UNIVERSIDAD NACIONAL AUTÓNOMA

DE MÉXICO

ESCUELA NACIONAL DE ESTUDIOS SUPERIORES
UNIDAD LEÓN

TEMA:

SEGUIMIENTO DE GRANDES DELECIONES EN ORFS 7A/8 EN SARS-COV-2 PARA EL CONSORCIO MEXICANO DE VIGILANCIA GENÓMICA.

MODALIDAD DE TITULACIÓN:

TESIS Y EXAMEN PROFESIONAL

QUE PARA OBTENER EL TÍTULO DE:

LICENCIADO EN CIENCIAS AGROGENÓMICAS

P R E S E N T A :

JOSE ABEL LOVACO FLORES

ENESUNAM
UNIDADLEÓN

TUTOR EXTERNO:
DRA. NELLY SELEM MOJICA

TUTOR INTERNO:
DR. JESÚS ABRAHAM AVELAR RIVAS

ASESOR:
DR. ISRAEL PICHARDO CASAS

León, Guanajuato, México **2022**

ÍNDICE

ÍNDICE DE FIGURAS

DEDICATORIA

Hace 5 años que inicié este viaje. Recuerdo que llegué con muchas expectativas, no tenía empleo, no conocía a nadie, pero de los que sí estaba seguro era de que quería estudiar. Durante estos cinco años fue difícil poder trabajar y estudiar, por eso esta tesis va dedicada a todos aquellos que tienen que trabajar y ser autosuficientes. Animo lo vas a lograr. Esta tesis también va para mi yo de hace 5 años, ¡LO LOGRAMOS! pudimos terminar esa carrera que tanto anhelamos. Todos esos sacrificios valieron la pena.

AGRADECIMIENTOS

Quiero agradecer a mi papá Jose Isabel y a mi mamá Florina por darme el apoyo moral de seguir con mis sueños y por estar en cada momento difícil, los amo mucho. También quiero agradecer a Edith y su negocio familiar que me dieron apoyo incondicional, su amistad y por darme trabajo. Gracias nuevamente Edith por ser parte de este sueño que ya se hizo realidad y por darme ese abrazo y familia postiza que necesite durante este largo camino.

A mis compañeros de carrera Aaron, Gema, Citla, Maria y Vane por las risas, las pláticas, los consejos, las fiestas, pero sobre todo por estar en el momento que más lxs necesite. Porque literal gracias a ustedes sigo vivo. También gracias a Aaron, por ser mi mejor amigo y por ayudarme a ser una mejor persona. ¡Los amo y quiero un montón!

Por último, a las personas que me ayudaron a culminar este sueño, al Dr. Jesús Abraham y la Dra. Nelly, gracias por todos esos consejos académicos y por siempre creer en mí aun cuando yo no lo hacía. Gracias por las risas, los viajes, por animarme, por las anécdotas que siempre nos hacen reír y apoyarme siempre.

Me gustaría agradecer a la Licenciatura en Ciencias Agrogenómicas por todo el conocimiento que adquirí durante mis estudios de licenciatura. También a todo el cuerpo docente que siempre estuvo atento durante mis estudios y por supuesto a la Universidad Nacional Autónoma de México.

Quiero expresar mi profundo agradecimiento al CoViGen-Mex por permitirme colaborar con los investigadores que participan en este gran proyecto. El consorcio es parcialmente financiado por el proyecto “Vigilancia Genómica del Virus SARS-CoV-2 en México-2022” (PP-F003; a CFA) y “Caracterización de la diversidad viral y bacteriana” (FORDECYT a JAV-P.) del Consejo Nacional de Ciencia y Tecnología-México (CONACyT), donación 057 de la “Secretaría de Educación, Ciencia, Tecnología e Innovación (SECTEI) de la Ciudad de México” (a CFA). Un especial agradecimiento a la Dra. Celia Boukadida por compartirnos sus hallazgos sobre deleciones y dejar que pudiera continuar con esta investigación. Así mismo, a la Dra. Blanca Taboada por responder mis dudas acerca del proyecto. También quiero agradecer al Dr. Alfredo Herrera Estrella, a las empresas hermanas C3-BetterLab por el financiamiento de este proyecto y a todo el equipo que labora en la empresa en especial a Noé García Chávez e Israel por animarme y apoyarme en las técnicas de secuenciación.

RESUMEN

El nuevo coronavirus SARS-CoV-2, que fue detectado en Wuhan en el año 2019, ha adquirido mutaciones que le han conferido múltiples ventajas evolutivas. Aunque la mayoría de las mutaciones detectadas son por la eliminación, sustitución o adición de un nucleótido, los análisis genómicos han encontrado que en SARS-CoV- 2 existen regiones que frecuentemente muestran deleciones. Una de las regiones donde más frecuentemente se han detectado deleciones largas es en los ORFs 7a/8, aunque aún se conoce poco de sus consecuencias tanto en el virus como en los contagiados. En México, el Consorcio Mexicano de Vigilancia Genómica detectó que las deleciones largas no eran bien detectadas por software estándar ya que las posibles deleciones podrían ser confundidas con errores en la secuenciación. Por esto, fue necesario crear un *pipeline* que detectara posibles deleciones largas en SARS-CoV-2. Como parte del Consorcio Mexicano de Vigilancia Genómica, en este trabajo se presenta un *pipeline* llamado Indel-Mex para detectar deleciones largas en alineamientos de SARS-CoV-2. La validez de sus resultados los confirmamos en una deleción de *Δ*411 nt en el linaje B.1.243 que fue detectada a final del año 2020 que fue confirmada mediante PCR y secuenciación. El *pipeline* nos mostró que dicha deleción solo círculo durante 3 meses. Además, el análisis de la secuenciación de 3805 secuencias nos permitió detectar múltiples deleciones entre el ORF7a y el ORF8 que circularon en el año 2021 y que no son de un tamaño específico ni tienen un origen en común. Entre ellas mostramos la identificación de una posible deleción de *Δ*222 nt en muestras de SARS-CoV-2 asignadas al linaje B.1.243. Ahora sabemos que han circulado múltiples deleciones de los ORFs 7a/8 que se han detectado en el año 2021. Nuestro *pipeline* permitirá una búsqueda de deleciones largas en todo el genoma y queda determinar si estas deleciones siguieron apareciendo los años próximos de la pandemia.

I. INTRODUCCIÓN

En la familia *Coronaviridae*, existen alrededor de 26 especies de coronavirus (CoVs), los cuales se clasifican en los siguientes géneros: alfa (α), beta (β), gamma (γ) y delta (δ) (Shereen et al., 2020; Yang et al., 2006). De estos CoVs a la fecha existen siete especies que infectan a los seres humanos, las cuales son 229E, NL63, OC43, HKU1, el virus causante del síndrome respiratorio de Oriente Medio (MERS-CoV), Síndrome Respiratorio Agudo Severo-CoV (SARS)-CoV, y el más reciente Síndrome Respiratorio Agudo Severo 2 causado por coronavirus (SARS-CoV-2) (Prajapat et al., 2020) perteneciente al género *Betacoronavirus*.

En el año 2003, en la provincia de Guangdong, China emergió el virus SARS-CoV que origina síntomas de infecciones respiratorias y en algunos casos diarrea. Este brote fue provocado por la zoonosis de animales a humanos (Shereen et al., 2020; Yang et al., 2006; Zhong et al., 2003). En este primer brote, SARS-CoV infectó a 8,098 personas con una tasa de mortalidad del 9%, en 26 países del mundo (Shereen et al., 2020).

Una década más tarde, en 2012, dos ciudadanos de Arabia Saudita iniciaron con síntomas de infecciones respiratorias y se realizaron pruebas clínicas confirmando la infección por un nuevo coronavirus. Esta enfermedad fue el síndrome respiratorio de Oriente Medio causado por coronavirus (MERS-CoV). La enfermedad se transmitió por una nueva zoonosis con camellos infectados (Shereen et al., 2020). La Organización mundial de la salud (OMS) informó que el coronavirus MERS infectó a más de 2,428 personas y dejó un saldo de 838 muertes.

El virus del SARS-CoV-2 es un nuevo coronavirus que se detectó en Wuhan, China en el año 2019 y se fue expandiendo por todo el mundo en poco tiempo. El 11 de marzo de 2020, la OMS por los alarmantes niveles de propagación de la enfermedad y por su gravedad de la enfermedad, declaró pandemia (*COVID-19*, s/f). En el área de urgencias de Wuhan entraron personas con síntomas respiratorios graves como: tos, fiebre y dificultad para respirar (Zhou et al., 2020). En México, el primer caso de COVID-19 se detectó el 28 de febrero del 2020. Hasta la fecha (21 de mayo de 2022) México tiene más de 6 Millones de casos confirmados (*COVID-19 Tablero México*, s/f).

Los virus de SARS-CoV-2 son de cadena de *RNA* de sentido positivo (30 kbs de longitud) y tienen un tamaño aproximado de 65-125 nm. Dentro de su envoltura, el genoma viral está encerrado en un núcleo de ribonucleoproteína (RNP) (Rossi et al., 2020). En la superficie del virus se detectan picos (receptores) que le dan forma de una corona y por eso recibe el

nombre de coronavirus (Prajapat et al., 2020; Shereen et al., 2020). Los CoVs tienen en su genoma 4 proteínas estructurales importantes **(Figura 1.1)**: la proteína espiga (S) por su nombre en inglés, la proteína de membrana (M), la proteína de envoltura (E) y la proteína nucleocápside (N) (Prajapat et al., 2020; Shereen et al., 2020).

En el trabajo de Roujian Lu et. al. Se compararon los genomas de SARS-CoV y SARS-CoV-2 y se encontró que tenían un porcentaje de identidad del 79% (R. Lu et al., 2020). También se encontró la presencia de la proteína 8a en SARS-CoV, pero esta proteína estaba ausente en el SARS-CoV-2 (Shereen et al., 2020; Wu et al., 2020). Mientras que la proteína 8b tiene un tamaño de 84 aminoácidos en SARS-CoV, en SARS-CoV-2 es de 121 aminoácidos (Wu et al., 2020).

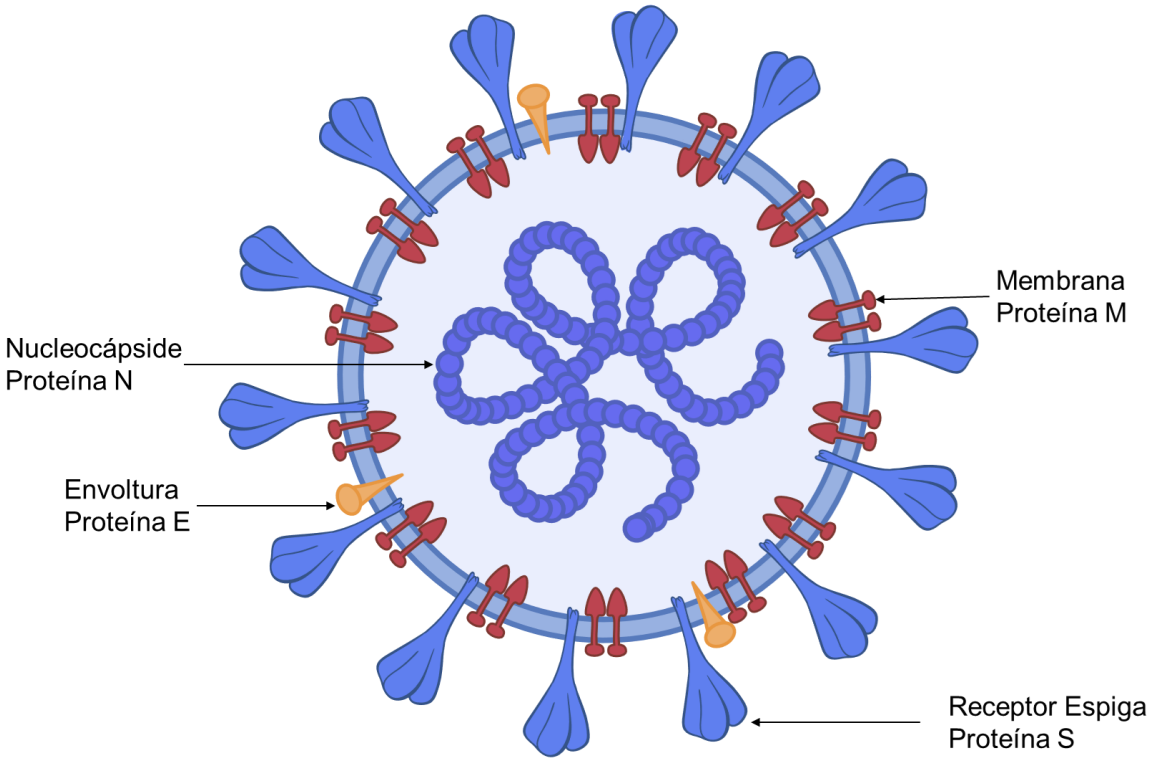


Figura 1.1 Estructura del SARS-CoV-2. La estructura del SARS-CoV-2 está compuesta por proteínas estructurales importantes como: la proteína Spike (estructuras con tres curvaturas aglomeradas de color azul), la Nucleócapside (circulos aglomerados azules), la Membrana (figuras rojo oscuro) y la Envoltura (figuras en forma de clavo naranja). Estas proteínas estructurales se encargan de dar la forma característica del coronavirus. Créditos de la figura:https://bioicons.com/Hanna Vega

El genoma de SARS-CoV-2 además codifica para genes no estructurales (ORFs 1ab-10) los cuales tienen una variedad de funciones accesorias, algunas de las cuales pueden ser no esenciales para la replicación o patogénesis viral (Yadav et al., 2021). Las proteínas virales están codificadas en Marcos de Lectura Abiertos (ORFs, *Open Reading Frames*, por su acrónimo en inglés) y se caracterizan por tener un inicio de traducción delimitado por un

codón o señal de paro al final del gen. En el coronavirus, algunos ORFs codifican varias proteínas en la misma región, mismas que son traducidas por los ribosomas de las células infectadas y procesadas por proteasas intracelulares y virales (*Marco abierto de lectura | NHGRI*, s/f). Hasta ahora no se conoce a profundidad el funcionamiento de los ORF 7b/8 de SARS-CoV-2. Sin embargo, en SARS-CoV han detectado que el gen ORF8 en SARS-CoV induce el estrés del retículo endoplasmático al dirigirse específicamente al factor de transcripción activador 6 (ATF6) y este facilita el plegamiento y procesamiento de proteínas (Sung et al., 2009).

Una mutación ocurre cuando uno o más nucleótidos son añadidos, eliminados o sustituidos por otro nucleótido en un genoma. En el virus del SARS-CoV-2, se observan múltiples mutaciones a lo largo de su genoma en comparación con su pariente más cercano, el SARS-CoV. En especial, las mutaciones identificadas en la proteína Espiga (S) parecen conferir al virus ventajas evolutivas, por ejemplo, una mejor capacidad de adherirse a los receptores de la enzima convertidora de angiotensina 2 (ACE2) de células humanas (Wang et al., 2020). Algunos reportes han mencionado que estas mutaciones en la proteína S tiene una mayor transmisibilidad (Peacock et al., 2021, p. 2). En este sentido, un linaje es un grupo genéticamente relacionado a las variantes de un virus derivadas de un ancestro común. Una variante tiene una o más mutaciones que la diferencian de otras variantes de los virus (CDC, 2020). Para tener un mejor entendimiento de estas variantes las OMS las clasificó en; variantes de interés (VOI) o variantes de preocupación (VOC), las cuales pueden ser asociadas con una reducción de la neutralización de anticuerpos, una mayor transmisibilidad ó gravedad de la enfermedad (CDC, 2020; *WHO | SARS-CoV-2 Variants*, s/f). Estas variantes pueden emerger y desaparecer mientras otras persisten en el tiempo (CDC, 2020). Tal es el caso de la variante 201/500 Y.V1 (B1.1.7), llamada alfa que se detectó en el territorio de Reino Unido y podría ser una variante que permita a este coronavirus ser más infectivo. También se reconocieron otras dos variantes de preocupación como es la 20H/501.V2 (B.1.351), denominada Beta, que se detectó en sudáfrica y la variante 20J/501 Y.V3 (P.1), Gamma encontrada en Brasil, donde la preocupación principal es la capacidad de provocar una re-infección en humanos (*Spike E484K Mutation in the First SARS-CoV-2 Reinfection Case Confirmed in Brazil, 2020 - SARS-CoV-2 Coronavirus / NCoV-2019 Genomic Epidemiology*, 2021). Por otro lado, la variante B.1.617.2, denominada Delta se detectó en India en octubre del 2020 y tomó gran relevancia debido a su capacidad de ser más transmisible e infectiva (Raman et al., 2021). Finalmente, la variante B.1.1.52, llamada ómicron, se detectó en Estados Unidos en noviembre del 2021 y el incremento de los casos fue causado por esta variante. Se cree que uno de los factores más importantes fue su capacidad de evadir el sistema inmune y

reducir la capacidad de neutralización de las vacunas de tal forma que es capaz de infectar a humanos previamente infectados o vacunados (*How Bad Is Omicron?*, 2021).

Además de mutaciones de un nucleótido, también se han detectado deleciones relevantes en SARS-CoV-2 y en SARS-CoV. De hecho, anteriormente se creía que una deleción de 29 nucleótidos (nt) en la región genómica del ORF8 en SARS-CoV había sido importante para el cambio de hospederos ya que la deleción de 29 nt en el ORF8 estaba ausente en genomas de coronavirus cercanos a SARS-CoV aislados de muestras en murciélagos, pero presente en algunos SARS-CoV que infectan a los humanos (Muth et al., 2018). Este indicio apuntaba a que la falta de este fragmento del ORF8 conducía a la adaptación del SARS-CoV con humanos. Ya en SARS-CoV-2, Su y colaboradores mostraron una deleción de 382 nt en el ORF 7b/8 (Su et al., 2020), la cual no fue persistente durante la pandemia del 2019. Incluso durante la realización de este trabajo se publicó que una variante con deleción estaba asociada a síntomas más leves en personas con SARS-CoV-2 (Young et al., 2020). Por estas razón es importante que el monitoreo incluya también la detección de indeles.

En México, el monitoreo de las variantes de este virus se ha efectuado en el Instituto de Diagnóstico y Referencia Epidemiológicos (InDRE), el cual es un centro de investigación gubernamental que ha diagnosticado y secuenciado de forma masiva, las muestras de pacientes infectados. Para realizar el diagnóstico de infección se han empleado pruebas como las de antígenos, las serológicas y las de Reacción en Cadena Polimerasa (PCR) (Sethuraman et al., 2020). La prueba de RT-PCR sintetiza fragmentos de genes blanco que son reconocidos por oligos de una región del genoma viral. El *test* de RT-PCR es considerado el *Gold standard* para la detección del virus que causa la enfermedad de la COVID-19 (*Nucleic Acid Amplification Tests (NAATs)* | *CDC*, s/f). Cuando se realiza la prueba de PCR mediante exudado nasofaríngeo la especificidad es de 99,5% y la sensibilidad oscila entre 85-90% (Langa et al., 2021).

La secuenciación de las muestras positivas sirve, entre otras cosas, para identificar que mutaciones podrían darle un beneficio al virus o perjuicio a pacientes. Esto ha impulsado la preocupación de científicos y médicos alrededor del mundo. En México se realizó un monitoreo constante de SARS-CoV-2 para conocer las mutaciones que el virus pudo generar. El monitoreo se ha logrado mediante la secuenciación masiva del material genético del coronavirus. La secuenciación masiva se puede ejecutar en cuatro etapas principales: la extracción de ácidos nucleicos, la preparación de las librerías, la secuenciación y el análisis bioinformático e interpretación de los resultados (H. Lu et al., 2016).

