sus hospederos, los humanos (Tosta, y otros, 2023). Por ello a nivel mundial se crearon bases de datos como el GISAID que han permitido compartir secuencias y notificar de manera oportuna la aparición de variantes de preocupación también conocidas como VOCs. Las VOCs son linajes de virus que comparten características biológicas que afectan su fitness viral, es decir su capacidad de replicación, transmisión y su patogenicidad; y que pueden llegar a afectar la inmunidad alcanzada ya sea por infecciones previas o vacunas (Konings, y otros, 2021).

Estas características biológicas se originarán a partir de eventos genéticos que afectarán la secuencia de ARN "original" del virus. Dentro de estos eventos genéticos tenemos sustituciones, inserciones y deleciones. Las sustituciones han sido ampliamente descritas por investigadores pues su monitoreo desde los primeros años de la pandemia fue crucial, ya que muchas de ellas como ejemplo D614G han afectado la efectividad de las vacunas (Cheng, y otros, 2021). Por otro lado, muy poco se ha monitoreado de las deleciones e inserciones, siendo fenómenos que han sido un tanto relegados, pese a que en los últimos años ha aumentado la evidencia de que estos tienen un rol dentro del éxito evolutivo de VOCs como Alfa (Akaishi & Fujiwara, 2023). Por lo que este proyecto busca identificar deleciones e inserciones que impacten positivamente el fitness viral. Esto mediante el análisis secundario con la plataforma NextClade de secuencias del virus del SARS CoV-2 de Ecuador dentro del periodo de octubre de 2023 a septiembre de 2024. También, y para tener un panorama más extenso de la situación en la región, también se realizó este análisis con secuencias de Colombia, México y Brasil. Sin dejar de lado que los cuatro países por su situación geográfica y social; brindará presiones selectivas diferentes al virus. Finalmente, el proyecto pretende demostrar la utilidad de herramientas de libre acceso robustas como NextClade de NextStrain para facilitar el acceso a la gran cantidad información genómica disponible.

1.1 Objetivos

1.1.1 Objetivo general

Identificar la diversidad de inserciones y deleciones de nucleótidos en los genomas de SARS CoV2 circulantes en Ecuador, Brasil, Colombia y México, entre el periodo de octubre 2023 a septiembre 2024, reportados en la Base GISAID.

1.1.2 Objetivos específicos

- Trabajar con las herramientas disponibles en NextStrain para el análisis secundario de secuencias de SARS CoV-2
- Acceder a las secuencias de los genomas de SARS CoV-2 en la base de datos de GISAID
- Identificar la diversidad de inserciones y deleciones en los genomas de SARS CoV-2 en Ecuador, Brasil, Colombia y México.
- Evaluar la correlación entre indels y el fitness del virus del SARS CoV-2

2 Justificación

El presente proyecto es relevante para la vigilancia genómica viral en Ecuador, ya que, aunque actualmente se realiza seguimiento al virus en el país, esta información continúa sin aprovecharse en el sentido de que todavía desconocemos como esta información puede ser relevante en la toma de decisiones a nivel de salud pública. A través de este estudio, se busca de manera modesta pero significativa, demostrar el potencial de los datos genómicos para comprender mejor la evolución del virus del SARS CoV-2 dentro del país, y su impacto en la población, utilizando herramientas sencillas y de libre acceso como NextStrain. El proyecto también aspira a contribuir en los constantes esfuerzos por entender globalmente la dinámica del virus.

3 Planteamiento del problema

Desde la aparición del virus del SARS-CoV-2 en 2019, la secuenciación ha sido fundamental en la caracterización del virus y para vigilar su co evolucionado con los humanos. Además, que es una herramienta que ha permitido generar alarmas tempranas ante la aparición de variantes que sean epidemiológicamente relevantes, ya sea porque tengan una mayor transmisibilidad, mayor gravedad en la enfermedad, reduzcan la eficiencia en los tratamientos disponibles o presenten evasión inmunitaria. Como consecuencia, actualmente, se cuenta con una gran cantidad de genomas del virus en las diferentes bases de datos disponibles (Farkas, Mella, Turgeon, & Haigh , 2021). Un ejemplo, y la base de datos con la que se realizó esto proyecto es el GISAID (*Global Initiative on Sharing All Influenza Data*), una iniciativa que inició en 2018 para almacenar información genómica de influenza, pero que en los últimos años a raíz de la pandemia se convirtió en la mayor base de datos de genomas de SARS CoV-2. Ecuador es miembro del GISAID y participa en esta, proporcionando secuencias y actualizaciones constantes sobre el estado epidemiológico del SARS CoV-2.

Para estudiar la evolución del virus, es importante identificar los eventos que generan diversidad genética y que, en el caso de virus de ARN como el SARS-CoV-2, se ven influenciados por altas tasas de mutación, replicación y recombinación. Las inserciones y deleciones serán originadas, por estos últimos eventos (Rogozin, y otros, 2024).

Y es importante correlacionar la presencia de estos eventos con ventajas evolutivas que han otorgado a los linajes una mayor propagación y prevalencia. Por lo que con este proyecto deseo caracterizar la diversidad de deleciones e inserciones de nucleótidos presentes en los genomas del virus del SARS CoV-2, circulantes en el periodo de tiempo desde octubre de 2023 a septiembre de 2024, en cuatro países: Ecuador, Colombia, Brasil y México. Además, y determinar sí existe una relación entre ciertos indels y el *fitness* del virus (Li, Yan, Wong , & Cui, 2023).

También, y conociendo lo complicado que es actualmente acceder a herramientas gratuitas para realizar análisis secundario de secuencias de SARS CoV-2 para personas no familiarizadas con bioinformática, utilizamos NextClade de NextStrain, para que su implementación en este proyecto pueda servir como referencia a otros investigadores.

4 Marco teórico

4.1 Antecedentes o marco referencial

El foco principal de este proyecto es el Ecuador, lo tomamos como punto de partida, ya que en un inicio nuestro objetivo era centrarnos en analizar la diversidad genética del SARS CoV-2, no obstante, la historia estaba incompleta, por lo que añadimos más sets de datos de Colombia, Brasil y México; para tratar de comprender un poco la dinámica del virus en este periodo de tiempo en estos territorios, y como se detalla más adelante, son "ecosistemas" muy particulares que brindaron escenarios interesantes al virus para su evolución. Sin embargo, solo en Ecuador ahondamos en temas relacionados con ministerios de salud, vigilancia, políticas, etc., ya que es la realidad con la que nos relacionamos diariamente y de la que tenemos un mayor conocimiento.

Aclarado esto, esta primera parte se va a concentrar en Ecuador, en el país la respuesta ante la pandemia del SARS CoV-2 estuvo encabezado por el Ministerio de Salud Pública el mismo que desarrollo el *Plan Fénix*. Un plan que inicio en 2021 y cuyo objetivo era responder (no hay datos sobre cuando concluyó) de manera integral ante el COVID-19, entre sus pilares podemos descartar: el fortalecimiento de los laboratorios nacionales y la prevención del control de infecciones. Ambos permitieron que entre 2021 y 2022, se implementará la infraestructura para realizar vigilancia epidemiológica del virus a nivel nacional y a través de los hospitales parte de la red de salud pública (Acosta, y otros, 2022).

Esto siguiendo un protocolo donde primero se realiza una evaluación por parte de un médico quien, al identificar los criterios clínicos afines a la enfermedad, indicará al paciente que se realice una prueba de RT-qPCR para confirmar el diagnóstico. Se cierra el caso si el paciente es negativo. Mientras que, si es positivo, se procederá con el tratamiento correspondiente, además que se dará seguimiento en el caso de que los síntomas sean más graves (Acosta, y otros, 2022).

Como es conocido, la alta variabilidad del virus representa también un reto importante al momento de diagnosticar y dar un tratamiento a los pacientes, por lo que fue necesario implementar un protocolo adicional para la vigilancia de variantes (Acosta, y otros, 2022). Es allí donde una institución pública jugó un rol protagónico, el INSPI (*Instituto Nacional de Investigación en Salud Pública*), con dos departamentos: El Centro de Referencia Nacional de Genómica, Secuenciación y Bioinformática (Con Sede Quito) y el Centro de Referencia Nacional de Influenza y otros Virus Respiratorios (Con Sede Guayaquil). Ambos realizando vigilancia genómica del virus en tiempo real, lo que permitió que Ecuador identificara a tiempo variantes relevantes como FLiRT (Ministerio de Salud Pública, 2024).

Regresando con el protocolo de secuenciación para vigilancia de variantes, este se sigue en pacientes que presenten ciertos síntomas como: reinfecciones, infección grave pese a contar con vacunación, fallecimiento pese a contar con vacuna, y pacientes menores de 50 años sin comorbilidades (Acosta, y otros, 2022). Se debe mencionar que estos criterios han cambiado considerando como las nuevas variantes han afectado la sintomatología "típica" del COVID-19.

En Ecuador, la vigilancia genómica de SARS CoV-2 es un ejemplo positivo del impacto que tuvo la pandemia en las políticas de salud pública, pues permitió que se equiparan con tecnología de secuenciación dos laboratorios en Quito y Guayaquil (en INSPI) orientados a la vigilancia de virus respiratorios como el SARS CoV-2, la influenza; de bacterias

multirresistentes y de otras enfermedades infecciosas como el Dengue. Sin mencionar, que existen también universidades que colaboran en este esfuerzo por realizar vigilancia del virus. Y es gracias a estas instituciones que los investigadores contamos con la información genómica para realizar proyectos de este tipo. Anteriormente se ha trabajado con esta información caracterizando variantes, no obstante, no se ha realizado una investigación con el enfoque que proponemos de identificar y caracterizar inserciones y deleciones, junto con caracterizarlas y correlacionar su efecto en el fitness.

4.2 Conociendo al virus

El virus del SARS CoV-2 (*Severe Acute Respiratory Syndrome Coronavirus 2*) es el virus causante de la enfermedad respiratoria de coronavirus de 2019 o ***COVID-19*****,** que fue el causante de la pandemia del COVID-19 (PAHO, s.f.). Este virus de ARN monocatenario de cadena positiva de la familia de los *Coronaviridae* se originó en un mercado en Wuhan donde se comercializa fauna silvestre, por lo que se cree el virus "migró" hacia los humanos, es decir es un virus zoonótico (Tosta, y otros, 2023).

La primera alerta epidemiología por parte de la OMS (*Organización mundial de la Salud*) se dio el 16 de enero de 2020 debido a la rápida propagación que este presentaba dentro de Asia y Europa, en un principio se especulaba que su transmisión era por vía aérea, un hecho que se confirmaría más adelante. El alto nivel de globalización e interconectividad del mundo moderno permitió que el virus se propagará en todo el mundo en cuestión de meses, declarándose de manera oficial el inicio de la pandemia el 11 de marzo de 2020 (Tosta, y otros, 2023).

Este virus de puso a prueba los sistemas de salud pública a nivel mundial. Pues al ser un virus novedoso, no se contaba con guías de tratamiento a seguir, esto sumado con la rápida propagación y las complicaciones en población en riesgo como las personas de la tercera edad, hizo que la demanda por camas, medicamentos sea insostenible para las entidades de salud, lo que llevó a su colapso. Todo esto sin detallar el gran drama humano, social y económico que trajo consigo la pandemia (Tosta, y otros, 2023).

En los primeros años de la pandemia, la comunidad científica se centró en buscar una vacuna que permitiera que se recuperará la tan necesaria normalidad, y para ello, la vigilancia genómica fue fundamental pues brindó la información para desarrollarla, además de los recursos para desarrollar pruebas diagnósticas que se ajustaran a los cambios que ocurrían en el genoma del virus (Tosta, y otros, 2023).

4.3 Evolución del virus

Se debe mencionar un hecho, y es que el virus del SARS CoV-2 es un virus altamente variable, y esto se debe a múltiples factores, que se detallan más adelante. Primero, el virus del SARS CoV-2, al igual que los virus de la influenza y del HIV, pertenecen a una variedad de "virus envueltos", que son virus recubiertos en una bicapa lipídica que contiene glicoproteínas que están glicosiladas en los grupos nitrógeno. Esta capa los hace más susceptibles al estrés ambiental como sequía, alcohol, y otros agentes; y por tanto presentarán una mayor plasticidad, respondiendo "evolutivamente" de manera acelerada (Rao, y otros, 2021).

Segundo, estos virus recurren a la polimerasa dependiente de ARN o RdRp, para su replicación, y esta es más susceptibles a errores lo que hará que exista una mayor aparición de mutaciones, hasta 2021 se reportó una tasa de 7.23 mutaciones por muestra en

promedio (Rao, y otros, 2021). Además, relacionado con este mecanismo al copiarse el ARN, en algunas ocasiones la RdRp se disociará de la cadena que está copiando, integrándose nuevamente a otro lugar del genoma, similar al que se estaba copiando y que, al unirse a un lugar no exactamente homólogo, se generarán dos situaciones. En una de ellas se generará una copia de más originando una inserción. O una deleción, omitiéndose una parte del genoma. Todo esto en un proceso llamado recombinación homóloga imperfecta (Sierra , y otros, 2021).

Tercero, los virus del SARS CoV-2 dependen de la maquinaria de la célula hospedera para la edición de su genoma como consecuencia ocurren mutaciones sustanciales y dirigidas (Markov, y otros, 2023).