La tecnología de secuenciación Illumina sigue abarcando la mayor parte de investigaciones genómicas. Illumina realiza la amplificación de los fragmentos genómicos mediante una reacción en cadena de polimerasa (PCR) de puente. Para la secuenciación se utilizan cuatro nucleótidos con terminadores reversibles y cada ciclo tiene lugar con los cuatro nucleótidos simultáneamente (*Next-Generation Sequencing (NGS)* | *Explore the technology*, s/f). Esta tecnología miniaturizada permite la generación de miles de millones de regiones clonales con una alta fidelidad (Martín et al., 2020).

La secuenciación de nanoporos ofrecida por *Oxford Nanopore Technologies (ONT)* detecta directamente los cambios en las corrientes generadas cuando los ácidos nucleicos pasan a través de nanoporo proteico. La velocidad de estas moléculas es rápida, ya que pasan a través de una proteína nanoporosa (≈450 bases S^{-1}) para DNA y para RNA (≈80 bases S^{-1}) (Wang et al., 2020). La tecnología de ONT tiene las características de; ser portátil; procesar fragmentos de cualquier longitud (*reads* largos); no requiere maquinaria costosa (Faria et al., 2016). Esto nos permitiría encontrar deleciones largas. La pandemia de SARS-CoV-2 uso a la secuenciación como herramienta para la toma de decisiones que coadyuven a controlar la propagación de la enfermedad (Pal et al., 2020; Resende et al., 2020).

Por otro lado, la secuenciación Sanger es un método que proporciona información sobre la identidad y el orden de las cuatro bases de nucleótidos en un segmento de DNA (Heather & Chain, 2016). Esta tecnología se dirige a una región específica de la plantilla de DNA utilizando un cebador de secuenciación de oligonucleótidos, que se une al DNA adyacente a la región de interés. Es una de las tecnologías que tiene una precisión del 99,99%, siendo el estándar de oro para la mayoría de las aplicaciones, tanto de investigación como clínicas (*What is Sanger sequencing?*, s/f).

Las plataformas de secuenciación que fueron descritas anteriormente pueden ser utilizadas para responder preguntas biológicas. Para el caso de SARS-CoV.2 , Illumina nos permite conocer la variabilidad genética del virus, mientras que ONT y Sanger nos permiten analizar posibles deleciones que por su tamaño podrían pasar desapercibidas en los reads cortos que produce Illumina. **(Figura 1.2).** Por supuesto, las cualidades de cada una de estas tecnologías pueden ser usadas como mejor le convenga a cada investigador.

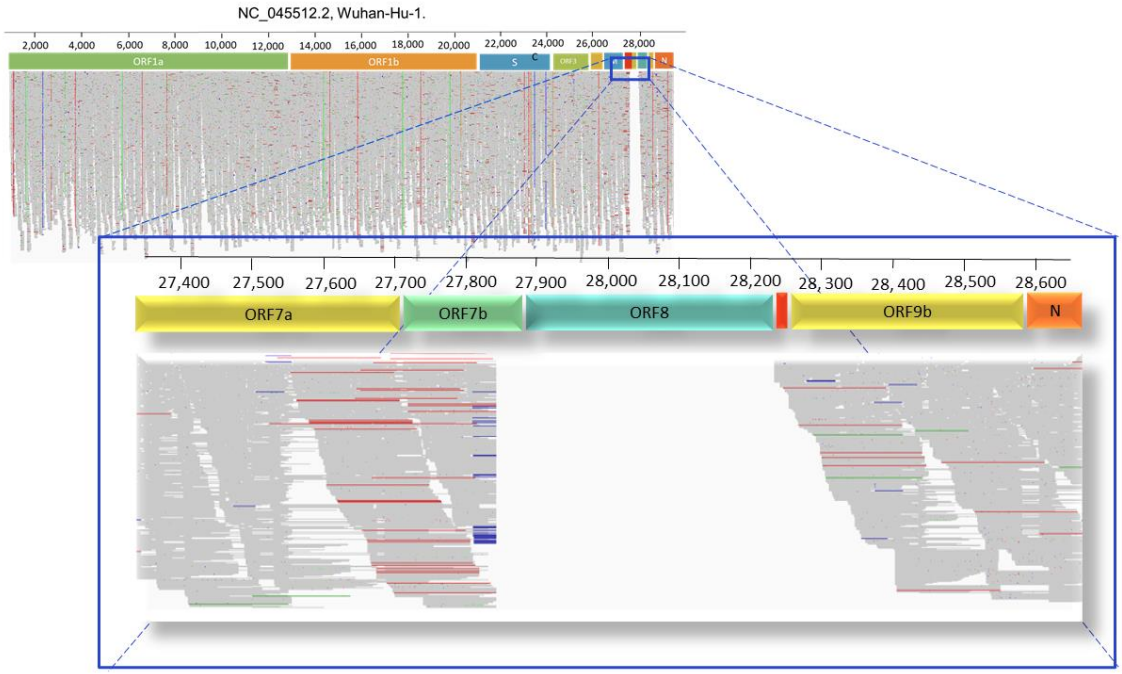


Figura 1.2. Alineamientos de *reads* de Illumina al genoma de SARS-CoV-2 y ubicación del posible fragmento faltante de 411nt en el linaje B.1.243. En la parte superior se esquematizan los genes virales contenidos en ~30,000 nt del genoma de SARS-CoV-2. Cada rectángulo representa un marco de lectura abierto y algunos genes estructurales como S, M y N se pueden distinguir. La mayoría de los genes no estructurales están contenidos en el ORF1 a/b. Un acercamiento hacia la región 3' se enmarca la deleción de Δ411 nt observada experimentalmente y que comprende los ORF7b y ORF8. Figura elaborada utilizando el visualizador de Next Strain (Hadfield et al., 2018).

II. ANTECEDENTES

En México, el monitoreo de las variantes de SARS-CoV-2 ha sido coordinado por el Instituto de Diagnóstico y Referencia Epidemiológicos (InDRE), el cual ha diagnosticado y secuenciado muestras de los pacientes infectados. También se creó el Consorcio Mexicano de Vigilancia Genómica (CoViGen-Mex), formado por el Instituto Mexicano del Seguro Social (IMSS), el Instituto Nacional de Enfermedades Respiratorias (INER), el Laboratorio Nacional de Genómica para la Biodiversidad (LANGEBIO) del CINVESTAV, el Centro de Investigación en Alimentación y Desarrollo A.C. (CIAD) y el Instituto de Biotecnología (IBT) de la UNAM. El Consorcio se consolidó para identificar en la República Mexicana variantes del virus SARS-CoV-2 que pudieran tener un comportamiento biológico de interés para la salud pública nacional (*MexCoV2*, s/f). En este sentido, muestras de RNA de pacientes positivos son obtenidas mediante un acuerdo entre el CoViGen-Mex y el IMSS. Semanalmente se entregan alrededor de 500 muestras, las cuales son procesadas en sedes autorizadas por el CoViGen-Mex. Las distintas sedes del consorcio realizan la secuenciación de muestras de SARS-CoV-2 utilizando la tecnología de Illumina. Esta tecnología utiliza el conjunto de oligos de ARTIC (Quick, 2020) para amplificar regiones de 150 pares de bases que en conjunto cubren todo el genoma.

La Dra Celia Boukadida del INER y miembro del CoViGen-Mex se percató de la ausencia de Δ411 nucleótidos en la región genómica de los ORFs 7b/8 cuando realizaba la secuenciación y los análisis de ensamblaje de genomas del SARS-CoV-2 del linaje B.1.243. Al analizar más genomas notaron la ausencia de esta región en otras muestras. Estas irregularidades fueron detectadas con más frecuencia. Posteriormente, en el INER decidieron realizar las síntesis de regiones de los ORFs 7b/8 utilizando tres pares de oligos para demostrar la deleción. Estos oligos tendrían tamaños esperados de amplificación: 680, 995, 1018 para muestras sin deleción. Los resultados fueron congruentes con los tamaños esperados de: 269, 584, 607 para muestras con deleción. En este punto quedó por determinar qué tan común era dicha deleción entre las otras secuencias obtenidas por el consorcio. También se realizó una búsqueda en la literatura y esta deleción fue reportada por primera vez en Estados Unidos en muestras de diciembre de 2020 con las secuencias disponibles en GISAID en enero de 2021. Mientras que en México sospechamos de esta deleción en muestras de febrero 2021 de la variante B.1.243 que fueron publicadas en GISAID en marzo de 2021. Ya que al inicio de la pandemia GISAID solicitaba que los genomas ensamblados tuvieran un cobertura del 97% del genoma. Algunas de nuestras muestras fueron rechazadas ya que la pérdida de 411 nt afectaba la cobertura del genoma. por lo que es necesario buscar y corregir los genomas que tengan esa deleción de Δ411 nt.

El *pipeline* para generar nuevos genomas de SARS-CoV-2 del CoViGen-Méx está compuesto por instrucciones que permiten correr secuencialmente diferentes softwares bioinformáticos (Taboada et al., 2020). Para el inicio de este *pipeline*, se necesitan reads de muestras de SARS-CoV-2 secuenciados por la tecnología Illumina. La salida de este *pipeline* es un archivo fasta que contiene la secuencia genómica SARS-CoV-2. Este genoma se obtiene mediante el alineamiento de todos los *reads que* pertenezcan a la muestra, contra una secuencia de referencia y la posterior generación de una secuencia consenso. Además de los 4 nucleótidos (A,C,G y T), el genoma consenso puede contener N 's por varias razones; la primera: el nucleótido no pudo ser descifrado correctamente debido a una falla en la secuenciación o falta de cobertura y la segunda es la existencia real de una deleción cuando se compara contra el genoma utilizado como referencia.

Este trabajo demuestra estrategias para diferenciar las deleciones reales de las fallas técnicas de la secuenciación. En las muestras secuenciadas por Illumina las fallas pueden deberse a la eficiencia de los oligos que estuvieran en esta región. Como ejemplo, nos centramos en la ausencia de 411 nt de muestras selectas de la variante B.1.243. Cuando estos genomas son visualizados se puede observar 411 N 's consecutivas en las regiones de los ORFs 7b/8, las cuales indican que los nucleótidos no pudieron ser descifrados correctamente. En este trabajo se demuestra que se trata de una deleción y no una secuenciación fallida. Para ello, se utilizaron estrategias tanto experimentales como bioinformáticas.

En primer lugar, se realizó la amplificación de las regiones genómicas de los ORFs 7b/8 de SAR-CoV-2 asignados al linaje B.1.243. Los amplicones fueron secuenciados utilizando las tecnologías Nanopore y Sanger. Esto con el fin de confirmar la deleción o su ausencia. El CoViGen-Mex ha secuenciado y analizado aproximadamente 19,471 muestras de SARS-CoV-2 en distintas sedes del país (*MexCoV2*, s/f). Se analizaron experimentalmente las muestras asignadas al linaje B.1.243, Sin embargo, el poder recuperar el material genético sería complicado, ya que algunas muestras pudieron degradarse con el tiempo. Como consecuencia, desarrollamos Indel-Mex un *pipeline* bioinformático programado con el lenguaje bash que sirve para detectar deleciones largas en una región de interés del genoma de SARS-CoV-2 definida *a priori*. Finalmente, se utilizó Indel-Mex para buscar otras posibles deleciones largas en las regiones de los ORFs a 7a/7b/8 de todos los genomas secuenciados en la sede de LANGEBIO de SARS-CoV-2.

III. OBJETIVOS

Objetivo General

- Determinar la prevalencia espacio-temporal de deleciones en la región de los ORFs 7a/8 de SARS-CoV-2 en genomas obtenidos por secuenciación de *reads* cortos mediante un *pipeline* bioinformático que analice las muestras secuenciadas del CoViGen-Mex para observar su recurrencia y persistencia entre los SARS-CoV-2 circulantes en México durante el año 2021.

Objetivo específico:

- Desarrollar un software para la búsqueda de deleciones largas en genomas secuenciados por la tecnología de Illumina.
- Identificar la ocurrencia de dichas deleciones en los datos del CoViGen-Méx
- Confirmar y analizar la deleción de Δ411 en el linaje B.1.243 utilizando diferentes tecnologías de secuenciación: Nanopore y Sanger.
- Comparar deleciones largas en las regiones del ORF 7a/8 en distintos linajes de SARS-CoV-2.
- Implementar en el Parque de Innovación Agrobioteg la secuenciación de genomas de SARS-CoV-2, utilizando la tecnología de secuenciación por Oxford Nanopore Technologies (ONT).

IV. JUSTIFICACIÓN

- El monitoreo genómico de SARS-CoV-2 ha permitido la detección de mutaciones puntuales, así como de inserciones o deleciones en diversas variantes del virus que le podrían estar otorgando una ventaja evolutiva. En cuanto a las deleciones, se han encontrado en las regiones de los ORFs 7b/8 donde aún se desconoce su funcionalidad. Adicionalmente, se han detectado mutaciones similares en distintas partes del mundo.

- El CoViGen-Mex reportó, mediante amplificación de las regiones de los ORFs 7b/8, una deleción de Δ411 nt. Sin embargo, había más muestras que posiblemente también tenían una deleción similar y que GISAID no aceptaba por que la deleción hacía parecer que tenían una cobertura menor al 97%. Para corregir estos y otros casos similares es conveniente tener una estrategia bioinformática ya que si esta deleción se buscará a gran escala hubiera sido tardado y costoso hacerlo por técnicas de biología molecular.

- Indelseek es un software para la detección de inserciones y deleciones complejas en datos que se generen con *Next Generation Sequencing* (Au et al., 2017). Sin embargo, cuando analizamos nuestros datos con dicho software no se detectaron deleciones en los ORFs 7b/8 con los parámetros estándar.

V. HIPÓTESIS

Existen deleciones largas en las regiones de los ORFs 7a/8 en muestras de SARS-CoV-2 del año 2021 que no han sido correctamente identificadas en México. Los orígenes de estas deleciones y la frecuencia de aparición cambiaron durante el monitoreo en 2021.

VI. MATERIALES Y MÉTODOS

Desarrollo de un *pipeline* para detectar deleciones largas en SARS-CoV-2.

Descripción del problema.

Las secuencias provienen de muestras obtenidas por el IMSS que previamente fueron identificadas como positivas para SARS-CoV-2 por RT-PCR. Cada mes el consorcio pedía estratégicamente muestras para monitorear todos los estados o para darle seguimiento a posibles VOIs o VOCs. Una vez al mes dichas muestras se secuenciaron por Illumina en el LANGEBIO y esas fueron las analizadas aquí. Cuando la Dra. Boukadida del INER realizaba los ensamblajes genómicos de *reads* largos y se percató de que algunas muestras carecían de ~400 nt. Esta información era consistente con las 11 muestras depositadas en GISAID con deleciones ~411 nt en genomas de SARS-CoV-2, los identificadores de acceso de estas muestras se pueden consultar en el anexo de **Tabla suplementaria 1**. Después de hacer un análisis manual y secuenciar amplicones el INER confirmó la presencia de una deleción de *Δ*411 nt. El INER compartió estos hallazgos con los miembros del consorcio, por lo que fue necesario buscar si las muestras de otras sedes también presentaban esta deleción. Las herramientas existentes en ese momento no identificaban estas deleciones largas en ensamblajes con *reads* cortos de Illumina, por lo que se decidió desarrollar una herramienta para el seguimiento de estas deleciones en las muestras del CoviGen-Mx. Finalmente, para corroborar la deleción de *Δ*411 nt se decidió realizar validación experimental y bioinformática de ambas con muestras de la sede LANGEBIO.

Ensamblado de *reads* de Illumina.

Se solicitaron al CoviGen-Mx los alineamientos que se generaron contra el genoma de referencia (archivos BAM/SAM) del año 2021. Los alineamientos y los *reads* se encuentran en resguardo en los servidores de CONACYT. El proceso de ensamblado utilizado por el CoviGen-Mx ya fue descrito por Taboada et. al. en su trabajo sobre análisis genómicos de SARS-CoV-2 (Taboada et al., 2020, 2021), así como por Zarate et. al. en su trabajo sobre cómo la variante alfa no llegó a ser dominante en México (Zárate et al., 2022). Para efectos de claridad de la tesis, a continuación se describen las principales herramientas de *software* utilizadas en estos trabajos: en primer lugar se utiliza fastp V0.20 que funciona para el análisis de los archivos FASTQ (recorte de adaptadores, calidad de la secuencia, etc.); después cd-hit V4.8.1 es un programa que agrupa y compara secuencias de proteínas o nucleótidos (Li & Godzik, 2006); bowtie2 V2.3.5.1 se utiliza para el alineamiento de secuencias contra un genoma de referencia (Langmead & Salzberg, 2012): samtools V1.9

es un conjunto de utilidades que manipulan alineaciones en los formatos SAM (Sequence Alignment/Map) (Danecek et al., 2021) y finalmente, con ivar V1.2 que es un paquete computacional para analizar amplicones virales, en este caso, se utilizó para generar la secuencia consenso de genomas virales (Grubaugh et al., 2018). Estas herramientas de análisis y los comandos secuenciales de cada uno se encuentran disponibles en el siguiente repositorio de Github: https://github.com/fabel134/Deleciones-Mex/tree/main/Mazorca.

Programación del *pipeline* Bioinformático para detectar deleciones largas.

Para corroborar que podía detectar la deleción Δ411nt que se observó experimentalmente, se desarrolló "Indel-Mex" un *pipeline* que detecta deleciones largas en los ORFs 7b/8 de SARS-CoV-2. El proceso de InDel-Mex busca *reads* que hayan alineado en dos regiones distintas del genoma de referencia y que su alineamiento se divida justo en la frontera de la posible deleción. Los *reads* pueden ser de dos tipos, *forward* es o *reverse* y esto será importante para el *pipeline* al momento de calcular las coordenadas y las intersecciones. Los *reads forward* son secuencias de nucleótidos que están en dirección 5' - 3' y los *reads reverse* son secuencias de nucleótidos que están en dirección 3'- 5'. Para los *read forward* el script los detecta porque *bi* < *bf* (Se muestran los posibles tipos de *reads* respecto a una posible deleción en la **Figura 6.1**). Indel-Mex consta de 7 pasos principales que se observan en la **Figura 6.2**. Indel-Mex requiere los siguientes parámetros: posición de inicio y posición final de la delección (**Figura 6.2, paso 1)**. Estas posiciones fueron tomadas de los *reads* alineados al genoma de Wuhan-1. Indel-Mex por default toma un pre-inicio (PI) que es -150 nt antes del inicio y un pos-final (PF)+150 nt después del final de la deleción. Se toma el valor de 150 ya que es el máximo tamaño que puede tener un *read* de Illumina (tamaño que fue elegido por un acuerdo del Consorcio). También necesita un archivo BAM/SAM (estos archivos tienen las características alineadas al genoma de referencia Wuhan-1). Se seleccionan todos los *reads* que se encuentren en la región de interés, es decir, entre PI y PF (**Figura 6.1).**

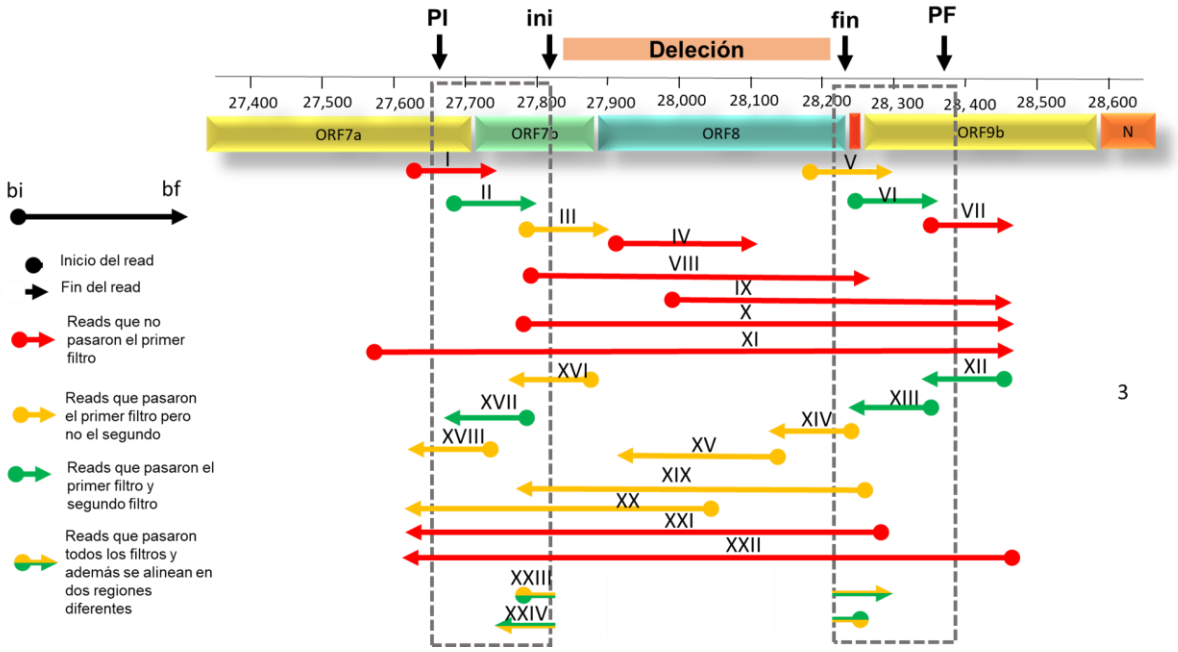


Figura 6.1. Clasificación de *reads* utilizando "Indel-Mex".
En la parte superior se encuentra la estructura genómica de SARS-CoV-2, así como, los parámetros iniciales para el *pipeline*. En los rectángulos punteados de color gris, es el area donde Inde-Méx clasifica a los reads alineados antes y después de la deleción. Las flechas rojas indican los reads que no pasaron el primer filtro, las amarillas indican que pasaron el primer filtro pero no el segundo, las verdes que pasaron el primer y segundo filtro y las flechas bicolores (amarillo y verde) son reads partidos. La dirección de la flecha es igual a la dirección del read, así como su inicio y el final.