Cuarto, la tasa de sustitución para el virus, en comparación a otros es muy elevada. Aquí es importante diferenciar entre tasa de sustitución y de mutación que no son lo mismo. Por un lado, la tasa de mutación es definida como la frecuencia a la cual ocurrirán nuevas mutaciones en el genoma dentro de cierto periodo de tiempo, siendo estas causadas por errores durante la replicación, o motivada por otros factores ambientales causantes de mutaciones como la radiación. Por otro lado, la tasa de sustitución describe a la frecuencia con la que las mutaciones que ya ocurrieron regresan a la versión anterior (Markov, y otros, 2023).

Entre mutaciones relevantes podemos mencionar a las que afectan a la proteína SPIKE, presente en la envoltura del virus y que es clave en el ingreso del virus a la célula hospedera, a través del receptor ACE2 (*angiostensin-converting enzyme*). Para ser más específicos, la sustitución D614G que se ha observado aumenta la transmisión del virus, permitiendo la expansión y prevalencia del virus (Cheng, y otros, 2021).

4.4 El SARS CoV-2 en América Latina

Este estudio se va a concentrar en lo ocurrido con el virus del COVID-19 en Ecuador; durante el periodo comprendido entre octubre del 2023 y septiembre del 2024. No obstante, para tener información contrastante se eligió tres países adicionales: Colombia, Brasil y México. Antes de ahondar en los motivos por los cuales se eligió cada país, es importante mencionar que se decidió analizar a cada uno de estos “ecosistemas” para comprender como el virus co-evoluciona con sus hospederos, pues cada país de maneras peculiares permitió que ocurriera lo siguiente:

4.4.1 Flujo genético:

El ingreso de nuevas variantes al acervo genético local. Que puede ser causado por fenómenos sociales como la migración, el comercio en fronteas y el turismo. Entonces al ingresar nuevas variantes al acervo genético, si la propagación del virus es lo suficientemente descontrolada ya sea por factores como una baja vacunación, alta movilidad interna o medidas sanitarias inadecuadas; estas variantes transmitirán internamente, llegando a ocurrir recombinaciones, lo que llevará a una alta diversidad genética, que a largo plazo tendrá consecuencias como aparición de nuevos linajes y eventualmente clados (LaRotta, y otros, 2023).

4.4.2 Selección natural:

Es decir que, por procesos como la vacunación o la diversidad de la respuesta inmune de su población, el virus se enfrente a diferentes escenarios a los que debe adaptarse. En el caso de Latinoamérica, la región presenta diversidad de poblaciones mestizas y migrantes,

lo que hace que la respuesta inmune en la región sea muy variada. Y a esto hay que sumarle que, como consecuencia de las diversas condiciones políticas y sociales de los países analizados, sus estrategias de vacunación fueron diferentes, y las vacunas utilizadas muy diversas (LaRotta, y otros, 2023).

Se detalla las características de cada país para dar un contexto del "ecosistema" de cada uno:

4.4.3 Ecuador:

Representa un escenario en el que durante los primeros meses la propagación fue descontrolada en zonas altamente urbanas como Guayaquil, la propagación comunitaria se frenó debido a la fuerte cuarentena y reducida actividad social y comercial que se observó en el primer año de la pandemia, algo imperativo para un sistema de salud como el de Ecuador. Este ecosistema también recibirá un flujo de migración de América Latina (ya que comparte su frontera más larga con Colombia), y turismo en baja medida desde América del Norte, Europa y Asia. Además, por sus regiones naturales y la gran heterogeneidad de su población, y diversidad de vacunas utilizadas para inmunizar a la población la presión evolutiva varía significativamente (Herrera, Troya, & Gaus, 2020). En resumen, Ecuador funcionará como receptor de variantes y propagador de nuevos linajes.

4.4.4 Colombia:

Es un escenario interesante para el flujo genético de nuevos linajes, más allá de ser un destino turístico atractivo a nivel mundial, es un país de tránsito importante en América Latina no solo aéreo (ya que tiene el segundo aeropuerto más transitado), también terrestre, considerando que aquí se encuentra la selva del Darién, que conecta Colombia con Panamá, un sitio de paso para migrantes. Por otra parte, al contener regiones naturales diferentes (andes, costa, amazonia, etc.), la respuesta y propagación del virus ocurrió de diversas formas. En resumen, se podría decir que Colombia funciona como un receptor y propagador de nuevos linajes (Ramírez, y otros, 2021).

4.4.5 Brasil:

Dentro de los fenómenos demográficos que generarán un mayor diversidad genética en este, el país más grande de Latinoamérica se debe mencionar que recibe un flujo migratorio y turístico importante de América Latina, Europa, Asia y América del norte. Y que posee un intenso movimiento migratorio interno que no cesó ni durante la cuarentena del 2020. Y hay recordar que por la baja cobertura de vacunación que hubo durante los primeros años de la pandemia, la tasa de transmisión comunitaria fue muy alta, algo agravado por la alta urbanidad que tiene el país. Lo que ha puesto a Brasil en múltiples olas de COVID, siendo la más relevante la segunda ola de COVID, durante la cual surgió el linaje Gamma, una variante caracterizada por su alta transmisibilidad y evasión inmune. (Giovanetti, y otros, 2022). Se podría considerar que este ecosistema en el comportamiento de su población se asemejó a lo observado en Estados Unidos por el movimiento antivacunas y un flujo intenso de migración internacional. En resumen, Brasil ofrece un ecosistema con un alto flujo genético, y diversidad de respuestas inmunes por una población heterogénea, que hace que funcione como sitio de origen de nuevas variantes.

4.4.6 México:

En este caso el país presentará un alto flujo genético, ya que comparte con Estados Unidos la frontera con mayor tránsito de América, además de concentrar un flujo migratorio masivo

proveniente de Latinoamérica y Centroamérica. También, presentará una alta diversidad de presión inmune de parte de la población, no solo por las diversas estrategias y vacunas utilizadas para inmunizar a la población, también por el desigual acceso al sistema de salud en las diferentes regiones de México (Flores, y otros, 2022). En resumen, México, funcionará como un sitio de recepción de nueva variantes, y por la presión inmune de la población al virus, permitirá que estas se diversifiquen y propagará nuevos linajes a nivel mundial.

De manera general, ecosistemas como México, Colombia y Ecuador, por factores demográficos, sociales y ambientales; en estos años han funcionado como propagadores y generadores de nuevos linajes de COVID-19. Mientras que Brasil, debido al tamaño de su población, y a las estrategias de vacunación utilizadas durante los primeros años (que permitieron un propagación comunitaria descontrolada), se ha comportado como un ecosistema que brinda una mayor diversidad genética, como ejemplo tenemos a la variante VOC (*variants of concern*) a la que dio origen.

5 Marco conceptual

5.1 SARS CoV-2 y su nomenclatura

La alta variabilidad genética característica de este virus dio origen a múltiples "variantes de COVID-19", con lo que fue necesario clasificarlas y nominarlas para que su seguimiento sea más sencillo. Aquí se debe aclarar varios conceptos, el término "variante" es una forma coloquial de referirse a lo que en taxonomía conocemos como *linajes,* que son un conjunto de virus que comparten algunas características genéticas y un origen común (Dellicour, y otros, 2021). Estos linajes a su vez se pueden agrupar en *clados*, dentro de los cuales se incluirá todos los descendientes y el ancestro común más próximo, compartiendo una característica genética clave. Aquí se debe aclarar que un clado puede contener uno o varios linajes, y que un linaje estará contenido dentro de un clado (Fauver, y otros, 2020).

Dentro los sistemas de nomenclatura que se utilizaron durante los primeros años de la pandemia existen cuatro que son relevantes por su popularidad estos son: OMS, NextStrain, Pango y GISAID. Más adelante, se va a revisar en detalle cada uno de estos sistemas, sin embargo, en este proyecto solo haremos referencia a los tres primeros, por su uso más difundido en plataformas de análisis genómico y en reportes de eventos genéticos relevantes.

5.1.1 Nomenclatura de la OMS

Esta nomenclatura fue desarrollada a finales del 2020, y entre sus finalidades tenemos crear una nomenclatura fácil de identificar por un público no técnico, y sobre todo liberar de estigma a los países dónde por primera vez se reportaba una variante. Para ello se utilizarían letras del alfabeto griego, para variantes de relevancia epidemiológica, es decir para variantes VOC (*variants of concern*) y VOI (*variants of interest*) (Konings, y otros, 2021). Ambas se basan en la presencia de mutaciones, o inserciones dentro de un linaje que tendrán afectaciones en: la transmisibilidad del virus, la severidad de la enfermedad y la eficiencia (ya sea del diagnóstico, la terapia o las vacunas) (Ahmad , Fawaz, & Aisha, 2022).

Las variantes VOI (variantes de interés), representan un riesgo a la salud pública ya que presentan eventos genéticos que se sospecha podrían afectar una de las tres características del virus es decir su transmisibilidad, la severidad o la eficiencia de los tratamientos disponibles para tratarla o prevenirla. Además, que este evento parece haber

causado una mayor tasa de propagación en el virus. Dentro de este criterio se ha agrupado a variantes como Mu, Lamba, Epsilon (Ahmad , Fawaz, & Aisha, 2022).

Por otro lado, las variantes VOC (variantes de preocupación), son aquellas en las que de manera efectiva, reiterada y sólida se demuestra que una o más de las tres características antes mencionadas del virus, sí han cambiado, junto con una mayor tasa de propagación del virus. En este grupo tenemos a las variantes: Alfa, Gamma, Delta y Ómicron (Ahmad , Fawaz, & Aisha, 2022).

5.1.2 Nomenclatura Pango

Esta nomenclatura se desarrolló en abril de 2020, y surgió como una respuesta al gran volumen de información genómica que surgía sobre el virus del SARS CoV-2, buscando no enfocarse en clados, como lo hacían las nomenclaturas científicas más usadas hasta el momento (NextStrain y GISAID=. La nomenclatura PANGO por sus siglas en inglés *Phylogenetic Assignment of Named Global Outbreak Lineages,* hila mucho más fino, pues utiliza Pangolin una herramienta computacional que asigna de manera automática el linaje más probable tomando en cuenta su genoma, y brindándole una nomenclatura dinámica (O'Toole, y otros, 2021). Podemos mencionar como ejemplo: linaje B.1.1.7 más tarde agrupado como variante Alpha (VOC).

5.1.3 Nomenclatura NextStrain

Esta nomenclatura designa a los clados de SARS CoV-2 buscando lo siguiente: asignar un nombre a aquellos que están genéticamente definidos y que han alcanzado una frecuencia y dispersión geográfica histórica. Además, que en caso de ser necesario designar un nombre temporal, a clados emergentes cuya nomenclatura se volverá "oficial" sí este llega a ganar presencia y relevancia epidemiológica. Finalmente, utilizar una nomenclatura fácil de entender sin perder rigor científico junto con que esta nomenclatura seguirá funcionando, aunque surjan nuevos clados, o que el virus mute (Bedford, y otros, 2025).

Por otro lado, esta nomenclatura se conformará por el año en el que se estima emergió el clado junto con una letra del alfabeto, no obstante, anualmente estos reinician lo que evita que lleguemos a usar todas las letras del alfabeto además que se evitará nombres largos o muy técnicos. Es importante mencionar que el hecho de que anualmente se reinicien las letras no implica que nuestros clados ya nombrados sean renombrados, aunque ya no circulen. Aquí es importante aclarar el concepto de clado principal, que ocurre cuando al menos un clado presenta dos mutaciones que lo distancia de su clado padre, tiene una frecuencia a nivel mundial del 20% y que sus secuencias se distribuyen de manera adecuada en el espacio y tiempo (Bedford, y otros, 2025). Por estas razones recurrimos a esta nomenclatura durante este proyecto, junto con la de OMS.

5.2 Fitness Viral

Durante esta investigación un concepto importante a entender es el fitness viral, que es un concepto muy amplio, recordemos que el fitness desde el punto de vista de la evolución de Darwin se refiere al éxito que tiene una especie para heredar sus genes a una siguiente generación, algo no muy lejano a lo que buscamos evaluar en los virus al momento de hablar de fitness viral, que se define como: "La capacidad del virus de producir progenie infecciosa en un ambiente dado". Evaluándose tres tipos de fitness: replicativo, de transmisión y epidemiológico (Wargo & Kurath, 2012).

5.2.1 Fitness Replicativo

Se evalúa mediante estudios *in vivo* e *in vitro*, donde se infecta modelos animales o celulares con dos o más variantes de un virus para determinar cuál se replica de manera exitosa. Esta medida trata de explicar la capacidad del virus para propagarse o replicarse eficientemente en una célula hospedera (Wargo & Kurath, 2012).

5.2.2 Fitness de Transmisión

Mide la efectividad que tiene un virus replicándose dentro de un hospedero y transmitiéndose a través de sus vectores, para de esta manera llegar a nuevos hospederos (Wargo & Kurath, 2012).

5.2.3 Fitness Epidemiológico

Es el que más nos interesa en este estudio, para entenderlo hay que analizar los cambios en la composición de la población viral (genotipos), junto con su distribución y prevalencia. Para ello se puede ejecutar estudios *in vivo*, pero también se puede recurrir a información como las estadísticas de los virus circulantes durante cierto periodo de tiempo (Wargo & Kurath, 2012), este último fue el enfoque que siguió este estudio.