Los reads previamente seleccionados se convierten en un archivo procesable (formato Fasta) para BLAST. Se realiza el *Basic Local Alignment Search Tool* (BLAST). Estos datos son alineados *read* por *read* al genoma Wuhan-1 utilizando la herramienta BLAST (**Figura 6.2, paso 3)**. El formato de salida es de tipo "outfmt 6", el cual nos muestra el nombre del *read*, Inicio del alineamiento con BLAST (bi) y el final del alineamiento con BLAST (bf) **(Figura 6.1)**.

Método: Reads alineados en dos regiones del genoma (REDOG)

Indel-Mex se queda con todos los *reads* (*forward* y *reverse*) que hayan alineado entre PI y PF. Después se crea una lista de los nombres de *reads* que están alineados entre el PI y el inicio de la deleción (ini) y *reads* que estén alineados entre final de la deleción (fin) y el PF. Por último, las listas generadas se comparan para detectar los *reads* que se encuentren en ambas listas. Si se encuentran *reads* comunes en ambas listas, se dice que estos *reads* están alineado en dos regiones cercanas a la deleción y por ende sugieren la deleción **Figura 6.2A, paso 4**.

Método: Número de nucleótidos en la deleción (NUNUDEL)

Este método toma el archivo de salida del proceso BLAST y lo siguiente que realiza es crear un archivo que incluya todos los *reads forward* entre la región PI y PF, es decir, *bi* < *bf*. Para los r*eads Reverse* se toman los *reads* que este entre la región PI y PF, es decir, *bi* > *bf* **Figura 6.1** y **Figura 6.2A, paso 5,** . Se calcula la intersección de cada *read* usando las siguientes fórmulas:

Tamaño de la intersección en los reads *forward*:

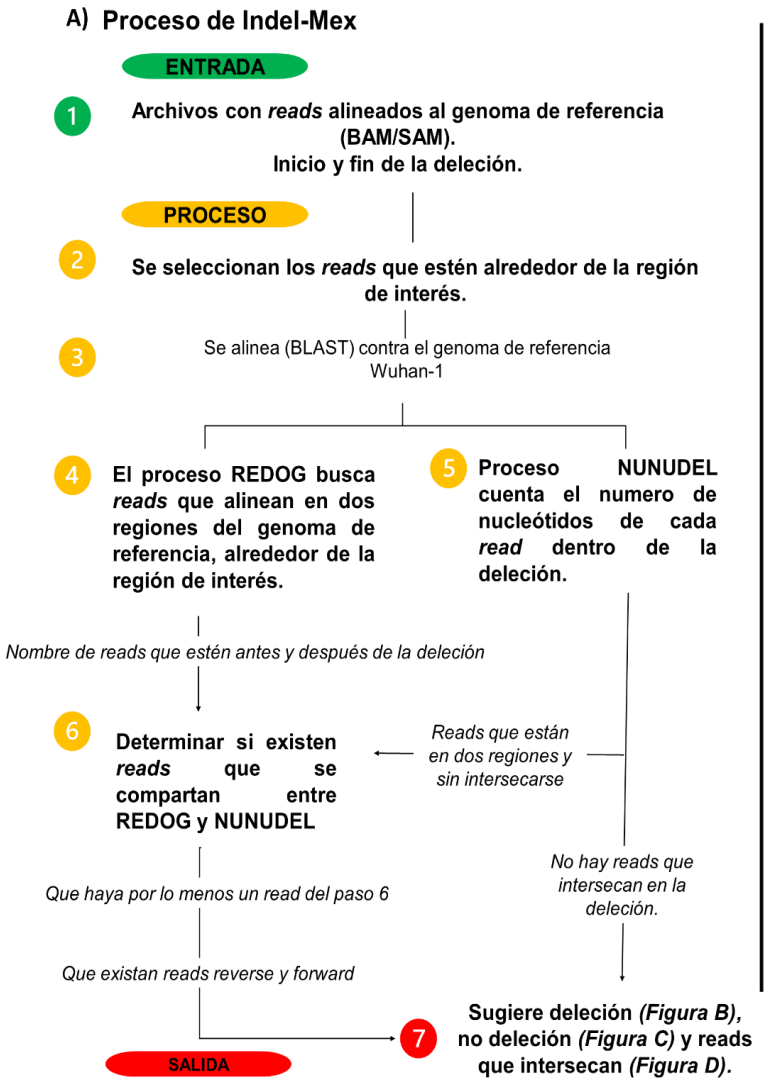
$$min\ (fin, bf) - max\ (ini, bi) \ldots (1)$$

Tamaño de la intersección en los reads *reverse*:

$$min\ (fin, bi) - max\ (ini, bf) \ldots (2)$$

Después de calcular el tamaño de intersección con la deleción de cada *read forward* y *reverse*. Indel-Mex busca los reads que alinean en dos regiones cercanas al inicio y al final de la deleción, es decir, el tamaño de la intersección debe estar en el intervalo -2, 2 **Figura 6.1 (*reads* verde-amarillo)**. Posteriormente, se buscan los *reads* que se encuentren dentro de la deleción entre *ini* y *fin*, es decir, el tamaño de intersección debe estar entre 3 y 150 que representan *reads* que provienen de genoma sin deleción. Como paso previo al final, se contabilizan de forma independiente los *reads* que alinean antes, después y dentro de la deleción.

Por último, Indel-Mex compara *REDOG y NUNUDEL,* contando el número de *reads* que fueron detectados por ambos métodos **Figura 6.3, paso 6**. Indel-Mex culmina mostrandos la existencia de una deleción al identificar las muestras con *reads* que alinean en las dos regiones que flanquean la frontera de la deleción (**Figura 6B**), no deleción ó muestras con *reads* dentro de la deleción **(Figura 6C**) y muestras con una combinación de las características descritas previamente (**Figura 6D**).



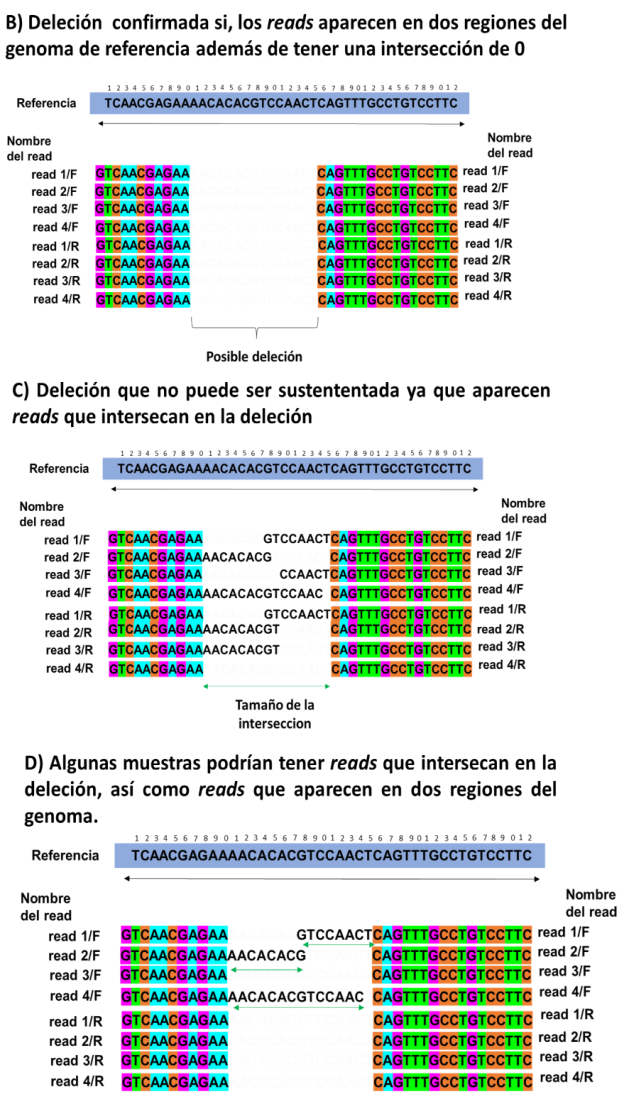


Figura 6.2. Proceso del Indel-mex en muestras de SARS-CoV-2.

Principales pasos para el procesamiento del *pipeline* en búsqueda de deleciones. (1). Para iniciar el *pipeline* necesita archivos en formato BAM/SAM. Además de también conocer la posición inicial y final de la posible deleción. (2) Así mismo, se hace un filtrado para obtener los *reads que* están ubicadas en la región de la posible deleción. (3) Se realiza un alineamiento (BLAST) contra el genoma de referencia. (4) (5) Cada muestra es analizada por el la sección que encuentra *Reads* Alineados a Dos lugares del Genoma (REDOG) alrededor de la deleción y Número de Nucleótidos del *read* que alineen en la Deleción NUNUDEL y determina si en estas hay *reads* compartidos (6) . (7) Indel-Mex te muestra posibles deleciones.

Confirmación de una deleción de Δ411 nt utilizando: PCR, ONT y Sanger.

Obtención de RNA viral de muestras con sospecha de deleción.

Las muestras fueron facilitadas por el IMSS al Consorcio Mexicano de Vigilancia Genómica. Estas muestras son de RNA de pacientes positivos confirmados por RT-PCR obtenido a partir de exudado nasofaríngeo. Los pacientes son de edad, sexo, y origen diferente, así como de sintomatología variada. La cuantificación de las muestras se hizo con el equipo NanoDrop 2000 (Thermo Scientific™).

Síntesis de cDNA para secuenciar por Nanopore.

La transcripción reversa se realizó utilizando el kit Protoscript II First Strand DNAc Synthesis (E6560S,New England). Siguiendo las recomendaciones del fabricante. Además, se agregó inhibidor de RNAsas (M03147, New England) para tener un mejor rendimiento. Para finalizar, estas reacciones se incubaron de la siguiente manera; 25°C 00:05:00 (incubación); 42°C 01:00:00 (incubación); 80 °C 00:05:00 (inactivación de la enzima).

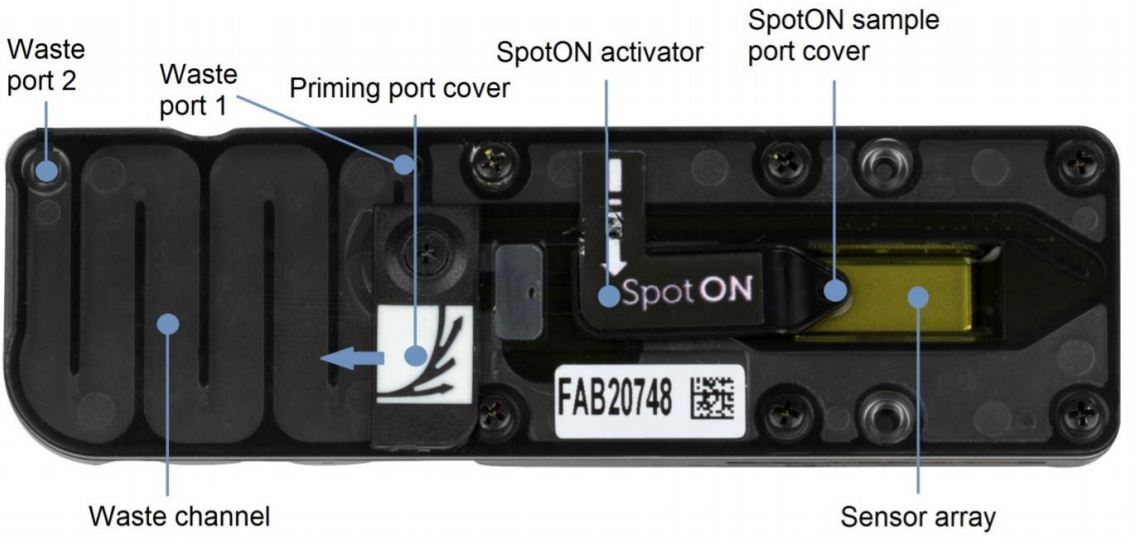
PCR de Muestras con posible deleción.

Para amplificar las regiones de las muestras con una posible deleción. Se seleccionaron un par de oligos de la red de Artic (Quick, 2020). El oligo nCoV-2019_92_LEFT *forward* (5'TTTGTGCTTTTTAGCCTTTCTGCT3') abarca desde la posición 27,785 a la 27,808 y el oligo nCoV-2019_93_RIGHT *reverse* (5'AGGTCTTCCTTGCCATGTTGAG3') abarca desde la posición 28,443 a la 28,464. Ambos oligos fueron sintetizados por la empresa T4 oligo, ubicada en Irapuato, Guanajuato. La concentración final para la reacción de PCR fue de 10 µM de cada oligo. La PCR se realizó utilizando la Q5® High-Fidelity 2X Master Mix así como las recomendaciones del fabricante. En el termociclador, las temperaturas fueron de la siguiente manera: desnaturalización de 98°C (30 seg); posteriormente 35 ciclos de 98°C (10 seg), 55°C (30 seg), 72°C (30 seg) y un ciclo de 72°C (2 min), por último, se conservó a 4° C (∞). Los amplicones fueron teñidos con SYBR Green en un gel de agarosa al 1%. El gel fue corrido a 80 V por 1 hora. Finalmente fueron visualizados en un fotodocumentador (Gel Doc™ XR, Bio-Rad).

Estandarización de secuenciación por nanopore.

Para poder comprobar la calidad de la celda se puede realizar una prueba rápida. Los reactivos que se utilizaron para esta secuenciación fueron: el Rapid Barcoding Kit (SQK-RBK004) para ligación del material genético con los adaptadores que son reconocidos por el nanoporo; el Control Expansion (EXP-CTL001) contiene el DNA control del fago lambda;

La R9.4.1 flow cells (FLO-MIN106), se debe observar que el sensor array esté libre de burbujas **Figura 6.3**. El protocolo de secuenciación de prueba se encuentra disponible en la documentación de Oxford Nanopore (*Ligation sequencing gDNA - Lambda control (SQK-LSK109)*, s/f). Es necesario lavar la celda de flujo con el Flow Cell Wash Kit (EXP-WSH004) para que posteriormente pueda ser utilizada. El protocolo de limpieza se encuentra disponible en la documentación de Oxford Nanopore (*Flow Cell Wash Kit (EXP-WSH004)*,



s/f).

Figura 6.3. Estructura de una celda de flujo de ONT. Esta imagen fue tomada del protocolo Flow Cell Wash Kit (EXP-WSH003) Version: WFC_9088_v1_revF_18Sep2019. La estructura de la celda está compuesta por múltiples orificios los cuales toman relevancia al momento de secuenciar. El *Waste port 1* funciona para extraer los desechos que son generados por la secuenciación. El *SpotON sample port Cover* por donde es introducida la librería que se va a secuenciar. El *Priming port cover* es el orificio por el cual se introduce la reacción para hacer el *priming* a la celda. El *Waste Channel* funciona como un almacén de desechos. Por último tenemos el *Sensor Array* que es donde se encuentran los nanoporos para para secuenciar DNA ó RNA. El sensor array debe estar libre de burbujas.

Secuenciación de amplicones con deleción: método nanoporos (ONT).

Los productos de PCR antes de la secuenciación fueron limpiados usando perlas AMPure XP (relación de perlas 1:1). Primero fue necesario reparar los extremos cohesivos de los amplicones utilizando el NEBNext Ultra II End repair / dA-tailing Module (E7546S). Después los amplicones ARTIC (~400 nt) son marcados usando el kit Native Barcoding Expansion (EXP-NBD114) los cuales son códigos de barras (secuencia de nucleótidos conocidas) que funcionan para identificar a las muestras. Además, se utilizó NEBNext® Ultra™ II Ligation Module (E7595S) para sintetizar los códigos de barras con los amplicones. Se usaron el Ligation Sequencing Kit (SQK-LSK109) y también el NEBNext® Quick Ligation Module (E6056S) que sirven para la ligación del material genético con los adaptadores que son reconocidos por el nanoporo. Para la preparación final de la librería se utilizaron el Sequencing Buffer (SQB) y el Loading Beads (LB), los cuales se encuentran en el Ligation Sequencing Kit (SQK-LSK109). Por último, utilizamos el Flow cell priming kit (EXP-FLP002)

para la configuración de la celda de secuenciación. Las cantidades de cada reactivo se utilizaron de acuerdo con el protocolo propuesto por Quick y las recomendaciones de Nanopore (*Ligation sequencing amplicons - native barcoding (SQK-LSK109 with EXP-NBD104 and EXP-NBD114)*, s/f; Quick, 2020).

Basecalling, filtrado y limpieza de adaptadores de secuencias de ONT.

Se realizó el Base-calling utilizando el software Guppy Basecalling Software, (C) Oxford nanopore Technologies, Limited. Versión 3.3.0+ef22818. Para conocer la calidad de los reads se utilizó el software FASTQC V0.11.5. Los filtrados y recortes de adaptadores se realizaron utilizando el software Porechop V0.2.4.

Secuenciación de amplicones con deleción: método Sanger.

Las muestras se secuenciaron en los Servicios Genómicos del Laboratorio Nacional de Genómica para la Biodiversidad. Se envió un *pool* de cada muestra a una concentración de 40 ng/ul (2.5 ul por reacción). Los electroferogramas fueron analizados con la ayuda de los softwares: BioEdit Sequence Alignment Editor versión 7.0.3 (Hall et al., 2011) y Finch TV versión 1.4.0 (Geospiza Inc. 2006).

Alineamientos al genoma de referencia de las secuencias obtenidas de ONT y Sanger.

Con el fin de verificar dónde alinean, las secuencias que fueron generadas por las plataformas de Nanopore y Sanger se alinearon usando el software MUSCLE V.3.8.31, usando los parámetros pre-configurados (Edgar, 2004). Para este alineamiento se utilizó el genoma de referencia NC_045512.2, el primer genoma completo de SARS-COV-2 detectado en Wuhan. Finalmente, el alineamiento de secuencias (Nanopore y Sanger) fue visualizado con el programa Jalview V2.11.2.2 (Waterhouse et al., 2009).

Secuenciación de un genoma completo de SARS-CoV-2 por ONT.