6 Metodología y Técnicas

A continuación, vamos a detallar la metodología utilizada durante este estudio:

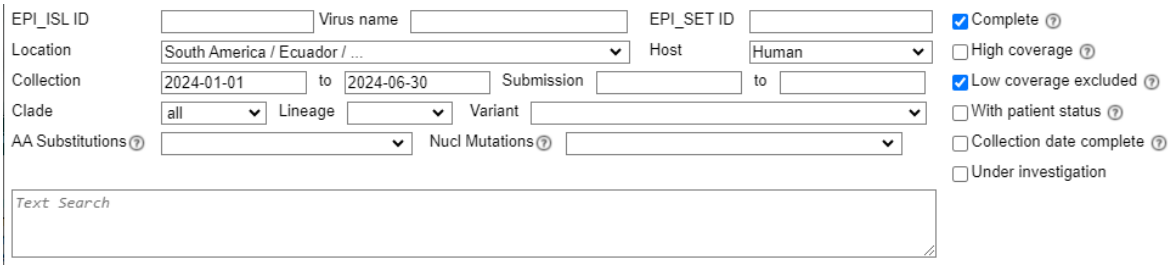
6.1.1 Muestreo

Para acceder a las muestras de interés, se creó una cuenta en la base de datos del GISAID, esto permitió acceso a secuencias de todo el mundo, pero los países de interés para este proyecto son: Ecuador, Colombia, Brasil y México. Considerando que la cuenta toma cerca de dos semanas en ser aprobada, y una vez que se contó con el acceso se procedió a familiarizarse con la página y las herramientas que esta tiene.

6.1.1.1 Búsqueda refinada de las secuencias

Para descargar las secuencias disponibles en la base de datos, se ingresa a la pestaña *Downloads,* y se elige la opción *Custom selection*, esto permite elegir las secuencias del país que nos interesa, desplegándose la ventana que se muestra abajo:

Imagen 1. Ventana de búsqueda personalizada GISAID



Descripción: En la imagen se puede observar los filtros a llenar dentro de la ventana de búsqueda personalizada en GISAID para descargar las secuencias de SARS CoV-2 con las que se va a realizar el proyecto.

A continuación, se detalla las configuraciones utilizadas en los filtros para los diferentes países, junto con comentarios y el total de secuencias recuperadas:

Tabla 1. Configuraciones utilizadas para descargar secuencias de cada país

Ubicación	South America/ Ecuador	South America/ Colombia	South America/ Brazil/Sao Paulo	North America/ Mexico/ Mexico City
Host	Human	Human	Human	Human
Collection date	2023-10-01 to 2024-09-30	2023-10-01 to 2024-09-30	2023-10-01 to 2024-10-31	2023-10-01 to 2024-10-31
Complete	Check	Check	Check	Check
Low coverage excluded	Check	Check	Check	Check
Comentarios	Se mantiene el filtro general "Ecuador" tomando en cuenta el bajo número de muestras de SARS CoV-2 que llegan a ser secuenciadas en todo el país	Se mantiene el filtro general "Colombia" tomando en cuenta el bajo número de muestras de SARS Cov-2 reportadas en la capital, y considerando la diversidad de nomenclaturas utilizadas por los investigadores.	Se colocó como ubicación filtro Sao Paulo, tomando en cuenta que este estado contiene a la ciudad más poblada de Brasil, Sao Paulo, importante capital económica de Brasil.	Se colocó como ubicación filtro Mexico Cityr, tomando en cuenta que es la ciudad más poblada de México, y su estatus de capital, hace que concentre una población muy heterogénea.
Total, secuencias recuperadas al momento*	831	313	1155	1039

**Se debe tomar en cuenta, que la base de datos de GISAID está en constante curación, por lo que los reportes de secuencias variarán con el tiempo, llegando a aumentar su número (sí se reportan nuevas) o disminuir (sí existe algún cambio en estas). Por lo que los números reportados aquí pueden variar.*

Descripción: En la tabla se registra las configuraciones utilizadas para cada país, en el caso de México y Brasil se explica por qué geográficamente se eligieron ciertas ciudades o estados. También el total de secuencias que en al momento de realizar el muestreo se recuperaron de GISAID.

Es importante mencionar novedades que ocurrieron durante el muestreo, en el caso de Ecuador se dejó el filtro de tiempo original desde el primero de octubre de 2023 al 30 de septiembre de 2024. Para Colombia el caso fue diferente, se colocó el mismo filtro de tiempo, pero no existían registros para octubre de 2023, ni para septiembre de 2024. Para México, se colocó el filtro, pero no había registros para septiembre y octubre de 2024. Finalmente, para Brasil, se mantuvo el filtro indicado en la tabla 1, es decir que se incluyó datos de un mes extra. Esto será muy evidente en los anexos 3 y 4. En el caso de la tendencia de las deleciones a lo largo del tiempo, nos limitamos a trabajar con los meses desde noviembre de 2023 hasta agosto de 2024, para todos los países. Pero en el caso de linajes a lo largo del tiempo, se conservaron los meses originales con los que se realizó el muestreo.

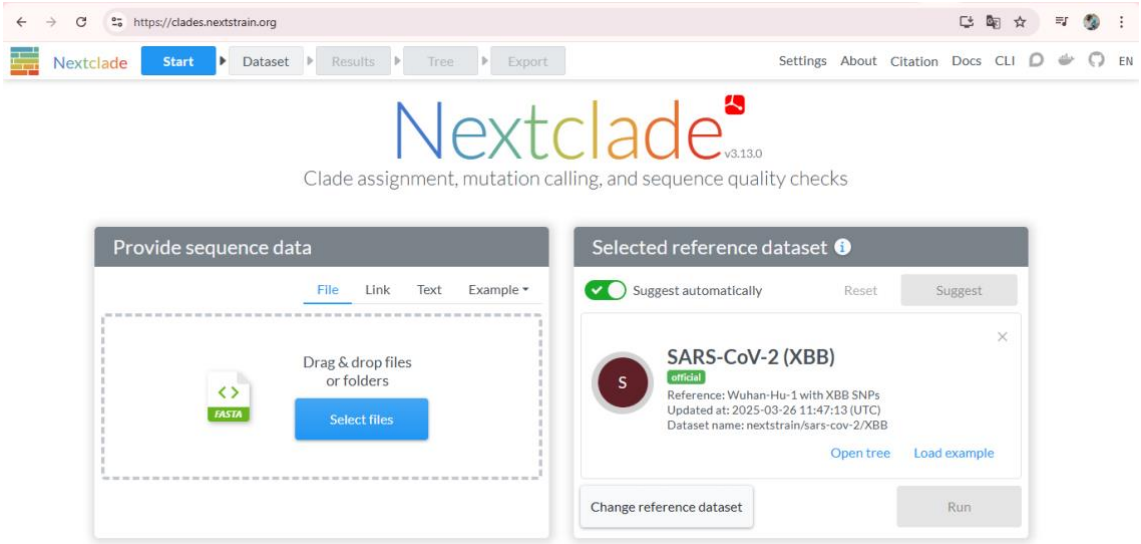
6.1.1.2 *Descarga de las secuencias en un solo archivo FASTA*

Una vez que se concluye la búsqueda con los filtros indicados anteriormente, se selecciona todas las secuencias a la vez y se los procede a descargar. Se generará un documento comprimido que contiene un documento FASTA con todas las secuencias de los virus y un archivo TSV con metadata de estos, en donde se nos indicará información como la fecha de recolección de la muestra, país, región de recolección, paciente, etc.

6.1.1.3 *Trabajando con la plataforma NextStrain*

Para realizar el análisis secundario de las secuencias (es decir alinear la secuencia para identificar variantes y asignar clados), se utilizó la plataforma NextStrain de Nextclade.

Imagen 2. Ventana Plataforma NextClade para análisis secundario



Descripción: En la imagen se observa la ventana de NextClade para el análisis secundario de SARS CoV-2, en la izquierda se observa el botón para la carga de las secuencias compiladas en un mismo documento FASTA. A la derecha se observa la selección de la secuencia de referencia, en este caso la secuencia SARS CoV-2 de Wuhan.

Se trabajó con el genoma de referencia de uno de los primeros virus del SARS CoV-2 aislado de Wuhan. Posteriormente, se cargó las secuencias de cada país, analizándolos por separados, ya que, al tratar de cargar cerca de 2000 secuencias, la plataforma se caía, esto porque al ser una plataforma anclada en la nube, utilizará los recursos computaciones de la PC con la que se accede y estos eran limitados. Razón por la cual tampoco se recurrió al CLI de NextClade (Aksamentov, Roemer, Hodcroft, & Neher, 2021).

6.1.1.4 *Alineación con el genoma de referencia*

La plataforma utiliza un algoritmo optimizado de Smith-Waterman, que realizará una alineación en parejas, donde se alineará la secuencia de interés (*query*) con la referencia o raíz (*en este caso la secuencia Wuhan*), utilizando un "alineamiento local con banda" que hace que el algoritmo sea más rápido. Revisándolo por partes, este algoritmo comenzará alineando dos secuencias (query y raíz), más adelante buscará fragmentos idénticos entre ambas secuencias, estos se los conocerá como *seads* (semillas), y a partir de ellos se identificará las regiones más grandes del genoma por las cuales se puede comenzar a comparar (bandas) (Neher, y otros, 2025).

Para identificar los seed utilizará el FM-INDEX una estrategia para hacerlo de manera más sensible ya que ignorará la tercera letra, esto debido a que muchas mutaciones silenciosas sin impacto en la proteína ocurren como consecuencia de un cambio en la tercera letra, de esta manera se identificará secuencias similares, aunque existan mutaciones genéticas. Entonces comenzará con la alineación dirigida a las regiones "banda" del genoma, aquellas dónde hay una mayor probabilidad de que las secuencias coincidan y con ello de los sitios en los que ocurre los cambio, en caso de que no lo consiga, ajusta los parámetros y vuelve a intentar. De la comparación del query y el root, NextClade identificará SNPs, sustituciones, deleciones, inserciones y sí está anotado en el genoma, cambios en las proteínas. Y en algunas ocasiones cuando no se tiene claro dónde ocurrió la mutación o dónde se ubicará el gap, el algoritmo buscará ubicarlo en dónde tenga más sentido biológico (Neher, y otros, 2025).

Como resultado, se obtiene un documento CSV con asignación de clado, llamada de variantes (sustituciones, deleciones e inserciones de nucleótidos), chequeo de calidad de las secuencias, cambios en aminoácidos (sustituciones, deleciones e inserciones de aminoácidos); y otros relevantes para construir árboles como la alineación de las secuencias. También en la plataforma de NextClade se puede observar un árbol filogenético que ha sido construido a partir de un árbol base con el que cuenta NextStrain y que contiene miles de secuencias globales. Para hacer la alineación, en este caso, no solo lo comparará con la secuencia original de Wuhan, sino que también lo compara con el ancestro común más próximo, esto permite entender mejor la historia evolutiva de los linajes que estamos estudiando (Bedford, y otros, 2025).

6.1.1.5 Trabajando con los llamados de variantes

Una vez que se obtuvo los CSV del análisis secundario de todos los países, se reservó solo los datos que tenían un *qc overall*: *Good,* este filtro indica que la secuencia:

- No presenta muchas Ns (es decir nucleótidos ambiguos)
- La secuencia query se alineó de manera satisfactoria a la referencia
- No presenta una cantidad inusual de mutaciones (lo que podría ser indicativo de errores de secuenciación).
- No presenta inserciones o deleciones grandes o poco comunes.

Tomado de: (Aksamentov, y otros, 2025).

En total, tras este análisis las secuencias que se recuperaron de cada país fueron:

Tabla 2. Secuencias totales que pasaron el filtro de QC

Ubicación	Ecuador	Colombia	Brasil	México
Secuencias recuperadas de GISAID	831	313	1155	1039
Secuencias que pasaron el QC	582	303	937	811

Descripción: En la tabla se registra el total de secuencias que se recuperaron y que pasaron el filtro de calidad después del análisis secundario con la plataforma NextClade.

De manera resumida, es una secuencia confiable, a esto hay que añadir que solo nos quedamos con los índices que se generaron al comparar la secuencia query con la raíz o secuencia de referencia Wuhan, no con los índices que también se nos presentaba al comparar la secuencia query con el ancestro común más próximo (Aksamentov, y otros,

2025). Como resultado, de las cerca de 500 columnas del CSV generadas del análisis secundario, solo se preservaron las siguientes:

- Index
- Date (esta se incluyó del metadata)
- SeqName
- Clade WHO
- Clade NextStrain
- Qc. Overall
- Total deleteions
- Total insertions
- Total frame
- Coverage
- Deletiones (Nucleótidos)
- Insertions (Nucleótidos)
- FrameShift

6.1.1.6 Analizando las deleciones e inserciones

Una vez que nos quedamos con las secuencias de interés, se revisó a las inserciones y deleciones que se llamaron en las secuencias en cada periodo de tiempo, por meses. Se identificó sesenta y siete tipos de deleciones y doce tipos de inserciones entre todas las secuencias de los cuatro países. Más adelante, se procedió a clasificarlas según los sitios dónde ocurrían dentro del genoma de SARS CoV-2, esto con el objetivo de identificar su potencial efecto dentro de la transcripción o replicación del virus (para revisar en mayor detalle ir al Anexo 1 y 2). Aquí, cabe mencionar que se utilizó el genoma del SARS CoV-2 Wuhan-Hu-1 (NCBI Reference: NC_045512.2), para poder construir las tablas y ubicar los eventos de deleción.

Después se tomaron decisiones con relación al manejo de información, para realizar análisis descriptivos sobre la tendencia y ocurrencia de estos eventos de INDELS nos centramos en aquellos que tenían una incidencia mayor, no fue una decisión arbitraria sino más bien utilitaria para más adelante revisar bibliografía sobre reportes de estos indels y efectos en el fitness (se detallará más adelante). Entonces de las 67 deleciones identificadas, se conservó 10; y de las 12 inserciones se conservó 2. En ambos casos se procedió a colocarles un "alias", que está conformado por una letra del alfabeto y la región genómica en la que ocurren para de manera más sencilla poder identificarlos.

Por otro lado, también al indagar en profundidad sobre estos eventos, se comprendió que no solo es importante analizar el evento de manera individual, sino en la perspectiva de la región genómica donde estos ocurren, dado que existirán regiones que por la proteína que generan o por la función que desempeñan en la regulación genética del virus tendrán una mayor o menor tolerancia a mutaciones, deleciones e inserciones. Es aquí donde sí recurrimos a la totalidad de los eventos ocurridos en cada región, para realizar un cálculo de frecuencias y una prueba estadística que nos permita identificar efectivamente sí existe una relación entre la ocurrencia y el sitio.

6.1.1.7 Análisis terciario de estas variantes

A continuación, después de recuperar las 10 deleciones y las 2 inserciones, y revisar sus reportes en internet sobre un efecto en el fitness, el proyecto tomó dos perspectivas diferentes. Sí bien ambas perspectivas partieron de buscar reportes en el internet de

eventos donde se ha reportado que dichas deleciones/inserciones afectan el fitness del virus, ya sea influyendo es la transmisibilidad, gravedad en la enfermedad, eficiencia en los tratamientos disponibles o respuesta inmune. Para las deleciones, sí había reportes de estos eventos de deleción ocurriendo en los nucleótidos, e incluso de su presencia en VOCs. Pero en el caso de las inserciones de nucleótidos, fue más complejo, pues por el momento no hay reportes de eventos nucleotídicos de inserción puntual con efecto en el fitness, pero sí de inserciones de aminoácidos causados por múltiples eventos de inserción de nucleótidos en un región determinada del genoma que sí presentaban efectos en el fitness, es decir que estos eventos presentaban relevancia cuando ocurren en una región. Esto se analiza más adelante en la discusión.

Posteriormente, y a partir de la perspectiva de que estos eventos de deleción podrían ocurrir o heredarse en conjunto, se procedió a analizar las tendencia de los eventos a lo largo del tiempo (Anexo 3), y observar tendencias. En este caso de observa que los eventos de deleción F, G, T y M presentan una tendencia similar, lo que sería un indicativo de que hay una alta probabilidad de que al heredarse en conjunto ofrezcan ventajas adaptativas al virus, ya que han prevalecido a lo largo del tiempo, y a pesar de las transiciones entre linajes algo que también se observa en el Anexo 4, dónde se evidencia que en la mayor parte de casos los linajes del 2023 (caracterizados por el prefijo 23) fueron reemplazados en su totalidad por los linajes del 2024, particularmente 24A. También, tomando en cuenta solo la información de las secuencias de Ecuador se procedió a analizar la ocurrencia en conjunto de las deleciones como se observa en el gráfico 3 con el mapa de calor. Aquí, también se identifica que con una frecuencia mayor a 300 secuencias los eventos de deleción mencionados anteriormente (F, G, T y M) ocurren en conjunto.

Otra tendencia que se observa es entre las deleciones Q, P, V y K; que al comparar con el mapa de calor tiene una frecuencia menor a 300, y que tenderán a ocurrir en gran frecuencia con F, G, T y M; pero no entre ellas; algo que podría indicar que su rol en el fitness no es claro.

Finalmente, I y S no presentan una tendencia clara de ocurrencia, al igual que su ocurrencia en conjunto con las otras deleciones es de menor a 100 eventos. En el caso de I, esta surge en 2024, lo que es muy interesante para un futuro, ya que puede ser una deleción que surgió al azar pero que sí presenta alguna ventaja se podría llegar a fijarse en la población. En el caso de S, su comportamiento un poco más errático con subidas y bajadas podría indicar que es neutral a la evolución.

6.1.1.8 Análisis de linajes

A la par que se recuperaba la información sobre las inserciones/deleciones en cada periodo de tiempo para cada país, se procedió a registrar la información de los linajes reportados (en nomenclatura NextStrain), para más adelante también poder realizar un análisis en tendencia de cómo estos iban cambiando a lo largo del tiempo. Finalmente, para poder entender cómo ocurría la transición entre linajes se realizó un árbol filogenético utilizando NextClade, ya que buscábamos una aproximación mejor, tomando en cuenta que este cuenta con un árbol base que trabaja con secuencias a nivel mundial.

6.1.1.9 Agrupación por regiones genómicas

Más adelante, se tomo un enfoque final, el de agrupar los eventos de deleción e inserción por regiones genómicas, para calcular un índice de frecuencia. Para ello se calculó lo siguiente: longitud, eventos totales y frecuencia.

$$Longitud = |Coordenada\ superior - coordenada\ inferior| + 1$$

Se suma 1 considerando que en teoría se cuenta nucleótido por nucleótido para determinar la longitud.

$$Eventos\ totales = (suma\ de\ los\ eventos\ reportados\ en\ la\ región\ de\ todos\ los\ países)$$

Para calcular la frecuencia utilizamos la siguiente fórmula sencilla:

$$Frecuencia = \frac{Eventos\ totales}{Longitud}$$

Esto nos permite obtener la frecuencia con la que ocurrirá un evento de deleción o inserción por región genómica, considerándose frecuencias altas, aquellas que son mayores a 1. En el caso de frecuencias, el índice fue alto para deleción S, N y 3UTR. Mientras que, para inserciones, sería S.

6.1.1.10 Análisis estadístico Kruskal Wallis

Se realizó el análisis no paramétrico estadístico Kruskal Wallis, ya que los datos con los que estamos trabajando (Tabla 5 y 6), no cumplen con los supuestos de ANOVA, independencia, homocedasticidad y distribución normal (para las frecuencias). También que permitirá comparar las medianas de las observaciones dentro de cada grupo independiente, en este caso de las regiones del genoma. Para de esta manera determinar sí existe una diferencia estadísticamente significativa entre las frecuencias de deleción para cada una de estas regiones (Amat, 2016).

6.1.1.11 Árbol filogenético

Para comprender cómo ocurría la evolución de linajes de SARS CoV-2 a lo largo del tiempo dentro de los cuatro países de interés. Se tomo al azar 250 linajes de Colombia, 250 de Brasil y 250 de México. Junto con todas las secuencias de Ecuador. Y se trabajó con NextStrain, porque su árbol base que cuenta con información global y de amplios periodos de tiempo que la hace ideal para realizar aproximaciones.

7 Resultados

Se analizó la información secundaria generada a partir de NextClade de las 2633 secuencias de SARS CoV-2 de los cuatro países en estudio, identificándose sesenta y siete eventos de deleción, estos se redujeron a 10 que hemos considerado relevantes por la alta frecuencia que presentan dentro de las poblaciones en análisis, y se los escogió para analizar a profundidad. Cabe mencionar que se les colocó alias que permiten una identificación más sencilla.

Tabla 3. Eventos relevantes de deleción nucleotídica por región genómica

Deleciones								
Región	**Coordenadas**		**Ubicación**	**Alias**	**Secuencias**			
					EC	**MX**	**BZ**	**CO**
5UTR	1	265						
ORF1a	266	13,468	11288-11296	F-ORF1A	582	811	937	293
ORF1b	13,468	21,555						
S	21,563	25,384	21633-21641	G-S	519	771	889	292
			21653-21655	I-S	22	100	84	41
			21765-21770	K-S	288	708	581	151
			21992-21994	M-S	579	768	921	291
			22194-22196	P-S	282	737	592	173
			23009-23011	Q-S	228	699	537	160
ORF3a	25,393	26,220						
E	26,245	26,472						
M	26,523	27,191						
ORF6	27,202	27,387						
ORF7a	27,394	27,759						
ORF7b	27,756	27,887						
ORF9	28,284	28,577	28254	S-ORF9	24	15	33	7
N	28,274	29,533	28362-28370	T-N	580	797	936	263
ORF10	29,558	29,674						
3UTR	29,675	29,903	29734-29759	V-3UTR	545	731	793	216

Descripción: En la tabla se puede apreciar las diez deleciones más frecuentes que superan las 20 secuencias por lo menos en dos países, junto con un alias que está compuesto por una letra del alfabeto y el gen dónde ocurren. Estas 10 deleciones se evaluarán bibliográficamente, y evidenciar sí efectivamente existen reportes sobre impacto en el fitness viral de SARS CoV-2.

En el caso de los eventos de inserción, se realizó lo mismo, quedándonos con aquellos eventos que presentan una alta frecuencia dentro de las poblaciones de estudio.

Tabla 4. Eventos relevantes de inserción nucleotídica por región genómica

	Inserciones								
Región	**Coordenadas**		**Ubicación**	**Longitud**	**Alias**	**Secuencias**			
						EC	**MX**	**BZ**	**CO**
5UTR	1	265							
ORF1a	266	13,468							
ORF1b	13,468	21,555							
S	21,563	25,384	21,608	12 bp	Ains-S	185	662	558	93
ORF3a	25,393	26,220							
E	26,245	26,472							
M	26,523	27,191							
ORF6	27,202	27,387							
ORF7a	27,394	27,759							
ORF7b	27,756	27,887							

ORF8	27,894	28,259							
ORF9b	28,284	28,577							
N	28,274	29,533							
3UTR	29,675	29,903							

Descripción: En la tabla se puede apreciar la única inserción relevante, ya que tiene una frecuencia elevada para los cuatro países.

Para que la información sea más digerible se procedió a calcular un índice de frecuencia de deleciones e inserciones para las regiones genómicas.

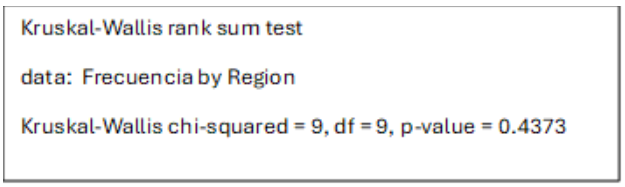
Tabla 5. Frecuencia deleciones por región

Deleciones					
Región	**Coordenadas**		**Longitud**	**Eventos totales**	**Frecuencia**
5UTR	1	265	265.00	0.00	0.0000
ORF1a	266	13,468	13,203.00	2,682.00	0.2031
S	21,563	25,384	3,822.00	10,446.00	2.7331
ORF3a	25,393	26,220	828.00	2.00	0.0024
E	26,245	26,472	228.00	0.00	0.0000
Genómico	26,473	26,522	50.00	2.00	0.0400
M	26,523	27,191	669.00	0.00	0.0000
ORF6	27,202	27,387	186.00	3.00	0.0161
ORF7a	27,394	27,759	366.00	4.00	0.0109
ORF7b	27,756	27,887	132.00	5.00	0.0379
ORF9	28,284	28,577	294.00	85.00	0.2322
N	28,274	29,533	1,260.00	2,580.00	2.0476
ORF10	29,558	29,674	117.00	0.00	0.0000
3UTR	29,675	29,903	229.00	2,295.00	10.0218

Descripción: La tabla presenta los índices de frecuencia de un evento de deleción por región genómica, lo que observamos es que existen solo 3 regiones que presentan una frecuencia alta S, N y UTR.

Para realizar el análisis de Kruskal Wallis, no se tomó en cuenta los valores de cero, es decir aquellos en los que no ocurrieron eventos de deleción, por dos motivos, primero porque hay una explicación biológica detrás de esto, y porque estos valores extremos pueden afectar el poder estadístico de la prueba.

Imagen 3. Resultado Análisis Kruskal-Wallis para frecuencia de deleciones



Descripción: En la imagen se observa el resultado del análisis de la prueba estadística no paramétrica Kruskal Wallis, ejecutada en R. En este caso no existe suficiente evidencia estadísticas para aceptar la hipótesis nula ya que el valor de p es mayor a 0.05. Por tanto, la prueba nos está indicando que no existe un diferencia estadísticamente significativa entre la frecuencia para las 10 regiones genómicas analizada. No obstante, el valor p alto podría ser un indicador de que la prueba

ha perdido poder estadístico debido a los valores tan extremos que presentamos para algunas regiones (Ej. ORF3a) que poseen pocos eventos de deleción, frente a otros como S que acumulan miles. Por tanto, se va a buscar el efecto biológico tras de estas observaciones.

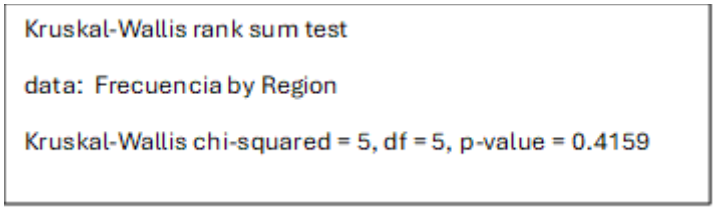
Tal y como se hizo con la frecuencia por regiones de deleción, para frecuencia por regiones de inserción, no se consideró valores outliners de cero, para poder realizar el cálculo de Kruskal- Wallis, quedándonos solo con seis valores para el análisis.