Se utilizaron dos *pools de primers*: el nCoV-2019_1, que contenía 110 oligos y el nCoV-2019_2, que contenía 108 oligos. En conjunto estos dos *pools* funcionan para amplificar todo el genoma de SARS-CoV-2. Estos oligos fueron donados por el CoViGen-Mex y estan disponibles en el protocolo de Quick V3 https://github.com/artic-network/artic-ncov2019/blob/master/primer_schemes/nCoV-2019/V3/nCoV-2019.tsv. Estos oligos fueron modificandose debido a las mutaciones que ocurrieron en el genoma de SARS-CoV-2 durante la pandemia. Los fragmentos que se generan durante la amplificación son de aproximadamente 400 nucleotidos que se sobrelapan para así cubrir todo el genoma. La PCR se realizó utilizando la Q5® High-Fidelity 2X Master Mix. Los procedimientos se realizaron según lo descrito por la red Artic (Quick, 2020).

Para la secuenciación se prepararon los amplicones que se generaron con el protocolo de ARTIC (~ 400 nt) usando el *kit Native Barcoding Expansion* (EXP-NBD114). Se usó el *kit Rapid Sequencing* (SQK-RAD004) para la preparación de la librería. Por último utilizamos el *Flow cell priming kit* (EXP-FLP002) para la configuración de la celda de secuenciación. Todos los kits fueron de ONT y se utilizaron de acuerdo con el protocolo del fabricante.

VII. RESULTADOS

Detección de una deleción de Δ411 nt en genomas de SARS-CoV-2 mediante PCR.

Durante la secuenciación para la vigilancia genómica de abril de 2021 el CoViGen-Mex discutió la falta de *reads* en las regiones de los ORFs 7b/8 en muestras ligadas al linaje B.1.243. La **Figura 7.1**, nos muestra alineamientos al genoma de referencia de *reads,* así como, la falta de *reads.* Al realizar un acercamiento notamos que la ausencia de *reads* inicia en la posición 27,825 y termina en la posición 28,233. Hasta ahora hemos explicado que cuando existe una deleción el tamaño del *read* se vuelve más corto y al menos un *read forward* o reverse debe mapear en dos regiones cercanas a la deleción. Otra característica de estos *reads* es que tienen que mapear antes y hasta la deleción, y el otro pedazo tiene que mapear inmediatamente después de la deleción.

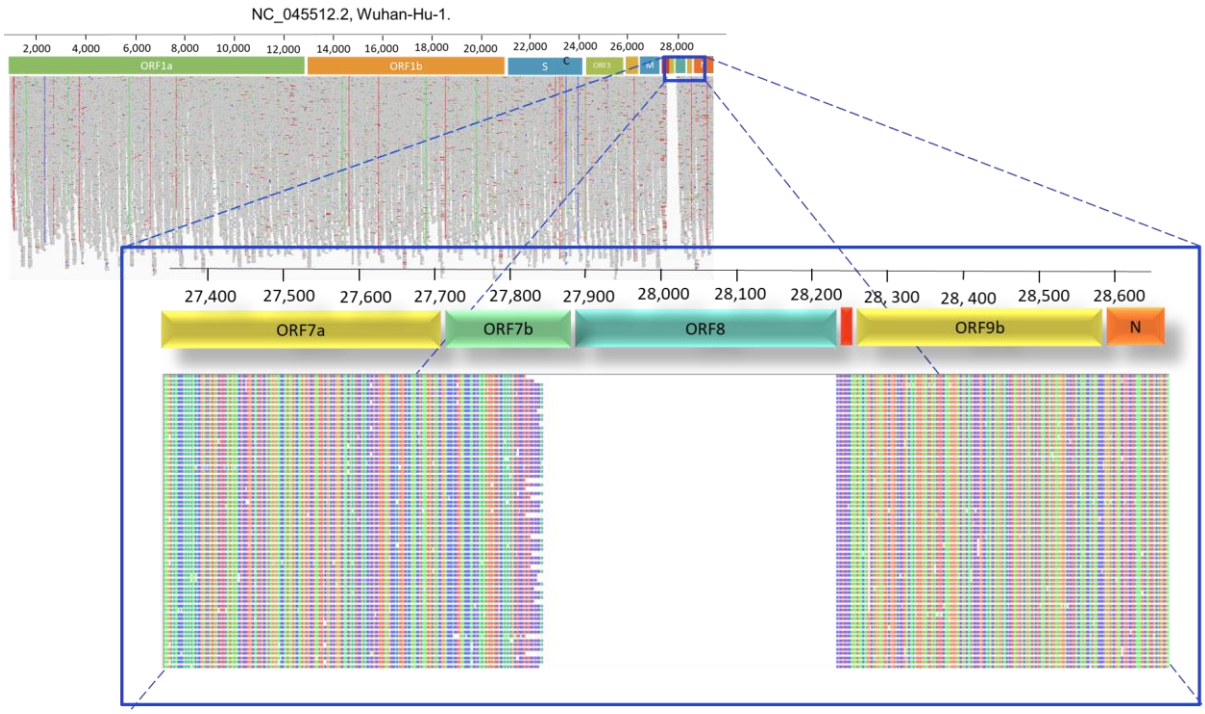


Figura 7.1. La muestra 590 mostró una deleción de 411 nt en el ORF7b/8
En la siguiente figura se observa los reads Illumina alineados al genoma de referencia. En la parte superior se observa la reconstrucción genética de SARS-CoV-2 y debajo los reads que alinean al genoma de referencia Wuhan-1. La imagen inferior nos muestra un acercamiento de las regiones de los ORFs 7b/8 y se observa una posible deleción de Δ411nt.

En una visualización manual de las muestras asignadas al linaje B.1.243 no se observaron reads en las regiones de los ORFs 7b/8 (visualización TABLET V.1.25) **Figura 7.1**. Estas muestras se analizaron con el software Indelseek que no logró mostrar a Δ411 (Au et al., 2017). Fue en entonces que decidimos investigar esta variante tanto experimental como

bioinformáticamente. Para ello, se tomó una muestra del linaje B.1.243 y se añadieron dos muestras correspondientes al linaje B.1.1.222 para utilizarlas como control negativo respecto a las deleciones.

Las mutaciones en SARS-CoV-2 por lo regular ocurren en unos pocos pares de bases (Singer et al., 2020). Sin embargo, datos publicados (noviembre, 2020) en GISAID sugieren una deleción de Δ411 nt en genoma de SARS-CoV-2, para verificar esta observación se sintetizaron dos oligos específicos que cubrieran la región ORF 7b y ORF 8. Estos oligos se seleccionaron a partir de los descritos en el protocolo ARTIC para secuenciar el genoma completo (Quick, 2020), *ver métodos*. El tamaño esperado del fragmento obtenido por PCR al amplificar las regiones (nCoV-2019_92_LEFT *forward* y nCoV-2019_93_RIGHT *reverse*) de los ORFs 7b/8 completos es de 680 nt, mientras que para los ORFs 7b/8 con deleción el tamaño del fragmento es de 280 nt **(Figura 7.2A-B).** Dada la explicación anterior, se espera que la muestra con deleción sea de menor tamaño **(Figura 7.2.C).**

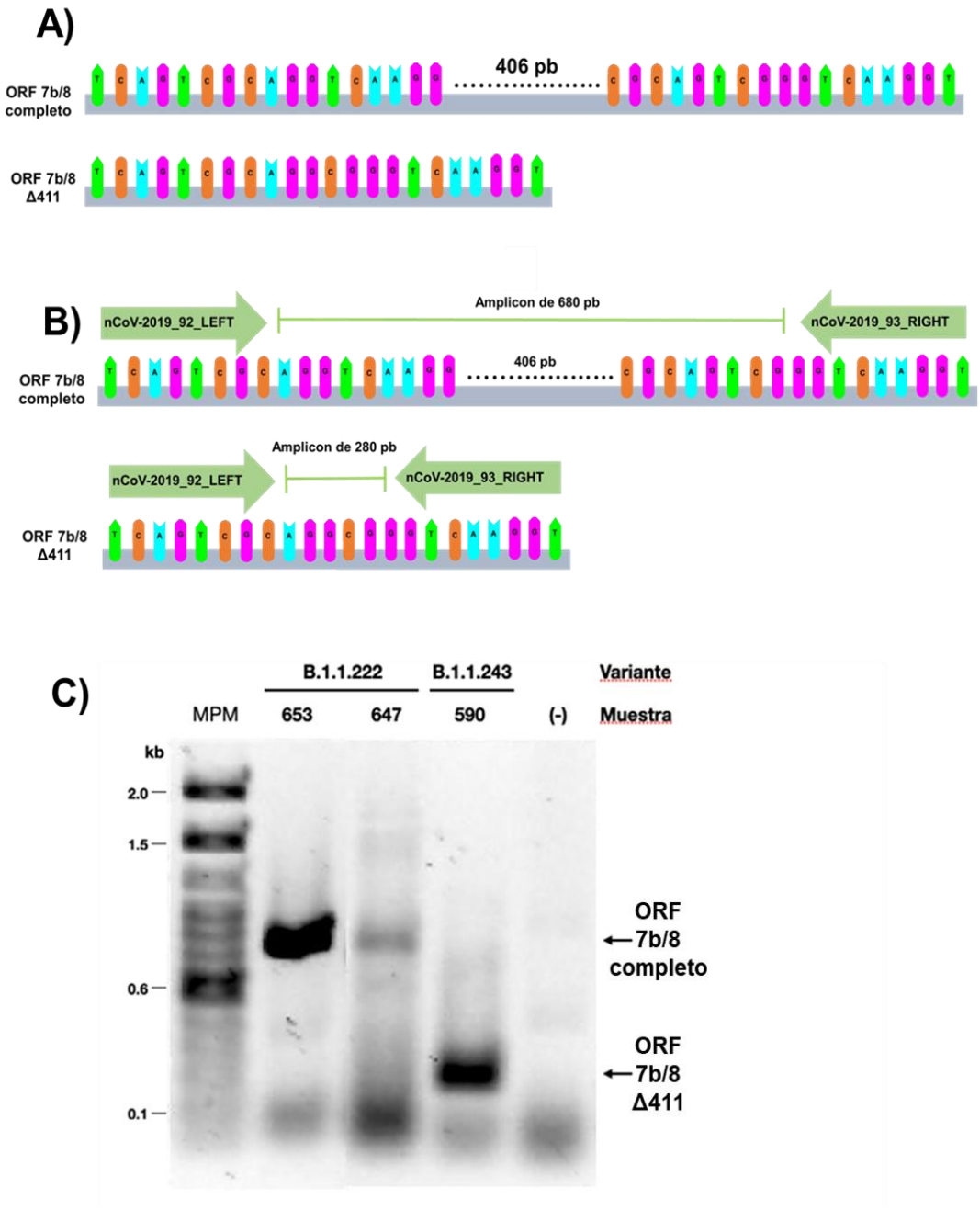


Figura 7.2. Confirmación de la deleción de Δ411 nt en SARS-CoV-2 mediante PCR. A) Esta figura es una simulación de las regiones genómicas de los ORFS 7b/8. El tamaño de la región con una deleción de 400 nt sería menor si es comparado con el tamaño de la región sin deleción**. B)** Se observa la simulación cuando los oligos sintetizan un amplicón de 680 nt de las regiones de los ORFs 7b/8, así mismo, se observa la síntesis de un amplicón de 280 nt cuando existe una deleción de 411 nt. **C)** Gel de agarosa al 1% teñido con Sybr Green. En el carril 1 podemos observar el marcador de peso molecular (MPM de 100 pb a 1kb). En el carril 2 y 3 son los controles positivos donde podemos observar los fragmentos de 680 nt para las muestras B.1.1.222 que no tienen deleción y en el carril 4 podemos observar el fragmento de 280 para una muestra asignada al linaje B.1.243 con deleción. En el carril 5 tenemos el control negativo.

En la visualización de los productos de PCR se puede observar que las muestras control tienen un tamaño aproximado de 600 nt, mientras que en la muestra (590) con deleción se observa un tamaño aproximado de 200 nt **(Figura 7.2.B)**. Esto sugiere una deleción en la muestra 590 que previamente había sido visualizada con Tablet **(Figura 7.1).** El tamaño esperado del amplicón para muestras con deleción (Linaje B.1.243) fue de 280 nt, mientras que para las muestras control (Linajes B.1.1.222) fue de 680 nt.

Confirmación de una deleción de Δ411 nt por secuenciación.

Los fragmentos obtenidos por PCR se secuenciaron por las tecnologías de ONT y de Sanger. Con la secuenciación de ONT se obtuvieron *reads* largos (~280 nt) que cubrieran esta región de Δ411 nt para verificar que correspondían a la deleción **(Figura 7.3)**. Mediante la secuenciación Sanger se obtuvo una única secuencia (~280 nt) de una muestra asignada al linaje B.1.243. Posteriormente, se realizó un alineamiento de las secuencias utilizando la herramienta BLAST y el software MUSCLE. Se observó que estas secuencias se alinearon en dos sitios distintos del genoma de referencia separados por 411 nt, la primera región de alineamiento corresponde a las posiciones 27,724-27,824 y la segunda a la región 28,239-28,465 Esta división de una única secuencia larga, como se mencionó en la introducción, sólo es posible si la muestra presenta una deleción de Δ411 nt **(Figura 7.3).** Por el método de secuenciación por nanoporos se obtuvieron 5 *reads* que pertenecen a la muestra 590 y que dos de ellas nos demostraron la división de una lectura en las mismas dos regiones distintas al genoma que las obtenidas por Sanger.

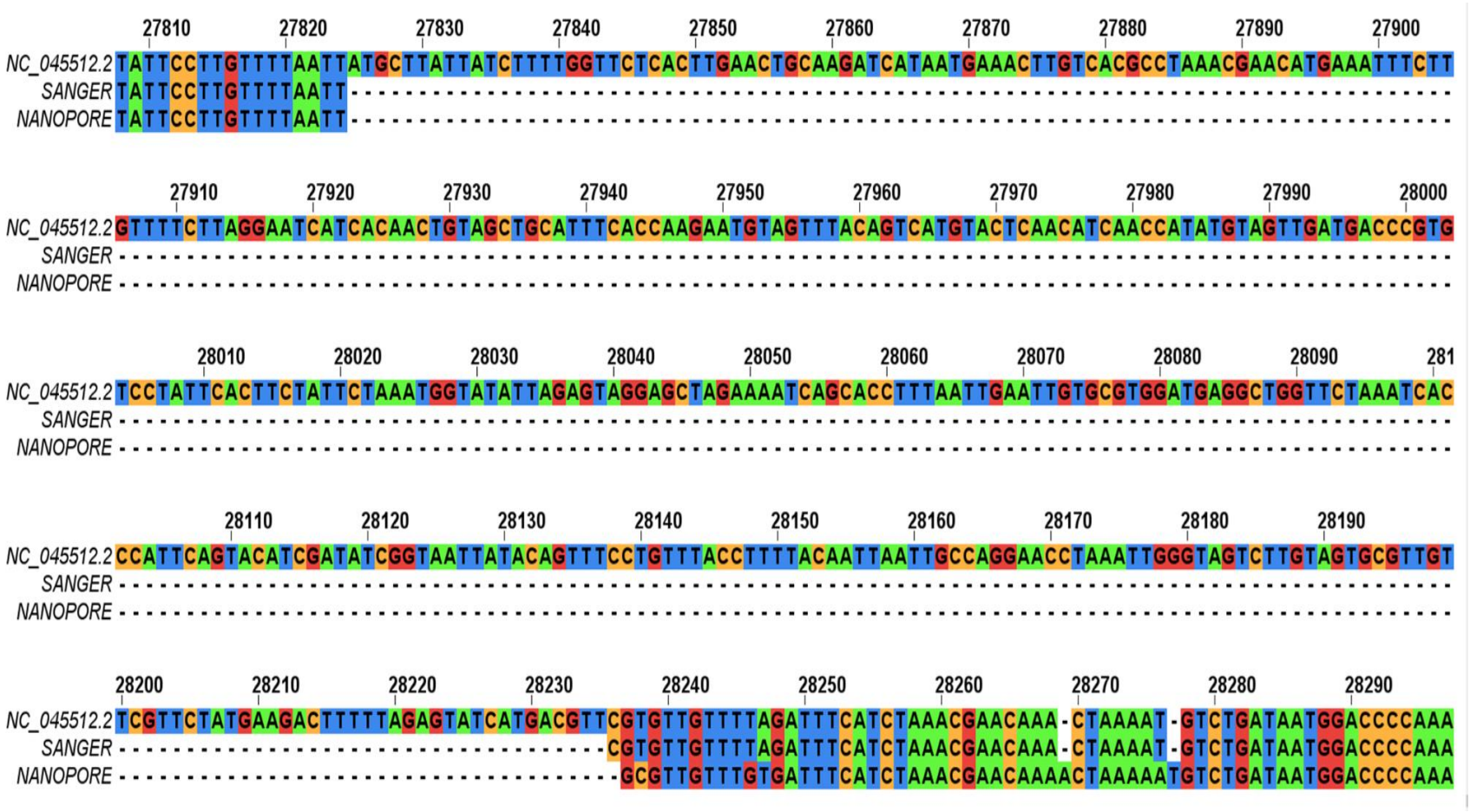


Figura 7.3. Confirmación de la deleción de Δ411 nt en SARS-CoV-2 usando secuenciación Sanger y Nanopore de la región de interés Nos muestra el resultado de un alineamiento por MUSCLE de una lectura de ONT y la secuenciación Sanger contra el genoma de referencia de Wuhan-1, donde podemos observar que las dos secuencias se dividen en dos regiones distintas al genoma de SARS-CoV-2. Esto solo es posible si existe una deleción.

Indel-Mex confirmó una una deleción de Δ411nt en muestras del linaje B.1.243.

El *pipeline* desarrollado fue capaz de detectar la deleción de Δ411 nt en muestras de SARS-CoV-2 asignadas al linaje B.1.243. Esta deleción fue detectada en los meses de abril, mayo y julio, donde se analizaron 384 muestras de cada mes. Después de analizar todo el año 2021 con Indel-Mex se demostró que las muestras con *Δ*411 nt en las regiones de los ORFs 7b/8 circularon los primeros dos meses (abril, mayo) y luego no volvieron a aparecer **(Figura 7.4A verde y azul)**. Sin embargo, Indel-Mex encontró muestras del linaje B.1.243 que sugiere una nueva deleción de 222 nt en SARS-CoV-2 **(Figura 7.4A, color salmón)**.

Para verificar molecularmente que existían otras muestras con deleción de Δ411 , se realizaron PCRs. Algunas muestras del linaje B.1.243 demostraron dos bandas tenues; una de tamaño aproximado de 600 nt; y la otra más o menos de 200 nt **(Figura 7.4B)**. Este resultado puede explicarse de dos maneras. Una primera posibilidad es la existencia simultánea de una variante con deleción (Ausencia de los ORFs 7b/8) y una variante sin deleción. La segunda posibilidad es una contaminación durante el proceso experimental. En el caso particular de la muestra 613 se encontró un *read* que alinea dentro de la deleción y 123 *reads* que alinean en ambas regiones que flanquean la frontera de la deleción. Aunque es posible la existencia de coinfecciones, es decir, la coexistencia en un hospedero de partículas virales de diferentes variantes, la baja abundancia de *reads* que alinean dentro de la deleción en la muestra 613 no nos permite discernir entre contaminación y la verdadera presencia de dos variantes en el paciente. En un caso similar de este linaje, el *pipeline* detectó en la muestra 413 cuatro *reads* alineando dentro de la región de la deleción y 719 *reads.* Estos datos nos sugieren que el hospedero podría albergar dos variantes (ORFs 7b/8 completos y Δ411). Finalmente, en la **Figura 7.4C** la cobertura de otras muestras del linaje B.1.243 nos enseña una deleción de Δ411 nt en los ORFs 7b/8, mientras que los controles positivos (B.1.1.222) no.