Tabla 6. Frecuencia inserciones por región

Inserciones					
Región	**Coordenadas**		**Longitud**	**Eventos Totales**	**Frecuencia**
5UTR	1	265	265	0.00	0.0000
ORF1a	266	13,468	13,203	0.00	0.0000
ORF1b	13,468	21,555	3,822	1.00	0.0003
S	21,563	25,384	828	1,525.00	1.8418
ORF3a	25,393	26,220	228	1.00	0.0044
E	26,245	26,472	50	0.00	0.0000
Genómico	26,473	26,522	669	0.00	0.0000
M	26,523	27,191	186	0.00	0.0000
ORF6	27,202	27,387	366	1.00	0.0027
ORF7a	27,394	27,759	132	0.00	0.0000
ORF7b	27,756	27,887	366	0.00	0.0000
ORF9	27,894	28,259	294	12.00	0.0095
N	28,274	29,533	117	0.00	0.0000
ORF10	29,558	29,674	117.00	0.00	0.0000
3UTR	29,675	29,903	229	2.00	0.0087

Descripción: La tabla presenta los índices de frecuencia de eventos de inserción por región genómica, lo que observamos es que existen solo 1 región que presentan una frecuencia alta S.

Imagen 4. Resultado Análisis Kruskal-Wallis para frecuencia de inserciones



Descripción: En la imagen se observa el resultado del análisis de la prueba estadística no paramétrica Kruskal Wallis, ejecutada en R. Ya que el valor de p es mayor a 0.05 no se puede aceptar la hipótesis de que existe una diferencia estadísticamente significativa entre la frecuencia de deleción para las 6 regiones genómicas. No obstante, este valor tan alto podría ser un indicador de que la prueba ha perdido poder estadística debido a los valores tan extremos que presentamos para algunas regiones (Ej. ORF9) que presentan eventos bajos de deleción, frente a otros como S que acumulan miles. Por tanto, se va a buscar el efecto biológico tras de estas observaciones.

También, para tener una representación gráfica de estas frecuencias de inserción y deleción se realizó un gráfico de barras, que nos permite visualizar de manera sencilla las regiones genómicas mencionadas anteriormente, y que registran altas frecuencias.

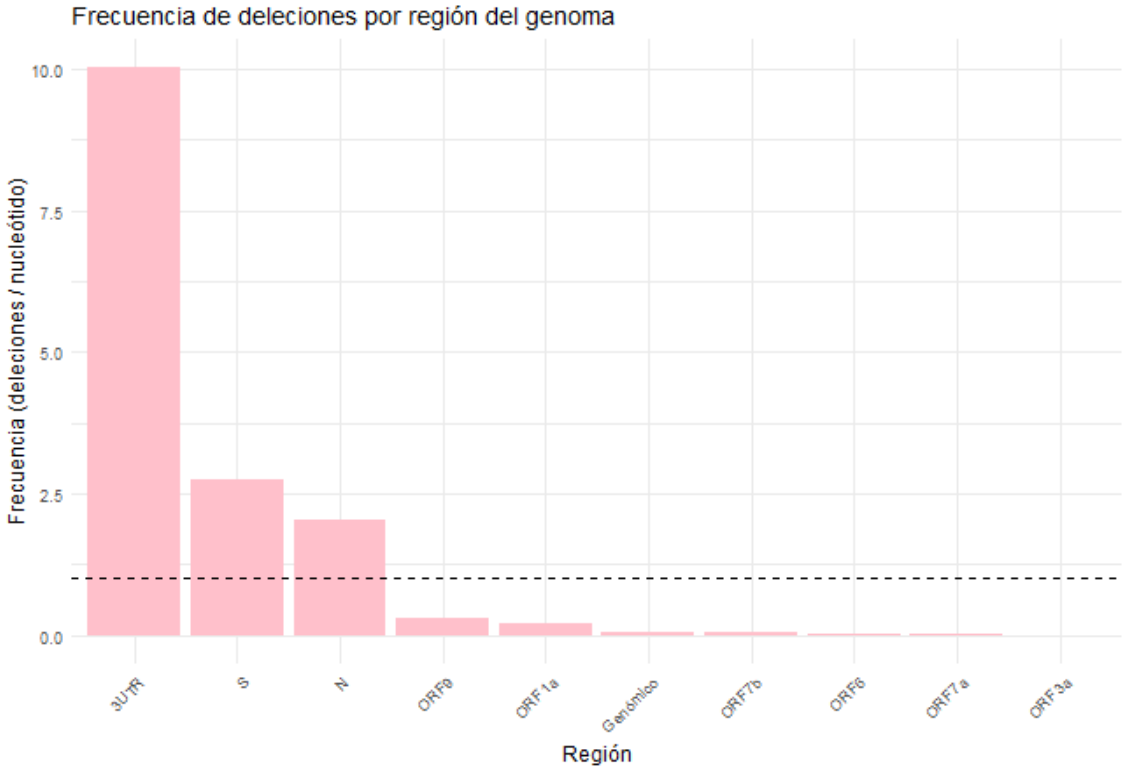


Figura 1. Frecuencia de deleciones por región: En el gráfico se aprecia las regiones genómicas que presentan una mayor frecuencia de deleción: 3UTR, gen S y gen N; estos presentan un índice de frecuencia mayor a 1, aquí representado con la línea punteada. En la discusión se va a buscar la explicación evolutiva detrás de esto.

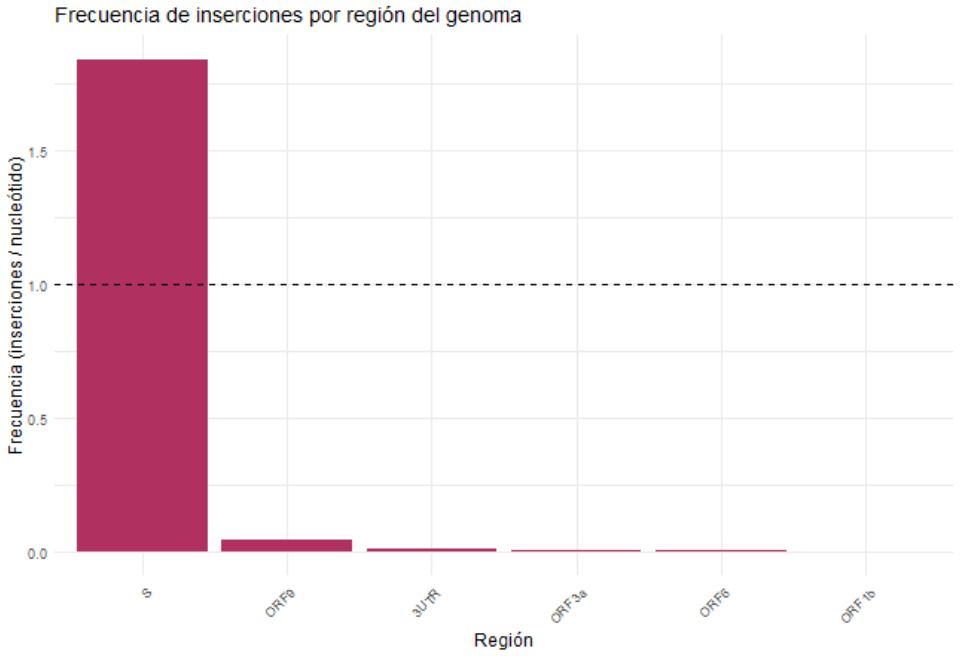


Figura 2. Frecuencia de inserciones por región: En el gráfico se aprecia la región genómica que presenta una mayor frecuencia de inserción: gen S. Este presenta un índice de frecuencia mayor a 1, aquí representado con la línea punteada. En la discusión se va a buscar la explicación evolutiva detrás de esto.

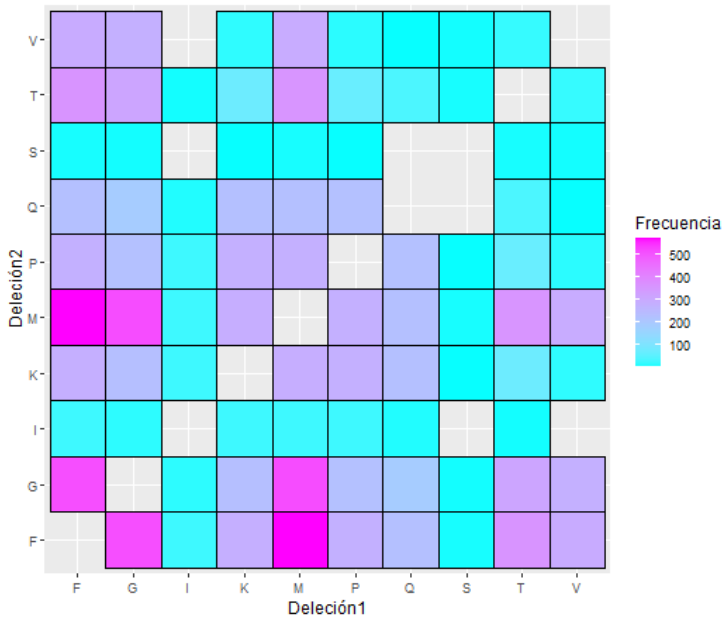


Figura 3. Mapa de calor ocurrencia de deleciones. En este gráfico se observa la ocurrencia a la deleciones, algo que se puede interpretar las veces que estas deleciones se heredaron en conjunto. Se puede observar que F, G, M y T; ocurren en conjunto con una frecuencia mayor a 400 secuencias. Mientras que Q, P, V y K si bien ocurren muy frecuentemente con F, G y M (aproximadamente 300 secuencias) no con T (menor o igual a 100 secuencias), ni entre ellas. Finalmente, I y S presentan una frecuencia baja menor a 100 secuencias, pero se las menciona por su presencia alta.

8 Discusión

Tras revisar los resultados (que también se extienden a los anexos), tenemos los siguientes hallazgos, que se discutirán a continuación. Primero, tras revisar los índices de frecuencia de deleción e inserción se logró identificar regiones genómicas que presentan una alta tolerancia a los eventos de deleción e inserción, mientras que otras una baja tolerancia y otras nula. Segunda, que de los diez eventos de deleción que se consideraron relevantes existen algunos que parecen heredarse en conjunto, y otros cuyo comportamiento es más complicado de comprender. A continuación, revisaremos las explicaciones biológicas detrás de estos hallazgos, ya que las prueba estadística utilizada no fue informativa.

Para comenzar, analizaremos la relación entre las regiones genómicas y la tolerancia a eventos de inserción y deleción, como se observa en el gráfico 1 y 2; las regiones mayor tolerancia a las deleciones son 3UTR (presenta en índice más alto), gen S y gen N; y tolerancia a inserciones el gen S. Mientras que las que presentan intolerancia a deleciones son 5UTR, gen E, gen M y ORF10 (Tabla 5). En el caso de las inserciones: 5UTR, ORF1A, gen E, gen M, región genómica, ORF7a/b, ORF10 y gen N.

Comenzaremos, analizando el rol de cuatro regiones que se repiten en ambos casos 5UTR, gen M, gen E y ORF10; y que parecen tener una tolerancia cero a eventos de deleción e inserción. En el caso del 5UTR, no lo podemos analizar sin su “contraparte” el 3UTR que en este caso presenta un comportamiento totalmente opuesto, con una alta tolerancia a las deleciones. Ambos no generarán productos proteicos, pues el rol que desempeñan dentro del virus es de regulación de la traducción y patogénesis. En el caso del 5UTR, está altamente conservado dentro de la familia de los Betacoronaviridae, ya que regula la transcripción de la poliproteína del ORF1a/b, fundamental para el ciclo viral. Por otro lado,

el 3UTR generará estructuras secundarias de tipo tallo y bucle que desempeñan roles importantes en la transcripción y replicación. Si bien a grandes rasgos parecería imposible que en esta región ocurran deleciones, hay estudios que identifican una región hipervariable en este UTR (común en los betacoronaviridae) (Verma, y otros, 2021). Además, que se ha identificado de una región s2m (Steam loop II motif) que, pese a que se la mutó o varió, la funcionalidad del 3UTR no se vio afectado (Jiang, y otros, 2023). Por lo que podría decirse que esta región tiene alta tolerancia a eventos de deleción ya que estos al ocurrir en sitios clave no lo afectarán. Sí tiene una ventaja evolutiva, la deleción en este sitio, se analizará más adelante.

Continuaremos con el gen E y M; la baja tolerancia de ambas se debe a que codificarán para proteínas estructurales fundamentales para el ciclo del virión del SARS CoV-2, ya que inducirán la formación de la membrana que se requiere para liberarlas y ensamblarlas; es decir que una deleción que llegue a afectar estas proteínas y por tanto este proceso tan delicado en el ciclo del virión, es mortal, por esto experimentan una presión evolutiva tan alta (Zhou, y otros, 2023).

Por otro lado, en el caso del ORF10, este codifica para una proteína que contiene 38 aminoácidos, cuya función no es del todo clara, pero que se creía desempeñaba un rol importante durante la infección primaria. En un estudio ejecutado por Li y otros en el 2022 efectivamente se identificó que induce la mitofagia lo que lleva a la degradación del MAVS (señalización antiviral mitocondrial) bloqueando la respuesta del interferón 1, y con ello inhibiendo la respuesta primaria del sistema inmune del hospedero. Por lo que una falla en esta proteína implicaría que el virus no realice una infección ni replicación exitosa (Li, y otros, 2022).