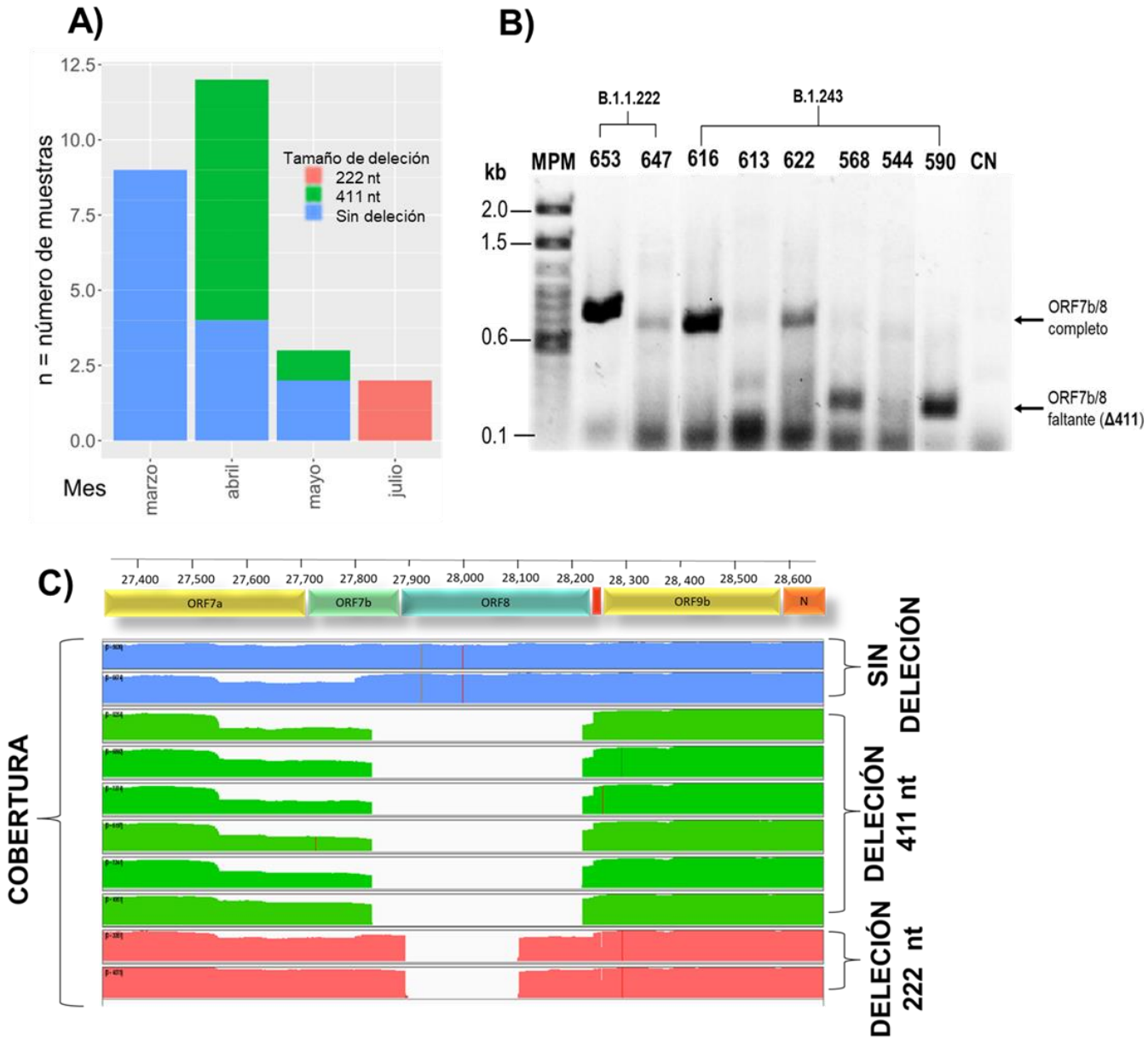


Figura 7.4. Confirmación bioinformática y molecular de la deleción de Δ411 nt en SARS-CoV-2 en otras muestras asignadas al linaje B.1.243. A) Muestras B.1.243 analizadas con Indel-Mex**.** Se observan las muestras con deleción que fueron encontradas para el linaje B.1.243. Así como, las muestras sin deleción (barras azules). En el eje de la X podemos observar el mes donde fueron detectados los linajes. En el eje de la Y podemos observar el número de muestras analizadas, las barras azules nos muestran los controles negativos (Linaje B.1.1.222). Las barras de colores salmon (*Δ*222 nt) y verde (Δ411 nt) nos muestran las diferentes deleciones detectadas. **B)** Gel de agarosa al 1% en este se observan los fragmentos de 680 nt para las muestras B.1.1.222 y B.1.243 que no tienen deleción y fragmentos de 280 para muestras B.1.243 que tienen una deleción. **C)** Cobertura de las muestras con deleción. En la parte superior se observa un acercamiento al genoma de SARS-CoV-2. En la parte inferior se encuentra la cobertura de cada muestra. Los colores verdes indican las muestras sin deleción, mientras que los colores verdes indican las muestras con una deleción de Δ411 nt y el color rosa indican las muestras con una deleción de 222 nt.

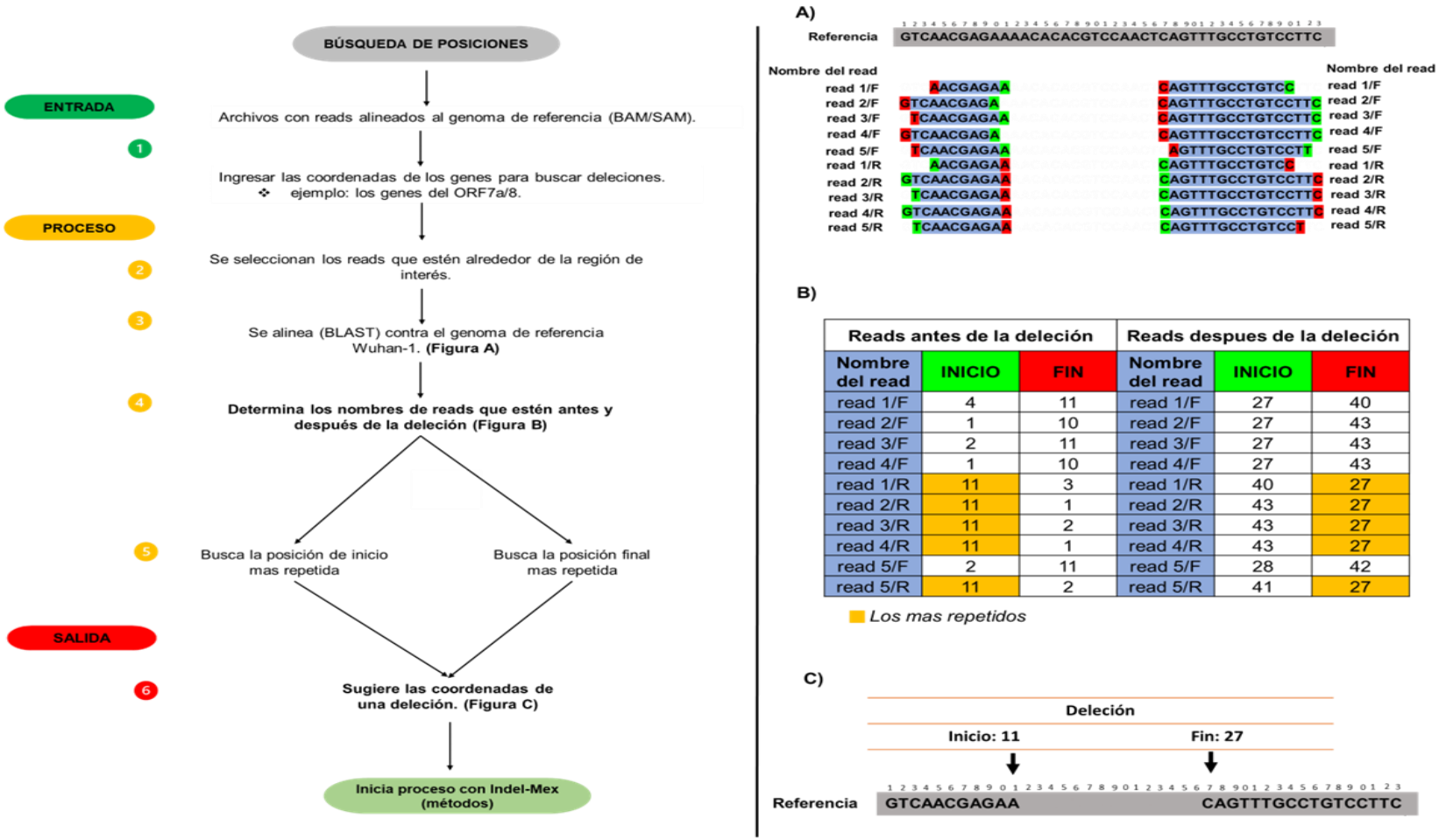
Indel-Mex detectó una deleción de Δ222 nt en los ORFs 7b/8.

Además de la confirmación de la deleción en 9 muestras en el mes de abril con una deleción de Δ411 nt, el *pipeline* detectó 2 genomas con una deleción de *Δ*222 en julio, asignadas al linaje B.1.243, **(Figura 7.4.A)**. Cabe mencionar que teníamos la hipótesis que todas las muestras asignadas al linaje B.1.243 podrían tener una deleción de Δ411 nt. Sin embargo, hay 15 muestras del B.1.243 que no tienen deleción, así como también otras muestras con una deleción de *Δ*222. Todas estas deleciones detectadas por Indel-Mex están en las regiones de los ORFs 7b/8 de genomas de SARS-CoV-2. La muestras que están asignadas al linaje B.1.1.222 se utilizaron como controles negativos en las cuales no se detectó *Δ*222 ni Δ411, **(Figura 7.4.A)**.

Indel-Mex encontró otras deleciones en otros linajes en los ORFS 7a/8.

Inicialmente solo se buscaba analizar la deleción de Δ411 nt, pero conforme se hacía este análisis, se reportaron deleciones en las regiones de los ORFs 7a/8 en diferentes linajes tales como A, B, Delta, B.1.1, entre otros (Gong et al., 2020; Mazur-Panasiuk et al., 2021; Panzera et al., 2021; Su et al., 2020). Después del primer análisis con Indel-Mex también se mostró una deleción en dos muestras de *Δ*222 asignadas al linaje B.1.243, **(Figura 7.4A)**. Este resultado nos llevó a generalizar el *pipeline* para que pudiera postular posibles posiciones de deleciones que se encuentren en la región de los ORFs 7a/8. Un caso particular fue que Indel-Mex al analizar la muestra 4216 se detectaron 1210 *reads* que alinearon en dos regiones del genoma y 102 *reads* que alinean dentro de la deleción. Lo cual podría apoyar que es posible la existencia de muestras tanto con ORFs 7a/8 completos y ORFs 7a/8 con deleción.

Para la búsqueda de estas posiciones en otras muestras, se agregó un análisis previo que puede ser visto en la **Figura 7.5.** Primero, se necesita indicar la región de interés, es decir unas coordenadas interés_inicio e interés_final en donde el *pipeline* realizará un barrido. En este caso la región de interés es la que comprende los ORFs 7a/8. Básicamente, el *pipeline* busca sin filtrar *reads* que alinean en dos regiones en toda muestra, **(Figura 7.5A)**. Luego, se enlistan las posiciones de alineamiento más repetidas en cada muestra, **(Figura 7.5B)**. Después de obtener las posiciones candidatas de las muestras, **(Figura 7.5C)**, se sigue el proceso explicado en la sección de métodos, **(Figura 6.3)** y se corre Indel-Mex en estas posiciones candidatas para identificar la presencia de deleciones, realizando de esta manera un barrido sobre toda la región que abarca de la posición 27,399 a la 28,259 (ORFS: 7a, 7b y 8)**.**



Reads antes de la deleción			Reads despues de la deleción		
Nombre del read	INICIO	FIN	Nombre del read	INICIO	FIN
read 1/F	4	11	read 1/F	27	40
read 2/F	1	10	read 2/F	27	43
read 3/F	2	11	read 3/F	27	43
read 4/F	1	10	read 4/F	27	43
read 1/R	11	3	read 1/R	40	27
read 2/R	11	1	read 2/R	43	27
read 3/R	11	2	read 3/R	43	27
read 4/R	11	1	read 4/R	43	27
read 5/F	2	11	read 5/F	28	42
read 5/R	11	2	read 5/R	41	27

Los mas repetidos

Figura 7.5. Proceso para la búsqueda de posiciones candidatas en muestras de SARS-CoV-2. A) Este diagrama nos muestra un proceso de 6 pasos que previamente es corrido para conocer posiciones iniciales y finales de una deleción. **B)** Podemos observar el BLAST que se hace con los reads hacia el genoma de referencia Wuhan-1. **C)** Este *pipeline* se basa en encontrar las posiciones de inicio y final más repetidas en la región de interés

Esta generalización del *pipeline* nos permite analizar todas las muestras que fueron secuenciadas a lo largo del 2021. Los resultados arrojaron que de abril a diciembre se detectaron al menos 1% de muestras totales con deleción de tamaños variados (**Figura 7.6A)**. Abril nos mostró más deleciones (5% de las muestras analizadas), mientras que en julio fue el porcentaje más bajo de los meses, apenas el 1.2% de las muestras analizadas **(Figura 7.6A)**. En marzo no se observaron deleciones, según nuestros criterios de filtración. Estos datos nos demuestran que en 9 de 10 meses analizados existieron deleciones y que estas deleciones se detectaron en gran parte de México, **(Figura 7.6 B)**.

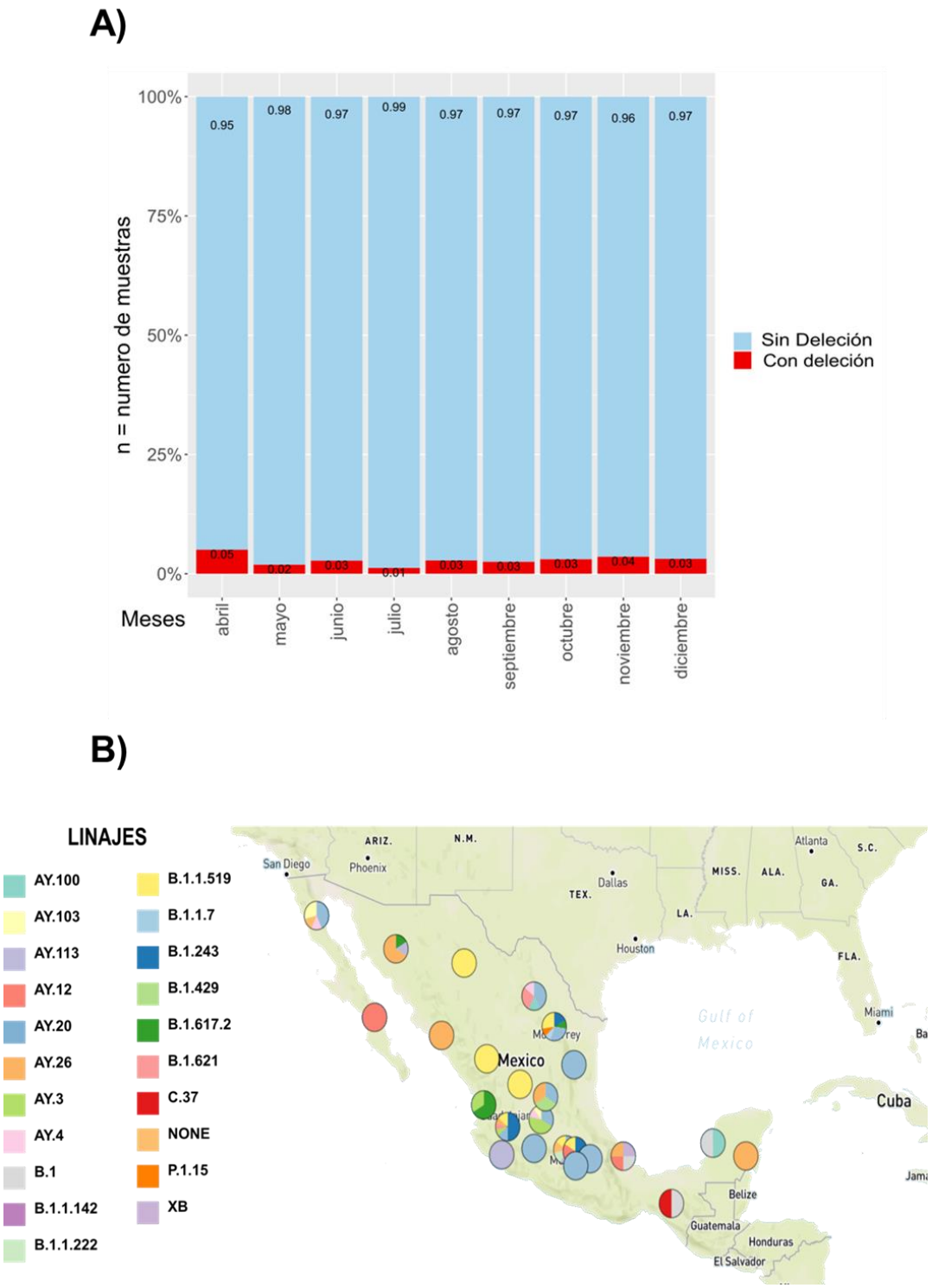


Figura 7.6. En el año 2021 se detectaron muestras con deleciones y origen geográfico de las muestras analizadas. A) Se observan los porcentajes de muestras que fueron analizadas en el 2021. El color rojo nos muestra el porcentaje de muestras con deleción positiva (no se distinguió entre linaje y tamaño) y el color azul claro nos dice el porcentaje de muestras sin deleción (no existen *reads* partidos). **B**) Se observan los linajes que circularon en el país de 97 muestras con deleción en los ORFs 7 y 8. Los círculos de color indican el linaje asignado a cada muestra.

Las deleciones surgieron de forma independiente.

Durante 2021, se detectaron diferentes linajes con una deleción en los ORFs 7b/8 del genoma de SARS-CoV-2, **(Figura 7.7B)**. El clado B fue el más frecuente en los meses de abril, mayo y junio. En julio y agosto el clado AY (Delta) desplazó a las otras variantes y para el resto de los meses la variante AY (Delta) fue la más prominente. En casi todo el año del 2021 surgieron deleciones en las regiones de los ORFs 7/8, no discriminado en función a los linajes, **(Figura 7.7B)**. Sin embargo, el tamaño de la deleción sí fue variado en los linajes, lo que indica que distintos eventos de deleciones tuvieron que haber ocurrido aún dentro del mismo linaje, **(Figura 7.7A).** El linaje AY.20 fue el que tuvo un mayor número de muestras con deleción, pero el tamaño de la deleción fue variado. Para el linaje B.1.1.243 no hay muestras asignadas con deleción, pero en julio esta deleción vuelve a reaparecer con una deleción de 222 nt para el linaje B.1.1.243, **(Figura 7.7A)**. Los linajes que no se muestran en esta figura se pueden observar en la parte de los *anexos* S1 y S2.