En el caso de las otras regiones genómicas, que poseen eventos ocasionales de deleción o inserción como el ORF6, ORF9 y ORF3a, estos eventos podrían ser neutrales a la evolución, es decir que ocurrieron al azar y que al no presentar un efecto ni positivo ni negativo para el virus no se fijaron pero que eventualmente se eliminarán (Jeronimo, Aksenen, Duarte, Lins , & Miyajima, 2024).

En el caso de las regiones con alta frecuencia de deleciones el 3UTR ya se revisó que ocurre, a continuación, se revisará lo que pasó en el gen S y gen N. El gen S, codifica para la proteína estructural spike, la más importante para el virus, ya que es la que permite la entrada hacia las células para que comience la infección. En el caso de los humanos, se une al receptor ACE2, para que se fusionen las membrana. Las inserciones/deleciones en este gen, aumentarán la transmisibilidad y la evasión al sistema inmune, un rasgo que ha sido muy característico de los linajes de Ómicron, entonces las deleciones alias G, K, M, P y Q; están relacionados con una mejora en el fitness, lo que explica porque su persistencia heredándose en conjunto como lo muestra el gráfico 3.

Continuando el hilo con este hallazgo mutaciones que parecen heredarse en conjunto, se revisó información que respalde esta presunción, y se encontró un artículo del 2023 que relata que existen siete deleciones que ocurren muy comúnmente en los linajes de Ómicron. Estas son 11,288/11,296 (F-ORF1a); 21,633/21,641 (G-S); 21,765/21,770 (K-S): 21,992/21,994 (M-S); 22,194/22,196 (P-S); 28,362/28,370 (T-N) y 29,734/29,759 (V-3UTR) (Akaishi & Fujiwara, 2023). También se procedió a analizar el comportamiento de las deleciones a lo largo del tiempo (Anexo 3.) para identificar alguna tendencia particular, y efectivamente se observa que F, G, T y M comparten tendencia en los cuatro países.

F es la deleción con un origen más antiguo aproximadamente 2020, se ha fijado desde las VOCs Gamma y Alpha. En el otro lado, tenemos a G que es una deleción de un origen más reciente (data del 2022) y que solo estará presente en los linajes derivados de Ómicron (Akaishi & Fujiwara, 2023). Entonces, podríamos decir que F y G efectivamente presentan una ventaja evolutiva para el virus. Para ello se revisó la deleción que ocasionarán dentro de la secuencia de aminoácidos F (del ORF1a:S3675:G3676:F3677) y G (del S:ΔH69/V70).

Por otro lado, en el caso de T este se originó junto con Ómicron, manteniéndose a lo largo de los linajes. También se revisó el cambio en aminoácidos que generaría esta deleción, este sería del N: E31:R32:S33. Y en el caso de M del S:N211. Hay un artículo en el que se reporta que estos eventos de deleción ocurren comúnmente en los linajes de Ómicron conocidos como JN.1 (Nomenclatura Pango) y que comprenden la mayor parte de linajes 23 y 24 en nomenclatura NextStrain (Kumar, 2024). Y que también ocurrirán junto con la inserción que identificamos como relevante en S, ya que esta genera un producto, la ins16MPLF, que funciona muy bien en conjunto, ya que compensará los efectos negativos que puedan tener las deleciones (Meng, y otros, 2021).

También, durante los primeros estudios que se hizo de esta inserción presente en el sublinaje de Omicrón, BA.2.86 conocida como pirola, se consideró que podría desempeñar un rol importante en la evasión inmunitaria (Rothstein, y otros, 2023). Además, que hay reportes de que esta inserción (incluso en la secuencia de aminoácidos) no es identificada por los pipelines bioinformáticos disponibles (WHO, 2024).

Por otro lado, las cuatro deleciones que presentan un patrón divergente de frecuencia, y que se alinea con lo que observamos con los linajes y su permanencia son: K, P, Q y V. En el caso de la deleción K, que, si bien ocurrirá con F, M y G, no lo hará con V y T, aquí es importante mencionar que esta guarda un origen evolutivo con alfa, es decir que esta deleción estará presente en los sublinajes de Ómicron identificados como 23A y 23F que en el caso de Ecuador (datos a los que corresponde el heatplot) presentó un pico en los meses de octubre a febrero de 2023 (un comportamiento que también se observa en los otros países). Este declive nos lleva a pensar que V si bien presentó en su momento una ventaja evolutiva, fue relegada por otros eventos. M se originó a la par con K, no obstante, esta muestra persistencia, lo que nos lleva a pensar que en el ambiente actual M, tiene una mayor ventaja evolutiva que K. En relación a V y T, estas se originaron con el clado Ómicron, pero su baja incidencia, nos lleva a pensar que su rol en el fitness todavía no es claro, lo mismo para Q (Akaishi & Fujiwara, 2023). Finalmente, I y S que son las deleciones con un comportamiento más atípico, no tienen reportes en literatura sobre impacto en el fitness, porque en base a su comportamiento, en el Anexo 3, se podría intuir que I es muy reciente por lo que falta datos para poder intuir su influencia en el fitness; mientras que S, al tener una tendencia más oscilante, se podría decir que su efecto en el fitness es neutro.

Algo que también llamó nuestra atención es la mayor prevalencia de eventos de deleción sobre eventos de inserción, y algunos investigadores sugieren que esto responde a una tendencia que se ha observado en el SARS CoV-2 y otros Betacoronavirus, de reducir el tamaño de su genoma para disminuir el gasto energético que representa la retro transcripción de estos elementos, algo común en genes codificantes para proteínas no estructurales como: ORF3, ORF6, ORF7a, ORF7b, ORF8a, y ORF8b (Jeronimo, Aksenen, Duarte, Lins , & Miyajima, 2024).

9 Conclusiones y recomendaciones

Se concluye que las regiones que presentan alta tolerancia a las deleciones/inserciones gen N y gen S, brindarán una mayor plasticidad al virus, y le permitirán adaptarse a ambientes altamente variables (como los cuatro países estudiados) pues al codificar proteínas que se unen a la superficie de la célula humana le permitirán al virus adaptarse a cambios generados por eventos como las vacunas, o la inmunidad adquirida. En el caso de 3UTR, este tiene una región altamente variable y otra que tolerará la ocurrencia de este tipo de deleciones y que por el momento se podría considerar como probablemente con impacto positivo en el fitness.

En cuanto a las deleciones de manera individual, la deleciones alias F, G, T y M junto con la inserción A, han demostrado que tienden a heredarse juntos dentro de los linajes de Ómicron, y en conjunto se ha evidenciado que tienen un impacto positivo en el fitness del virus, razón por la cual han permanecido desde los linajes 23A/23F/23G hasta los más recientes 24A. Por otro, lado el rol de las deleciones I y S, no es claro, para I al ser muy reciente es difícil de determinar, y en el caso de S su tendencia divergente, nos lleva a creer que tiene un impacto neutro dentro de la evolución. Por otro lado, para las deleciones alias K, V, P y Q, parece que también se heredan en conjunto, pero al tener una frecuencia menor es difícil de comprender su impacto en el fitness del virus.

Referente al uso de NextClade para realizar análisis secundario, se concluye que su uso es altamente recomendado, pues es una herramienta con amplias funcionalidades que pueden facilitar estudios con relación al virus del SARS CoV-2 y su diversidad genética. También, al ser de código abierto y trabajar con la base de datos global de GISAID, nos permiten tener mejores aproximaciones en cuanto al comportamiento global del virus. Se recomienda que para análisis más complejos o más específicos, se puede trabaje con la versión CLI de NextClade, la misma que será más poderosa al trabajar con línea de comando.

Para futuros estudios, se recomienda también incluir la información de las deleciones en la secuencia de aminoácidos ya que, trabajar con la deleción en la secuencia de nucleótidos representó un reto muy importante, esto ya que dependiendo del genoma del virus del SARS CoV-2 que se tome de referencia, las coordenadas para las deleciones pueden cambiar y variar muchísimo, lo que nos puede llevar a hallazgos no esperados, o a no encontrar información que pueda ser relevante, pues la mayor parte de investigadores toman la secuencia de aminoácidos (Rothstein, y otros, 2023). Esto tomando en cuenta que su relación con la funcionalidad de la proteína es más fácil de vislumbrar trabajando con la secuencia de aminoácidos. También, y otra recomendación, esta información nos permitirá realizar simulaciones y llevar a predicciones en relación con cómo se verá afectada la funcionalidad de las proteínas por estas deleciones.

Otra recomendación es indagar sobre otras pruebas estadísticas que sean menos sensibles a valores extremos, y que sean más adecuados para este tipo de eventos genéticos. Ya que en este caso Kruskal Wallis, parece a ver sido afectado por los valores extremos.

10 Bibliografía

Acosta, P., Escobar, M., Castillo , X., Flores , H., Lovato, R., Granda , J., . . . Narváez, A. (Enero de 2022). *Lineamiento de Vigilancia Integrada para COVID-19 y otros virus respiratorios*. Obtenido de Minsiterio de Salud Pública: https://www.salud.gob.ec/wp-content/uploads/2022/01/Lineamiento-vigilancia-COVI-19-Enero-2022-_.pdf

Ahmad , A., Fawaz, M., & Aisha, A. (2022). A comparative overview of SARS-CoV-2 and its variants of concern. *Infez Med, 30*(3), 328-343. doi:10.53854/liim-3003-2

Akaishi, T., & Fujiwara, K. (2023). Insertion and deletion mutations preserved in SARS-CoV-2 variants. *Archives of microbiology, 205*(4), 154. Obtenido de https://doi.org/10.1007/s00203-023-03493-0

Aksamentov, I., Bedford, T., Neher, R., Anderson, J., Andrews, K., Chang, J., & Haldfield, J. (2025). *Quality Control* . Obtenido de NextClade: https://docs.nextstrain.org/projects/nextclade/en/stable/user/algorithm/06-quality-control.html

Aksamentov, I., Roemer, C., Hodcroft, E. B., & Neher, R. A. (2021). Nextclade: clade assignment, mutation calling and quality control for viral genomes. *Journal of Open Source Software, 6*(67), 3773. doi:https://doi.org/10.21105/joss.03773

Amat, J. (Enero de 2016). *Test Kruskal-Wallis*. Obtenido de RPubs: https://rpubs.com/joaquin_ar/219504

Bedford, T., Neher, R., Aksamentov, I., Anderson, J., Andrews, K., Chang, J., & Haldfield, J. (2025). *Clade Naming & Definitions*. Obtenido de NextStrain: https://docs.nextstrain.org/projects/ncov/en/latest/reference/naming_clades.html

Cheng, Y., Chao, T., Li, C., Wang, S., Kao, H., Tsai, Y., . . . Yeh, S. (2021). D614G Substitution of SARS-CoV-2 Spike Protein Increases Syncytium Formation and Virus Titer via Enhanced Furin-Mediated Spike Cleavage. *mBio, 12*(4). doi:10.1128/mBio.00587-21

Dellicour, S., Hong, S. L., Vrancken, B., Chaillon, A., Gill, M., Maurano, M., . . . Duerr, R. (2021). Dispersal dynamics of SARS-CoV-2 lineages during the first epidemic wave in New York City. *PLOS Pathogens, 17*(5), e1009571. doi:https://doi.org/10.1371/journal.ppat.1009571

Farkas, C., Mella, A., Turgeon, M., & Haigh , J. (2021). A Novel SARS-CoV-2 Viral Sequence Bioinformatic Pipeline Has Found Genetic Evidence That the Viral 3' Untranslated Region (UTR) Is Evolving and Generating Increased Viral Diversity. *Front Microbiol*(12). Obtenido de https://doi.org/10.3389/fmicb.2021.665041

Fauver, J., Petrone , M., Hodcroft, E., Neher, R., Ko, A., Grubaugh, N., & Vogels, C. (2020). Coast-to-Coast Spread of SARS-CoV-2 during the Early Epidemic in the United States. *Cell, 181*(5), 990-996. Obtenido de https://www.cell.com/cell/fulltext/S0092-8674(20)30484-0?_returnURL=https%3A%2F%2Flinkinghub.elsevier.com%2Fretrieve%2Fpii%2FS0092867420304840%3Fshowall%3Dtrue

Flores, A., Delgado, G., Espinosa-Camacho, L., Rodríguez-Gómez, F., Cruz-Rangel, A., Sander-Miranda, L., . . . Morales-Espinosa, R. (2022). Two Years of Evolutionary Dynamics of SARS-CoV-2 in Mexico, With Emphasis on the Variants of Concern. *Frontiers in microbiology, 13*, 886585. doi:https://doi.org/10.3389/fmicb.2022.886585

Giovanetti, M., Slavov, S., Fonseca, V., Wilkinson, E., Tegally, H., James, E., . . . Demarchi, L. (2022). Genomic epidemiology of the SARS-CoV-2 epidemic in Brazil. *Nature Microbiology, 7*, 1490-1500. Obtenido de https://doi.org/10.1038/s41564-022-01191-z

Herrera, D., Troya, C., & Gaus, D. (17 de November de 2020). COVID-19 in Ecuador: Imported Control Strategies without Context in a Challenged Healthcare System. *Am J Trop Med Hyg., 194*(2), 414-415. doi:https://doi.org/10.4269/ajtmh.20-1347

Jeronimo, M., Aksenen, C., Duarte, I., Lins , R., & Miyajima, F. (2024). Evolutionary deletions within the SARS-CoV-2 genome as signature trends for virus fitness and adaptation. *Journal of virology, 98*(1), e0140423. doi:https://doi.org/10.1128/jvi.01404-23

Jiang, H., Joshi, A., Gan, T., Janowski, A., Fujii, C., Bricker, T., . . . Harastani, H. (2023). The Highly Conserved Stem-Loop II Motif Is Dispensable for SARS-CoV-2. *Journal of Virology, 97*(6), e0063523. doi:10.1128/jvi.00635-23

Konings, F., Perkins, M., Kuhn, J., Pallen, M., Alm, E., Archer, B., . . . Bhiman, J. (2021). SARS-CoV-2 Variants of Interest and Concern naming scheme conducive for global discourse. *Nature Microbiology, 6*, 821-823. doi:https://doi.org/10.1038/s41564-021-00932-w

Kumar, A. (2024). Coronaviruses have reached at Pre-elimination Stage with Nine Amino Acid Spike Deletions and Forty-nine Nucleotide 3'-UTR Deletions. *Int J Clin Virol, 8*(2), 31-44. Obtenido de https://dx.doi.org/10.29328/journal.ijcv.1001060.