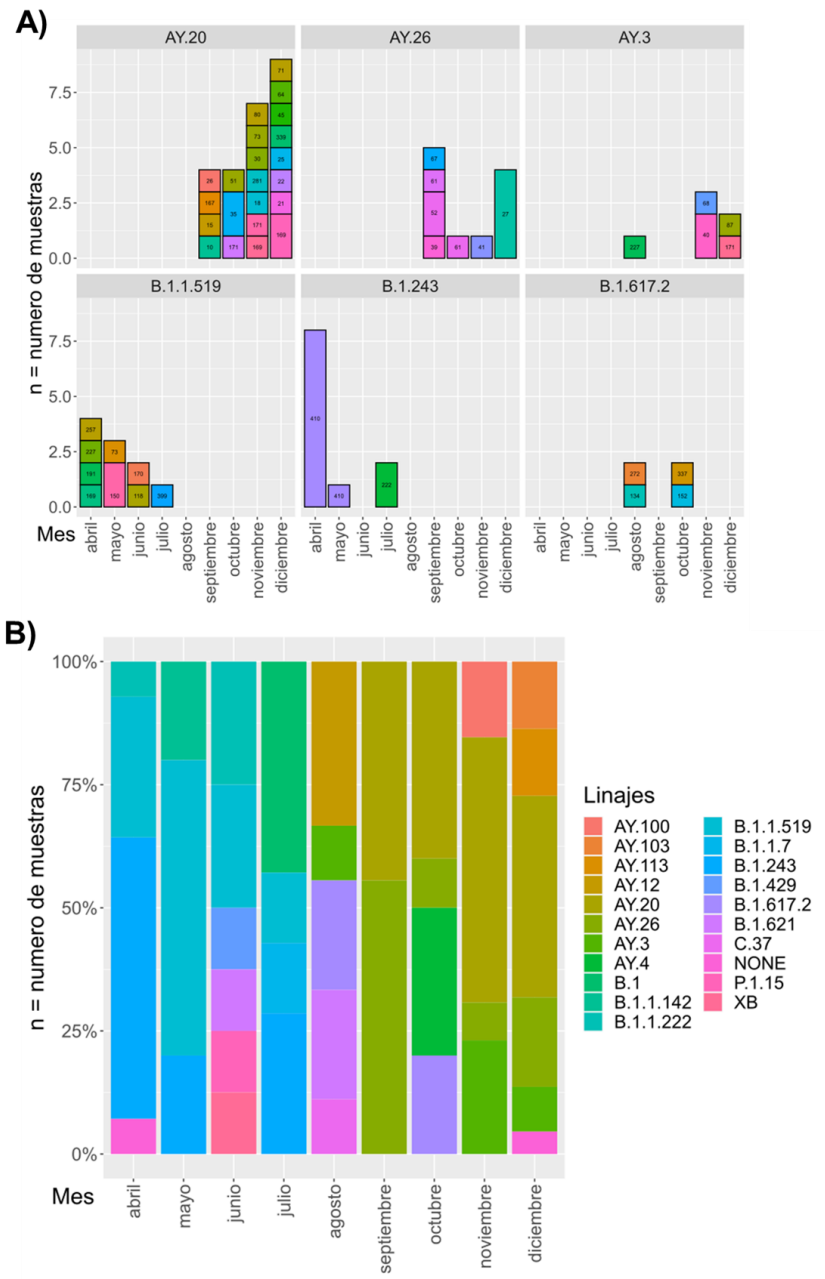


Figura 7.7. Deleciones y tamaño en las regiones de los ORFs 7a/7b/8 de SARS-CoV-2.
A) En el eje de las X podemos observar los meses y en el eje de las Y podemos observar el número de muestras. Los números que se encuentran en cada barra nos indican el tamaño de la delegación, así como el color de cada muestra nos dice el tamaño de la deleción. **B)** En el eje de las X podemos observar los meses y en el eje de las Y podemos observar el porcentaje de las muestras. Los colores de relleno nos indica el linaje que se detectaron con deleciones positivas en 2021.

Los datos sugieren que las deleciones largas en ORF7b/8 no tienen influencia en el estatus del paciente.

Nuestros datos sugieren que esta deleción es capaz de transmitirse de persona a persona. Observamos que tres muestras de SARS-CoV-2 con deleción podrían tener algún tipo de relación, es decir, podrían de ser de un núcleo muy cercano. Al analizar los metadatos las fechas de recolección coinciden entre ellas. También los genomas de estas muestras fueron

asignados a la misma variante (AY.20) y nos hace pensar que hubo transmisión de este virus entre personas. La deleción de *Δ*382 en la región del ORF8 se ha asociado a una enfermedad más leve y menos inflamación sistémica en pacientes con COVID-19 (Fong et al., 2022). Se han detectado diversas mutaciones de SARS-CoV-2 en la proteína del ORF8 las cuales se relacionan con una enfermedad más leve (Muth et al., 2018; Young et al., 2020) y una menor incidencia a hipoxia (Young et al., 2020). Para (Trieu & Trieu, 2021) estos datos sugerían que ORF8 es importante en la virulencia del virus.

Al realizar un análisis Fisher no encontramos relación significativa entre hospitalizaciones y presencia de la deleción. Los metadatos no muestran relevancia en la sintomatología y la edad de los hospederos. Aunque falta realizar un estudio estadístico más profundo que apoyé nuestra hipótesis. La investigación de Young et al. señalaba que las infecciones con deleciones estaban ligadas a tener menos hipoxia en pacientes con deleciones en el ORF8. Los metadatos disponibles en este experimento no son comparables con los de Young porque no se tiene el dato cuantitativo de medición de hipoxia, sino sólo la decisión médica hospitalizado vs ambulatorio. Por esta razón no podemos evaluar si los pacientes portadores de virus con deleción tienen menos hipoxia, pero nuestros resultados no sugieren enfermedad más leve con deleción entre nuestras muestras a diferencia de la observación de Young *et. al* donde muestran que dichas personas presentan síntomas más leves (Young et al., 2020). Las deleciones de SARS-CoV-2 en los ORF 7*a*/8 parecen no estar asociadas con el estatus de salud de los pacientes al nivel de hospitalizado/ambulatorio, **(Tabla 1).** Se realizó una prueba estadística de Fisher donde se obtuvo el p-valor de 0,6059. El resultado no es significativo para el umbral $p < 0,05$. El mes que tuvo menos casos de pacientes hospitalizados fue el mes de julio con el 14.29%. Los demás meses oscilaron entre 30-50% de pacientes hospitalizados, **(Figura 7.8).**

Tabla 1. Estatus del paciente con infecciones con deleción y sin deleción

Paciente	Con deleción	Sin deleción	Total
Hospitalizado	**46**	**1637**	**1683**
Ambulatorio	**51**	**2031**	**2082**
Total	**97**	**3668**	**3765**

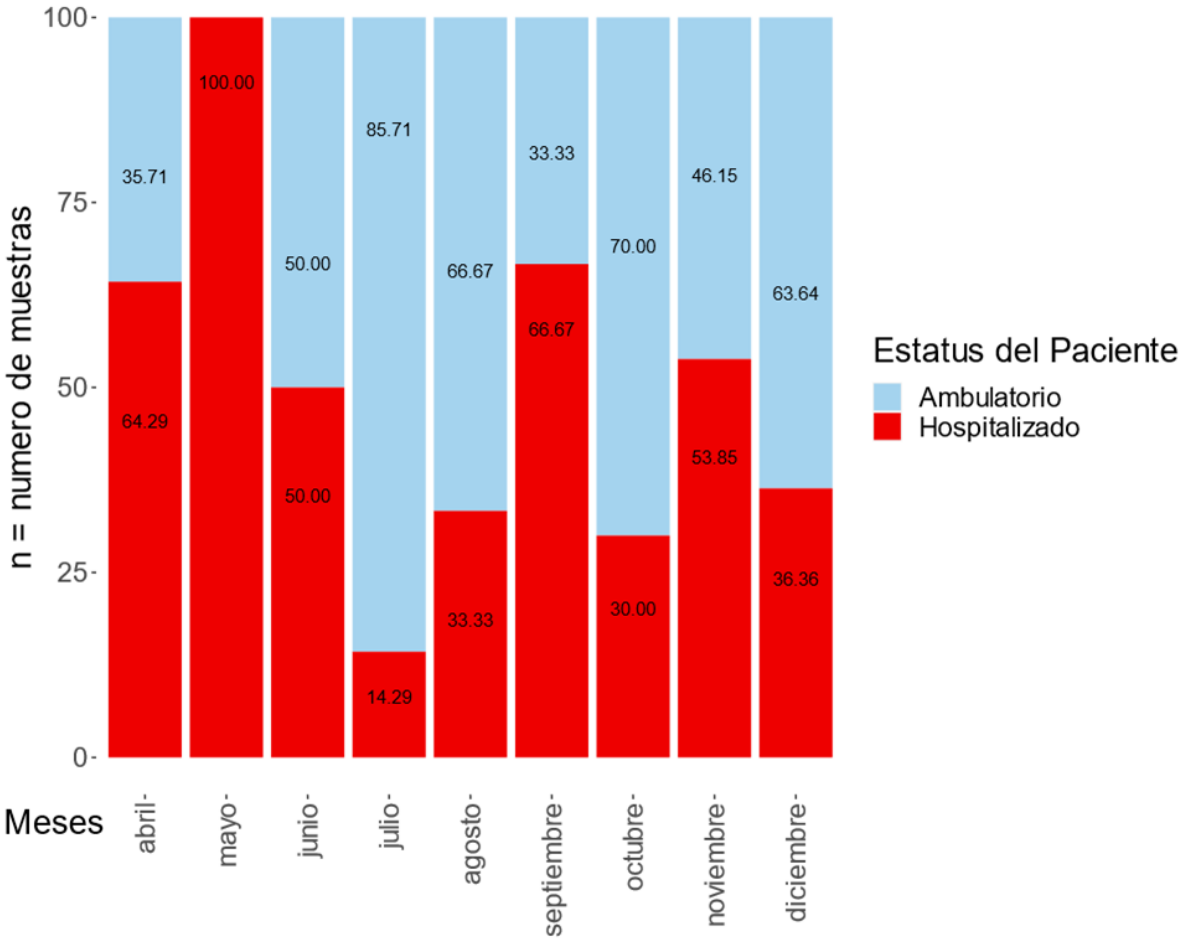


Figura 7.8. El estatus del paciente parece no estar influenciado por deleciones largas. Porcentajes del estatus de los pacientes infectados con variantes que poseen una deleción en el 2021. El color rojo nos muestra el porcentaje de pacientes hospitalizados y el color azul claro nos muestra el porcentaje de pacientes ambulatorios.

Secuenciación del genoma completo en tiempo real mediante la tecnología *Oxford Nanopore*.

La tecnología Nanopore se utilizó para secuenciar un genoma de SARS-CoV-2 con un total de 4,346 *reads.* El tiempo de corrida fue de 6 horas y el tamaño de los *reads* fue de 400 nt como se esperaría por la utilización de oligos ARTIC. En la **Figura 7.8,** se pueden observar los *reads* alineados al genoma de referencia de SARS-CoV-2. La profundidad promedio del genoma obtenido fue de 57X, es decir en promedio cada base es leída 57 veces y una cobertura que abarca 88.8% de todo el genoma. Utilizamos la técnica del protocolo ARTIC (Quick, 2020) para caracterizar genomas virales y es parte del programa de desarrollo de productos del startup BetterLab.

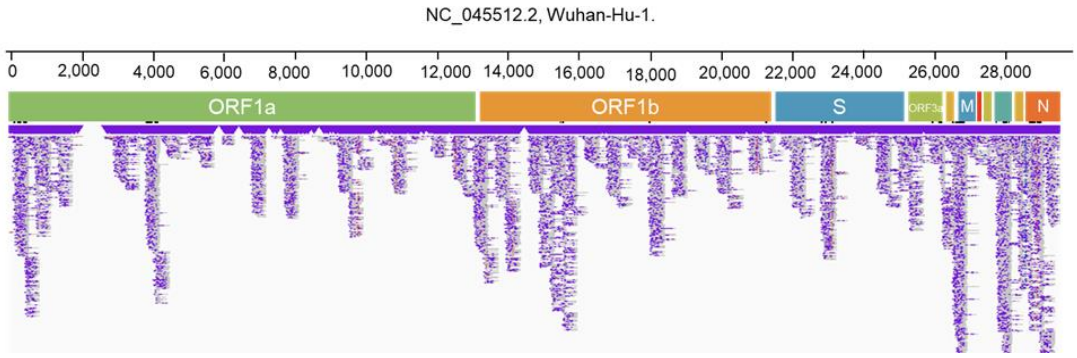


Figura 7.9. Genoma completo de SARS-CoV-2 secuenciado con la plataforma de Oxford Nanopore. En la parte superior se encuentra la numeración correspondiente al genoma de referencia Wuhan-1. Los genes de SARS-CoV-2 están representados en rectángulos de colores. En la parte inferior, de color morado, se muestran los reads obtenidas por la secuenciación por nanoporos, empalmadas y alineadas al genoma de referencia.

VIII. DISCUSIÓN

Se han detectado Indeles (inserciones y deleciones) en genomas de SARS-CoV-2 en los genes nsp6, S, ORF8 y N. En las secuencias recopiladas de todo el mundo se revelaron una cantidad notable de indeles en todo el genoma del virus. (Liu et al., 2021). En la región de los ORFs 7a/8 se han detectado largas deleciones >50 nt (Liu et al., 2021). Liu y colaboradores creen que la ocurrencia de estos tipos de Indel es múltiple e independiente (Liu et al., 2021). En el año 2018 se publicó una deleción de 29 nucleótidos (nt) en la región genómica del ORF8 en SARS-CoV (Muth et al., 2018). La deleción de 29 nt en el ORF8 estaba ausente en genomas de coronavirus cercanos a SARS-CoV aislados de muestras en murciélagos, pero presente en algunos SARS-CoV que infectan a los humanos (Muth et al., 2018). Este indicio apuntaba a que la falta de este fragmento del ORF8 conducía a la adaptación del SARS-CoV con humanos. Años más tarde, Su y colaboradores reportaron la presencia de una deleción de 382 nt en la región del ORF 7b/8 de SARS-CoV-2 (Su et al., 2020), la cual no fue persistente durante la pandemia del 2019. Así pues, se han presentado recurrentemente deleciones en los diferentes virus de SARS-coronavirus. Por esa razón en este trabajo se desarrolló un *pipeline* para detectar deleciones largas en la región 7a/8 de SARS-CoV-2 que permitiera identificarlas bioinformáticamente para el CoViGen-Mex.

Entre las primeras variantes que se estudiaron en el consorcio en relación a las deleciones está la B.1.243 gracias a la observación de la doctora Boukadida. En GISAID existían depositados genomas de esta variante provenientes de Arizona, USA por lo que posiblemente, esta fue la entrada de dicha variante a México (Skidmore et al., 2021). Cuando la deleción en este linaje fue observada en el CoviGen-Mx la variante mostraba un 3.52% de prevalencia en México (*MexCoV2*, s/f). Aún si Δ411 estuviera relacionada con un mejor *fitness* replicativo como sucede con Δ382 nt en el ORF8, es posible que la tasa de replicación de B.1.243 sea menor en comparación con la variante delta, ya que B.1.243 fue desplazada por ella.

Indel-Mex es capaz de detectar deleciones largas, siempre y cuando le ingresamos las coordenadas correctas de la región de interés. Es por ello que, para amplificar las posibilidades de Indel-Mex hicimos un *pipeline* que se aplica previamente, donde se barre una región en busca de una lista de posiciones que indiquen una posible deleción. Con estas dos herramientas, primero, se analizaron las muestras ligadas al linaje B.1.243 y se encontró que algunas de ellas contenían a Δ411, además dos muestras de julio nos demostraron una nueva *Δ*222 en SARS-CoV-2. Este resultado nos hace pensar que existen

deleciones largas son de tamaño distinto y que estas deleciones podrían surgir de manera independiente.

La secuenciación por Nanopore y Sanger confirmó la falta de los ORFs 7b/8 en muestras selectas del genoma de SARS-CoV-2, lo que constituye una deleción de Δ411 nt en el linaje B.1.243. Esta deleción elimina completamente el ORF7b y suprime parte del ORF8. Se han publicado otras deleciones del ORF8 de menor tamaño en Bangladesh (345 nucleótidos), Australia (138 nucleótidos) y España (62 nucleótidos) (Gong et al., 2020; Su et al., 2020; Young et al., 2020) detectadas entre Enero-Marzo 2020, mientras que en México, Δ411 se detectó durante el período Febrero-Marzo 2021, un año después de las anteriores en otras partes del mundo.

Cuando se realizó el análisis masivo de las muestras del 2021, logramos detectar 97 muestras con deleción en los ORFs 7a/7b/8 de SARS-CoV-2. Ninguna de estas deleciones estaba ligadas con un patrón en común, por ejemplo tamaño del deleción o linaje asignado, (**Figura 7.6)**. A lo largo del tiempo y en distintas partes del mundo se han detectado genomas de SARS-CoV-2 con deleciones en los ORFs 7a/7bb/8 donde se elimina gran parte de estas regiones. Los tamaños son variados y no tienen un origen en común, se podría pensar que de alguna manera favorecen al SARS-CoV-2, pero estas son desplazadas por otras variantes más prominentes, al menos en México se observa esta tendencia.

Este trabajo apoya a trabajos anteriores donde demuestran que existen múltiples deleciones en SARS-CoV-2 que surgieron en los ORFs 7a/8. Nuestro *pipeline* es capaz de analizar y detectar deleciones candidatas en regiones de los ORFs 7a/8. Como lo explicaron Mazur y colaboradores cualquier tamaño de la deleción puede conducir a un cambio estructural o funcional en la proteína. Estos cambios pueden verse reflejados en en el fenotipo del virus, infectividad, gravedad de la enfermedad o respuesta inmunitaria del huésped. Ellos detectaron una deleción de 872 nt donde se pierden 3 genes ORF7a/b y ORF8 en muestras ligadas a la variante Delta (Mazur-Panasiuk et al., 2021).

Es necesario hacer un estudio más profundo para conocer las consecuencias de estas deleciones en SARS-CoV-2 y su funcionamiento. Aún no se ha estudiado con profundidad la función que tienen los ORF 7, 8 en SARS-CoV-2, sin embargo, en SARS-CoV está mejor caracterizado. Y es que según los estudios que realizaron en 2007 encontraron que ORF8a mejora la replicación de manera eficiente. Así como, induce la apoptosis cuando se expresa ORF8a. Entonces ellos concluyen que ORF8a regula la producción de radicales libres y provoca la hiperpolarización del potencial de membrana. Este estudio concluye que el ORF8a desempeña un papel importante en la patogénesis de infección humana por SARS-CoV (Chen et al., 2007). Al final de la epidemia del 2003/2004 se describió una mutación de *Δ*29 nt en el ORF8 del genoma de SARS-CoV (SARS), un coronavirus cercano a SARS-CoV-2 (Muth et al., 2018). Se comprobó que esta deleción causa una atenuación en la tasa de replicación de SARS-CoV (Muth et al., 2018). En SARS-CoV-2 las deleciones de *Δ*382 nt detectadas en Singapur muestran mejor *fitness* replicativo *in vitro* que la variante de tipo silvestre (Su et al., 2020).

La evasión inmune es una de las características únicas del SARS-CoV-2 y se atribuye a la proteína ORF8 (Trieu & Trieu, 2021). La proteína accesoria ORF8 es una de las proteínas de betacoronavirus que evoluciona rápidamente (Trieu & Trieu, 2021). Sin embargo Indel-Mex ha logrado detectar otras deleciones de tamaños y en linajes distintos. Esto nos hace pensar que los ORFs 7a/8 podrían estarse adaptando al huésped, pero debido al *fitness* de otras variantes (como Delta) no ha sido posible fijarse. Estas mutaciones han surgido en el paso del 2020 y 2021, los cual era de esperarse, por la actual pandemia que se vive. Sin embargo, estas deleciones largas siguen surgiendo.

Aunque no se sabe concretamente la función de los ORFs 7a/8. Algunos datos muestran que los pacientes con la deleción de 382 nt en el ORF8 mostraban menos probabilidad de mostrar hipoxia que los infectados con la variante silvestre (Young et al., 2020). Otro estudio encontró una interacción de la proteína viral de SARS-CoV-2 codificada por ORF8 con el

Complejo mayor de histocompatibilidad de clase I del huésped (MHC-I). Las células que expresan el ORF8 degradan selectivamente a MHC-I y esta disminución atenúa la actividad antiviral de los linfocitos T CD8 citotóxicos (Zhang et al., 2020). Finalmente, en cuanto a pruebas diagnósticas serológicas, un estudio sugiere que la respuesta inmune a la presencia conjunta de ORF8 y ORF3 identifica un 96.5% de muestras con SARS-CoV-2 tanto en etapas tempranas como tardías de la infección (Hachim et al., 2020). La pérdida del ORF8 reduciría la efectividad de estas pruebas para el linaje B.1.243 y otros linajes. Una vez descrita la importancia del ORF8 pensamos que se requiere más investigación para entender el rol de las deleciones que aparecen en las regiones del ORF7a/8 en SARS-CoV-2, poniendo especial énfasis en la predicción de la proteína que se genera cuando existe esta deleción para saber si hay algún cambio fenotípico en el virus (Su et al., 2020).