LaRotta, J., Escobar , O., Ávila-Aguero, M., Torres , J., Sini de Almeida, R., Morales, G., & Srivastava, A. (2023). COVID-19 in Latin America: A Snapshot in Time and the Road Ahead. *Infectious diseases and therapy, 12*(2), 389-410. Obtenido de https://doi.org/10.1007/s40121-022-00748-z

Li, X., Hou, P., Wenqing, M., Xuefeng, W., Hongmei, W., Zhangping, Y., . . . Tiecheng, W. (2022). SARS-CoV-2 ORF10 suppresses the antiviral innate immune response by degrading MAVS through mitophagy. *Cellular & Molecular Immunology, 19*, 67-78. Obtenido de https://doi.org/10.1038/s41423-021-00807-4

Li, X., Yan, H., Wong , G., & Cui, J. (2023). Identifying featured indels associated with SARS-CoV-2 fitness. *Microbiology spectrum, 11*(5). Obtenido de https://doi.org/10.1128/spectrum.02269-23

Markov, P., Ghafari, M., Beer, M., Lythgoe, K., Simmonds, P., Stilianakis, N., & Katzourakis, A. (2023). The evolution of SARS-CoV-2. *Nat Rev Microbiol*(21), 361-379. Obtenido de https://doi.org/10.1038/s41579-023-00878-2

Meng, B., Kemp, S., Papa, G., Datir, R., Ferreira, I., Marelli, S., . . . Lytras, S. (2021). Recurrent emergence of SARS-CoV-2 spike deletion H69/V70 and its role in the

Alpha variant B.1.1.7. *Cell Reports, 35*(13), 109292. doi:doi: 10.1016/j.celrep.2021.109292. Epub 2021 Jun 8. PMID: 34166617; PMCID: PMC8185188.

Ministerio de Salud Pública. (3 de Julio de 2024). *Ecuador registra dos casos de FLiRT, la nueva variante de COVID*. Obtenido de Ministerio de Salud Pública: https://www.salud.gob.ec/ecuador-registra-dos-casos-de-flirt-la-nueva-variante-de-covid/

Neher, R., Bedford, T., Aksamentov, I., Anderson, J., Amdrews, K., Chang, J., . . . Sibley, T. (2025). *Sequence alignment*. Obtenido de NextClade: https://docs.nextstrain.org/projects/nextclade/en/stable/user/algorithm/01-sequence-alignment.html

O'Toole, Á., Scher, E., Underwood, A., Jackson, B., Hill, V., McCrone, J., . . . Ruis, C. (2021). Assignment of epidemiological lineages in an emerging pandemic using the pangolin tool. *Virus Evolution, 7*(2), veab064. doi:https://doi.org/10.1093/ve/veab064

PAHO. (s.f.). *RESVIGEN (Respiratory Virus Genomic Surveillance Regional Network)*. Obtenido de Pan American Health Organization: https://www.paho.org/en/topics/influenza-sars-cov-2-rsv-and-other-respiratory-viruses/resvigen-respiratory-virus-genomic

Rao, R., Ahsan, N., Xu, C., Su, L., Verburgt, J., Fornelli, L., . . . Xu, D. (2021). Dynamics of Indels in SARS-CoV-2 Spike Glycoprotein. *Evol Bioinform Online, 17*. doi:10.1177/11769343211064616.

Rogozin, I., Saura, A., Poliakov, E., Bykova, A., Roche-Lima, A., Pavlov, Y., & Yurchenko, V. (2024). Properties and Mechanisms of Deletions, Insertions, and Substitutions in the Evolutionary History of SARS-CoV-2. *International Journla of Molecular Science, 25*(7). Obtenido de https://doi.org/10.3390/ijms25073696

Rothstein, A., Qiu, X., Robinson, K., Collins, S., Muir, G., Lu, B., . . . Philipson, C. (2023). Bayesian phylogenetics on globally emerging SARS-CoV-2 variant BA.2.86 suggest global distribution and rapid evolution. *BioRxiv*. doi:https://doi.org/10.1101/2023.09.08.556912

Sierra , B., Paskov, K., Stockham, N., Tabatabel, K., Jung, J., Washington, P., . . . Wall, D. (2021). Indels in SARS-CoV-2 occur at template-switching hotspots. *BioData Mining, 14*(20). doi:https://doi.org/10.1186/s13040-021-00251-0

Tosta, S., Moreno, K., Schuab, G., Fonseca , V., Cardozo, F., Kashima, S., . . . Giovanetti, M. (2023). Global SARS-CoV-2 genomic surveillance: What we have learned (so far). *Infect Genet Evol*(105405). doi:10.1016/j.meegid.2023.105405

Verma, R., Saha, S., Kumar, S., Mani, S., Maiti, T., & Surjit, M. (2021). RNA-Protein Interaction Analysis of SARS-CoV-2 5' and 3' Untranslated Regions Reveals a Role of Lysosome-Associated Membrane Protein-2a during Viral Infection. *mSystems, 6*(4), e0064321. doi:https://doi.org/10.1128/mSystems.00643-21

Wargo, A., & Kurath, G. (2012). Viral fitness: definitions, measurement, and current insights. *Curr Opin Virol, 2*(5), 538 -545. doi:10.1016/j.coviro.2012.07.007

WHO. (26 de April de 2024). *Statement on the antigen composition of COVID-19 vaccines*. Obtenido de WHO: https://www.who.int/news/item/26-04-2024-statement-on-the-antigen-composition-of-covid-19-vaccines

Zhou, S., Lv, P., Li, M., Xin H, Chen, Z., Reilly, S., & Zhang, X. (2023). SARS-CoV-2 E protein: Pathogenesis and potential therapeutic development. *Biomedicine & pharmacotherapy*. Obtenido de https://doi.org/10.1016/j.biopha.2023.114242

11 Anexo 1. Eventos de deleción totales

Tabla 7. Eventos de deleción totales identificados en los cuatro países

Deleciones							
Región	**Coordenadas**		**Ubicación**	**Secuencias**			
				EC	**MX**	**BZ**	**CO**
5UTR	1	265					
ORF1a	266	13,468	425-427	1			
			506-511			1	
			509-520				1
			509-523		2	6	1
			510-518		4	1	
			512-517			1	
			515-520	1	3	6	2
			516-518	1			
			518-520			2	
			519-524			1	
			521-523		1	1	
			686-694	2	4	6	3
			1431-1433			1	
			3161-3163		2		
			3333-3335	1			
			4880-4882				1
			6656-6679			1	
			8562-8564			1	
			9761-9763			1	
			11288-11296	582	811	937	293
ORF1b	13,468	21,555					
S	21,563	25,384	21632-21640		1		
			21633-21641	519	771	889	292
			21635-21643	1		2	
			21653-21655	22	100	84	41
			21655-21657	14			
			21765-21770	288	708	581	151
			21990-21995	1			
			21992-21994	579	768	921	291
			21993-21998		1		
			21995-21997			1	
			22101-22115	6			
			22115-22120	2			1
			22115-22129		1		1
			22194-22196	282	737	592	173
			23004-23006		1		

			23009-23011	228	699	537	160
ORF3a	25,393	26,220	26155-26157		1		
			25423-25431			1	
E	26,245	26,472					
			26513-26520	2			
M	26,523	27,191					
ORF6	27,202	27,387	27373				1
			27265-27291			1	
			27293			1	
ORF7a	27,394	27,759	27601-27625		1		
			27700-27702			1	
			27571-27594			1	
			27568-27580			1	
ORF7b	27,756	27,887	27792-27794		1		2
			27879-27880			1	
			27795-27797			1	
ORF8	27,894	28,259	28254	24	15	33	7
			28090-28095			5	
			28151-28152			1	
N	28,274	29,533	28240-28245			1	
			28242-28251		1		
			28344-28352		1		
			28362-28370	580	797	936	263
			28365-28370	1			
ORF10	29,558	29,674	29564-29565		1		
3UTR	29,675	29,903	29680				
			29682-29683		3		
			29711			1	
			29726-29729		1		
			29728-29776	1			
			29728-29766			1	
			29733-29758		1	1	
			29734-29759	545	731	793	216
			29819-29821		1		

Descripción: En la tabla se puede observar todas las sesenta y siete deleciones identificadas en todas las secuencias de SARS CoV-2 de los cuatro países. Se observa la región genómica (junto con el gen) en la que ocurre, su ubicación dentro del genoma y la secuencias identificadas para cada una de estas.

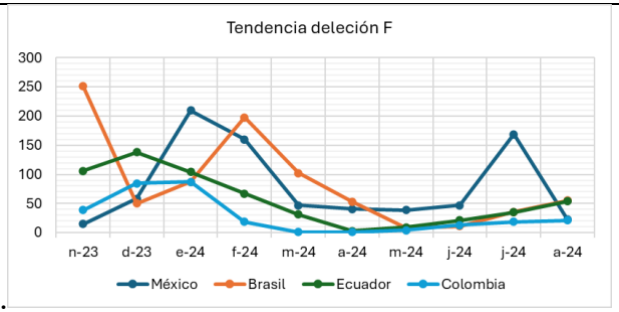
12 Anexo 2. Eventos de inserción totales

Inserciones							
Región	**Coordenadas**		**Ubicación**	**Secuencias**			
				EC	**MX**	**BZ**	**CO**
ORF1a	266	13,468					
ORF1b	13,468	21,555	14,376		1		
S	21,563	25,384	21,608	185	662	558	93
			21,610		2		
			21,613	22			
			22,204			1	
			22,304		2		
ORF3a	25,393	26,220	25,701			1	
E	26,245	26,472					
M	26,523	27,191					
ORF6	27,202	27,387	27,298			1	
ORF7a	27,394	27,759					
ORF7b	27,756	27,887					
ORF8	27,894	28,259	28,065			1	
			28,250	3	1	7	
ORF9b	28,284	28,577					
N	28,274	29,533					
			29,758			1	
			29,889	1			

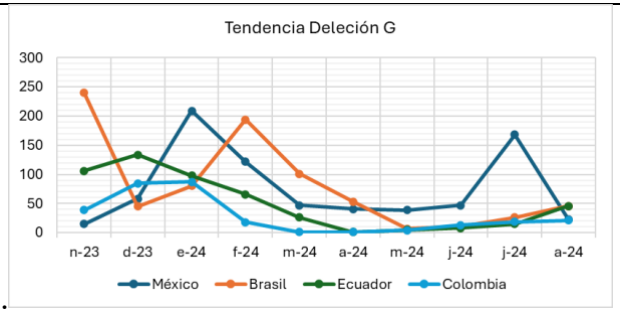
Descripción: En la tabla se puede observar todas las doce inserciones identificadas en todas las secuencias de SARS CoV-2 de los cuatro países. Se observa la región genómica (junto con el gen) en la que ocurre, su ubicación dentro del genoma y la secuencias identificadas para cada una de estas.

13 Anexo 3. Tendencia a lo largo del tiempo deleciones más frecuentes

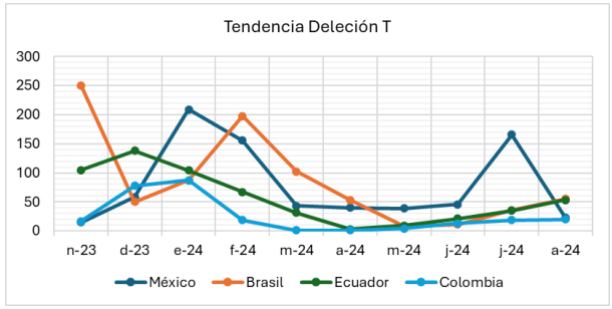
1.



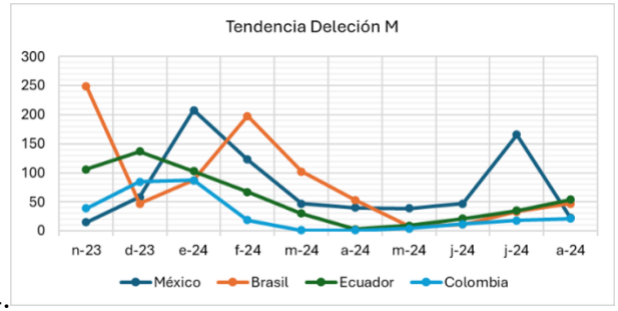
2.



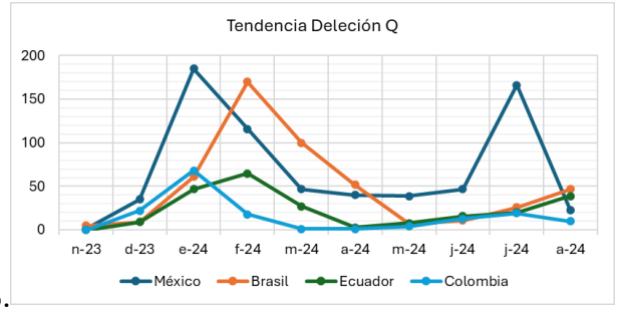
3.



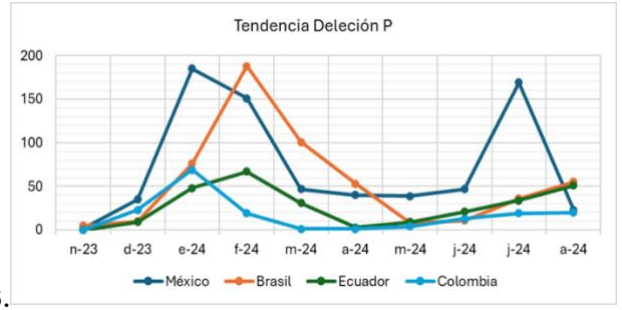
4.

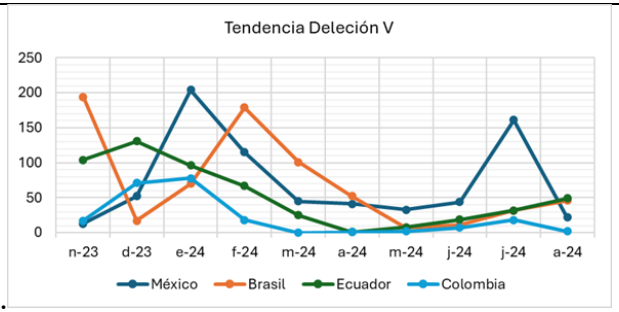


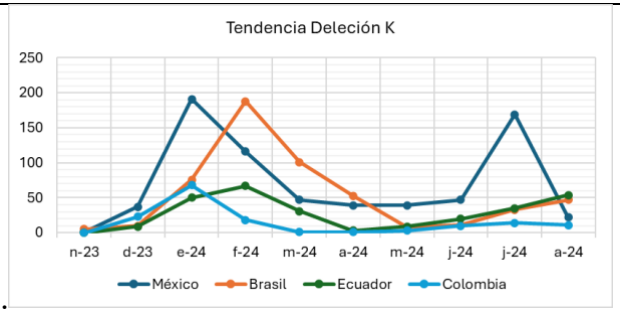
5.

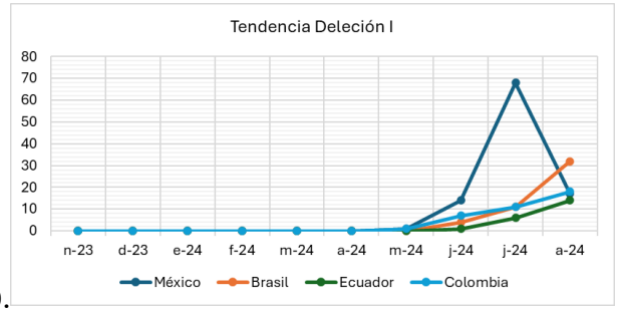


6.



7. 

8. 

9. 

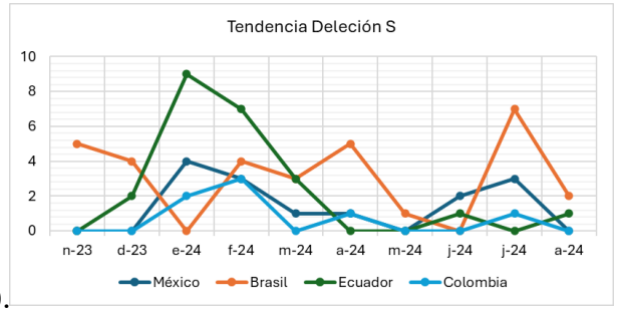
10. 

Figura 4. Tendencia de deleciones a lo largo del tiempo. En el caso de la deleción F, G, T y M se observa una tendencia similar de ocurrencia: 1,2,3 y 4. Lo mismo para las deleciones Q, P, V y K como se observa en 5, 6, 7 y 8. Por otro lado la deleción I (9) este recién aparecerá entre mayo y agosto de 2024, con un pico que similar al que se observa tanto en el primer grupo como en el segundo. Finalmente, la deleción S (10) no muestra una tendencia similar a ninguno de los otros grupos, lo que se va a analizar en la discusión.

14 Anexo 4. Tendencia de linajes a lo largo del tiempo

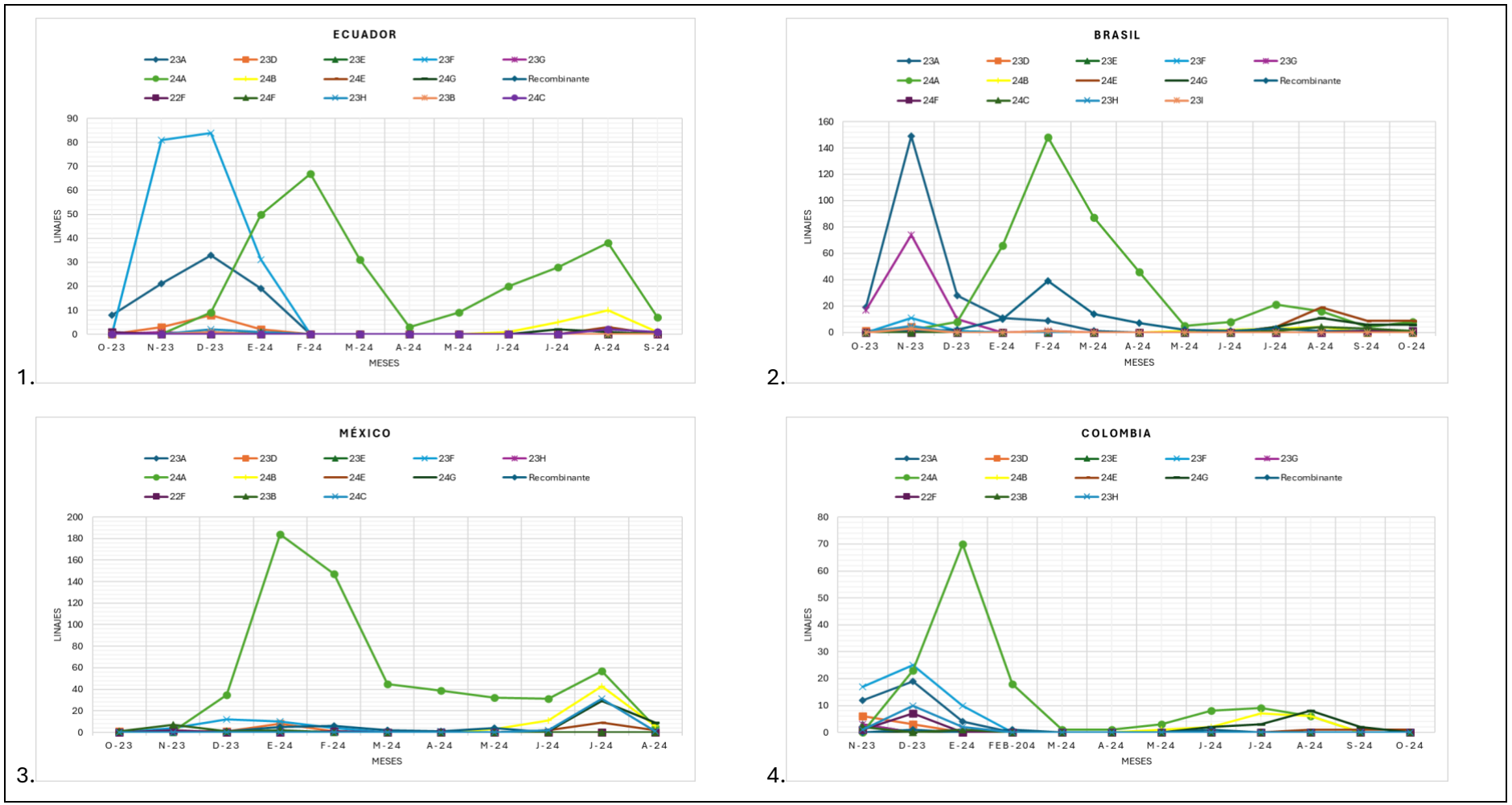


Figura 5. Tendencia de linajes a lo largo del tiempo. En este gráfico observamos la tendencia de ocurrencia de los linajes para cada país. En 1, la tendencia en Ecuador se observa una transición entre 23F/23A y 24A; con pequeños picos de 24B. En 2, Brasil, se observa un cambio entre 23A/23G con 24A/24E. En 3, México, se observa predominancia de 24A, con picos delante de 24B, 24C y 24G. Finalmente, en 4, Colombia, se observa predominancia de 24A, una ligera transición desde 23F, y picos en 2024 de 24B y 24G.

15 Anexo 5. Árbol filogenético SARS CoV-2

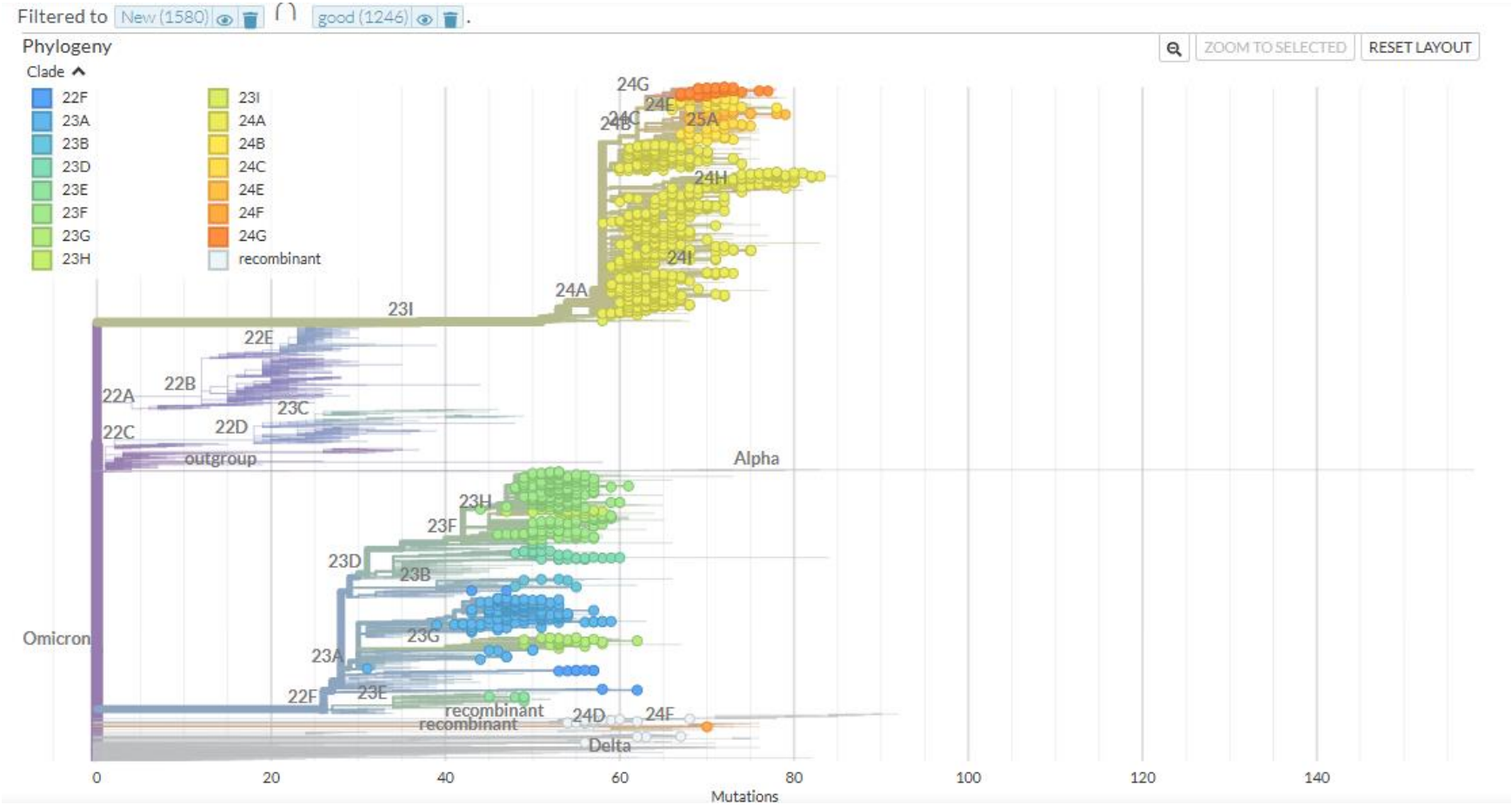


Figura 6. Árbol filogenético de secuencias de SARS CoV-2. En el árbol filogenético generado por NextStrain, se observa un clado monofilético que surge a partir de 24A, y engloba a todos los linajes que se originaron en 2024 y 2025. Y el 24A parte de un clado 23I, con el que compartirán más similitud. Mientras que linajes de Ómicron 23A/ 23F y 23G comparten un clado, y una mayor similitud con el VOC Alfa. Algo que se revisará en mayor detalle en la discusión.