Finalmente, en el análisis de los resultados de PCR encontramos que en algunas muestras se podían observar dos bandas tenues donde el tamaño de las bandas corresponde a la existencia de dos tipos de fragmento diferentes en una misma muestra, posiblemente provenientes de dos variantes, una con deleción y otra sin deleción. Indel-Mex detectó en algunas muestras *reads* que alinean justo dentro de la deleción, esta podría ser otra prueba a favor de la posible existencia de dos variantes. Creemos que Indel-Mex podría incrementar su utilidad si fuera capaz de detectar la presencia de dos variantes. Se deben realizar análisis para detectar variantes *intrahost* y conocer si estos pacientes tienen infecciones mixtas. Sería muy interesante conocer si existe alguna interacción entre las variantes que beneficie o perjudique el desarrollo del virus en el hospedero. Por otro lado, se tienen que realizar estudios filogenéticos más exhaustivos para que así podamos conocer el origen de esta deleción. Frecuentemente en GISAID, se publican genomas que tienen alguna deleción en las regiones que se estudiaron. Aún sigue abierta la pregunta ¿Por qué frecuentemente ocurren diferentes deleciones en los ORFs 7a/8 de SARS-CoV-2?.

IX. CONCLUSIONES.

- Se confirmó por PCR, secuenciación y bioinformática la presencia de una deleción de Δ411 nt en los ORFs 7b/8 del genoma de SARS-CoV-2.
- Se creó un *pipeline* capaz de identificar deleciones largas en muestras secuenciadas por Illumina de genomas de SARS-CoV-2.
- Las deleciones en la región ORF 7a/8 surgen de forma independiente.
- Los SARS-CoV-2 con deleción en los ORFs 7b/8 son capaces de infectar a su hospedero.

X. PERSPECTIVAS

Indel-Mex puede detectar una deleción larga, pero no puede detectar dos deleciones en la misma región genómica. Se detectaron 3 muestras (1487, 1488 y 633) de SARS-CoV-2 asignadas al linaje B.1.1.222, estas muestras mostraban una posible deleción de 346 nt en la región del ORF7b/8. Sin embargo, al ser visualizada en *Tablet*, notamos que sí estaba la deleción, pero había otra de menor tamaño, esta deleción menor no fue detectada por Indel-Mex. Sin duda, es necesario que Indel-Mex pueda expandir sus opciones para que pueda buscar deleciones en todo el genoma de SARS-CoV-2.

Por otro lado, Indel-Mex es una herramienta poderosa, la cual puede ser mejorada para detectar deleciones en otras regiones genómicas. Por ahora falta determinar si estas deleciones en los ORFs 7a/8 siguen apareciendo en lo que resta de la pandemia.

XI. BIBLIOGRAFÍA

Au, C. H., Leung, A. Y. H., Kwong, A., Chan, T. L., & Ma, E. S. K. (2017). INDELseek: Detection of complex insertions and deletions from next-generation sequencing data. *BMC Genomics*, *18*(1), 16. https://doi.org/10.1186/s12864-016-3449-9

CDC. (2020, febrero 11). *Coronavirus Disease 2019 (COVID-19)*. Centers for Disease Control and Prevention. https://www.cdc.gov/coronavirus/2019-ncov/variants/about-variants.html

Chen, C.-Y., Ping, Y.-H., Lee, H.-C., Chen, K.-H., Lee, Y.-M., Chan, Y.-J., Lien, T.-C., Jap, T.-S., Lin, C.-H., Kao, L.-S., & Chen, Y.-M. A. (2007). Open Reading Frame 8a of the Human Severe Acute Respiratory Syndrome Coronavirus Not Only Promotes Viral Replication but Also Induces Apoptosis. *The Journal of Infectious Diseases*, *196*(3), 405–415. https://doi.org/10.1086/519166

COVID-19: Cronología de la actuación de la OMS. (s/f). Recuperado el 21 de junio de 2022, de https://www.who.int/es/news/item/27-04-2020-who-timeline---covid-19

COVID-19 Tablero México. (s/f). COVID - 19 Tablero México. Recuperado el 26 de mayo de 2022, de https://datos.covid-19.conacyt.mx/index.php

Danecek, P., Bonfield, J. K., Liddle, J., Marshall, J., Ohan, V., Pollard, M. O., Whitwham, A., Keane, T., McCarthy, S. A., Davies, R. M., & Li, H. (2021). Twelve years of SAMtools and BCFtools. *GigaScience*, *10*(2), giab008. https://doi.org/10.1093/gigascience/giab008

Edgar, R. (2004). MUSCLE: a multiple sequence alignment method with reduced time and space complexity BMC Bioinformatics 5: 113. *Google Scholar There is no corresponding record for this reference*.

Faria, N. R., Sabino, E. C., Nunes, M. R., Alcantara, L. C. J., Loman, N. J., & Pybus, O. G. (2016). Mobile real-time surveillance of Zika virus in Brazil. *Genome medicine*, *8*(1), 1–4.

Flow Cell Wash Kit (EXP-WSH004). (s/f). Nanopore Community. Recuperado el 3 de junio de 2022, de https://community.nanoporetech.com/protocols/flow-cell-wash-kit-exp-wsh004/v/wfc_9120_v1_revd_08dec2020

Fong, S.-W., Yeo, N. K.-W., Chan, Y.-H., Goh, Y. S., Amrun, S. N., Ang, N., Rajapakse, M. P., Lum, J., Foo, S., Lee, C. Y.-P., Carissimo, G., Chee, R. S.-L., Torres-Ruesta, A., Tay, M. Z., Chang, Z. W., Poh, C. M., Young, B. E., Tambyah, P. A., Kalimuddin, S., … Ng, L. F. P. (2022). Robust Virus-Specific Adaptive Immunity in COVID-19 Patients with SARS-CoV-2 Δ382 Variant Infection. *Journal of Clinical Immunology*, *42*(2), 214–229. https://doi.org/10.1007/s10875-021-01142-z

Gong, Y.-N., Tsao, K.-C., Hsiao, M.-J., Huang, C.-G., Huang, P.-N., Huang, P.-W., Lee, K.-

M., Liu, Y.-C., Yang, S.-L., & Kuo, R.-L. (2020). SARS-CoV-2 genomic surveillance in Taiwan revealed novel ORF8-deletion mutant and clade possibly associated with infections in Middle East. *Emerging microbes & infections*, *9*(1), 1457–1466.

Grubaugh, N. D., Gangavarapu, K., Quick, J., Matteson, N. L., Jesus, J. G. D., Main, B. J., Tan, A. L., Paul, L. M., Brackney, D. E., Grewal, S., Gurfield, N., Rompay, K. K. V., Isern, S., Michael, S. F., Coffey, L. L., Loman, N. J., & Andersen, K. G. (2018). *An amplicon-based sequencing framework for accurately measuring intrahost virus diversity using PrimalSeq and iVar* (p. 383513). bioRxiv. https://doi.org/10.1101/383513

Hachim, A., Kavian, N., Cohen, C. A., Chin, A. W. H., Chu, D. K. W., Mok, C. K. P., Tsang, O. T. Y., Yeung, Y. C., Perera, R. A. P. M., Poon, L. L. M., Peiris, J. S. M., & Valkenburg, S. A. (2020). ORF8 and ORF3b antibodies are accurate serological markers of early and late SARS-CoV-2 infection. *Nature Immunology*, *21*(10), 1293–1301. https://doi.org/10.1038/s41590-020-0773-7

Hadfield, J., Megill, C., Bell, S. M., Huddleston, J., Potter, B., Callender, C., Sagulenko, P., Bedford, T., & Neher, R. A. (2018). Nextstrain: Real-time tracking of pathogen evolution. *Bioinformatics*, *34*(23), 4121–4123. https://doi.org/10.1093/bioinformatics/bty407

Hall, T., Biosciences, I., & Carlsbad, C. (2011). BioEdit: An important software for molecular biology. *GERF Bull Biosci*, *2*(1), 60–61.

Heather, J. M., & Chain, B. (2016). The sequence of sequencers: The history of sequencing DNA. *Genomics*, *107*(1), 1–8. https://doi.org/10.1016/j.ygeno.2015.11.003

How bad is Omicron? Some clues are emerging, and they're not encouraging. (2021). [Data set]. https://doi.org/10.1126/science.acx9789

Langa, L. S., Sallent, L. V., & Díez, S. R. (2021). Interpretación de las pruebas diagnósticas de la COVID-19. *Fmc*, *28*(3), 167–173. https://doi.org/10.1016/j.fmc.2021.01.005

Langmead, B., & Salzberg, S. L. (2012). Fast gapped-read alignment with Bowtie 2. *Nature Methods*, *9*(4), 357–359. https://doi.org/10.1038/nmeth.1923

Li, W., & Godzik, A. (2006). Cd-hit: A fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics*, *22*(13), 1658–1659.

Ligation sequencing amplicons—Native barcoding (SQK-LSK109 with EXP-NBD104 and EXP-NBD114). (s/f). Nanopore Community. Recuperado el 3 de junio de 2022, de https://community.nanoporetech.com/protocols/native-barcoding-amplicons/v/nba_9093_v109_revk_12nov2019

Ligation sequencing gDNA - Lambda control (SQK-LSK109). (s/f). Nanopore Community. Recuperado el 3 de junio de 2022, de https://community.nanoporetech.com/protocols/lambda-control-sqk-

lsk109/v/cde_9062_v109_revag_14aug2019

Liu, X., Guo, L., Xu, T., Lu, X., Ma, M., Sheng, W., Wu, Y., Peng, H., Cao, L., Zheng, F., Huang, S., Yang, Z., Du, J., Shi, M., & Guo, D. (2021). A comprehensive evolutionary and epidemiological characterization of insertion and deletion mutations in SARS-CoV-2 genomes. *Virus Evolution*, *7*(2), veab104. https://doi.org/10.1093/ve/veab104

Lu, H., Giordano, F., & Ning, Z. (2016). Oxford Nanopore MinION Sequencing and Genome Assembly. *Genomics, Proteomics & Bioinformatics*, *14*(5), 265–279. PubMed. https://doi.org/10.1016/j.gpb.2016.05.004

Lu, R., Zhao, X., Li, J., Niu, P., Yang, B., Wu, H., Wang, W., Song, H., Huang, B., Zhu, N., Bi, Y., Ma, X., Zhan, F., Wang, L., Hu, T., Zhou, H., Hu, Z., Zhou, W., Zhao, L., … Tan, W. (2020). Genomic characterisation and epidemiology of 2019 novel coronavirus: Implications for virus origins and receptor binding. *Lancet (London, England)*, *395*(10224), 565–574. https://doi.org/10.1016/S0140-6736(20)30251-8

Marco abierto de lectura | *NHGRI*. (s/f). Genome.gov. Recuperado el 18 de agosto de 2021, de https://www.genome.gov/es/genetics-glossary/Marco-abierto-de-lectura

Martín, J. M. V., Ortigosa, F., & Pendon, R. A. C. (2020). Métodos de secuenciación: Segunda generación. *Encuentros en la Biología*, *13*(174), 17–23.

Mazur-Panasiuk, N., Rabalski, L., Gromowski, T., Nowicki, G., Kowalski, M., Wydmanski, W., Szulc, P., Kosinski, M., Gackowska, K., Drweska-Matelska, N., Grabowski, J., Piotrowska-Mietelska, A., Szewczyk, B., Bienkowska-Szewczyk, K., Swadzba, J., Labaj, P., Grzybek, M., & Pyrc, K. (2021). Expansion of a SARS-CoV-2 Delta variant with an 872 nt deletion encompassing ORF7a, ORF7b and ORF8, Poland, July to August 2021. *Eurosurveillance*, *26*(39), 2100902. https://doi.org/10.2807/1560-7917.ES.2021.26.39.2100902

MexCoV2. (s/f). Recuperado el 16 de junio de 2021, de http://mexcov2.ibt.unam.mx:8080/COVID-TRACKER/

Muth, D., Corman, V. M., Roth, H., Binger, T., Dijkman, R., Gottula, L. T., Gloza-Rausch, F., Balboni, A., Battilani, M., Rihtarič, D., Toplak, I., Ameneiros, R. S., Pfeifer, A., Thiel, V., Drexler, J. F., Müller, M. A., & Drosten, C. (2018). Attenuation of replication by a 29 nucleotide deletion in SARS-coronavirus acquired during the early stages of human-to-human transmission. *Scientific Reports*, *8*(1), 15177–15177. PubMed. https://doi.org/10.1038/s41598-018-33487-8

Next-Generation Sequencing (NGS) | *Explore the technology*. (s/f). Recuperado el 18 de agosto de 2021, de https://www.illumina.com/science/technology/next-generation-sequencing.html

Nucleic Acid Amplification Tests (NAATs) | *CDC*. (s/f). Recuperado el 27 de mayo de 2022, de https://www.cdc.gov/coronavirus/2019-ncov/lab/naats.html

Pal, M., Berhanu, G., Desalegn, C., & Kandi, V. (2020). Severe Acute Respiratory Syndrome Coronavirus-2 (SARS-CoV-2): An Update. *Cureus*, *12*(3), e7423–e7423. PubMed. https://doi.org/10.7759/cureus.7423

Panzera, Y., Ramos, N., Frabasile, S., Calleros, L., Marandino, A., Tomás, G., Techera, C., Grecco, S., Fuques, E., Goñi, N., Ramas, V., Coppola, L., Chiparelli, H., Sorhouet, C., Mogdasy, C., Arbiza, J., Delfraro, A., & Pérez, R. (2021). A deletion in SARS-CoV-2 ORF7 identified in COVID-19 outbreak in Uruguay. *Transboundary and Emerging Diseases*, 10.1111/tbed.14002. https://doi.org/10.1111/tbed.14002

Peacock, T. P., Sheppard, C. M., Brown, J. C., Goonawardane, N., Zhou, J., Whiteley, M., de Silva, T. I., Barclay, W. S., & PHE Virology Consortium. (2021). The SARS-CoV-2 variants associated with infections in India, B. 1.617, show enhanced spike cleavage by furin. *BioRxiv*.

Prajapat, M., Sarma, P., Shekhar, N., Avti, P., Sinha, S., Kaur, H., Kumar, S., Bhattacharyya, A., Kumar, H., Bansal, S., & Medhi, B. (2020). Drug targets for corona virus: A systematic review. *Indian Journal of Pharmacology*, *52*(1), 56–65. PubMed. https://doi.org/10.4103/ijp.IJP_115_20

Quick, J. (2020). NCoV-2019 sequencing protocol. *protocols.io*.

Raman, R., Patel, K. J., & Ranjan, K. (2021). COVID-19: Unmasking Emerging SARS-CoV-2 Variants, Vaccines and Therapeutic Strategies. *Biomolecules*, *11*(7), 993. https://doi.org/10.3390/biom11070993

Resende, P. C., Motta, F. C., Roy, S., Appolinario, L., Fabri, A., Xavier, J., Harris, K., Matos, A. R., Caetano, B., & Orgeswalska, M. (2020). *SARS-CoV-2 genomes recovered by long amplicon tiling multiplex approach using nanopore sequencing and applicable to other sequencing platforms*.

Rossi, G. A., Sacco, O., Mancino, E., Cristiani, L., & Midulla, F. (2020). Differences and similarities between SARS-CoV and SARS-CoV-2: Spike receptor-binding domain recognition and host cell infection with support of cellular serine proteases. *Infection*, 1–5.

Sethuraman, N., Jeremiah, S. S., & Ryo, A. (2020). Interpreting Diagnostic Tests for SARS-CoV-2. *JAMA*, *323*(22), 2249–2251. https://doi.org/10.1001/jama.2020.8259

Shereen, M. A., Khan, S., Kazmi, A., Bashir, N., & Siddique, R. (2020). COVID-19 infection: Origin, transmission, and characteristics of human coronaviruses. *Journal of Advanced Research*.

Singer, J., Gifford, R., Cotten, M., & Robertson, D. (2020). *CoV-GLUE: A Web Application for Tracking SARS-CoV-2 Genomic Variation*. https://doi.org/10.20944/preprints202006.0225.v1

Skidmore, P. T., Kaelin, E. A., Holland, L. A., Maqsood, R., Wu, L. I., Mellor, N. J., Blain, J.

M., Harris, V., LaBaer, J., Murugan, V., & Lim, E. S. (2021). Genomic Sequencing of SARS-CoV-2 E484K Variant B.1.243.1, Arizona, USA. *Emerging Infectious Diseases*, *27*(10), 2718–2720. https://doi.org/10.3201/eid2710.211189

Spike E484K mutation in the first SARS-CoV-2 reinfection case confirmed in Brazil, 2020—SARS-CoV-2 coronavirus / nCoV-2019 Genomic Epidemiology. (2021, enero 10). Virological. https://virological.org/t/spike-e484k-mutation-in-the-first-sars-cov-2-reinfection-case-confirmed-in-brazil-2020/584

Su, Y. C., Anderson, D. E., Young, B. E., Linster, M., Zhu, F., Jayakumar, J., Zhuang, Y., Kalimuddin, S., Low, J. G., & Tan, C. W. (2020). Discovery and genomic characterization of a 382-nucleotide deletion in ORF7b and ORF8 during the early evolution of SARS-CoV-2. *MBio*, *11*(4), e01610-20.

Sung, S.-C., Chao, C.-Y., Jeng, K.-S., Yang, J.-Y., & Lai, M. M. C. (2009). The 8ab protein of SARS-CoV is a luminal ER membrane-associated protein and induces the activation of ATF6. *Virology*, *387*(2), 402–413. https://doi.org/10.1016/j.virol.2009.02.021

Taboada, B., Vazquez-Perez, J. A., Muñoz-Medina, J. E., Ramos-Cervantes, P., Escalera-Zamudio, M., Boukadida, C., Sanchez-Flores, A., Isa, P., Mendieta-Condado, E., Martínez-Orozco, J. A., Becerril-Vargas, E., Salas-Hernández, J., Grande, R., González-Torres, C., Gaytán-Cervantes, F. J., Vazquez, G., Pulido, F., Araiza-Rodríguez, A., Garcés-Ayala, F., … Arias, C. F. (2020). Genomic Analysis of Early SARS-CoV-2 Variants Introduced in Mexico. *Journal of Virology*, *94*(18), e01056-20. https://doi.org/10.1128/JVI.01056-20

Taboada, B., Zárate, S., Iša, P., Boukadida, C., Vazquez-Perez, J. A., Muñoz-Medina, J. E., Ramírez-González, J. E., Comas-García, A., Grajales-Muñiz, C., Rincón-Rubio, A., Matías-Florentino, M., Sanchez-Flores, A., Mendieta-Condado, E., Verleyen, J., Barrera-Badillo, G., Hernández-Rivas, L., Mejía-Nepomuceno, F., Martínez-Orozco, J. A., Becerril-Vargas, E., … Arias, C. F. (2021). Genetic Analysis of SARS-CoV-2 Variants in Mexico during the First Year of the COVID-19 Pandemic. *Viruses*, *13*(11), 2161. https://doi.org/10.3390/v13112161

Trieu, G., & Trieu, V. N. (2021). *Mutational analysis of SARS-CoV-2. ORF8 and the evolution of the Delta and Omicron variants* (p. 2021.12.19.21268069). medRxiv. https://doi.org/10.1101/2021.12.19.21268069

Wang, M., Fu, A., Hu, B., Tong, Y., Liu, R., Liu, Z., Gu, J., Xiang, B., Liu, J., & Jiang, W. (2020). Nanopore Targeted Sequencing for the Accurate and Comprehensive Detection of SARS-CoV-2 and Other Respiratory Viruses. *Small*, *16*(32), 2002169.

Waterhouse, A. M., Procter, J. B., Martin, D. M. A., Clamp, M., & Barton, G. J. (2009). Jalview Version 2—A multiple sequence alignment editor and analysis workbench. *Bioinformatics*, *25*(9), 1189–1191. https://doi.org/10.1093/bioinformatics/btp033

What is Sanger sequencing? - MX. (s/f). Recuperado el 12 de agosto de 2021, de //www.thermofisher.com/mx/es/home/life-science/sequencing/sequencing-learning-center/capillary-electrophoresis-information/what-is-sanger-sequencing.html

WHO | SARS-CoV-2 Variants. (s/f). WHO; World Health Organization. Recuperado el 10 de mayo de 2021, de http://www.who.int/csr/don/31-december-2020-sars-cov2-variants/en/

Wu, A., Peng, Y., Huang, B., Ding, X., Wang, X., Niu, P., Meng, J., Zhu, Z., Zhang, Z., Wang, J., Sheng, J., Quan, L., Xia, Z., Tan, W., Cheng, G., & Jiang, T. (2020). Genome Composition and Divergence of the Novel Coronavirus (2019-nCoV) Originating in China. *Cell Host & Microbe*, *27*(3), 325–328. https://doi.org/10.1016/j.chom.2020.02.001

Yadav, R., Chaudhary, J. K., Jain, N., Chaudhary, P. K., Khanra, S., Dhamija, P., Sharma, A., Kumar, A., & Handu, S. (2021). Role of Structural and Non-Structural Proteins and Therapeutic Targets of SARS-CoV-2 for COVID-19. *Cells*, *10*(4), 821. https://doi.org/10.3390/cells10040821

Yang, H., Bartlam, M., & Rao, Z. (2006). Drug design targeting the main protease, the Achilles' heel of coronaviruses. *Current pharmaceutical design*, *12*(35), 4573–4590.

Young, B. E., Fong, S.-W., Chan, Y.-H., Mak, T.-M., Ang, L. W., Anderson, D. E., Lee, C. Y.-P., Amrun, S. N., Lee, B., Goh, Y. S., Su, Y. C. F., Wei, W. E., Kalimuddin, S., Chai, L. Y. A., Pada, S., Tan, S. Y., Sun, L., Parthasarathy, P., Chen, Y. Y. C., … Ng, L. F. P. (2020). Effects of a major deletion in the SARS-CoV-2 genome on the severity of infection and the inflammatory response: An observational cohort study. *The Lancet*, *396*(10251), 603–611. https://doi.org/10.1016/S0140-6736(20)31757-8

Zárate, S., Taboada, B., Muñoz-Medina, J. E., Iša, P., Sanchez-Flores, A., Boukadida, C., Herrera-Estrella, A., Selem Mojica, N., Rosales-Rivera, M., Gómez-Gil, B., Salas-Lais, A. G., Santacruz-Tinoco, C. E., Montoya-Fuentes, H., Alvarado-Yaah, J. E., Molina-Salinas, G. M., Espinoza-Ayala, G. E., Enciso-Moreno, J. A., Gutiérrez-Ríos, R. M., Loza, A., … Arias, C. F. (2022). The Alpha Variant (B.1.1.7) of SARS-CoV-2 Failed to Become Dominant in Mexico. *Microbiology Spectrum*, *10*(2), e02240-21. https://doi.org/10.1128/spectrum.02240-21

Zhang, Y., Zhang, J., Chen, Y., Luo, B., Yuan, Y., Huang, F., Yang, T., Yu, F., Liu, J., & Liu, B. (2020). The ORF8 protein of SARS-CoV-2 mediates immune evasion through potently downregulating MHC-I. *BioRxiv*.

Zhong, N., Zheng, B., Li, Y., Poon, L., Xie, Z., Chan, K., Li, P., Tan, S., Chang, Q., & Xie, J. (2003). Epidemiology and cause of severe acute respiratory syndrome (SARS) in Guangdong, People's Republic of China, in February, 2003. *The Lancet*, *362*(9393), 1353–1358.

Zhou, P., Yang, X.-L., Wang, X.-G., Hu, B., Zhang, L., Zhang, W., Si, H.-R., Zhu, Y., Li, B., Huang, C.-L., Chen, H.-D., Chen, J., Luo, Y., Guo, H., Jiang, R.-D., Liu, M.-Q., Chen, Y., Shen, X.-R., Wang, X., … Shi, Z.-L. (2020). A pneumonia outbreak associated with a new coronavirus of probable bat origin. *Nature*, *579*(7798), 270–273. https://doi.org/10.1038/s41586-020-2012-7

XII. ANEXOS

Tabla suplementaria 1. Datos de muestras con deleción de 411 nt depositados en GISAID y detectados en USA con fecha de envío entre 07/11/2020 – 03/05/2021.

Accession ID	Collection date	Submission date	Location
EPI_ISL_1094327	14/01/2021	27/02/2021	North America / USA / Arizona
EPI_ISL_1094325	19/01/2021	27/02/2021	North America / USA / California
EPI_ISL_1824306	02/02/2021	01/05/2021	North America / USA / California
EPI_ISL_977899	09/11/2020	11/02/2021	North America / USA / California
EPI_ISL_984437	22/12/2020	13/02/2021	North America / USA / California
EPI_ISL_977898	07/11/2020	11/02/2021	North America / USA / California
EPI_ISL_1094326	29/01/2021	27/02/2021	North America / USA / Indiana
EPI_ISL_1094324	27/01/2021	27/02/2021	North America / USA / Montana

EPI_ISL_776668	06/12/2020	07/01/2021	North America / USA / Washington
EPI_ISL_776669	12/12/2020	07/01/2021	North America / USA / Washington
EPI_ISL_776670	12/12/2020	07/01/2021	North America / USA / Washington

Además, para fomentar la repetibilidad la primera versión de los datos y los programas utilizados en este trabajo, con la que se participó en el concurso de salud del estado de Guanajuato fueron hechos públicos de la siguiente manera:

1. Microreact: En esta liga se encuentran almacenados los metadatos de cada muestra, así como la localización de cada una de ellas. https://microreact.org/project/foKsfc8cCbBikQ1E4wAaDU
2. GitHub: En este repositorio se encuentran los scripts que se utilizaron para analizar las muestras. https://github.com/fabel134/deleciones-Mex.git
3. Zenodo: En este enlace se encuentran los datos disponibles para esta investigación. https://doi.org/10.5281/zenodo.5226240

Figuras Suplementarias.

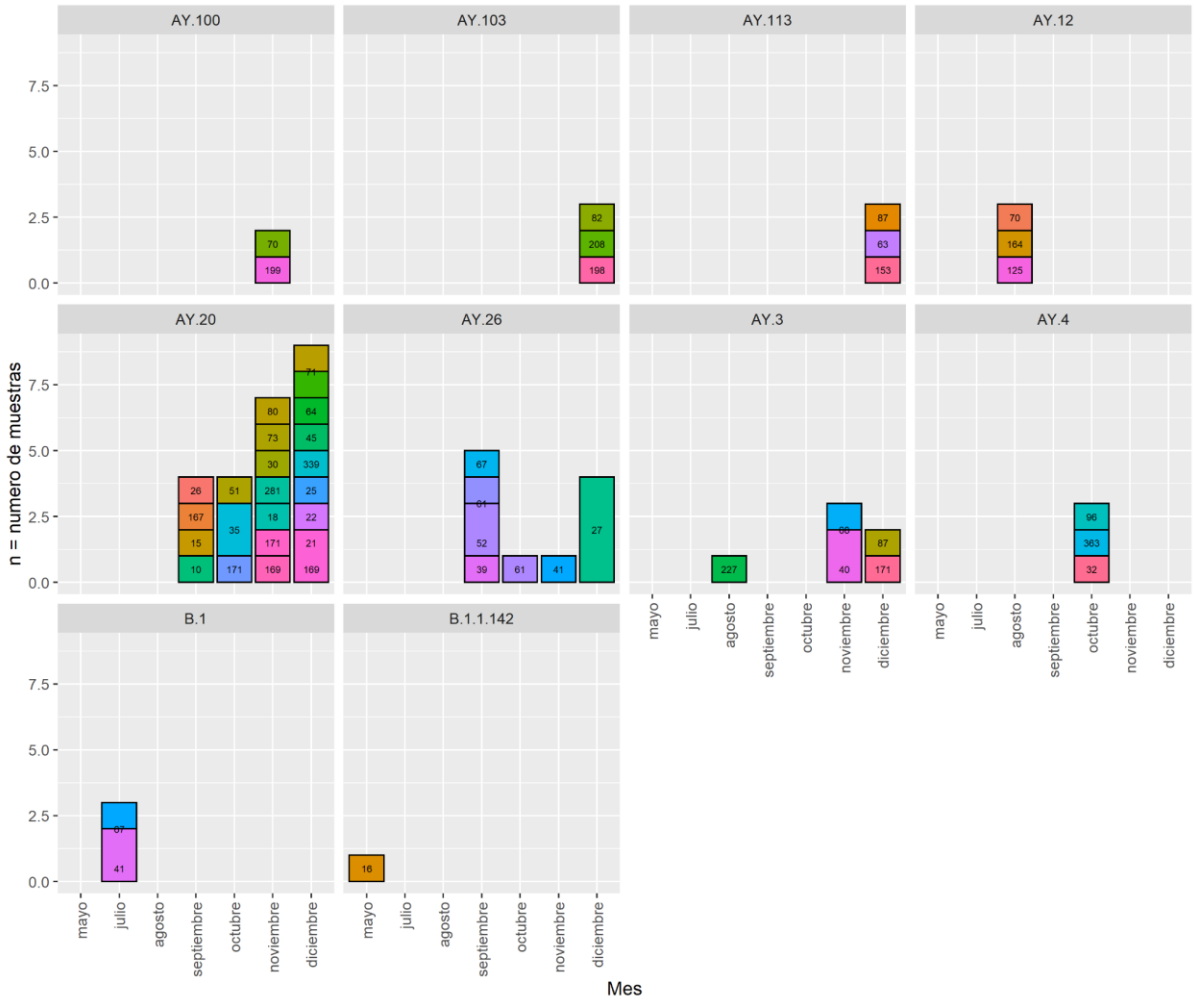


Figura S1. Tamaño de deleción de acuerdo a su linaje en SARS-CoV-2. El siguiente gráfico nos muestra los gráficos de las muestras en SARS-CoV-2. En el eje de las X podemos observar los meses y en el eje de las Y podemos observar el número de muestras. Los números que se encuentran en cada barra nos indican el tamaño de la deleción, así como el color de cada muestra nos dice el tamaño de la deleción.

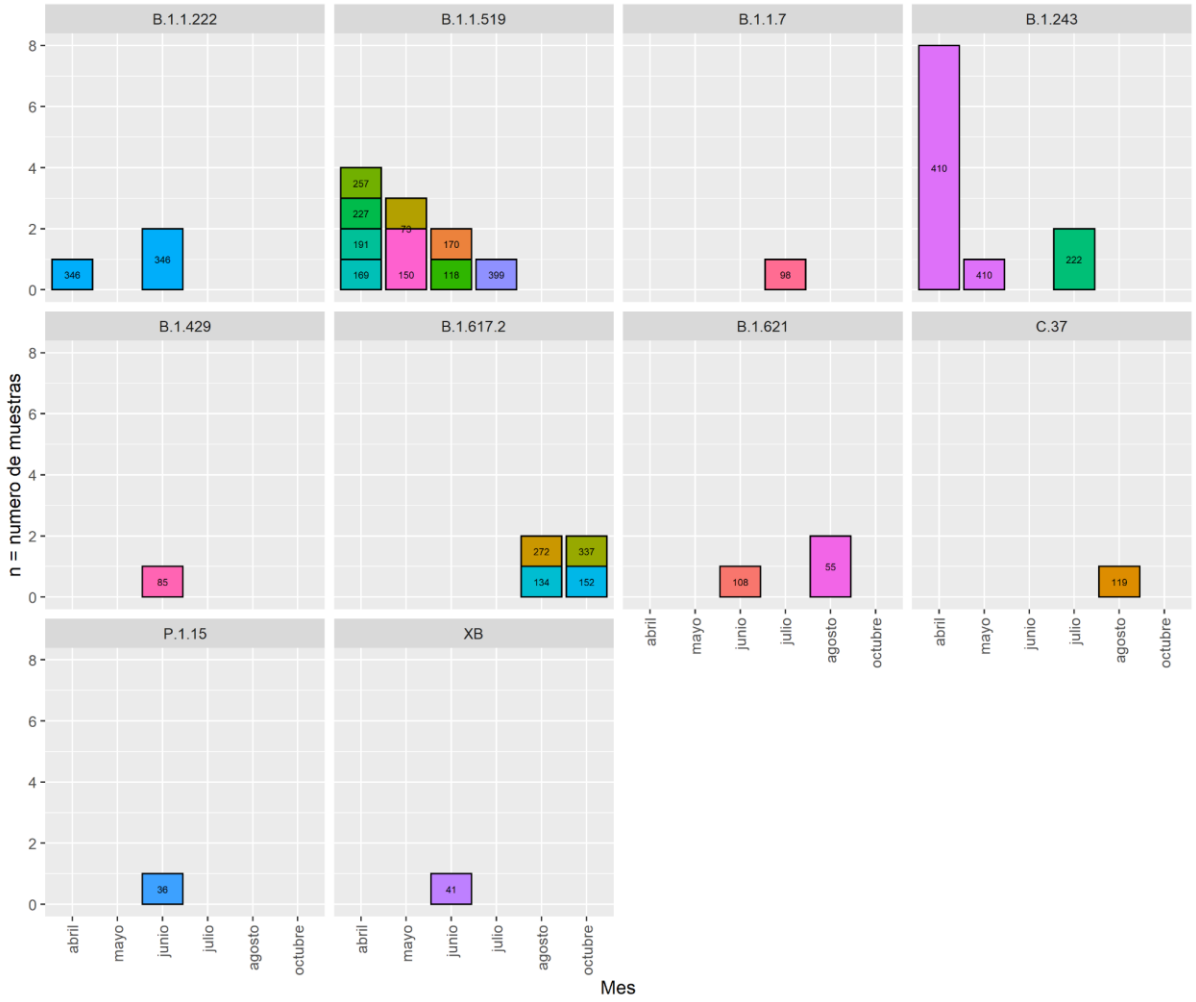


Figura S2. Tamaño de deleción de acuerdo a su linaje en SARS-CoV-2. El siguiente gráfico nos muestra los gráficos de las muestras en SARS-CoV-2. En el eje de las X podemos observar los meses y en el eje de las Y podemos observar el número de muestras. Los números que se encuentran en cada barra nos indican el tamaño de la deleción, así como el color de cada muestra nos dice el tamaño de la deleción.

Síntesis de CDNA

Se necesita una concentración de 100 ng/ul final para cada muestra. En un tubo eppendorf 1.5 mL, se agregan los reactivos como se enlistan a continuación. Todos los reactivos vienen contenidos en el siguiente kit *Protoscript II First Strand cDNA Synthesis* (E6560S,New England). Además, se agregó inhibidor de *RNAsas* (M 03147, New England) para tener un mejor rendimiento.

Tabla 1. Síntesis de cDNA

Síntesis de cDNA	
H2O	*variable
RNA (muestra con una concentración 100 ng/ul)	*variable

Oligo d(T)VN	1ul
Random primer Mix	1ul
Total	20 ul

La cantidad del H_2O depende de la cantidad de RNA para obtener un volumen final de 20 ul. Se le da una mezcla con la pipeta y se lleva al termociclador a 65° C por 5 min e inmediatamente se pone en hielo. Tomamos la muestra y se le agregan los siguientes reactivos

Protoscript Reaction Mix	10ul
Protoscript Enzyme Mix	2ul

para finalizar estas reacciones se lleva al termociclador

25°C 00:05:00
42°C 01:00:00
80 °C 00:05:00
4°C infinito
almacenar a -20° C

Tabla 2. Oligos que se utilizaron para comprobar la deleción.

Nombre del oligo	**Secuencia**	**Posición en la que mapean**
nCoV-2019_92_LEFT forward	TTTGTGCTTTTTAGCCTTTCTGCT	27785 .. 27808
nCoV-2019_93_RIGHT reverse	AGGTCTTCCTTGCCATGTTGAG	28443 .. 28464

A) Cuando existe una sola deleción en el genoma de SARS Indel-Mex es capaz de detectar lecturas que alinean en dos regiones diferentes de SARS-CoV-2.

B) Cuando existe dos deleciones cercanas en el genoma de SARS Indel-Mex solo es capaz de detectar lecturas que alinean en dos regiones diferentes de SARS-CoV-2, pero para una de ellas, ya que las posiciones de inicio y de fin son importantes para que se pueda realizar un buen diagnostico

C) Esto sucede porque InDel-Mex, esta construido para confirmar deleciones en una región bien definida. Indel-Mex solo cuenta las lecturas que se encuentra en dos regiones distintas del genoma y en la región donde esta definida esta deleción.

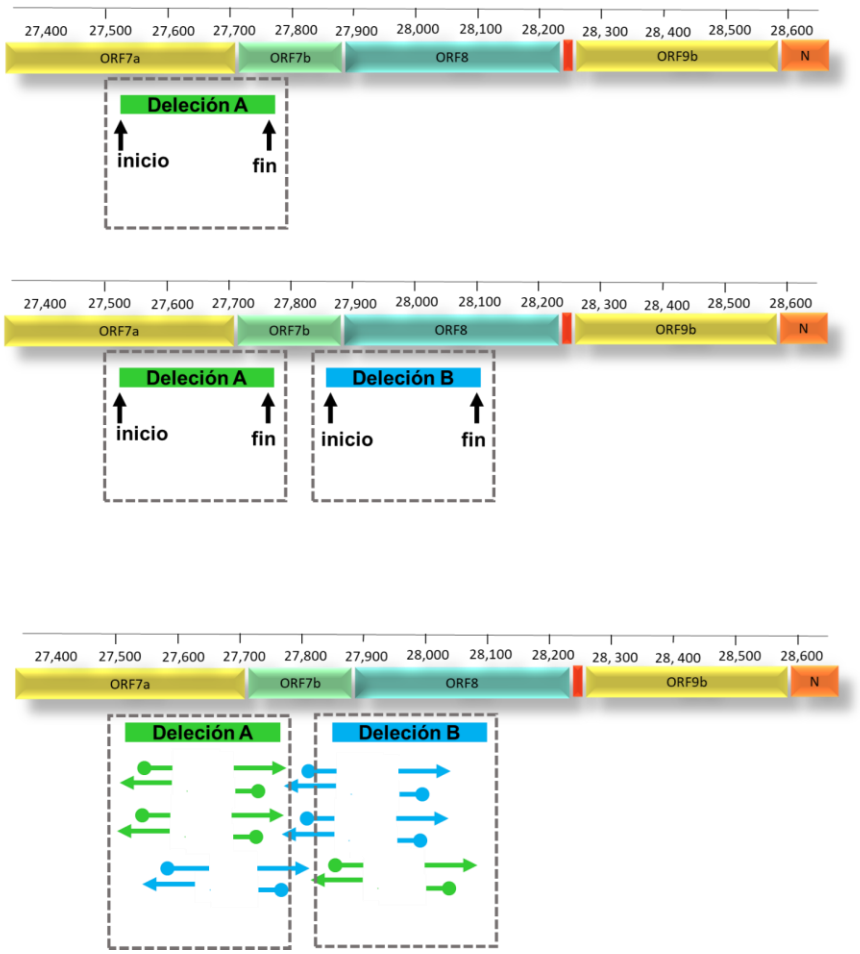


Figura 3. Indel-Mex no detecta dos deleciones que se encuentren en el mismo ORF. A) Se puede observar una deleción en los ORFs, el cual Indel-Mex es capaz confirmar la deleción. B) Indel Mex necesita posiciones específicas. C) si la muestra tiene dos o más deleciones en los ORFs, solo puede detectar a una de ellas.